INNOVATION, ENTREPRENEURSHIP AND MANAGEMENT SERIES

BIG DATA, ARTIFICIAL INTELLIGENCE AND DATA ANALYSIS SET

Volume 2

Data Analysis and Applications 1

Clustering and Regression, Modeling-estimating, Forecasting and Data Mining

Edited by
Christos H. Skiadas
James R. Bozeman

ISTE
WILEY

Data Analysis and Applications 1

Big Data, Artificial Intelligence and Data Analysis Set

coordinated by
Jacques Janssen

Volume 2

Data Analysis and Applications 1

Clustering and Regression, Modeling-estimating, Forecasting and Data Mining

Edited by

Christos H. Skiadas
James R. Bozeman

ISTE

WILEY

First published 2019 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc.

ISTE Ltd
27-37 St George's Road
London SW19 4EU
UK

www.iste.co.uk

John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030
USA

www.wiley.com

Library of Congress Control Number: 2018965155

British Library Cataloguing-in-Publication Data
A CIP record for this book is available from the British Library
ISBN 978-1-78630-382-0

Contents

Preface

Thanks to significant work by the authors and contributors, we have developed these two volumes on new and classical approaches to *Data Analysis and Applications*.

The data analysis field has been continuously growing over recent decades following the wide applications of computing and data collection along with new developments in analytical tools. Therefore, the need for publications is evident. New publications appear as printed or e-books covering the need for information from all fields of science and engineering, thanks to the wide applicability of data analysis and statistics packages.

These two volumes thus present collected material in over 30 chapters, divided into seven parts, in a form that will provide the reader with theoretical and applied information on data analysis methods, models and techniques along with appropriate applications.

In addition to the chapters, we include in both volumes an excellent introductory and review paper entitled "50 Years of Data Analysis: From Exploratory Data Analysis to Predictive Modeling and Machine Learning" by Gilbert Saporta, a leading expert in the field. The paper was based on his speech given at the celebration of his 70th birthday at the ASMDA2017 International Conference in London (held in De Morgan House of the London Mathematical Society).

Volume 1 contains the following three parts:

Part 1, Clustering and Regression, includes four chapters: "Cluster Validation by Measurement of Clustering Characteristics Relevant to the User" by Christian Hennig; "Histogram-Based Clustering of Sensor Network Data" by Antonio Balzanella and Rosanna Verde; "The Flexible Beta Regression Model" by Sonia

Migliorati, Agnese M. Di Brisco and Andrea Ongaro; and "*S*-Weighted Instrumental Variables" by Jan Amos Visek.

Part 2, Models and Modeling, includes seven chapters: "Grouping Property and Decomposition of Explained Variance in Linear Regression" by Henri Wallard; "On GARCH Models with Temporary Structural Changes" by Norio Watanabe and Okihara Fumiaki; "A Note on the Linear Approximation of TAR Models" by Francesco Giordano, Marcella Niglio and Cosimo Damiano Vitale; "An Approximation to Social Well-Being Evaluation Using Structural Equation Modeling" by Leonel Santos-Barrios, Monica Ruiz-Torres, William Gómez-Demetrio, Ernesto Sánchez-Vera, Ana Lorga da Silva and Francisco Martínez-Castañeda; "An SEM Approach to Modeling Housing Values" by Jim Freeman and Xin Zhao; "Evaluation of Stopping Criteria for Ranks in Solving Linear Systems" by Benard Abola, Pitos Biganda, Christopher Engström and Sergei Silvestrov; and "Estimation of a Two-Variable Second-Degree Polynomial via Sampling" by Papatsouma Ioanna, Farmakis Nikolaos and Ketzaki Eleni.

Part 3, Estimators, Forecasting and Data Mining, includes five chapters: "Displaying Empirical Distributions of Conditional Quantile Estimates: An Application of Symbolic Data Analysis to the Cost Allocation Problem in Agriculture" by Dominique Desbois; "Frost Prediction in Apple Orchards Based upon Time Series Models" by Monika A. Tomkowicz and Armin O. Schmitt; "Efficiency Evaluation of Multiple-Choice Questions and Exams" by Evgeny Gershikov and Samuel Kosolapov; "Methods of Modeling and Estimation in Mortality" by Christos H. Skiadas and Konstantinos Zafeiris; and "An Application of Data Mining Methods to the Analysis of Bank Customer Profitability and Buying Behavior" by Pedro Godinho, Joana Dias and Pedro Torres.

Volume 2 continues with a further four parts as follows:

Part 1, Applications, includes six chapters: "Context-specific Independence in Innovation Studies" by Federica Nicolussi and Manuela Cazzaro; "Analysis of the Determinants and Outputs of Innovation in the Nordic Countries" by Catia Rosario, Antonio Augusto Costa and Ana Lorga da Silva; "Bibliometric Variables Determining the Quality of a Dentistry Journal" by Pilar Valderrama, Manuel Escabias, Evaristo Jiménez-Contreras, Mariano J. Valderrama and Pilar Baca; "Analysis of Dependence among Growth Rates of GDP of V4 Countries Using 4-Dimensional Vine Copulas" by Jozef Komornik, Magda Komornikova and Tomas Bacigal; "Monitoring the Compliance of Countries on Emissions Mitigation Using Dissimilarity Indices" by Eleni Ketzaki, Stavros Rallakis, Nikolaos Farmakis and Eftichios Sartzetakis; and "Maximum Entropy and Distributions of Five-Star Ratings" by Yiannis Dimotikalis.

Part 2, The Impact of the Economic and Financial Crisis in Europe, contains one chapter about credit: “Access to Credit for SMEs after the 2008 Financial Crisis: The Northern Italian Perspective” by Cinzia Colapinto and Mariangela Zenga. This is followed by two chapters on the labor market: “Gender-Based Differences in the Impact of the Economic Crisis on Labor Market Flows in Southern Europe”, and “Measuring Labor Market Transition Probabilities in Europe with Evidence from the EU-SILC, both by Maria Symeonaki, Maria Karamessini and Glykeria Stamatopoulou.

Part 3, Student Assessment and Employment in Europe, has an article concerning university students who are about to graduate and hence are close to employment that is related to Part 4: “Almost Graduated, Close to Employment? Taking into Account the Characteristics of Companies Recruiting at a University Job Placement Office” by Franca Crippa, Mariangela Zenga and Paolo Mariani, followed by a paper on how students are assessed: “How Variation of Scores of the Programme for International Student Assessment Can be Explained through Analysis of Information” by Valérie Girardin, Justine Lequesne and Olivier Thévenon.

Part 4, Visualization, examines this topic in computing: “A Topological Discriminant Analysis” by Rafik Abdesselam, followed by “Using Graph Partitioning to Calculate PageRank in a Changing Network” by Christopher Engström and Sergei Silvestrov, and in politics: “Visualizing the Political Spectrum of Germany by Contiguously Ordering the Party Policy Profiles by Andranik Tangian.

We deeply thank the authors and contributors to this book. We pass on our sincere appreciation to the referees for their hard work and dedication in providing an improved book form. Finally, we express our thanks to the secretariat and, of course, the publishers.

December 2018

Christos H. SKIADAS, Athens, Greece
James R. BOZEMAN, Bormla, Malta

Introduction

50 Years of Data Analysis: From Exploratory Data Analysis to Predictive Modeling and Machine Learning

In 1962, J.W. Tukey wrote his famous paper "The Future of Data Analysis" and promoted exploratory data analysis (EDA), a set of simple techniques conceived to let the data speak, without prespecified generative models. In the same spirit, J.P. Benzécri and many others developed multivariate descriptive analysis tools. Since that time, many generalizations occurred, but the basic methods (SVD, *k*-means, etc.) are still incredibly efficient in the Big Data era.

On the other hand, algorithmic modeling or machine learning is successful in predictive modeling, the goal being accuracy and not interpretability. Supervised learning proves in many applications that it is not necessary to understand, when one needs only predictions.

However, considering some failures and flaws, we advocate that a better understanding may improve prediction. Causal inference for Big Data is probably the challenge of the coming years.

It is a little presumptuous to want to make a panorama of 50 years of data analysis, while David Donoho (2017) has just published a paper entitled "50 Years of Data Science". But 1968 is the year when I began my studies as a statistician and I would very much like to talk about the debates of the time and the digital revolution that profoundly transformed statistics and which I witnessed. The terminology followed this evolution–revolution: from data analysis to data mining

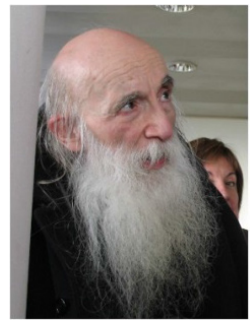
Chapter written by Gilbert SAPORTA.

and then to data science while we went from a time when the asymptotics began to 30 observations with a few variables in the era of Big Data and high dimension.

I.1. The revolt against mathematical statistics

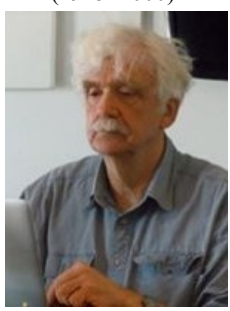
Since the 1960s, the availability of data has led to an international movement back to the sources of statistics ("let the data speak") and to sometimes fierce criticisms of an abusive formalization. Along with to John Tukey, who was cited above, here is a portrait gallery of some notorious protagonists in the United States, France, Japan, the Netherlands and Italy (for a color version of this figure, see www.iste.co.uk/skiadas/data1.zip).

John Wilder Tukey
(1915–2000)



Jean-Paul Benzécri
(1932–)

Chikio Hayashi
(1918–2002)



Jan de Leeuw
(1945–)

J. Douglas Carroll
(1939–2011)

Carlo Lauro
(1943–)

And an anthology of quotes:

> He (Tukey) seems to identify statistics with the grotesque phenomenon generally known as mathematical statistics and find it necessary to replace statistics by data analysis. (Anscombe 1967)

> Statistics is not probability, under the name of mathematical statistics was built a pompous discipline based on theoretical assumptions that are rarely met in practice. (Benzécri 1972)
>
> The models should follow the data, not vice versa. (Benzécri 1972)
>
> Use the computer implies the abandonment of all the techniques designed before of computing. (Benzécri 1972)
>
> Statistics is intimately connected with science and technology, and few mathematicians have experience or understand of methods of either. This I believe is what lies behind the grotesque emphasis on significance tests in statistics courses of all kinds; a mathematical apparatus has been erected with the notions of power, uniformly most powerful tests, uniformly most powerful unbiased tests, etc., and this is taught to people, who, if they come away with no other notion, will remember that statistics is about significant differences […]. The apparatus on which their statistics course has been constructed is often worse than irrelevant – it is misleading about what is important in examining data and making inferences. (Nelder 1985)

Data analysis was basically descriptive and non-probabilistic, in the sense that no reference was made to the data-generating mechanism. Data analysis favors algebraic and geometrical tools of representation and visualization.

This movement has resulted in conferences especially in Europe. In 1977, E. Diday and L. Lebart initiated a series entitled Data Analysis and Informatics, and in 1981, J. Janssen was at the origin of biennial ASMDA conferences (Applied Stochastic Models and Data Analysis), which are still continuing.

The principles of data analysis inspired those of data mining, which developed in the 1990s on the border between databases, information technology and statistics. Fayaad (1995) is said to have the following definition: “Data Mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data”. Hand *et al*. precised in 2000, “I shall define Data Mining as the discovery of interesting, unexpected, or valuable structures in large data sets”.

The metaphor of data mining means that there are treasures (or nuggets) hidden under mountains of data, which may be discovered by specific tools. Data mining is generally concerned with data which were collected for another purpose: it is a secondary analysis of databases that are collected not primarily for analysis, but for the management of individual cases. Data mining is not concerned with efficient

methods for collecting data such as surveys and experimental designs (Hand *et al.* 2000).

I.2. EDA and unsupervised methods for dimension reduction

Essentially, exploratory methods of data analysis are dimension reduction methods: unsupervised classification or clustering methods operate on the number of statistical units, whereas factorial methods reduce the number of variables by searching for linear combinations associated with new axes of the space of individuals.

I.2.1. *The time of syntheses*

It was quickly realized that all the methods looking for eigenvalues and eigenvectors of matrices related to the dispersion of a cloud (total or intra) or of correlation matrices could be expressed as special cases of certain techniques.

Correspondence analyses (single and multiple) and canonical discriminant analysis are particular principal component analyses. It suffices to extend the classical Principal Components Analysis (PCA) by weighting the units and introducing metrics. The duality scheme introduced by Cailliez and Pagès (1976) is an abstract way of representing the relationships between arrays, matrices and associated spaces. The paper by De la Cruz and Holmes (2011) brought it back to light.

From another point of view (Bouroche and Saporta 1983), the main factorial methods PCA, Multiple Correspondence Analysis (MCA), as well as multiple regression are particular cases of canonical correlation analysis.

Another synthesis comes from the generalization of canonical correlation analysis to several groups of variables introduced by J.D. Carroll (1968). Given p blocks of variables $\mathbf{X}_j$, we look for components $\mathbf{z}$ maximizing the following criterion: $\sum_{j=1}^{p} R^2\left(\mathbf{z}, \mathbf{X}_j\right)$.

The extension of this criterion in the form $Max_Y \sum_{j=1}^{p} \Phi(Y, X_j)$, where Φ is an adequate measure of association, leads to the maximum association principle (Tenenhaus 1977; Marcotorchino 1986; Saporta 1988), which also includes the case of *k*-means partitioning.

The PLS approach to structural equation modeling also provides a global framework for many linear methods, as has been shown by Tenenhaus (1999) and Tenenhaus and Tenenhaus (2011).

Criterion	Analysis
$\max \sum_{j=1}^{p} r^2(c, x_j)$ with x_j numerical	PCA
$\max \sum_{j=1}^{p} \eta^2(c, \boldsymbol{x}_j)$ with $\boldsymbol{x}_j$ categorical	MCA
$\max \sum_{j=1}^{p} R^2(c, \mathbf{X}_j)$ with $\mathbf{X}_j$ data set	GCA (Carroll)
$\max \sum_{j=1}^{p} Rand(Y, \boldsymbol{x}_j)$ with Y and $\boldsymbol{x}_j$ categorical	Central partition
$\max \sum_{j=1}^{p} \tau(y, x_j)$ with rank orders	Condorcet aggregation rule

Table I.1. *Various cases of the maximum association principle*

I.2.2. *The time of clusterwise methods*

The search for partitions in k classes of a set of units belonging to a Euclidean space is most often done using the k-means algorithm: this method converges very quickly, even for large sets of data, but not necessarily toward the global optimum. Under the name of dynamic clustering, Diday (1971) has proposed multiple extensions, where the representatives of classes can be groups of points, varieties, etc. The simultaneous search for k classes and local models by alternating k-means and modeling is a geometric and non-probabilistic way of addressing mixture problems. Clusterwise regression is the best-known case: in each class, a regression model is fitted and the assignment to the classes is done according to the best model. Clusterwise methods allow for non-observable heterogeneity and are particularly useful for large data sets where the relevance of a simple and global model is questionable. In the 1970s, Diday and his collaborators developed "typological" approaches for most linear techniques: PCA, regression (Charles 1977), discrimination. These methods are again the subject of numerous publications in association with functional data (Preda and Saporta 2005), symbolic data (de Carvalho *et al*. 2010) and in multiblock cases (De Roover *et al*. 2012; Bougeard *et al*. 2017).

I.2.3. *Extensions to new types of data*

I.2.3.1. *Functional data*

Jean-Claude Deville (1974) showed that the Karhunen–Loève decomposition was nothing other than the PCA of the trajectories of a process, opening the way to functional data analysis (Ramsay and Silverman 1997). The number of variables being infinitely not countable, the notion of linear combination to define a principal component is extended to the integral $\xi = \int_0^T f(t)X_t dt$, $f(t)$ being an eigenfunction of the covariance operator $\int_0^T C(t,s)f(s)ds = \lambda f(t)$.

Deville and Saporta (1980) then extended functional PCA to correspondence analysis of trajectories of a categorical process.

The dimension reduction offered by PCA makes it possible to solve the problem of regression on trajectories, a problem that is ill posed since the number of observations is smaller than the infinite number of variables. PLS regression, however, is better adapted in the latter case and makes it possible to deal with supervised classification problems (Costanzo *et al*. 2006).

I.2.3.2. *Symbolic data analysis*

Diday is at the origin of many works that have made it possible to extend almost all methods of data analysis to new types of data, called symbolic data. This is the case, for example, when the cell i, j of a data table is no longer a number, but an interval or a distribution. See Table I.2 for an example of a table of symbolic data (from Billard and Diday 2006).

w_u	Court Type	Player Weight	Player Height	Racket Tension
w_1	Hard	[65, 86]	[1.78, 1.93]	[14, 99]
w_2	Grass	[65, 83]	[1.80, 1.91]	[26, 99]
w_3	Indoor	[65, 87]	[1.75, 1.93]	[14, 99]
w_4	Clay	[68, 84]	[1.75, 1.93]	[24, 99]

Table I.2. *An example of interval data*

I.2.3.3. *Textual data*

Correspondence analysis and classification methods were, very early, applied to the analysis of document-term and open-text tables (refer to Lebart *et al*. 1998 for a full presentation). Text analysis is now part of the vast field of text mining or text analytics.

I.2.4. *Nonlinear data analysis*

Dauxois and Pousse (1976) extended principal component analysis and canonical analysis to Hilbert spaces. By simplifying their approach, instead of looking for linear combinations of maximum variance like in PCA $\max V\left(\sum_{j=1}^{p} a_j x^j\right)$ subject to $\|\mathbf{a}\| = 1$, we look for separate nonlinear transformations Φ_j of each variable maximizing $V\left(\sum_{j=1}^{p} \Phi_j\left(x^j\right)\right)$. This is equivalent to maximize the sum of the squares of the correlation coefficients between the principal component c and the transformed variables $\sum_{j=1}^{p} \rho^2\left(c, \Phi_j\left(x^j\right)\right)$, which is once again an illustration of the maximum association principle.

With a finite number of observations *n*, this is an ill-posed problem, and we need to restrict the set of transformations Φ_j to finite dimension spaces. A classical choice is to use spline functions as in Besse (1988).

The search for optimal transformations has been the subject of work by the Dutch school, summarized in the book published by Gifi (1999).

Separate transformations are called semilinear. A different attempt to obtain "truly" nonlinear transformations is kernelization. In line with the work of V. Vapnik, Schölkopf *et al*. (1998) defined a nonlinear PCA in the following manner where the entire vector $\mathbf{x} = (x^1, x^2, \ldots, x^p)$ is transformed. Each point of the space of the individual E is transformed into a point in a space $\Phi(E)$ called extended space (or feature space) provided with a dot product. The dimension of $\Phi(E)$ can be very large and the notion of variable is lost. A metric multidimensional scaling is then performed on the transformed points according to the Torgerson method, which is equivalent to the PCA in $\Phi(E)$. Everything depends on the choice of the scalar product in $\Phi(E)$: if we take a scalar product that is easily expressed according to the scalar product of E, it is no longer necessary to know the transformation Φ, which is then implicit. All calculations are done in dimension n. This is the "kernel trick".

Let $k(\mathbf{x}, \mathbf{y})$ be a dot product in $\Phi(E)$ and $< \mathbf{x}, \mathbf{y} >$ the dot product of E. We then replace the usual Torgerson's matrix $\mathbf{W}$ by a matrix where each element is $k(\mathbf{x}, \mathbf{y})$, then doubly center W in rows and columns: its eigenvectors are the principal components in $\Phi(E)$.

Once the kernel-PCA was defined, many works followed, "kernelizing" by various methods, such as Fisher discriminant analysis by Baudat and Anouar (2000) found independently under the name of LS-SVM by Suykens and Vandewalle (1999), the PLS regression of Rosipal and Trejo (2001), the unsupervised classification with kernels *k*-means already proposed by Schölkopf *et al*. and canonical analysis (Fyfe and Lai 2001). It is interesting to note that most of these developments came not from statisticians but from researchers of artificial intelligence or machine learning.

I.2.5. *The time of sparse methods*

When the number of dimensions (or variables) is very large, PCA, MCA and other factorial methods lead to results that are difficult to interpret: how to make sense of a linear combination of several hundred or even thousands of variables? The search for the so-called "sparse" combinations limited to a small number of variables, that is, with a large number of zero coefficients, has been the subject of the attention of researchers for about 15 years. The first attempts requiring that the coefficients be equal to –1, 0 or 1, for example, lead to non-convex algorithms that are difficult to use.

The transposition to PCA of the LASSO regression de Tibshirani (1996) allowed exact and elegant solutions. Recall that the LASSO consists of performing a regression with an L^1 penalty on the coefficients, which makes it possible to easily manage the multicollinearity and the high dimension.

$$\hat{\boldsymbol{\beta}}_{lasso} = \arg\min_{\beta}\left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right).$$

Zou *et al*. (2006) proposed modifying one of the many criteria defining the PCA of a table $\mathbf{X}$: principal components $\mathbf{z}$ are such that:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \|\mathbf{z} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 .$$

The first constraint in an L^2 norm only implies that the loadings have to be normalized; the second constraint in an L^1 norm tunes the sparsity when the Lagrange multiplier λ_1 varies. Computationally, we get the solution by alternating an SVD $\boldsymbol{\beta}$ being fixed, to get the components $\mathbf{z}$ and an elastic-net to find $\boldsymbol{\beta}$ when $\mathbf{z}$ is fixed until convergence.

The positions of the null coefficients are not the same for the different components. The selection of the variables is therefore dimension by dimension. If

the interpretability increases, the counterpart is the loss of characteristic properties of PCA, such as the orthogonality of the principal components and/or the loadings. Since then, sparse variants of many methods have been developed, such as sparse PLS by Chun and Keleş (2009), sparse discriminant analysis by Clemmensen *et al*. (2011), sparse canonical analysis by Witten *et al*. (2009) and sparse multiple correspondence analysis by Bernard *et al*. (2012).

I.3. Predictive modeling

A narrow view would limit data analysis to unsupervised methods to use current terminology. Predictive or supervised modeling has evolved in many ways into a conceptual revolution comparable to that of the unsupervised. We have moved from a model-driven approach to a data-driven approach where the models come from the exploration of the data and not from a theory of the mechanism generating observations, thus reaffirming the second principle of Benzécri: "the models should follow the data, not vice versa".

The difference between these two cultures (generative models versus algorithmic models, or models to understand versus models to predict) has been theorized by Breiman (2001), Saporta (2008), Shmueli (2010) and taken up by Donoho (2015). The meaning of the word model has evolved: from that of a parsimonious and understandable representation centered on the fit to observations (*predict the past*), we have moved to black-box-type algorithms, whose objective is to forecast the most precisely possible new data (*predict the future*). The success of machine learning and especially the renewal of neural networks with deep learning have been made possible by the increase in computing power, but also and above all by the availability of huge learning bases.

I.3.1. *Paradigms and paradoxes*

When we ask ourselves what a good model is, we quickly arrive at paradoxes.

A generative model that fits well with collective data can provide poor forecasts when trying to predict individual behaviors. The case is common in epidemiology. On the other hand, good predictions can be obtained with uninterpretable models: targeting customers or approving loans does not require a consumer theory. Breiman remarked that simplicity is not always a quality:

> Occam's Razor, long admired, is usually interpreted to mean that simpler is better. Unfortunately in prediction, accuracy and simplicity (interpretability) are in conflict.

> Modern statistical thinking makes a clear distinction between the statistical model and the world. The actual mechanisms underlying the data are considered unknown. The statistical models do not need to reproduce these mechanisms to emulate the observable data. (Breiman 2001)

Other quotes illustrate these paradoxes:

> Better models are sometimes obtained by deliberately avoiding to reproduce the true mechanisms. (Vapnik 2006)

> Statistical significance plays a minor or no role in assessing predictive performance. In fact, it is sometimes the case that removing inputs with small coefficients, even if they are statistically significant, results in improved prediction accuracy. (Shmueli 2010)

In a Big Data world, estimation and tests become useless, because everything is significant! For instance, a correlation coefficient equal to 0.002 when the number of observations is 10^6 is significantly different from 0, but without any interest. Usual distributional models are rejected since small discrepancies between model and data are significant. Confidence intervals have zero length. We should keep in mind the famous sentence of George Box: "All models are wrong, some are useful".

I.3.2. *From statistical learning theory to empirical validation*

One of the major contributions of the theory of statistical learning developed by Vapnik and Cervonenkis was to give the conditions of generalizability of the predictive algorithms and to establish inequalities on the difference between the empirical error of adjustment of a model to observed data and the theoretical error when applying this model to future data from the same unknown distribution. If the theory is not easy to use, it has given rise to the systematization of the practice of dividing data into three subsets: learning, testing, validation (Hastie *et al*. 2001).

There had been warnings in the past, like that of Paul Horst (1941), who said, "the usefulness of a prediction procedure is not established when it is found to predict adequately on the original sample; the necessary next step must be its application to at least a second group. Only if it predicts adequately on subsequent samples can the value of the procedure be regarded as established" and the finding of cross-validation by Lachenbruch and Mickey (1968) and Stone (1974). But it is only recently that the use of validation and test samples has become widespread and has become an essential step for any data scientist. However, there is still room for improvement if we go through the publications of certain areas where prediction is rarely checked on a hold-out sample.

I.3.3. *Challenges*

Supervised methods have become a real technology governed by the search for efficiency. There is now a wealth of methods, especially for binary classification: SVM, random forests, gradient boosting, neural networks, to name a few. Ensemble methods are superimposed to combine them (see Noçairi *et al*. 2016). Feature engineering consists of constructing a large number of new variables functions of those observed and choosing the most relevant ones. While in some cases the gains over conventional methods are spectacular, this is not always the case, as noted by Hand (2006).

Software has become more and more available: in 50 years, we have moved from the era of large, expensive commercial systems (SAS, SPSS) to the distribution of free open source packages like R and ScikitLearn. The benefits are immense for the rapid dissemination of new methods, but the user must be careful about the choice and often the lack of validation and quality control of many packages: it is not always clear if user-written packages are really doing what they claim to be doing. Hornik (2012) has already wondered if there are not too many R packages.

Ten years ago, in a resounding article, Anderson (2008) prophesied the end of theory because "the data deluge makes the scientific method obsolete". In a provocative manner, he wrote "Petabytes allow us to say: 'Correlation is enough.' We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot". This was, of course, misleading, and the setbacks of Google's epidemic influenza forecasting algorithm brought a denial (Lazer *et al*. 2014). Correlation is not causality and drawing causal inference from observational data has always been a tricky problem. As Box *et al*. (1978) put it, "[t]o find out what happens when you change something, it is necessary to change it." The best way to answer causal questions is usually to run an experiment. Drawing causal inference from Big Data is now a hot topic (see Bottou *et al*. 2013; Varian 2016).

Quantity is not quality and massive data can be biased and lead to unfortunate decisions reproducing *a priori* that led to their collection. Many examples have been discovered related to discrimination or presuppositions about gender or race. More generally, the treatment of masses of personal data raises ethical and privacy issues when consent has not been gathered or has not been sufficiently explained. Books for the general public such as Keller and Neufeld (2014) and O'Neil (2016) have echoed this.

I.4. Conclusion

The past 50 years have been marked by dramatic changes in statistics. The ones that will follow will not be less formidable. The Royal Statistical Society is not afraid to write in its Data Manifesto "What steam was to the 19th century, and oil has been to the 20th, data is to the 21st".

Principles and methods of data analysis are still actual, and exploratory (unsupervised) and predictive (supervised) analysis are two sides of the same approach. But as correlation is not enough, causal inference could be the new frontier and could go beyond the paradox of predicting without understanding by going toward understanding to better predict, and act to change.

As the job of the statistician or data scientist becomes more exciting, we believe that it will have to be accompanied by an awareness of social responsibility.

I.5. References

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. http://www.wired.com/2008/06/pb-theory/.

Baudat, G., Anouar, F. (2000). Generalized discriminant analysis using a kernel approach. *Neural Comput.,* 12(10), 2385–2404.

Bernard, A., Guinot, C., Saporta, G. (2012). Sparse principal component analysis for multiblock data and its extension to sparse multiple correspondence analysis, In: *Proc. of 20th Int. Conference on Computational Statistics (COMPSTAT 2012)*, Colubi, A., Fokianos, K., Gonzalez-Rodriguez, G., Kontoghiorghes, E. (eds). International Statistical Institute (ISI), 99–106.

Besse, P. (1988). Spline functions and optimal metric in linear principal components analysis. In: *Components and Correspondence Analysis*, Van Rijckevorsel *et al*., (eds). John Wiley & Sons, New York.

Billard, L., Diday, E. (2012). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. John Wiley & Sons, Chichester.

Bottou, L. *et al*. (2013). Counterfactual reasoning and learning systems: The example of computational advertising. *J. Machine Learn. Res*., 14, 3207–3260.

Bougeard, S., Abdi, H., Saporta, G., Niang Keita, N. (2018). Clusterwise analysis for multiblock component methods. *Advances in Data Analysis and Classification*, 12(2), 285–313.

Box, G., Hunter, J.S, Hunter, W.G. (1978). *Statistics for Experimenters*, John Wiley & Sons, New York.

Breiman, L. (2001) Statistical modeling: The two cultures, *Statist. Sci.*, 16(3), 199–231.

Cailliez, F., Pagès, J.P. (1976). *Introduction à l'analyse des données*, Smash, Paris.

Carroll, J.D. (1968). Generalisation of canonical correlation analysis to three or more sets of variables. *Proc. 76th Annual Convention Am. Psychol. Assoc*., 3, 227–228.

Chun, H. , Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Statist. Soc. B,* 72, 3–25.

Clemmensen, L., Hastie, T., Ersboell, K. (2011). Sparse discriminant analysis. *Technometrics*, 53(4), 406–413.

Costanzo, D., Preda, C., Saporta, G. (2006). Anticipated prediction in discriminant analysis on functional data for binary response. In: *COMPSTAT'06*, A. Rizzi (ed.) Physica-Verlag, 821–828.

De Roover, K., Ceulemans, E., Timmerman, M.E., Vansteelandt, K., Stouten, J., Onghena, P. (2012). Clusterwise simultaneous component analysis for analyzing structural differences in multivariate multiblock data. *Psychol Methods*, 17(1), 100–119.

De la Cruz, O., Holmes, S.P. (2011). The Duality Diagram in Data Analysis: Examples of Modern Applications, *Ann. Appl. Statist.,* 5(4), 2266–2277.

Deville J.C., (1974). Méthodes statistiques et numériques de l'analyse harmonique, *Ann. l'INSEE,* 15, 3–101.

Deville J.C., Saporta, G. (1980). Analyse harmonique qualitative. In: *Data Analysis and Informatics*, E. Diday (ed.), North-Holland, Amsterdam, 375–389.

Diday, E. (1974). Introduction à l'analyse factorielle typologique, *Revue Statist. Appl.*, 22(4), 29–38.

Donoho, D. (2017). 50 Years of Data Science, *J. Comput. Graph. Statist.*, 26(4), 745–766.

Friedman, J.H. (2001). The Role of Statistics in the Data Revolution?, *Int. Statist. Rev.,* 69(1), 5–10.

Fyfe, C., & Lai, P. L. (2001). Kernel and nonlinear canonical correlation analysis. *Int. J. Neural Syst.*, 10, 365–374.

Gifi, A. (1990). *Non-linear multivariate analysis*, John Wiley & Sons, New York.

Hand, D., Blunt, G., Kelly, M., Adams, N. (2000). Data mining for fun and profit, *Statist. Sci.*, 15(2), 111–126.

Hand, D. (2006). Classifier Technology and the Illusion of Progress, *Statist. Sci.*, 21(1), 1–14.

Hastie,T., Tibshirani, R., Friedman, J. (2001). *The Elements of Statistical Learning*, Springer, New York.

Keller, M., Neufeld, J. (2014). *Terms of Service: Understanding Our Role in the World of Big Data*, Al Jazeera America, "http://projects.aljazeera.com/2014/terms-of-service/" http://projects.aljazeera.com/2014/terms-of-service/#1.

Hornik, K. (2012). Are There Too Many R Packages? *Aust. J. Statist.*, 41(1), 59–66.

Lazer, D., Kennedy, R., King, G., Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis, *Science*, 343(6176), 1203–1205.

Lebart, L., Salem, A., Berry, L. (1998). *Exploring Textual Data*, Kluwer Academic Publisher, Dordrecht, The Netherlands.

Marcotorchino, F. (1986). Maximal association as a tool for classification, in *Classification as a tool for research,* Gaul & Schader (eds), North Holland, Amstedam, 275–288.

Nelder, J.A. (1985) discussion of Chatfield, C., The initial examination of data, *J. R. Statist. Soc. A*, 148, 214–253.

Noçairi, H., Gomes,C., Thomas, M., Saporta, G. (2016). Improving Stacking Methodology for Combining Classifiers; Applications to Cosmetic Industry, *Electronic J. Appl. Statist. Anal.*, 9(2), 340–361.

O'Neil, C. (2016) *Weapons of Maths Destruction*, Crown, New York.

Ramsay, J.O., Silverman, B. (1997). *Functional data analysis*, Springer, New York.

Rosipal, A., Trejo, L. (2001). Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space, *J. Machine Learn. Res.,* 2, 97–123.

Schölkopf, B., Smola,A., Müller, K.L. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem, *Neural Comput.*, 10(5), 1299–1319.

Suykens, J.A.K.; Vandewalle, J. (1999). Least squares support vector machine classifiers, *Neural Process. Lett.*, 9(3), 293–300.

Saporta, G. (1988). About maximal association criteria in linear analysis and in cluster analysis. In: *Classification and Related Methods of Data Analysis*, H.H. Bock (ed.), 541–550, North-Holland, Amsterdam.

Saporta, G. (2008). Models for understanding versus models for prediction, In P. Brito (ed.), *Compstat Proceedings*, Physica Verlag, Heidelberg, 315–322.

Shmueli, G. (2010). To explain or to predict? *Statist. Sci.,* 25, 289–310.

Tenenhaus, M. (1977). Analyse en composantes principales d'un ensemble de variables nominales ou numériques, *Revue Statist. Appl.*, 25(2), 39–56.

Tenenhaus, M. (1999). L'approche PLS, *Revue Statist. Appl.*, 17(2), 5–40.

Tenenhaus, A., Tenenhaus, M. (2011). Regularized Generalized Canonical Correlation Analysis, *Psychometrika*, 76(2), 257–284.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc. B*, 58, 267–288.

Tukey, J.W. (1962). The Future of Data Analysis, *Ann. Math. Statist*., 33(1), 1–67.

Vapnik, V. (2006). *Estimation of Dependences Based on Empirical Data*, 2nd edition, Springer, New York.

Varian, H. (2016). Causal inference in economics and marketing, *Proc. Natl. Acad. Sci.*, 113, 7310–7315.

Witten, D., Tibshirani, R., Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics,* 10(3), 515–534.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.*, 15, 265–286.

PART 1

Clustering and Regression

1

Cluster Validation by Measurement of Clustering Characteristics Relevant to the User

There are many cluster analysis methods that can produce quite different clusterings on the same data set. Cluster validation is about the evaluation of the quality of a clustering; "relative cluster validation" is about using such criteria to compare clusterings. This can be used to select one of a set of clusterings from different methods, or from the same method ran with different parameters such as different numbers of clusters.

There are many cluster validation indexes in the literature. Most of them attempt to measure the overall quality of a clustering by a single number, but this can be inappropriate. There are various different characteristics of a clustering that can be relevant in practice, depending on the aim of clustering, such as low within-cluster distances and high between-cluster separation.

In this chapter, a number of validation criteria will be introduced that refer to different desirable characteristics of a clustering, and that characterize a clustering in a multidimensional way. In specific applications the user may be interested in some of these criteria rather than others. A focus of the chapter is on methodology to standardize the different characteristics so that users can aggregate them in a suitable way specifying weights for the various criteria that are relevant in the clustering application at hand.

1.1. Introduction

The aim of this chapter is to present a range of cluster validation indexes that provide a multivariate assessment covering different complementary aspects of cluster validity. Here, I focus on "internal" validation criteria that measure the quality of a clustering without reference to external information such as a known "true" clustering. Furthermore, I am mostly interested in comparing different clusterings on the same

Chapter written by Christian HENNIG.

Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining,
First Edition. Edited by Christos H. Skiadas and James R. Bozeman.

data, which is often referred to as "relative" cluster validation. This can be used to select one of a set of clusterings from different methods, or from the same method ran with different parameters such as different numbers of clusters.

In the literature (for an overview, see Halkidi *et al.* 2016), many cluster validation indexes are proposed. Usually, these are advertised as measures of global cluster validation in a univariate way, often under the implicit or explicit assumption that for any given data set there is only a single best clustering. Mostly, these indexes are based on contrasting a measure of within-cluster homogeneity and a measure of between-clusters heterogeneity such as the famous index proposed by Calinski and Harabasz (1974), which is a standardized ratio of the traces of the pooled within-cluster covariance matrix and the covariance matrix of between-cluster means.

In Hennig (2016; see also Hennig 2015), I have argued that depending on the subject matter background and the clustering aim, different clusterings can be optimal on the same data set. For example, clustering can be used for data compression and information reduction, in which case it is important that all data are optimally represented by the cluster centroids, or clustering can be used for recognition of meaningful patterns, which are often characterized by clear separating gaps between them. In the former situation, large within-cluster distances are not desirable, whereas in the latter situation large within-cluster distances may not be problematic as long as data objects occur with high density and without gap between the objects between which the distance is large. See Figure 1.1 for two different clusterings on an artificial data set with three clusters that may be preferable for these two different clustering aims.

Given a multivariate characterization of the validity of a clustering, for a given application a user can select weights for the different characteristics depending on the clustering aim and the relevance of the different criteria. A weighted average can then be used to choose a clustering that is suitable for the specific application. This requires that the criteria measuring different aspects of cluster validity and normalized in such a way that their values are comparable when doing the aggregation. Although it is easy in most cases to define criteria in such a way that their value range is $[0, 1]$, this is not necessarily enough to make their values comparable, because within this range the criteria may have very different variation. The idea here is that the expected variation of the criteria can be explored using resampled random clusterings ("stupid K-centroids", "stupid nearest neighbor clustering") on the same data set, and this can be used for normalization and comparison.

The approach presented here can also be used for benchmarking cluster analysis methods. Particularly, it does not only allow to show that methods are better or worse on certain data sets. It also allows to characterize the specific strengths and weaknesses of clustering algorithms in terms of the properties of the found clusters.

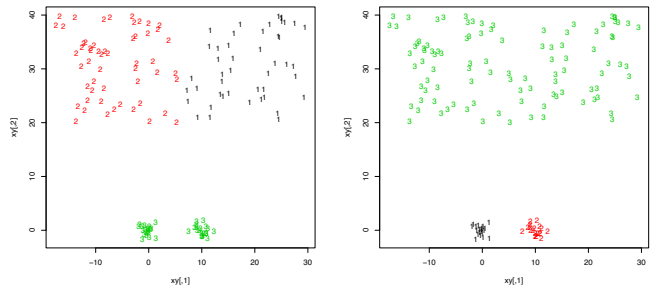


Figure 1.1. *Artificial data set. Left side: clustering by 3-means. Right side: clustering by single linkage with three clusters. For a color version of this figure, see www.iste.co.uk/skiadas/data.zip*

Section 1.2 introduces the general setup and defines notation. In section 1.3, all the indexes measuring different relevant aspects of a clustering are presented. Section 1.4 defines an aggregated index that can be adapted to practical needs. The indexes cannot be suitably aggregated in their raw form, and section 1.5 introduces a calibration scheme using randomly generated clusterings. Section 1.6 applies the methodology to two data sets, one illustrative artificial one and a real data set regarding species delimitation. Section 1.7 concludes the chapter.

1.2. General notation

Generally, cluster analysis is about finding groups in a set of objects $\mathcal{D} = \{x_1, \ldots, x_n\}$. There is much literature in which the objects $x_1, \ldots, x_n$ are assumed to be from the Euclidean space $\mathbb{R}^p$, but in principle they could be from any space $\mathcal{X}$.

A clustering is a set $\mathcal{C} = \{C_1, \ldots, C_K\}$ with $C_j \subseteq \mathcal{D}$, $j = 1, \ldots, K$. The number of clusters K may be fixed in advance or not. For $j = 1, \ldots, K$, let $n_j = |C_j|$ be the number of objects in C_j. Obviously not every such $\mathcal{C}$ qualifies as a "good" or "useful" clustering, but what is demanded of $\mathcal{C}$ differs in the different approaches of cluster analysis. Here, $\mathcal{C}$ is required to be a partition, e.g. $j \neq k \Rightarrow C_j \cap C_k = \emptyset$ and $\bigcup_{j=1}^{K} C_j = \mathcal{D}$. For partitions, let $\gamma : \{1, \ldots, n\} \mapsto \{1, \ldots, K\}$ be the assignment function, i.e. $\gamma(i) = j$ if $x_i \in C_j$. Some of the indexes introduced below could also by applied to clusterings that are not partitions (particularly objects that are not a member of any cluster could just be ignored), but this is not treated here to keep things simple. Clusters are here also assumed to be crisp rather than fuzzy, i.e. an object is either a full member of a cluster or not a member of this cluster at all. In case of probabilistic clusterings, which give as output probabilities p_{ij} for object i to be member of cluster

j, it is assumed that objects are assigned to the cluster j maximizing p_{ij}; in case of hierarchical clusterings, it is assumed that the hierarchy is cut at a certain number of clusters K to obtain a partition.

Most of the methods introduced here are based on dissimilarity data. A dissimilarity is a function $d : \mathcal{X}^2 \mapsto \mathbb{R}_0^+$ so that $d(x,y) = d(y,x) \geq 0$ and $d(x,x) = 0$ for $x, y \in \mathcal{X}$. Many dissimilarities are distances, i.e. they also fulfill the triangle inequality, but this is not necessarily required here. Dissimilarities are extremely flexible. They can be defined for all kinds of data, such as functions, time series, categorical data, image data, text data, etc. If data are Euclidean, obviously the Euclidean distance can be used. See Hennig (2016) for a more general overview of dissimilarity measures used in cluster analysis.

1.3. Aspects of cluster validity

In this section, I introduce measurements for various aspects of cluster validity.

1.3.1. *Small within-cluster dissimilarities*

A major aim in most cluster analysis applications is to find homogeneous clusters. This often means that all the objects in a cluster should be very similar to each other, although it can in principle also have different meanings, for example that a homogeneous probability model (such as the Gaussian distribution, potentially with large variance) can account for all observations in a cluster.

The most straightforward way to formalize that all objects within a cluster should be similar to each other is the average within-cluster distance:

$$I_{withindis}(\mathcal{C}) = \frac{2}{\sum_{j=1}^{K} n_j(n_j-1)} \sum_{j=1}^{K} \sum_{x \neq y \in C_j} d(x,y).$$

Smaller values are better. Knowing the data but not the clustering, the minimum possible value of $I_{withindis}$ is zero and the maximum is $d_{max} = \max_{x,y \in \mathcal{D}} d(x,y)$, so $I^*_{withindis}(\mathcal{C}) = 1 - \frac{I_{withindis}(\mathcal{C})}{d_{max}} \in [0,1]$ is a normalized version. When different criteria are aggregated (see section 1.4), it is useful to define them in such a way that they point in the same direction; I will define all normalized indexes so that larger values are better. For this reason, $\frac{I_{withindis}(\mathcal{C})}{d_{max}}$ is subtracted from 1.

There are alternative ways of measuring whether within-cluster dissimilarities are overall small. All of these operationalize cluster homogeneity in slightly different ways. The objective function of K-means clustering can be written as a constant

times the average of all squared within-cluster Euclidean distances (or more general dissimilarities), which is an alternative measure, giving more emphasis to the biggest within-cluster dissimilarities. Most radically, one could use the maximum within-cluster dissimilarity. On the other hand, one could use quantiles or trimmed means in order to make the index less sensitive to large within-cluster dissimilarities, although I believe that in most applications in which within-cluster similarity is important, these should be avoided and the index should therefore be sensitive against them.

1.3.2. *Between-cluster separation*

Apart from within-cluster homogeneity, the separation between clusters is most often taken into account in the literature on cluster validation (most univariate indexes balance separation against homogeneity in various ways). Separation as it is usually understood cannot be measured by averaging all between-cluster dissimilarities, because it refers to what goes on "between" the clusters, that is, the smallest between-cluster dissimilarities, whereas the dissimilarities between pairs of farthest objects from different clusters should not contribute to this.

The most naive way to measure separation is to use the minimum between-cluster dissimilarity. This has the disadvantage that with more than two clusters it only looks at the two closest clusters, and also in many applications there may be an inclination to tolerate the odd very small distance between clusters if by and large the closest points of the clusters are well separated.

I propose here an index that takes into account a portion p, say $p = 0.1$, of objects in each cluster that are closest to another cluster.

For every object $x_i \in C_j$, $i = 1, \ldots, n$, $j \in \{1, \ldots, K\}$, let $d_{j:i} = \min_{y \notin C_j} d(x_i, y)$. Let $d_{j:(i)} \leq \ldots \leq d_{j:(n_j)}$ be the values of $d_{j:i}$ for $x_i \in C_j$ ordered from the smallest to the largest, and let $\lfloor pn_j \rfloor$ be the largest integer $\leq pn_j$. Then the p-separation index is defined as

$$I_{p-sep}(\mathcal{C}) = \frac{1}{\sum_{j=1}^{K} \lfloor pn_j \rfloor} \sum_{j=1}^{K} \sum_{i=1}^{\lfloor pn_j \rfloor} d_{j:(i)}.$$

Obviously, $I_{p-sep}(\mathcal{C}) \in [0, d_{max}]$ and large values are good, therefore $I^*_{p-sep}(\mathcal{C}) = \frac{I_{p-sep}(\mathcal{C})}{d_{max}} \in [0, 1]$ is a suitable normalization.

1.3.3. *Representation of objects by centroids*

In some applications clusters are used for information reduction, and one way of doing this is to use the cluster centroids for further analysis rather than the full data

set. It is then relevant to measure how well the observations in a cluster are represented by the cluster centroid. The most straightforward method to measure this is to average the dissimilarities of all objects to the centroid of the cluster they are assigned to. Let $c_1, \ldots, c_K$ be the centroids of clusters $C_1, \ldots, C_K$. Then,

$$I_{centroid}(\mathcal{C}) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, c_{\gamma(i)}).$$

Some clustering methods such as K-means and partitioning around medoids (PAM; see Kaufman and Rousseeuw 1990) are centroid based, that is, they compute the cluster centroids along with the clusters. Centroids can also be defined for the output of non-centroid-based methods, most easily as

$$c_j = \arg\min_{x \in C_j} \sum_{\gamma(i)=j} d(x_i, x),$$

which corresponds to the definition of PAM. Again, there are possible variations. K-means uses squared Euclidean distances, and in case of Euclidean data the cluster centroids do not necessarily have to be members of $\mathcal{D}$, they could also be mean vectors of the observations in the clusters.

Again, by definition, $I_{centroid}(\mathcal{C}) \in [0, d_{max}]$. Small values are better, and therefore $I^*_{centroid}(\mathcal{C}) = 1 - \frac{I_{centroid}(\mathcal{C})}{d_{max}} \in [0, 1]$.

1.3.4. *Representation of dissimilarity structure by clustering*

Another way in which the clustering can be used for information reduction is that the clustering can be seen as a more simple summary or representation of the dissimilarity structure. This can be measured by correlating the vector of pairwise dissimilarities $\mathbf{d} = \text{vec}\left([d(x_i, x_j)]_{i<j}\right)$ with the vector of a "clustering induced dissimilarity" $\mathbf{c} = \text{vec}\left([c_{ij}]_{i<j}\right)$, where $c_{ij} = 1(\gamma(i) \neq \gamma(j))$, and $1(\bullet)$ denotes the indicator function. With r denoting the sample Pearson correlation,

$$I_{Pearson\Gamma}(\mathcal{C}) = r(\mathbf{d}, \mathbf{c}).$$

This index goes back to (Hubert and Schultz 1976), see also (Halkidi *et al.* 2016) for alternative versions. $I_{Pearson\Gamma} \in [-1, 1]$, and large values are good, so it can be normalized by $I^*_{Pearson\Gamma} = \frac{I_{Pearson\Gamma}+1}{2} \in [0, 1]$.

1.3.5. *Small within-cluster gaps*

The idea that a cluster should be homogeneous can mean that there are no "gaps" within a cluster, and that the cluster is well connected. A gap can be characterized as a split of a cluster into two subclusters so that the minimum dissimilarity between the two subclusters is large. The corresponding index measures the "length" (dissimilarity) of the widest within-cluster gap (an alternative would be to average widest gaps over clusters):

$$I_{widestgap}(\mathcal{C}) = \max_{C\in\mathcal{C},D,E:\ C=D\cup E}\ \min_{x\in D,y\in E} d(x,y).$$

$I_{widestgap} \in [0, d_{max}]$ and low values are good, so it is normalized as $I^*_{widestgap} = 1 - \frac{I_{widestgap}}{d_{max}} \in [0,1]$.

A version of this taking into account density values is defined in section 1.3.6. Widest gaps can be found computationally by constructing the within-cluster minimum spanning trees; the widest distance occurring there is the widest gap.

1.3.6. *Density modes and valleys*

A very popular idea of a cluster is that it corresponds to a density mode, and that the density within a cluster goes down from the cluster mode to the outer regions of the cluster. Correspondingly, there should be density valleys between different clusters.

The definition of indexes that measure such a behavior is based on a density function h that assigns a density value $h(x)$ to every observation. For Euclidean data, standard density estimators such as kernel density estimators can be used. For general dissimilarities, I here propose a simple kernel density estimator. Let $q_{d,p}$ be the p-quantile of the vector of dissimilarities $\mathbf{d}$, e.g., for $p = 0.1$, the 10% smallest dissimilarities are $\leq q_{d,0,1}$. Define the kernel and density as

$$k(d) = \left(1 - \frac{1}{q_{d,p}}d\right) 1(d \leq q_{d,p}), \qquad h(x) = \sum_{i=1}^{n} k(d(x,x_i)).$$

These can be normalized to take a maximum of 1:

$$h^*(x) = \frac{h(x)}{\max_{y\in\mathcal{D}} h(y)}.$$

Alternatively, $h_{k-nn}(x) = \frac{1}{d^k(x)}$ with $d^k(x)$ being the dissimilarity to the kth nearest neighbor would be another simple dissimilarity-based density estimator, although this has no trivial upper bound (h, even before normalizing by its

within-cluster maximum, is bounded by n). One could also standardize h by the within-cluster maxima if clusters with generally lower densities should have the same weight as high-density clusters, but lower density values rely on fewer observations and are therefore less reliable.

Three different aspects of density-based clustering are measured by three different indexes:

1) The density should decrease within a cluster from the density mode to the "outskirts" of the cluster ($I_{densdec}$).

2) Cluster boundaries should run through density "valleys", that is, high-density points should not be close to many points from other clusters ($I_{densbound}$).

3) There should not be a big gap between high-density regions within a cluster ($I_{highdgap}$; gaps as measured by $I_{widestgap}$ may be fine in the low-density outskirts of a cluster).

The idea for $I_{densdec}$ is as follows. For every cluster, starting from the cluster mode, i.e. the observation with the highest density, construct a growing sequence of observations that eventually covers the whole cluster by always adding the closest observation that is not yet included. Optimally, in this process, the within-cluster density of newly included points should always decrease. Whenever actually the density goes up, a penalty of the squared difference of the densities is incurred. The index $I_{densdec}$ aggregates these penalties. The following algorithm computes this, and it also constructs a set T that collects information about high dissimilarities between high-density observations and is used for the definition of $I_{highdgap}$ in the following:

Initialization $I_{d1} = 0$, $T = \emptyset$. For $j = 1, \ldots, K$:

Step 1 $S_j = \{x\}$, where $x = \arg\max_{y \in C_j} h^*(y)$.

Step 2 Let $R_j = C_j \setminus S_j$. If $R_j = \emptyset$: $j = j + 1$, if $j \leq K$ go to step 1, if $j = K + 1$ then go to step 5. Otherwise:

Step 3 Find $(x, y) = \arg\min_{(z_1,z_2): z_1 \in R_j, z_2 \in S_j} d(z_1, z_2)$. $S_j = S_j \cup \{x\}$, $T = T \cup \{\max_{z \in R_j} h^*(z) d(x, y)\}$.

Step 4 If $h^*(x) > h^*(y)$: $I_{d1} = I_{d1} + (h^*(x) - h^*(y))^2$. In any case, back to step 2.

Step 5 $I_{densdec}(\mathcal{C}) = \sqrt{\frac{I_{d1}}{n}}$.

$I_{densdec}$ collects the penalties from increases in the within-cluster densities during this process.

The definition of $I_{densdec}$ does not take into account whether the neighboring observations that produce high-density values $h^*(x)$ for x are in the same cluster as x. But this is important, because it would otherwise be easy to achieve a good value of $I_{densdec}$ by cutting through high-density areas and distributing a single high-density area to several clusters.

A second index can be defined that penalizes a high contribution of points from different clusters to the density values in a cluster (measured by h_o below), because this means that the cluster border cuts through a high-density region.

$$\text{For } x_i,\ i = 1, \ldots, n:\ h_o(x_i) = \sum_{k=1}^{n} k(d(x_i, x_k))1(\gamma(k) \neq \gamma(i)).$$

Normalizing:

$$h_o^*(x) = \frac{h_o(x)}{\max_{y \in \mathcal{D}} h(y)}.$$

A penalty is incurred if for observations with a large-density $h^*(x)$, there is a large contribution $h_o^*(x)$ to that density from other clusters:

$$I_{densbound}(\mathcal{C}) = \frac{1}{n} \sum_{j=1}^{K} \sum_{x \in C_j} h^*(x) h_o^*(x).$$

Both $I_{densdec}$ and $I_{densbound}$ are by definition ≥ 0. Also, the maximum contribution of any observation to any of $I_{densdec}$ and $I_{densbound}$ is $\frac{1}{n}$, because the normalized h^*-values are ≤ 1. These are penalties, so low values are good, and normalized versions are defined as

$$I_{densdec}^*(\mathcal{C}) = 1 - I_{densdec}(\mathcal{C}),\ I_{densbound}^*(\mathcal{C}) = 1 - I_{densbound}(\mathcal{C}).$$

An issue with $I_{densdec}$ is that it is possible that there is a large gap between two observations with high density, which does not incur penalties if there are no low-density observations in between. This can be picked up by a version of $I_{widestgap}$ based on the density-weighted gap information collected in T above. This is suggested instead of $I_{widestgap}$ if a density-based cluster concept is of interest:

$$I_{highdgap}(\mathcal{C}) = \max T.$$

$I_{highdgap}(\mathcal{C}) \in [0, d_{max}]$ and low values are good, so it is normalized as $I_{highdgap}^*(\mathcal{C}) = 1 - \frac{I_{highdgap}(\mathcal{C})}{d_{max}} \in [0, 1]$.

1.3.7. *Uniform within-cluster density*

Sometimes different clusters should not (only) be characterized by gaps between them; overlapping regions in data space may be seen as different clusters if they have different within-cluster density levels, which in some applications could point to different data generating mechanisms behind the different clusters, which the researcher would like to discover. Such a cluster concept would require that densities within clusters are more or less uniform.

This can be characterized by the coefficient of variation CV of either the within-cluster density values or the dissimilarities to the kth nearest within-cluster neighbor $d_w^k(x)$ (say $k = 4$). The latter is preferred here because, as opposed to the density values, $d_w^k(x)$ is clean from the influence of observations from the other clusters. Define for $j = 1, \ldots, k$, assuming $n_j > k$:

$$m(C_j; k) = \frac{1}{n_j} \sum_{x \in C_j} d_w^k(x), \qquad \mathrm{CV}(C_j) = \frac{\sqrt{\frac{1}{n_j - 1} \sum_{x \in C_j} (d_w^k(x) - m(C_j; k))^2}}{m(C_j; k)}.$$

Using this,

$$I_{cvdens}(\mathcal{C}) = \frac{\sum_{j=1}^{K} n_j \mathrm{CV}(C_j) 1(n_j > k)}{\sum_{j=1}^{K} n_j 1(n_j > k)}.$$

Low values are good. The maximum value of the coefficient of variation based on n observations is $\sqrt{n}$ Katsnelson and Kotz (1957), so a normalized version is $I^*_{cvdens}(\mathcal{C}) = 1 - \frac{I_{cvdens}(\mathcal{C})}{\sqrt{n}}$.

1.3.8. *Entropy*

In some clustering applications, particularly where clustering is done for "organizational" reasons such as information compression, it is useful to have clusters that are roughly of the same size. This can be measured by the entropy:

$$I_{entropy}(\mathcal{C}) = -\sum_{j=1}^{K} \frac{n_j}{n} \log\left(\frac{n_j}{n}\right).$$

Large values are good. The entropy is maximized for fixed K by $e_{max}(K) = -\log\left(\frac{1}{K}\right)$, so it can be normalized by $I^*_{entropy}(\mathcal{C}) = \frac{I_{entropy}(\mathcal{C})}{e_{max}(K)}$.

1.3.9. *Parsimony*

In case that there is a preference for a lower number of clusters, one could simply define

$$I^*_{parsimony} = 1 - \frac{K}{K_{max}},$$

(already normalized) with K_{max} the maximum number of clusters of interest. If in a given application there is a known nonlinear loss connected to the number of clusters, this can obviously be used instead, and the principle can be applied also to other free parameters of a clustering method, if desired.

1.3.10. *Similarity to homogeneous distributional shapes*

Sometimes the meaning of "homogeneity" for a cluster is defined by a homogeneous probability model, for example, Gaussian mixture model based clustering models all clusters by Gaussian distributions with different parameters, requiring Euclidean data. Historically, due to the Central Limit Theorem and Quetelet's "elementary error hypothesis", measurement errors were widely believed to be normally/Gaussian distributed (see Stigler 1986). Under such a hypothesis, it makes sense in some situations to regard Gaussian distributed observations as homogeneous, and as pointing to the same underlying mechanism; this could also motivate to cluster observations together that look like being generated from the same (approximate) Gaussian distribution. Indexes that measure cluster-wise Gaussianity can be defined (e.g. Lago-Fernandez and Corbacho 2010). One possible principle is to compare a one-dimensional function of the observations within a cluster to its theoretical distribution under the data distribution of interest, e.g. (Coretto and Hennig 2016) compare the Mahalanobis distances of observations to their cluster center with their theoretical χ^2-distribution using the Kolmogorow distance. This is also possible for other distributions of interest.

1.3.11. *Stability*

Clusterings are often interpreted as meaningful in the sense that they can be generalized as substantive patterns. This at least implicitly requires that they are stable. Stability in cluster analysis can be explored using resampling techniques such as bootstrap and splitting the data set, and clustering from different resampled data sets can be compared. This requires to run the clustering method again on the resampled data sets and I will not treat this here in detail, but useful indexes have been defined using this principle (e.g. Tibshirani and Walther 2005, Fang and Wang 2012).

1.3.12. *Further Aspects*

Hennig (2016) lists further potentially desirable characteristics of a clustering, for which further indexes could be defined:

– areas in data space corresponding to clusters should have certain characteristics such as being linear or convex;

– it should be possible to characterize clusters using a small number of variables;

– clusters should correspond well to an externally given partition or values of an external variable (this could, for example, imply that clusters of regions should be spatially connected);

– variables should be approximately independent within clusters.

1.4. Aggregation of indexes

The required cluster concept and therefore the way the validation indexes can be used depends on the specific clustering application. The users need to specify what characteristics of the clustering are desired in the application. The corresponding indexes can then be aggregated to form a single criterion that can be used to compare different clustering methods, different numbers of clusters and other possible parameter choices of the clustering.

The most straightforward aggregation is to compute a weighted mean of s selected indexes $I_1, \ldots, I_s$ with weights $w_1, \ldots, w_s > 0$ expressing the relative importance of the different methods:

$$A(\mathcal{C}) = \sum_{k=1}^{s} w_k I_k. \qquad [1.1]$$

Assuming that large values are desirable for all of $I_1, \ldots, I_s$, the best clustering for the application in question can be found by maximizing A. This can be done by comparing different clusterings from conventional clustering methods, but in principle it would also be an option to try to optimize A directly.

The weights can only be chosen to directly reflect the relative importance of the various aspects of a clustering if the values (or, more precisely, their variations) of the indexes $I_1, \ldots, I_s$ are comparable, and give the indexes equal influence on A if all weights are equal. In section 1.3, I proposed tentative normalizations of all indexes, which give all indexes the same value range $[0, 1]$. Unfortunately, this is not good enough to ensure comparability; on many data sets some of these indexes will cover almost the whole value range, whereas other indexes may be larger than 0.9 for all

clusterings that any clustering method would come up with. Therefore, section 1.5 will introduce a new computational method to standardize the variation of the different criteria.

Another issue is that some indexes by their very nature favor large numbers of clusters K (obviously large within-cluster dissimilarities can be more easily avoided for large K), whereas others favor small values of K (separation is more difficult to achieve with many small clusters). The method introduced in section 1.5 will allow to assess the extent to which the indexes deliver systematically larger or smaller values for larger K. Note that this can also be an issue for univariate "global" validation indexes from the literature (see Hennig and Lin 2015).

If the indexes are to be used to find an optimal value of K, the indexes in A should be chosen in such a way that indexes that systematically favor larger K and indexes that systematically favor smaller K are balanced.

The user needs to take into account that the proposed indexes are not independent. For example, good representation of objects by centroids will normally be correlated with having generally small within-cluster dissimilarities. Including both indexes will assign extra weight to the information that the two indexes have in common (which may sometimes but not always be desired).

There are alternative ways to aggregate the information from the different indexes. For example, one could use some indexes as side conditions rather than involving them in the definition of A. For example, rather than giving entropy a weight for aggregation as part of A, one may specify a certain minimum entropy value below that clusterings are not accepted, but not use the entropy value to distinguish between clusterings that fulfill the minimum entropy requirement. Multiplicative aggregation is another option.

1.5. Random clusterings for calibrating indexes

As explained above, the normalization in section 1.3 does not provide a proper calibration of the validation indexes. Here is an idea for doing this in a more appropriate way. The idea is that random clusterings are generated on $\mathcal{D}$ and index values are computed, in order to explore what range of index values can be expected on $\mathcal{D}$, so that the clusterings of interest can be compared to these. So in this section, as opposed to conventional probability modeling, the data set is considered as fixed but a distribution of index values is generated from various random partitions.

Completely random clusterings (i.e. assigning every observation independently to a cluster) are not suitable for this, because it can be expected that indexes formalizing desirable characteristics of a clustering will normally give much worse values for them than for clusters that were generated by a clustering method. Therefore, I

propose two methods for random clusterings that are meant to generate clusterings that make some sense, at least by being connected in data space. The methods are called "stupid K-centroids" and "stupid nearest neighbors"; "stupid" because they are versions of popular clustering methods (centroid-based clustering like K-means or PAM, and single linkage/nearest neighbor) that replace optimization by random decisions and are meant to be computable very quickly. Centroid-based clustering normally produces somewhat compact clusters, whereas single linkage is notorious for prioritizing cluster separation totally over within-cluster homogeneity, and therefore one should expect these two approaches to explore in a certain sense opposite ways of clustering the data.

1.5.1. *Stupid K-centroids clustering*

Stupid K-centroids works as follows. For fixed number of cluster K draw a set of K cluster centroids $Q = \{q_1, \ldots, q_K\}$ from $\mathcal{D}$ so that every subset of size K has the same probability of being drawn. $\mathcal{C}_{K-stupidcent}(Q) = \{C_1, \ldots, C_k\}$ is defined by assigning every observation to the closest centroid:

$$\gamma(i) = \underset{j\in\{1,\ldots,K\}}{\arg\min}\ d(x_i, q_j),\ i = 1, \ldots, n.$$

1.5.2. *Stupid nearest neighbors clustering*

Again, for fixed number of cluster K draw a set of K cluster initialization points $Q = \{q_1, \ldots, q_K\}$ from $\mathcal{D}$ so that every subset of size K has the same probability of being drawn. $\mathcal{C}_{K-stupidnn}(Q) = \{C_1, \ldots, C_k\}$ is defined by successively adding the not yet assigned observation closest to any cluster to that cluster until all observations are clustered:

Initialization Let $Q^* = Q$. Let

$$\mathcal{C}^*(Q) = \mathcal{C}^*(Q^*) = \{C_1^*, \ldots, C_L^*\} = \{\{q_1\}, \ldots, \{q_K\}\}.$$

Step 1 Let $R^* = \mathcal{D} \setminus Q^*$. If $R^* \neq \emptyset$, find $(x, y) = \underset{(z,q):z\in R^*, q\in Q^*}{\arg\min}\ d(z, q)$, otherwise stop.

Step 2 Let $Q^* = Q^* \cup \{x\}$. For the $C^* \in \mathcal{C}^*(Q^*)$ with $y \in C^*$, let $C^* = C^* \cup \{x\}$, updating $\mathcal{C}^*(Q^*)$ accordingly. Go back to step 1.

At the end, $\mathcal{C}_{K-stupidnn}(Q) = \mathcal{C}^*(Q^*)$.

1.5.3. *Calibration*

The random clusterings can be used in various ways to calibrate the indexes. For any value K of interest, $2B$ clusterings:

$$\begin{aligned}\mathcal{C}_{K-collection} &= (\mathcal{C}_{K:1},\ldots,\mathcal{C}_{K:2B}) \\ &= (\mathcal{C}_{K-stupidcent}(Q_1),\ldots,\mathcal{C}_{K-stupidcent}(Q_B), \\ &\quad \mathcal{C}_{K-stupidnn}(Q_1),\ldots,\mathcal{C}_{K-stupidnn}(Q_B))\end{aligned}$$

on $\mathcal{D}$ are generated, say $B = 100$.

As mentioned before, indexes may systematically change over K and therefore may show a preference for either large or small K. In order to account for this, it is possible to calibrate the indexes using stupid clusterings for the same K, i.e. for a clustering $\mathcal{C}$ with $|\mathcal{C}| = K$. Consider an index I^* of interest (the normalized version is used here because this means that large values are good for all indexes). Then,

$$I^{cK}(\mathcal{C}) = \frac{I^*(\mathcal{C}) - m^*(\mathcal{C}_{K-collection})}{\sqrt{\frac{1}{2B-1}\sum_{j=1}^{2B}\left(I^*(\mathcal{C}_{K:j}) - m^*(\mathcal{C}_{K-collection})\right)^2}}, \qquad [1.2]$$

where $m^*(\mathcal{C}_{K-collection}) = \frac{1}{2B}\sum_{j=1}^{2B} I^*(\mathcal{C}_{K:j})$. A desired set of calibrated indexes can then be used for aggregation in [1.1].

An important alternative to [1.2] is calibration by using random clusterings for all values of K together. Let $\mathcal{K} = \{2,\ldots,K_{max}\}$ be the numbers of clusters of interest (most indexes will not work for $K = 1$), $\mathcal{C}_{collection} = \{C_{K:j} : K \in \mathcal{K}, j = 1,\ldots,2B\}$, $m^*(\mathcal{C}_{collection}) = \frac{1}{2B(K_{max}-1)}\sum_{K=2}^{K_{max}}\sum_{j=1}^{2B} I^*(\mathcal{C}_{K:j})$. With this,

$$I^c(\mathcal{C}) = \frac{I^*(\mathcal{C}) - m^*(\mathcal{C}_{collection})}{\sqrt{\frac{1}{2B(K_{max}-1)-1}\sum_{K=2}^{K_{max}}\sum_{j=1}^{2B}\left(I^*(\mathcal{C}_{K:j}) - m^*(\mathcal{C}_{collection})\right)^2}}. \qquad [1.3]$$

I^c does not correct for potential systematic tendencies of the indexes over $\mathcal{K}$, but this is not a problem if the user is happy to use the uncalibrated indexes directly for comparing different values of K; a potential bias toward large or small values of K in this case needs to be addressed by choosing the indexes to be aggregated in [1.1] in a balanced way. This can be checked by computing the aggregated index A also for the random clusterings and check how these change over the different values of K.

Another alternative is to calibrate indexes by using their rank value in the set of clusterings (random clusterings and clusterings to compare) rather than a mean/standard deviation based standardization. This is probably more robust but comes with some loss of information.

1.6. Examples

1.6.1. *Artificial data set*

The first example is the artificial data set shown in Figure 1.1. Four clusterings are compared (actually many more clusterings with K between 2 and 5 were compared on these data, but the selected clusterings illustrate the most interesting issues).

The clusterings were computed by K-means with $K = 2$ and $K = 3$, single linkage cut at $K = 3$ and PAM with $K = 5$. The K-means clustering with $K = 3$ and the single linkage clustering are shown in Figure 1.1. The K-means clustering with $K = 2$ puts the uniformly distributed widespread point cloud on top together in a single cluster, and the two smaller populations are the second cluster. This is the most intuitive clustering for these data for $K = 2$ and also delivered by most other clustering methods. PAM does not separate the two smaller (actually Gaussian) populations for $K = 2$, but it does so for $K = 5$, along with splitting the uniform point cloud into three parts.

	K-means-2	**K-means-3**	**Single linkage-3**	**PAM-5**
$I^*_{withindis}$	0.654	0.799	0.643	0.836
$I^*_{0.1-sep}$	0.400	0.164	0.330	0.080
$I^*_{centroid}$	0.766	0.850	0.790	0.896
$I^*_{PearsonΓ}$	0.830	0.900	0.781	0.837
$I^*_{widestgap}$	0.873	0.873	0.901	0.901
$I^*_{densdec}$	0.977	0.981	0.981	0.985
$I^*_{densbound}$	1.000	0.999	1.000	0.997
$I^*_{highdgap}$	0.879	0.879	0.960	0.964
I^*_{cvdens}	0.961	0.960	0.961	0.959
$I^*_{entropy}$	0.863	0.993	0.725	0.967

Table 1.1. *Normalized index values for four clusterings on artificial data*

Table 1.1 shows the normalized index values for these clusterings. Particularly comparing 3-means and single linkage, the different virtues of these clusterings are clear to see. 3-means is particularly better for the homogeneity-driven $I^*_{withindis}$ and $I^*_{centroid}$, whereas single linkage wins regarding the separation-oriented $I^*_{0.1-sep}$ and $I^*_{widestgap}$, with 3-means ignoring the gap between the two Gaussian populations. $I^*_{PearsonΓ}$ tends toward 3-means, too, which was perhaps less obvious, because it does not like too big distances within clusters. It is also preferred by $I^*_{entropy}$ because of

joining two subpopulations that are rather small. The values for the indexes, $I^*_{densdec}$, $I^*_{densbound}$, $I^*_{highdgap}$ and I^*_{cvdens}, illustrate that the naive normalization is not quite suitable for making the value ranges of the indexes comparable. For the density-based indexes, many involved terms are far away from the maximum used for normalization, so the index values can be close to 0 (close to 1 after normalization). This is amended by calibration.

Considering the clusterings with $K = 2$ and $K = 5$, it can be seen that with $K = 5$ it is easier to achieve within-cluster homogeneity ($I^*_{withindis}$, $I^*_{centroid}$), whereas with $K = 2$ it is easier to achieve separation ($I^*_{0.1-sep}$).

	K-means-2	**K-means-3**	**Single linkage-3**	**PAM-5**
$I^{cK}_{withindis}$	0.906	1.837	−0.482	0.915
$I^{cK}_{0.1-sep}$	1.567	0.646	3.170	−0.514
$I^{cK}_{centroid}$	1.167	1.599	0.248	1.199
$I^{cK}_{PearsonΓ}$	1.083	1.506	0.099	0.470
$I^{cK}_{widestgap}$	1.573	1.156	1.364	0.718
$I^{cK}_{densdec}$	1.080	1.191	1.005	1.103
$I^{cK}_{densbound}$	0.452	0.449	0.519	0.647
$I^{cK}_{highdgap}$	1.317	0.428	2.043	1.496
I^{cK}_{cvdens}	1.153	0.836	0.891	0.286
$I^{cK}_{entropy}$	0.246	1.071	−0.620	0.986

Table 1.2. *Calibrated index values (using random clusterings with same K) for four clusterings on artificial data*

Table 1.2 shows the index values I^{cK} calibrated against random clustering with the same K. This is meant to account for the fact that some indexes differ systematically over different values of K. Indeed, using this calibration, PAM with $K = 5$ is no longer optimal for $I^{cK}_{centroid}$ and $I^{cK}_{withindis}$, and 2-means is no longer optimal for $I^{cK}_{0.1-sep}$. It can now be seen that 3-means is better than single linkage for $I^{cK}_{densdec}$. This is because density values show much more variation in the widely spread uniform subpopulation than in the two small Gaussian ones, so splitting up the uniform subpopulation is better for creating densities decreasing from the modes, despite the gap between the two Gaussian subpopulations. On the other hand, 3-means has to cut through the uniform population, which gives single linkage, which only cuts through clear gaps, an advantage regarding $I^{cK}_{densbound}$, and particularly 3-means incurs a large distance between the two Gaussian high-density subsets within one of its clusters, which makes single linkage much better regarding $I^{cK}_{highdgap}$. Ultimately, the user needs to decide here whether small within-cluster dissimilarities and short dissimilarities to centroids are more important than separation and the absence of within-cluster gaps. The $K = 5$-solution does not look very attractive regarding most criteria (although calibration with the same K makes it

look good regarding $I^{cK}_{densbound}$); the $K = 2$-solution only looks good regarding two criteria that may not be seen as the most important ones here.

	K-means-2	K-means-3	Single linkage-3	PAM-5
$I^c_{withindis}$	−0.483	1.256	−0.607	1.694
$I^c_{0.1-sep}$	2.944	0.401	2.189	−0.512
$I^c_{centroid}$	−0.449	0.944	−0.059	1.712
$I^c_{PearsonΓ}$	0.658	1.515	0.058	0.743
$I^c_{widestgap}$	0.939	0.939	1.145	1.145
$I^c_{densdec}$	−0.279	0.832	0.697	1.892
$I^c_{densbound}$	0.614	0.551	0.609	0.417
$I^c_{highdgap}$	0.464	0.464	1.954	2.025
I^c_{cvdens}	0.761	0.692	0.748	0.615
$I^c_{entropy}$	0.208	1.079	−0.720	0.904

Table 1.3. *Calibrated index values (using all random clusterings) for four clusterings on artificial data*

Table 1.3 shows the index values I^{cK} calibrated against all random clusterings. Not much changes regarding the comparison of 3-means and single linkage, whereas a user who is interested in small within-cluster dissimilarities and centroid representation in absolute terms is now drawn toward PAM with $K = 5$ or even much larger K, indicating that these indexes should not be used without some kind of counterbalance, either from separation-based criteria ($I^c_{0.1-sep}$ and $I^c_{densbound}$) or taking into account parsimony. A high-density gap within a cluster is most easily avoided with large K, too, whereas $K = 2$ achieves the best separation, unsurprisingly.

As this is an artificial data set and there is no subject-matter information that could be used to prefer certain indexes, I do not present specific aggregation weights here.

1.6.2. *Tetragonula bees data*

Franck *et al.* (2004) published a data set giving genetic information about 236 Australasian tetragonula bees, in which it is of interest to determine the number of species. The data set is incorporated in the package "fpc" of the software system R (www.r-project.org) and is available on the IFCS Cluster Benchmark Data Repository http://ifcs.boku.ac.at/repository. Bowcock *et al.* (1994) defined the "shared allele dissimilarity" formalizing genetic dissimilarity appropriately for species delimitation, which is used for the present data set. It yields values in $[0, 1]$. See also Hausdorf and Hennig (2010) and Hennig (2013) for earlier analyses of this data set including a discussion of the number of clusters problem. Franck *et al.* (2004) provide nine "true" species for these data, although this manual classification (using morphological information besides genetics) comes with its own problems and may not be 100% reliable.

In order to select indexes and to find weights, some knowledge about species delimitation is required, which was provided by Bernhard Hausdorf, Museum of Zoology, University of Hamburg. The biological species concept requires that there is no (or almost no) genetic exchange between different species, so that separation is a key feature for clusters that are to be interpreted as species. For the same reason, large within-cluster gaps can hardly be tolerated (regardless of the density values associated with them); in such a case one would consider the subpopulations on two sides of a gap separate species, unless a case can be made that potentially existing connecting individuals could not be sampled. Gaps may also occur in regionally separated subspecies, but this cannot be detected from the data without regional information. On the other hand, species should be reasonably homogeneous; it would be against biological intuition to have strongly different genetic patterns within the same species. This points to the indexes $I_{withindis}$, $I_{0.1-sep}$ and $I_{widestgap}$. The shape of the within-cluster density is not a concern here, and neither are representation of clusters by centroids, entropy and constant within-cluster variation. The index $I_{PearsonΓ}$ is added to the set of relevant indexes, because one can interpret the species concept as a representation of genetic exchange as formalized by the shared allele dissimilarity, and $I_{PearsonΓ}$ measures the quality of this representation. All these four indexes are used in [1.1] with weight 1 (one could be interested in stability as well, which is not taken into account here).

	AL-5	**AL-9**	**AL-10**	**AL-12**	**PAM-5**	**PAM-9**	**PAM-10**	**PAM-12**
$I^{cK}_{withindis}$	0.68	−0.04	1.70	1.60	1.83	2.45	2.03	1.80
$I^{cK}_{0.1-sep}$	1.79	2.35	2.00	2.42	0.43	1.59	2.12	0.94
$I^{cK}_{PearsonΓ}$	1.86	2.05	1.92	2.28	1.43	1.84	1.75	0.61
$I^{cK}_{widestgap}$	0.45	4.73	4.90	4.86	−1.03	0.41	0.42	−0.09
$A(\mathcal{C})$	4.78	9.09	10.51	11.13	2.66	6.30	6.32	3.30
ARI	0.53	0.60	0.95	0.94	0.68	0.84	0.85	0.64

Table 1.4. *Calibrated index values (using random clusterings with same K) for eight clusterings on tetragonula bees data with aggregated index and adjusted Rand index*

Again I present a subset of the clusterings that were actually compared for illustrating the use of the approach presented in this paper. Typically, clusterings below $K = 9$ were substantially different from the ones with $K \geq 9$; clusterings with $K = 10$ and $K = 11$ from the same method were often rather similar to each other, and I present clusterings from Average Linkage and PAM with $K = 5, 9, 10$, and 12. Table 1.4 shows the four relevant index values I^{cK} calibrated against random clustering with the same K along with the aggregated index $A(\mathcal{C})$. Furthermore, the adjusted Rand index (ARI; (Hubert and Arabie 1985)) comparing the clusterings from the method with the "true" species is given (this takes values between -1 and 1 with 0 expected for random clusterings and 1 for perfect agreement). Note that despite $K = 9$ being the number of "true" species, clusterings with $K = 10$ and

$K = 12$ yield higher ARI values than those with $K = 9$, so these clusterings are preferable (it does not help much to estimate the number of species correctly if the species are badly composed). Some "true" species in the original data set are widely regionally dispersed with hardly any similarity between subspecies.

The aggregated index $A(\mathcal{C})$ is fairly well related to the ARI (over all 55 clusterings that were compared, the correlation between $A(\mathcal{C})$ and ARI is about 0.85). The two clusterings that are closest to the "true" one also have the highest values of $A(\mathcal{C})$. The within-cluster gap criterion plays a key role here, preferring average linkage with 9–12 clusters clearly over the other clusterings. $A(\mathcal{C})$ assigns its highest value to AL-12, whereas the ARI for AL-10 is very slightly higher. PAM delivers better clusterings regarding small within-cluster dissimilarities, but this advantage is dwarfed by the advantage of average linkage regarding separation and within-cluster gaps.

	AL-5	**AL-9**	**AL-10**	**AL-12**	**PAM-5**	**PAM-9**	**PAM-10**	**PAM-12**
$I^c_{withindis}$	0.10	0.59	1.95	2.00	0.83	2.13	2.17	2.16
$I^c_{0.1-sep}$	1.98	1.54	1.05	1.02	0.53	1.01	1.13	0.21
$I^c_{PearsonΓ}$	1.79	1.87	1.86	1.87	1.38	1.71	1.73	0.72
$I^c_{widestgap}$	0.39	5.08	5.08	5.08	−1.12	0.39	0.39	−0.08
$A(\mathcal{C})$	4.26	9.08	9.93	9.97	1.62	5.24	5.41	3.01
ARI	0.53	0.60	0.95	0.94	0.68	0.84	0.85	0.64

Table 1.5. *Calibrated index values (using all random clusterings) for eight clusterings on tetragonula bees data with aggregated index and adjusted Rand index*

Table 1.5 shows the corresponding results with calibration using all random clusterings. This does not result in a different ranking of the clusterings, so this data set does not give a clear hint which of the two calibration methods is more suitable, or, in other words, the results do not depend on which one is chosen.

1.7. Conclusion

The multivariate array of cluster validation indexes presented here provides the user with a detailed characterization of various relevant aspects of a clustering. The user can aggregate the indexes in a suitable way to find a useful clustering for the clustering aim at hand.

The indexes can also be used to provide a more detailed comparison of different clustering methods in benchmark studies, and a better understanding of their characteristics.

The methodology is currently partly implemented in the "fpc" package of the statistical software system R and will soon be fully implemented there.

Most indexes require $K \geq 2$ and the approach can therefore not directly be used for deciding whether the data set is homogeneous as a whole ($K = 1$). The individual indexes as well as the aggregated index could be used in a parametric bootstrap scheme as proposed by Hennig and Lin (2015) to test the homogeneity null hypothesis against a clustering alternative.

Research is still required in order to compare the different calibration methods and some alternative versions of indexes. A theoretical characterization of the indexes is of interest as well as a study exploring the strength of the information overlap between some of the indexes, looking at, e.g., correlations over various clusterings and data sets. Random clustering calibration may also be used together with traditional univariate validation indexes. Further methods for random clustering could be developed and it could be explored what collection of random clusterings is most suitable for calibration (some work in this direction is currently done by my PhD student Serhat Akhanli).

1.8. Acknowledgment

This work was supported by EPSRC Grant EP/K033972/1.

1.9. References

Bowcock, A.M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J.R., Cavalli-Sforza, L.L. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, 368, 455–457.

Calinski, T., Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3, 1–27,

Coretto, P., Hennig, C. (2016). Robust improper maximum likelihood: tuning, computation, and a comparison with other methods for robust Gaussian clustering. *Journal of the American Statistical Association* 111, 1648–1659.

Fang, Y., Wang, J. (2012). Selection of the number of clusters via the bootstrap method. *Computational Statistics and Data Analysis*, 56, 468–477.

Franck, P., Cameron, E., Good, G., Rasplus, J.-Y., Oldroyd, B.P. (2004). Nest architecture and genetic differentiation in a species complex of Australian stingless bees. *Molecular Ecology*, 13, 2317–2331.

Halkidi, M., Vazirgiannis, M., Hennig, C. (2016). Method-independent indices for cluster validation and estimating the number of clusters. In *Handbook of Cluster Analysis*, Hennig, C., Meila, M., Murtagh, F., Rocci, R. (eds), CRC/Chapman & Hall, Boca Raton. 595–618.

Hausdorf, B., Hennig, C. (2010). Species delimitation using dominant and codominant multilocus markers. *Systematic Biology*, 59, 491–503.

Hennig, C. (2013). How many bee species? A case study in determining the number of clusters. In *Data Analysis, Machine Learning and Knowledge Discovery*, Spiliopoulou, M., Schmidt-Thieme, L., Janning, R. (eds), Springer, Berlin, 41–49.

Hennig, C. (2015). What are the true clusters? *Pattern Recognition Letters*, 64, 53–62.

Hennig, C. (2016). Clustering strategy and method selection. In *Handbook of Cluster Analysis*, Hennig, C., Meila, M., Murtagh, F., Rocci, R. (eds), CRC/Chapman & Hall, Boca Raton, 703–730.

Hennig, C., Lin, C.-J. (2015). Flexible parametric bootstrap for testing homogeneity against clustering and assessing the number of clusters. *Statistics and Computing*, 25, 821–833.

Hubert, L.J., Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2, 193–218.

Hubert, L.J., Schultz, J. (1976). Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical and Statistical Psychology*, 29, 190–241.

Kaufman, L., Rousseeuw, P.J. (1990). *Finding Groups in Data*, Wiley, New York.

Katsnelson, J., Kotz, S. (1957). On the upper limits of some measures of variability. *Archiv für Meteorologie, Geophysik und Bioklimatologie, Series B*, 8, 103–107.

Lago-Fernandez, L.F., Corbacho, F. (2010). Normality-based validation for crisp clustering. *Pattern Recognition*, 43, 782–795.

Stigler, S. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press, Cambridge.

Tibshirani, R., Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14, 511–528.

2

Histogram-Based Clustering of Sensor Network Data

In this chapter, we assume that a sensor network is used for monitoring, over time, a physical phenomenon. Each sensor performs repeated measurements at a very high frequency so that it is not possible to store the whole amount of data into some easy to access media. We propose a clustering strategy that processes online the incoming observations in order to find groups of sensors that behave similarly over time. The proposed strategy is made by two phases: the online phase aims at summarizing the incoming data; the offline phase provides the partitioning of the streams into clusters. In the online phase, the incoming observations are split into batches. Our proposal consists of summarizing each subsequence in the batch by a histogram. Finally, a fast clustering algorithm is performed on these summaries in order to get a local partitioning of the subsequences. The offline step finds a consensus partition starting from the achieved local partitions. Through an application on real data, we show the effectiveness of our strategy in finding homogeneous groups of data streams.

2.1. Introduction

Massive data sets, having the form of continuous streams with no fixed length, are becoming very common due to the availability of sensor networks that can perform, at a very high frequency, repeated measurements of some variable. We can think, for instance, of real-time data recorded by surveillance systems, of electricity consumption recording and of the monitoring of environmental variables. The statistical analysis in these applications is a very challenging task since the data cannot be stored or it is archived in a database-management system of a data warehouse, making access very expensive and time consuming.

The data stream mining framework offers a wide range of specific tools for dealing with these potentially infinite and online arriving data. An overview of recent contributions is available in the study of Garofalakis *et al.* (2016).

Chapter written by Antonio BALZANELLA and Rosanna VERDE.

Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining,
First Edition. Edited by Christos H. Skiadas and James R. Bozeman.

In general, data stream mining methods share a set of stringent design criteria (Ganguly *et al.* 2009):

– time required for processing the incoming observations has to be small and constant;

– the allowed memory resources are orders of magnitude smaller than the total size of input data;

– algorithms have to perform only one scan of the data;

– knowledge about data should be available at any point in time or on user demand.

The use of concise synopses/summaries of the streaming data is a typical way for meeting these design criteria. The general idea is to update and/or generate synopsis data structures every time a new element or a batch of elements is collected and, then, to extract the knowledge starting from the summaries rather than directly from the observations.

With few exceptions, synopsis construction is an inherently lossy compression process; thus, it imposes a trade-off between the accuracy of the data mining procedure and the computational and storage constraints.

Since data streams naturally carry a temporal dimension, synopses should support the updating over time without exceeding the small available memory resources. It is still desirable that synopses can be built independently on different parts of the stream and composed/merged in a simple fashion.

The literature on synopses for data stream processing is fairly wide. Several proposals focus on drawing samples from infinite data streams or from data streams split into windows (Al-Kateb *et al.* 2007, Aggarwal 2006). Other synopses address the problem of quantiles estimation (Alsabti *et al.* 1997, Arasu and Manku 2004). Some paper provide solutions to the online computation of histograms (Ganguly *et al.* 2009, Gibbons *et al.* 2003). Finally, techniques aiming at representing streaming time series into reduced space are proposed in the study of Balzanella *et al*. (2010) and Lin *et al*. (2005).

Data stream mining methods use synopses for performing, online or offline, usual data mining tasks such as frequent pattern mining, association rules detection, classification and clustering.

This chapter focuses on clustering of data streams. In general, the challenge addressed by classic clustering algorithms is to partition a set of objects into groups, such that objects in the same group are more similar to each other than to those in other groups. The restrictions imposed by the specific nature of data streams make the demand of specific methods.

According to Barbará (2002), data stream clustering methods should provide a compact representation of the data, process new examples fast and incrementally, do only one (or less) pass over the entire data set, have an answer available at any time and operate inside a limited RAM buffer. Since data streams are usually highly evolving, a further desirable feature is the capability to adapt the clustering structure to such evolution as well as to be able to detect changes in the data.

The literature on data stream clustering provides a distinction according to the data object to cluster. Most of the methods deal with the analysis of a single univariate or multivariate data stream, with the objective of revealing a partition of the observations it is composed of. Some methods analyze multiple data streams generated by several sources, such as sensor networks, with the aim of finding a partition of sources through the analysis, over time, of the proximity among the streams. Some authors refer to this second clustering problem as attribute clustering (Silva *et al.* 2013) or as clustering of streaming time series (Beringer and Hüllermeier 2006, Rodriguess and Pedroso 2008).

The literature on data stream clustering is quite extensive, however, most of the methods address the problem of clustering the observations of a single univariate or multivariate data stream. A state of art, in this sense, is available in Ghesmoune *et al.* (2016). The authors split clustering algorithms into five categories: growing neural gas based methods, hierarchical stream methods, partitioning stream methods, density-based stream methods and grid-based stream methods.

In the first category, there are several algorithms, such as the one proposed in (Ghesmoune *et al.* 2014), based on representing a high dimensional input space in a low dimensional feature map through incremental self-organizing approaches. The second category includes methods, such as BIRCH (Zhang *et al.* 1996) and E-Stream (Udommanetanakit *et al.* 2007), which extend the classic hierarchical clustering algorithms for stocked data to data streams. The third category gathers algorithms such as CluStream (Aggarwal *et al.* 2003) and StreamKM++ (Ackermann *et al.* 2012), which find a partition of the stream observations into k spherical clusters. The fourth category deals with extending density-based clustering algorithms, such as the well-known DBSCAN (Ester *et al.* 1996), to data streams. The main contributions in this category have been proposed in Cao *et al.* (2006), and Wang *et al.* (2013). Finally, the fifth category includes methods such as D-Stream (Chen and Tu 2007), in which the data space is quantized into finite number of cells that form the grid structure and perform clustering on the grids.

The literature on clustering streaming time series is far more limited. Some interesting proposals have been introduced in the following studies: (Beringer and Hüllermeier 2006, Dai *et al.* 2006, Rodriguess and Pedroso 2008, Balzanella *et al.* 2011). The first one is an extension of the *k-means* algorithm performed on time series to the data streams. The second proposal, named clustering on demand (COD), is based on online scan of the data in order to provide a dimensionality reduction.

The reduced time series are indexed by suitable hierarchies that are processed, off-line, by a suitable clustering algorithm that retrieves the partitioning structure of the streams. The third mentioned approach is a top-down strategy named online divisive-agglomerative clustering (ODAC) where a hierarchy is built online, according to a dissimilarity measure based on the correlation among the streams. The last method is a consensus-based strategy in which there is the online updating of a proximity graph and, then, the clustering of the graph in order to get the final partition of the streams.

This chapter focuses on the clustering of time series data streams. We consider each data stream as a univariate time series that acquires, online, new observations. Our strategy provides a partition of the streams available anytime and a set of summaries to give an overview of the monitored phenomenon.

We assume that data streams are highly evolving, that is, observations cannot be considered as generated by a stable distribution but there are several concepts, each one corresponding to a different data generation process. Our strategy addresses this issue through the use of histograms. Particularly, we use histograms as tool for summarizing the different concepts in the data so that streams expressing similar concepts, over time, will be allocated to the same cluster.

The processing schema we propose for clustering the streaming time series is shown in Figure 2.1. There are five main steps. The steps 1–4 are performed online, while step 5 is performed off-line, on user demand. Step 1 splits the incoming data streams into non-overlapping time windows. Each subsequence, framed by the time window, is summarized by a histogram in step 2. In this way, we get, for each time window, a set of histograms that become the input of a local clustering procedure. The aim of step 3 is to get a set of synopses that summarize the data behavior in each window and to provide a local partition of the data. Then, in step 4 there is the updating of a proximity matrix that records the proximity among the streams. As we will see in the following, we update the proximity among the streams only using information coming from the local partitions. Finally, step 5 provides the final partition of the time series data streams through the clustering of the online updated proximity matrix.

2.2. Time series data stream clustering

Let $Y = \{Y_1, \ldots, Y_i, \ldots, Y_n\}$ be a set of n data streams $Y_i = \{(y_i^1, t_1), ..., (y_i^j, t_j), ..., \}$ represented by real-valued observations y_i^j on a discrete time grid $T = \{t_1, ..., t_j, ...\}$, with $t_j \subseteq \Re$ and $t_j > t_{j-1}$.

The objective is to get a partition P of Y in a predefined number C of homogeneous clusters C_k, with $k = 1, ..., C$.

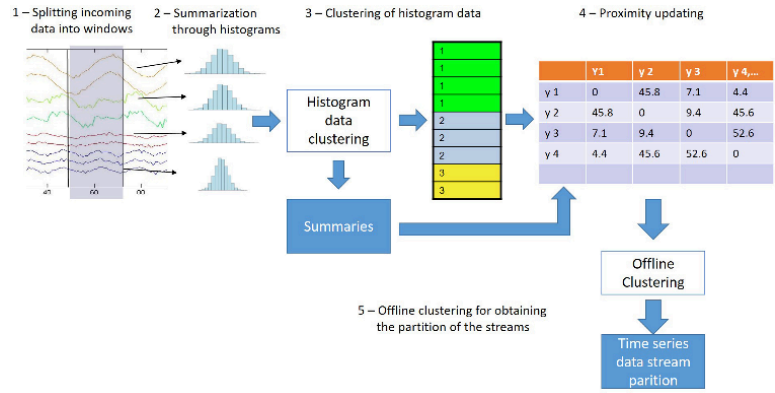


Figure 2.1. *Processing path. For a color version of this figure, see www.iste.co.uk/skiadas/data1.zip*

We assume that the potentially infinite data are recorded online so that we can keep into memory only subsets of the streams. Thus, the analysis is performed using the observations in the most recent batch and some synopsis of the old data, no longer available.

Following the processing scheme in Figure 2.1, we split the incoming data streams into non-overlapping windows, identified by $w = 1, \ldots, \infty$. A window is an ordered subset of T, having size b, which frames a data batch $Y^w = \{Y_1^w, \ldots, Y_i^w, \ldots, Y_n^w\}$, where $Y_i^w = \left\{(y_i^j, t_j), \ldots, (y_i^{j+b}, t_{j+b})\right\}$ is a subsequence of Y_i.

Every time a new batch of data is available; the support $D_i^w = [\underline{y}_i; \overline{y}_i]$ of the subsequence Y_i^w is partitioned into a set of non-overlapping intervals, or bins, such that:

$$D_i^w = \left\{I_{i,1}^w, \ldots, I_{i,l}^w, \ldots, I_{i,L}^w\right\}, where I_{i,l}^w = \left[\underline{y}_{i,l}, \overline{y}_{i,l}\right) \quad [2.1]$$

For each $I_{i,l}^w$, the relative frequencies $\pi_{i,l}^w$ (for $l = 1, \ldots, L$) are computed.

The histogram H_i^w that summarized the subsequence Y_i^w of observations is expressed as:

$$H_i^w = \left\{(I_{i,1}^w, \pi_{i,1}^w), \ldots, (I_{i,l}^w, \pi_{i,l}^w), \ldots, (I_{i,L}^w, \pi_{i,L}^w)\right\} \quad [2.2]$$

It is constituted by L weighted $I_{i,l}^w$ intervals (bins) with the associated $\pi_{i,l}^w$ the relative frequencies.

Histograms are fast to compute and, as we will see in the following, they are able to record much more information about the characteristics of the data distribution (position, variability and shape) than simple means and standard deviations or linear interpolation that are common summaries of the subsequences of streams. Furthermore, they support the use of some recent techniques developed for histogram data analysis, mostly based on a suitable metric to compare distributions: the Wasserstein metric.

The next step is the clustering of the histograms H_i^w (with $i = 1, \ldots, n$) in each window.

2.2.1. *Local clustering of histogram data*

The aim of this step is to recover and represent concisely the concepts in a time window. We use the set of histograms H_i^w (with $i = 1, \ldots, n$) as input of a fast clustering algorithm that provides a partition of the histograms and a set of representatives. The algorithm we propose for addressing this issue is based on the study of Zhang *et al*. (1996).

The intuition that underlies the method is to perform a single scan of the input data in order to get a partitioning into a high number of low variability clusters. The algorithm is based on two main concepts:

– the L^2-Wasserstein metric;

– the histogram micro-cluster (HMC).

Since our strategy is based on representing the concepts in the data through histograms, we need an appropriate distance function to perform their comparison. To this aim, we use the L^2-Wassserstein distance.

Let f and g be two $pdfs$. The L^2 Wasserstein distance can be expressed using the quantile functions F^{-1} and G^{-1} associated, respectively, to the cdf F and G as follows:

$$d_W(f,g) := \sqrt{\int_0^1 (F^{-1}(\xi) - G'^{-1}(\xi))^2 d\xi} \qquad [2.3]$$

In (Irpino and Verde 2015), an exact and efficient way to compute the L^2 Wasserstein distance is developed when data are histograms. Moreover, the authors provide suitable measures of variability as well as the definition of the mean histogram consistently with the L^2 Wasserstein metric. In Irpino and Verde (2006)

and Balzanella *et al.* (2013), there are proposals of its use for addressing specific clustering problems.

The HMC is the data structure used for summarizing the items allocated to a cluster. In particular, the algorithm keeps updated, on each window, a different set of data structures $\mu C^w = \{\mu C_1^w, \ldots, \mu C_k^w, \ldots, \mu C_K^w\}$, named HMC.

An HMC μC_k^w records the following information:

– $\overline{H_k^w}$: histogram centroid;

– n_k^w: number of allocated items;

– σ_k^w: L^2-Wassserstein-based standard deviation;

– Ids_k^w: identifiers of the allocated histograms.

It is an adaptation of the data structure proposed in the BIRCH algorithm for dealing with histogram data. Especially, given a set of allocated histograms, the centroid is the average histogram computed coherently with the L^2-Wasserstein metric (basic statistics based on the Wasserstein metric are widely presented in Irpino and Verde (2015)). Similarly, σ_k^w records the standard deviation for a set of histograms computed according to the same metric. The data field Ids_k^w records the identifier of the histograms allocated to μC_k^w.

Whenever a new window w of data is available and the corresponding set of histograms $H_1^w, \ldots, H_i^w, \ldots, H_n^w$ is built, the clustering algorithm scans the histograms from $i = 1$ to $i = n$ in order to allocate H_i^w to an existing microcluster or to generate a new one. The first preference is to assign H_i^w to a currently existing microcluster.

Especially, H_i^w is allocated to the microcluster μC_k^w such that:

$$d_W(H_i^w, \overline{H_k^w}) < d_W(H_i^w, \overline{H_{k'}^w}) \qquad [2.4]$$

(with $k \neq k'$ and $k = 1, \ldots, K$), if $d_W(H_i^w, \overline{H_k^w}) < u$, where $\overline{H_k^w}$ and $\overline{H_{k'}^w}$ are the prototypes or centroids, of the microclusters μC_k^w and $\mu C_{k'}^w$ respectively.

The threshold value u allows to control if H_i^w falls within the maximum boundary of the microcluster, which is defined as a factor of the standard deviation of the histograms in μC_k^w.

The allocation of a histogram to a microcluster involves an updating of its information. The first update is the increasing by 1 of n_k^w and the addition of the identifier i of the allocated histogram H_i^w to Ids_k^w. Then, it is necessary to update the microcluster centroid and standard deviation.

If H_i^w is outside the maximum boundary of any microcluster, a new microcluster is initialized setting the H_i^w as centroid and $n_k^w = 1$. The standard deviation σ_k^w is defined in a heuristic way by setting it to the squared L^2-Wasserstein distance to the closest cluster. With the creation of a new microcluster, it is necessary to check if $|\mu C^w| > K^*$, that is to evaluate if the number of microclusters for the window w is higher than the available memory resources. In such a case, the number of other clusters must be reduced by one in order to release memory space. This can be achieved by merging the two nearest microclusters.

The updating/generation of histogram microclusters, through a single scan of the histograms in a window, allows to get very quickly a partition into low variability clusters. Similarly to k-means-like methods, the algorithm can be performed iteratively, keeping only the centroids of the previous iteration, in order to improve the clusters homogeneity. While this is recommended in traditional clustering algorithms where K is low, when compared to the size of the data set to process, our aim is to get a summarization of the concepts in the data window. Due to the high number of microclusters and to the threshold on their variability, in most of cases a single scan is enough.

2.2.2. *Online proximity matrix updating*

In order to get a partition P of the streams into homogeneous clusters, we must collect information about their proximity.

Our idea is to evaluate the proximity between pairs of streams through the dissimilarity between the concepts they express over time. Since we have used the histogram H_i^w for describing the concept of the stream Y_i at the window w, two data streams Y_i and Y_m are very similar if, for each window w, the histograms H_i^w and H_m^w are very similar.

Due to the huge amount of data to process, traditional approaches for measuring the distance among the data, through the computation/updating of pairwise distances, are not feasible. To address this issue, we provide a strategy that updates the proximities, starting from the information recorded in the histogram microclusters.

We introduce a matrix $A^w = [a^w(i,m)]$ (with $i,m = 1,\ldots,n$), which records the status of the proximities between each couple Y_i, Y_m at the window w, using the distance between the histograms H_i^w and H_m^w. Thus, at each window we have to update each cell $a^w(i,m)$ of A^w in order to reflect the status of the proximity between the streams.

According to our strategy, the dissimilarities $d(Y_i, Y_m)$ for each $i,m = 1,\ldots,n$ are computed as follows:

– if $(H_i^w, H_m^w) \in \mu C_k^w \rightarrow a^w(i,m) = a^w(i,m) + \sigma_k^w$;

– if $H_i^w \in \mu C_k^w \wedge H_m^w \in \mu C_{k'}^w \rightarrow a^w(i,m) = a^w(i,m) + \frac{d_W(H_m^w;\overline{H_k^w}) + d_W(H_i^w;\overline{H_{k'}^w})}{2}$.

If the histograms H_i^w and H_m^w are allocated to the same histogram microcluster of the local partition, we update the proximity matrix by adding the value of the standard deviation of C_k^w. If H_i^w and H_m^w are allocated to different histogram microclusters, we propose to update the cell $a^w(i,m)$ of the proximity matrix by adding the average of two distances: the first one is the distance between H_i^w and the centroid of the microcluster where H_m^w is allocated; the second one is the distance between H_m^w and the centroid of the microcluster where H_i^w is allocated.

Through such updating strategy, we can use the information stored in the microclusters rather than performing further distance computations.

2.2.3. *Off-line partitioning through the dynamic clustering algorithm for dissimilarity tables*

In this section, we describe the approach we propose for discovering a partition of the stream using the distances recorded in A^w.

Usually, in data stream mining framework we are interested in getting the clustering structure over time periods defined at query time. That is, let t_1, t_2 be two time stamps, we should provide a partition P of $Y = \{Y_1, \ldots, Y_i, \ldots, Y_n\}$, only for the subsequences $Y_i = \left\{(y_i^1, t_1), \ldots, (y_i^j, t_2)\right\}$ (for each $i = 1, \ldots, n$).

Since the time series observations are not available anymore, we can use the proximities stored in A^w. Indeed, we can recover the proximities of a time period $[t_1, t_2]$, by $A^* = A^{w''} - A^{w'}$, where $A^{w''}$ is the proximity matrix at the time t_2 and $A^{w'}$ is the proximity matrix at the time t_1.

In order to get the partition P from A^*, we use the dynamic clustering algorithm on a dissimilarity table (DCLUST) (Diday 2008).

The aim of the DCLUST is to partition a set of elements into a fixed number of homogeneous classes on the basis of the proximities between pairs of elements. The optimized criterion is based on the sum of the dissimilarities between elements belonging to the same cluster. Because the dissimilarities between pair of streams (Y_i and Y_m) are the values in the cells $a^*(i,m)$ of A^*, the DCA criterion can be expressed as:

$$\Delta(P, L) = \sum_{k=1}^{C} \sum_{i,m \in C_k} a^*(i,m) \qquad [2.5]$$

According to the schema of DCLUST, the prototypes of the clusters corresponds to the streams Y_{m^*}: $m^* = argmin_m(\sum_{i \in C_k} d(Y_i, Y_m)) = argmin_m(\sum_{i \in C_k} a^*(i,m))$ with $Y_m \in C_k$ (for $k = 1, \ldots, C$).

The DCLUST algorithm schema is the following:

1) *Initialization*: the vector of prototypes L is initialized by a set of random elements of S;

2) *Allocation step*: a stream Y_i is allocated to the cluster C_k if $a^*(i,m) < a^*(i,j)$ with Y_m the prototype of C_k and Y_j the prototype of C_l (for all $k, l = 1, \ldots, C$);

3) *Representation step*: for each $k = 1, \ldots, C$, the prototype Y_m representing the class C_k is the stream $Y_{m^*} \in C_k$.

Steps 2 and 3 are repeated until convergence.

It is easy to prove that the DCLUST on the dissimilarity matrix A, choosing as prototypes of the clusters the elements (streams) to the minimum distance from the other elements of the clusters, converges to a stationary value.

2.3. Results on real data

The strategy proposed in this chapter has been tested on a real data set with the aim of evaluating its capability to cluster streaming time series. In particular, we want to evaluate if updating the proximity among the streams using the statistics stored in the histogram microclusters allows to get results which compare favorably with the use of a distance matrix computed offline on the histograms.

The test data set collects the records of 54 sensors placed at the Intel Berkeley Research lab between February 28 and April 5, 2004. Mica2Dot sensors with weather boards collected timestamped topology information, along with humidity, temperature, light and voltage values once every 31 s. Data were collected using the TinyDB in-network query processing system, built on the TinyOS platform. The data set includes the x and y coordinates of sensors (in meters relative to the upper right corner of the lab).

We have analyzed the temperature records of each sensor so that we have a set of 54 time series each one made by 65,000 observations.

In order get the clustering results, we have to set some input parameters. The first one is the size b of each time window in terms of number of observations. We choose $b = 116$, where each window collects 1 h of sensor recordings. The second parameter to set is the number of bin of each histogram. We set the value $L = 10$ since it is a good

compromise between the accuracy of the distribution approximation and compactness of the representation.

The next parameter is the threshold value u that controls the size of each microcluster in terms of maximum accepted standard deviation. A high threshold value involves a reduced number of microclusters with a strong data compression; however, it worsens the clustering performance. On the contrary, a low value for u improves the clustering performances but reduces the compression rate. In Figure 2.2, we show how the threshold value impacts on the data compression by plotting, for $0.001 \leq u \leq 0.01$ (step 0.001), the representation of the number of microclusters generated over the windows.

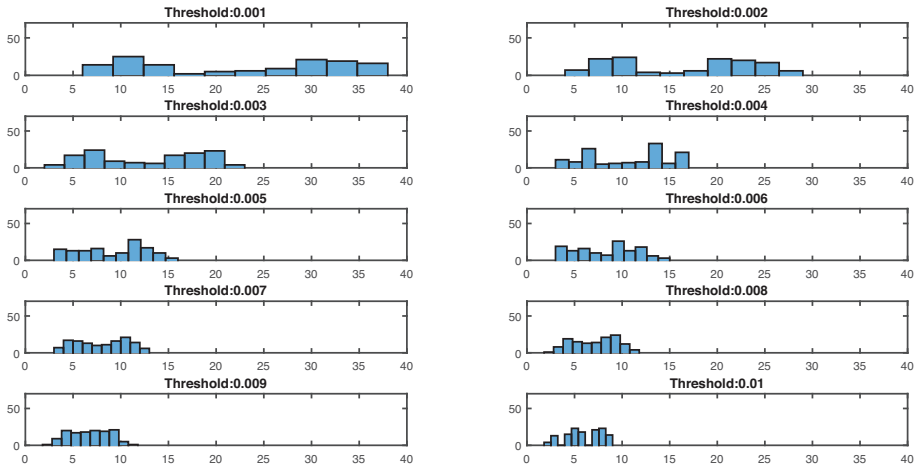


Figure 2.2. *Graphical representation of the number of microclusters generated at each time window according to the value of the threshold u*

It is evident that with the growing of the threshold value, the number of generated microcluster decreases.

Now, we compare the DCLUST on the proximity matrix updated using the information recorded in the microclusters, with the DCLUST performed on the distance matrix obtained by computing the pairwise distances between the histograms. The latter corresponds to use 0 as threshold value, that is the number of microclusters is equal to the number of streams so that no data compression is performed.

We make the comparison using the classic rand index (Hubert and Arabie 1985) to measure the consensus between the partitions obtained by the two approaches for the different levels of u.

Threshold value	Rand index
0.001	0.99
0.002	0.99
0.003	0.97
0.004	0.96
0.005	0.96
0.006	0.96
0.007	0.94
0.008	0.93
0.009	0.91
0.01	0.89

Table 2.1. *Rand index for* $0.001 \leq u \leq 0.01$

While the consensus is very high in all the cases, we can note that the increasing in u worsen the results, as expected.

2.4. Conclusions

In this chapter, we have introduced a new strategy that deals with the problems of clustering data streams generated by sensors through the evaluation of the proximity between the concepts in the data. We have proposed to represent the concepts through histograms and to compare them by a specific metric for distribution data. Due to online nature of data and to its potentially infinite size, the final data partition is obtained by clustering, at first, non-overlapping data batches and then running an appropriate clustering algorithm on the dissimilarity matrix updated using the outputs of the local clustering. An application on real data has confirmed that our approach for dealing with the streaming nature of data allows to get an effective data compression with a limited performance degradation.

2.5. References

Ackermann, M.R., Märtens, M., Raupach, C., Swierkot, K., Lammersen, C., Sohler, C. (2012). StreamKM++: a clustering algorithm for data streams. *ACM J. Exp. Algorithmics*. 17(1), 173–187.

Aggarwal, C.C. (2006). On biased reservoir sampling in the presence of stream evolution. *VLDB'06*, September 12–15, Seoul, Korea.

Aggarwal, C.C., Han, J., Wang, J., Yu, P.S. (2003). A framework for clustering evolving data streams. In *VLDB 2003: Proceedings of the 29th International Conference on Very Large Data Bases*, VLDB Endowment, 81–82.

Al-Kateb, M., Lee, B.S., Wang, X.S. (2007). Adaptive-size reservoir sampling over data streams. *Scientific and Statistical Database Management, 19th International Conference on Scientific and Statistical Database Management*.

Alsabti, K., Ranka, S., Singh, V. (1997). A one-pass algorithm for accurately estimating quantiles for disk-resident data. In *Proceedings of the 23rd International Conference on Very Large Data Bases*, Jarke, M., *et al.* (ed.), Los Altos, CA 94022, Morgan Kaufmann, SanMateo, USA, 346–355.

Arasu, A., Manku, G.S. (2004). Approximate counts and quantiles over sliding windows. *Proceedings of the 23rd ACM Symposium on Principles of Database Systems*, 286–296, France, Paris.

Balzanella, A., Rivoli, L., Verde, R. (2013). Data stream summarization by histograms clustering. In *Statistical Models for Data Analysis*, Giudici, P., Ingrassia, S., Vichi, M. (eds), Springer International Publishing, 27–35.

Balzanella, A., Irpino, A., Verde, R. (2010). Dimensionality reduction techniques for streaming time series: a new symbolic approach. *Classification as a Tool for Research, Studies in Classification, Data Analysis, and Knowledge Organization*, Springer Berlin Heidelberg, 381–89.

Balzanella, A., Lechevallier, Y., Verde, R. (2011). Clustering multiple data streams. *New Perspectives in Statistical Modeling and Data Analysis*. Springer Berlin Heidelberg.

Barbará, D. (2002). Requirements for clustering data streams. *SIGKDD Explorations – Special Issue on Online, Interactive, and Anytime Data Mining*. 3, 23–27.

Beringer, J., Hüllermeier, E. (2006). Online clustering of parallel data streams. *Data Knowledge Engineering*, 58, 180–204, https://doi.org10.1016/j.datak.2005.05.009.

Cao, F., Ester, M., Qian, W., Zhou, A. (2006). Density-based clustering over an evolving data stream with noise. *SDM. SIAM: 2006*. 328–39.

Chen, Y., Tu, L. (2007). Density-based clustering for real-time stream data. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 133–42, ACM.

Dai, B-R., Huang, J.-W., Yeh, M.-Y., Chen, M.-S. (2006). Adaptive clustering for multiple evolving streams. *IEEE Transactions On Knowledge And Data Engineering*, 18(9).

Diday E., Noirhomme-Fraiture, M. (2008). *Symbolic Data Analysis and the SODAS Software*. Wiley.

Ester, M., Kriegel, H.P., Jörg, S., Xiaowei, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, AAAI Press, 226–231.

Ganguly, A.R., Gama, J., Omitaomu, O.A., Gaber, M.M., Vatsavai, R.R. (2009). *Knowledge discovery from sensor data*. CRC Press.

Garofalakis, M., Gehrke, J., Rastogi, R. (2016). *Data Stream Management: Processing High-Speed Data Streams*, Springer, New York.

Ghesmoune, M., Lebbah, M., Azzag, H. (2016). State-of-the-art on clustering data streams. *Big Data Analytics*. 1(13), https://doi.org/10.1186/s41044-016-0011-3.

Ghesmoune, M., Azzag, H., Lebbah, M. (2014). G-stream: growing neural gas over data stream. *Neural Information Processing – 21st International Conference, ICONIP 2014*, Kuching, Malaysia, November 3–6.

Gibbons, P.B., Matias, Y., Poosala, V. (2003). Fast incremental maintenance of approximate histograms. *ACM Trans. Database Syst.* 27(3), 261–298.

Hubert, L., Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 193–18.

Irpino, A., Verde, R. (2006). Dynamic clustering of histograms using Wasserstein metric. *COMPSTAT 2006 – Advances in computational statistics*, Rizzi, A., Vichi, M. (eds), Heidelberg, Physica-Verlag, 869–876.

Irpino, A., Verde, R. (2015). Basic statistics for distributional symbolic variables: a new metric-based approach. *Adv. Data Analysis and Classification*, 9(2), 143–175.

Lin, J., Keogh, E., Lonardi, S., Chiu, B. (2005). A symbolic representation of time series, with implications for streaming algorithms. *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*. San Diego, CA. June 13.

Rodriguess, P.P., Pedroso, J.P. (2008). Hierarchical Clustering of Time Series Data Streams. *IEEE Transactions on Knowledge and Data Engineering*, 20(5).

Silva, J.A., Faria, E.R., Barros, R.C., Hruschka, E.R., de Carvalho, A.C.P.L.F., Gama, J. (2013). *Data stream clustering: A survey*. ACM Comput. Surv. 46(1).

Udommanetanakit, K., Rakthanmanon, T., Waiyamai, K. (2007). E-stream: evolution-based technique for stream clustering. In: *Advanced Data Mining and Applications. ADMA 2007*, Alhajj R., Gao H., Li J., Li X., Zaïane O.R. (eds), Lecture Notes in Computer Science, 4632, 605–615. Springer, Berlin, Heidelberg.

Wang, C., Lai, J., Huang, D., Zheng, W. (2013). SVStream: a support vector-based algorithm for clustering data streams. *IEEE Trans Knowl Data Eng*. 25(6), 1410–1424.

Zhang, T., Ramakrishnan, R., Livny, M. (1996). Birch: an efficient data clustering method for very large databases. *SIGMOD Conference*. ACM, New York, 103–14.

3

The Flexible Beta Regression Model

A relevant problem in applied statistics concerns modeling rates, proportions or, more generally, continuous variables restricted to the interval (0,1). Aim of this contribution is to study the performances of a new regression model for continuous variables with bounded support that extends the well-known beta regression model (Ferrari and Cribari-Neto 2004). Under our new regression model, the response variable is assumed to have a flexible beta (FB) distribution, a special mixture of two beta distributions that can be interpreted as the univariate version of the flexible Dirichlet distribution (Ongaro and Migliorati 2013). In many respects, the FB can be considered as the counterpart on (0,1) to the well-established mixture of normal distributions sharing a common variance. The FB guarantees a greater flexibility than the beta distribution for modeling bounded responses, especially in terms of bimodality, asymmetry and heavy tails. The peculiar mixture structure of the FB makes it identifiable in a strong sense and guarantees a bounded likelihood and a finite global maximum on the assumed parameter space. In the light of these good theoretical properties, the new model results to be very tractable from a computational perspective, in particular with respect to posterior computation. Therefore, we provide a Bayesian approach to inference and, in order to estimate its parameters, we propose a new mean-precision parameterization of the FB that guarantees a variation-independent parametric space. Interestingly, the FB regression (FBR) model can be understood itself as a mixture of regression models. The strength of our new FBR model is illustrated by means of application to a real data set. To simulate values from the posterior distribution, we implement the Gibbs sampling algorithm through the BUGS software.

3.1. Introduction

To implement standard linear regression models for continuous variables restricted to the interval $(0, 1)$, one has to transform the response variable so that its support becomes the real line. Despite having been the preferred method for a long time, such an approach has two relevant drawbacks: first, the difficulty in interpreting the estimated parameters with respect to the original response variable (Ferrari and Cribari-Neto 2004), and second, the failure of the assumptions of normality

Chapter written by Sonia MIGLIORATI, Agnese M. DI BRISCO and Andrea ONGARO.

Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining,
First Edition. Edited by Christos H. Skiadas and James R. Bozeman.

(proportions typically show asymmetric distributions) and homoscedasticity (Paolino 2001).

To overcome these drawbacks, many researchers have developed regression models assuming a beta distributed response variable on the original restricted space (Ferrari and Cribari-Neto 2004). Since the beta distribution is not a dispersion-exponential family (McCullagh and Nelder 1989), inference requires an ad hoc maximum likelihood estimation approach (Ferrari and Cribari-Neto 2004) or alternatively a Bayesian approach (Branscum *et al*. 2007).

The beta distribution can show very different shapes (unimodal, monotone and U-shaped) but it does not provide enough flexibility to model a wide range of phenomena, including heavy tailed responses with a bounded support (Bayes *et al*. 2012, García *et al*. 2011) and bimodality. A first attempt to handle greater flexibility is due to Hahn (2008) who introduced the beta rectangular (BR) distribution, which is defined as a mixture of a uniform and a beta distribution. Later, Bayes *et al*. (2012) defined a BR regression (BRR) model for both mean and dispersion parameters by considering a Bayesian approach. The authors showed that the model enables heavier tails and is robust in the presence of outliers. With the purpose of achieving even greater flexibility, we may consider a generic mixture of beta distributions. Nevertheless, despite mixture distributions provide accurate data fit and robustness (Markatou 2000, Gelman *et al*. 2014), a generic beta mixture may be hard to treat because of its lack of invariance under relabeling of the mixture components. Such an issue, well-known as the label switching problem, determines undesirable effects on posterior distributions, especially in case of overlapping components. To handle the trade-off between flexibility and tractability, we propose a new regression model based on a special mixture of beta distributions. To this end, we introduce the FB distribution (univariate version of the flexible Dirichlet distribution, see Ongaro and Migliorati 2013), which is a special mixture of two beta distributions with arbitrary means and common variance. The FB distribution enables a great variety of density shapes in terms of tail behavior, asymmetry and multimodality. Nevertheless, its peculiar mixture structure avoids the label switching problem, making the FB very tractable from a computational perspective, for example with respect to posterior computation in Bayesian inference.

The rest of the chapter is organized as follows. In section 3.2, we introduce the FB distribution and we propose a reparameterization that is designed for this regression context and enables a very clear interpretation of the new parameters. In section 3.3, we define the FBR model and we also interpret it as mixture of regression models (Frühwirth-Schnatter 2006). In section 3.4, we provide details concerning Bayesian inference and the Gibbs sampling algorithm specifically designed for mixture models. In order to evaluate the performance of the FBR model and compare it with the BR and beta regression ones, we perform an illustrative application on a real data set (section 3.5).

3.2. The FB distribution

3.2.1. *The beta distribution*

The beta is the preferred distribution for modeling a continuous response variable bounded on $(0, 1)$. Let us define a random variable beta distributed $Y \sim Beta(\mu\phi, (1-\mu)\phi)$, according to the mean-precision parameterization (see Ferrari and Cribari-Neto 2004 for details), that is, with a probability density function (pdf):

$$f_B^*(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \qquad 0 < y < 1, \qquad [3.1]$$

where $0 < \mu < 1$ and $\phi > 0$. The parameter μ identifies the mean of Y, whereas the parameter ϕ is interpreted as a *precision* parameter being:

$$Var[Y] = \frac{\mu(1-\mu)}{\phi+1}.$$

3.2.2. *The FB distribution*

The FB distribution is the univariate version of the flexible Dirichlet one, first proposed by Ongaro and Migliorati (2013) as a generalization of the Dirichlet distribution. While the latter has been shown to be inadequate to model compositional data because of its rigid structure, the flexible Dirichlet distribution allows for considerably greater flexibility still preserving a remarkable tractability (see Ongaro and Migliorati 2013 and Migliorati *et al*. 2017 for a detailed analysis of its properties and statistical potential).

The FB distribution is defined as a special mixture of two beta distributions with a common precision parameter ϕ and arbitrary (but distinct) means $\lambda_1 > \lambda_2$. Its pdf for $0 < y < 1$ results equal to:

$$f_{FB}^*(y; \lambda_1, \lambda_2, \phi, p) = p f_B^*(y; \lambda_1, \phi) + (1-p) f_B^*(y; \lambda_2, \phi), \qquad [3.2]$$

where $0 < \lambda_2 < \lambda_1 < 1$, $\phi > 0$, $0 < p < 1$ and f_B^* is the mean-precision parameterized beta [3.1]. The first two moments of the FB are equal to:

$$\mathbb{E}(Y) = p\lambda_1 + (1-p)\lambda_2$$

$$Var(Y) = \frac{\mathbb{E}(Y)(1-\mathbb{E}(Y)) + \phi(\lambda_1-\lambda_2)^2 p(1-p)}{\phi+1} \qquad [3.3]$$

The special mixture structure of the FB distribution greatly extends the variety of shapes of the beta mainly in terms of bimodality, asymmetry and tail behavior, as illustrated in Figure 3.1. In addition, it ensures that each component is distinguishable, avoiding the label switching problem. Interestingly, this property makes the FB distribution computationally very tractable, as we will point out in section 3.4.

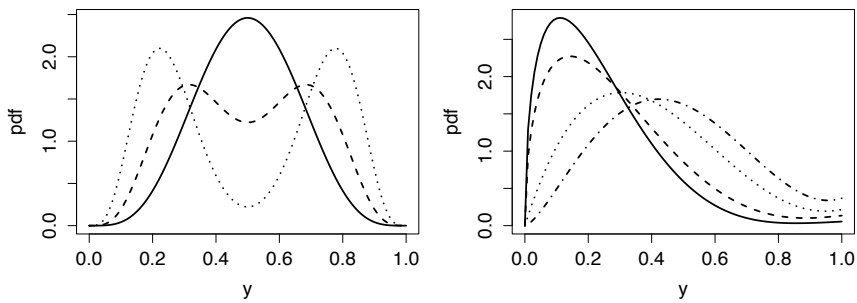


Figure 3.1. *Some examples of FB distributions. Left panel:* $p = 0.5$ *and* $\{\lambda_1 = 0.54, \lambda_2 = 0.45, \phi = 11\}$ *in solid line,* $\{\lambda_1 = 0.4, \lambda_2 = 0.3, \phi = 15\}$ *in dashed line and* $\{\lambda_1 = 0.3, \lambda_2 = 0.25, \phi = 20\}$ *in dotted line. Right panel:* $\{\lambda_1 = 0.85, \lambda_2 = 0.23, \phi = 6.5, p = 0.01\}$ *in solid line,* $\{\lambda_1 = 0.82, \lambda_2 = 0.27, \phi = 5.5, p = 0.03\}$ *in dashed line,* $\{\lambda_1 = 0.81, \lambda_2 = 0.38, \phi = 5.3, p = 0.05\}$ *in dotted line and* $\{\lambda_1 = 0.63, \lambda_2 = 0.45, \phi = 5.6, p = 0.08\}$ *in dashed-dotted line*

3.2.3. *Reparameterization of the FB*

With the aim of defining a regression model with an FB-distributed response variable, we propose a reparameterization that explicitly includes the mean, complemented with other three clearly interpretable parameters:

$$\begin{cases} \mu = \mathbb{E}(Y) = p\lambda_1 + (1-p)\lambda_2 \\ \phi = \phi \\ \tilde{w} = \lambda_1 - \lambda_2 \\ p = p \end{cases} \qquad [3.4]$$

where μ is the mean parameter, $\tilde{w}$ is a measure of distance between the two mixture components, p is the mixing proportion and ϕ plays the role of a precision parameter since $Var(Y)$ is a decreasing function of ϕ.

In the parametric space so far defined, ϕ is free to move in $\mathbb{R}^+$, whereas μ, p and $\tilde{w}$ are linked by some constraints. We then decide to require a variation-independent parametric space, first to properly adopt a Bayesian approach to inference through

Gibbs sampling (Albert 2009), as illustrated in section 3.4, and second to separately model any parameter as a function of covariates.

To this purpose, we choose to leave μ and p free to assume values in (0,1), and to properly normalize $\tilde{w}$ to make it free to move on the range (0,1) as well. Having fixed μ and p, the constraints $0 < \lambda_2 < \lambda_1 < 1$ imply that $\tilde{w}$ takes values between 0 and $\min\left\{\frac{\mu}{p}, \frac{1-\mu}{1-p}\right\}$.

Therefore, we normalize $\tilde{w}$ accordingly:

$$w = \frac{\tilde{w}}{\min\left\{\frac{\mu}{p}, \frac{1-\mu}{1-p}\right\}}. \qquad [3.5]$$

The chosen reparameterization guarantees a variation-independent parameter space where p, μ and w vary in $(0, 1)$ and $\phi > 0$ without penalizing the interpretability of the parameters.

3.3. The FB regression model

Given a vector of independent responses $\mathbf{Y}^T = (Y_1, \ldots, Y_i, \ldots, Y_n)$ that assume values in the unit interval $(0, 1)$, in accordance to the GLM methodology (McCullagh and Nelder 1989), a regression model for the mean can be defined as

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} \qquad i = 1, \ldots, n, \qquad [3.6]$$

where μ_i is the mean of Y_i, $\mathbf{x}_i^T = (x_{i0}, x_{i1}, \ldots, x_{ik})$ is a vector of covariates, $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_k)$ is a vector of regression parameters and $g(\cdot)$ is an adequate link function, strictly monotone and twice differentiable. The most popular link function is the logit, $logit(\mu_i) = \log(\mu_i/(1 - \mu_i))$, which allows to interpret the regression coefficients in terms of odds ratios.

If Y_i is assumed to follow a beta distribution, then the beta regression model is obtained (Ferrari and Cribari-Neto 2004). If the response variable is assumed to follow a BR distribution, having $Y \sim BR(\mu, \phi, \alpha)$ where $\mu = E(Y)$, $\alpha = p/(1 - |2\mu - 1|)$ and ϕ is the precision parameter of the beta component, then the BRR model is achieved (see Bayes *et al*. 2012 for further details).

Here, we define the FBR model by assuming that each Y_i is independently distributed as a FB under the parameterization given in section 3.2.3: $Y_i \sim FB(\mu_i, \phi, w, p)$.

Note that none of the above distributions belongs to the dispersion-exponential family (McCullagh and Nelder 1989), and therefore none of the above models is of the GLM type.

Although the regression model [3.6] concerns only the mean parameters, it naturally induces a form of heteroscedasticity since the response variances are functions of the corresponding means (see formula [3.3]). However, in some cases, it may be desirable to independently model the variance as a function of covariates. Many authors have proposed extensions in this direction (Paolino 2001, Smithson and Verkuilen 2006, Ferrari *et al*. 2011). This can be easily achieved in the FBR too, as the parameters μ and ϕ do not share any constraint. The regression for the dispersion can be defined as

$$h(\phi_i) = \mathbf{z}_i^T \boldsymbol{\delta}, \qquad [3.7]$$

where $h(\cdot)$ is an appropriate link function, $\mathbf{z}_i^T = (z_{i0}, z_{i1}, \ldots, z_{il})$ is a vector of covariates and $\boldsymbol{\delta}^T = (\delta_0, \delta_1, \ldots, \delta_l)$ is a vector of regression parameters. In the study of Paolino (2001), logarithm as a proper link function (strictly monotone and double differentiable) has been proposed as $\phi > 0$.

As a further extension, a regression function can be defined also for the remaining parameters of the FB distribution, namely w and p, since the parametric space is variation independent.

It is of interest to observe that the FBR model can be easily understood as a mixture of regression models (Frühwirth-Schnatter 2006). In these models, the regression function is not fixed over all realizations, but different groups of observations may display different dependencies of the means on covariates. In this respect, the FBR is a special mixture of two beta regression models, with common precision and means given by

$$\begin{cases} \lambda_{1i} = \mu_i + (1-p)\tilde{w}_i \\ \lambda_{2i} = \mu_i - p\tilde{w}_i \end{cases} \qquad [3.8]$$

where $\tilde{w}_i = w \min\left\{\frac{\mu_i}{p}, \frac{1-\mu_i}{1-p}\right\}$. Note that such means are piecewise increasing linear functions of μ_i, varying from 0 to 1. The underlying assumption here is that there are two groups, one of which displays a greater mean than the other, for any given value of covariates. The parameter w retains the meaning of distance between the regression functions of the two groups. Thus, the FBR structure can also be usefully employed to model two distinct groups that have special regression function patterns.

3.4. Bayesian inference

Given a sample of n independent observations $\mathbf{y}^T = (y_1, \ldots, y_i, \ldots, y_n)$, the likelihood function for the FBR model [3.6] results equal to:

$$L(\boldsymbol{\eta}|\mathbf{y}) = \prod_{i=1}^{n} f_{FB}^{*}(y_i|\mu_i, \phi, w, p), \qquad [3.9]$$

where $\boldsymbol{\eta} = (\boldsymbol{\beta}, \phi, w, p)$, $\mu_i = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})$, and $f^*_{FB}(y|\mu, \phi, w, p)$ is given by [3.2] with

$$\begin{cases} \lambda_1 = \mu + (1-p)\tilde{w} \\ \lambda_2 = \mu - p\tilde{w} \end{cases} \qquad [3.10]$$

where $\tilde{w} = w \min\left\{\frac{\mu}{p}, \frac{1-\mu}{1-p}\right\}$.

A mixture model can be seen as an incomplete data problem (Dempster *et al.* 1977) since the allocation of each ith observation to one of the mixture components is unknown. Since no explicit solution to the estimation problem exists, we propose here to adopt MCMC techniques such as data augmentation (Tanner and Wong 1987) and Gibbs sampling (Gelfand and Smith 1990), which are well suited to cope with incomplete data.

Formally, let us define an n-dimensional vector of latent variables $\mathbf{v}$, such that $v_i = 1$ if the ith observation belongs to the first mixture component and $v_i = 0$ otherwise. Having identified these latent variables as missing data, we define a Gibbs sampling algorithm that is split into two steps: one for the parameter simulation conditional on $\mathbf{v}$, and the other for the classification of the observations (i.e. updating $\mathbf{v}$) conditional on knowing the parameter. The posterior distribution $\pi(\boldsymbol{\eta}|\mathbf{y})$ is computed by marginalizing the "complete-data" posterior distribution $\pi(\boldsymbol{\eta}, \mathbf{v}|\mathbf{y})$. To compute the "complete-data" posterior distribution, the complete-data likelihood $L_{CD}(\boldsymbol{\eta}|\mathbf{y}, \mathbf{v})$ is needed, i.e. the likelihood based on both observed ($\mathbf{y}$) and missing ($\mathbf{v}$) data. More precisely:

$$\pi(\boldsymbol{\eta}, \mathbf{v}|\mathbf{y}) \propto L_{CD}(\boldsymbol{\eta}|\mathbf{y}, \mathbf{v})\pi(\boldsymbol{\eta}) \qquad [3.11]$$

with

$$L_{CD}(\boldsymbol{\eta}|\mathbf{y}, \mathbf{v}) = \prod_{i=1}^{n}[pf^*_B(y_i; \lambda_{1i}, \phi)]^{\{v_i\}}[(1-p)f^*_B(y_i; \lambda_{2i}, \phi)]^{\{1-v_i\}}, \qquad [3.12]$$

where λ_{1i} and λ_{2i} are given by [3.8], f^*_B is defined by [3.1] and $\pi(\boldsymbol{\eta})$ is an appropriate prior distribution. With respect to the prior distribution, we assumed a priori independence, which is a usual choice when no prior information is available. Since the parametric space is variation independent, the joint prior distribution can be factorized as:

$$\pi(\boldsymbol{\eta}) = \pi(\boldsymbol{\beta})\pi(\phi)\pi(w)\pi(p).$$

Moreover, we decided to adopt flat priors, so as to generate the minimum impact on the posteriors (e.g. Albert 2009). Specifically, we selected the usual multivariate normal prior $\boldsymbol{\beta} \sim N_{k+1}(\mathbf{a}, \mathbf{B})$ for the regression parameters with $\mathbf{a} = \mathbf{0}$ for the mean, and a diagonal covariance matrix with "large" values for the variances in $\mathbf{B}$. For the remaining parameters, we chose a gamma distribution $Ga(g, g)$ for ϕ, which is a rather standard choice for precision parameters (e.g. Branscum *et al*. 2007), and we selected non-informative uniform priors for the remaining parameters $w \sim U(0, 1)$ and $p \sim U(0, 1)$. The estimation procedure described above can be easily extended to deal with cases in which the precision parameter is modeled as a function of the covariates (see [3.7]). It is enough to replace the prior for ϕ with a convenient multivariate normal prior for the regression coefficients $\boldsymbol{\delta} \sim N_{l+1}(\mathbf{c}, \mathbf{D})$. Analogous considerations hold for the parameters w and p. We implemented the Gibbs sampling algorithm through the BUGS software (Thomas 1994, Lunn *et al*. 2000) in order to generate a finite set of values from the posterior distribution, and further analyzed the results through the R software (R Core Team 2016). We iterated the algorithm until convergence by burning-in the first B simulated values (for different values of B in the various contexts) to avoid the influence of the chains' initial values. Furthermore, to properly treat autocorrelations, we also set a thinning interval, say L, such that only the first generated values in every batch of L iterations were kept. Finally, we checked for convergence of the algorithm through several statistical tests, with a focus on diagnostic tests for stationarity (Geweke and Heidel diagnostics) and for the level of autocorrelation (Raftery diagnostic) (Mengersen *et al*. 1999, Ntzoufras 2011).

To the purpose of comparing the FBR model with other competing models, we take into consideration some comparison criteria. Typically, these criteria favor models with a better fit while simultaneously penalize more complex models. To quantify the lack of fit of a model, we shall consider the deviance that is defined as a function of the likelihood, $L(\boldsymbol{\eta}|\mathbf{y})$ (see [3.9] for the FBR model), and can be interpreted as the residual information in data, given the parameters:

$$D(\boldsymbol{\eta}) = -2log[L(\boldsymbol{\eta}|\mathbf{y})].$$

Given the MCMC output, the deviance can simply be estimated by taking the posterior mean of the deviance $\overline{D}$, i.e. the mean of the deviances of the MC sample.

To evaluate model complexity, we may consider different measures. By way of example, the deviance information criterion (DIC) (Spiegelhalter *et al*. 2002):

$$DIC = \overline{D} + p_D$$

penalizes the complexity of the model via:

$$p_D = \overline{D} - D(\bar{\boldsymbol{\eta}}),$$

where $\bar{\boldsymbol{\eta}}$ is the vector of posterior means of the parameters. Alternatively, one can penalize model complexity the same way as is done by the well-known AIC (Akaike 1998) and BIC (Schwarz 1978) criteria, thus obtaining the corresponding Bayesian counterparts EAIC and EBIC (Brooks 2002), that is,

$$EAIC = \overline{D} + 2p$$
$$EBIC = \overline{D} + p\log(n),$$

where p is the number of the model parameters and n is the sample size.

Clearly, the smaller the values of DIC, EAIC and EBIC, the better the model.

When dealing with mixture models, the values of such criteria are not implemented by default in BUGS. Nevertheless, they can be easily computed from the MCMC output.

3.5. Illustrative application

In this section, we show how the FBR can be successfully applied, comparing it with the two competing models, namely the BRR and the beta ones.

We consider a data set about the gasoline yield data (Prater 1956) (the data set *GasolineYield* is included in the R library *betareg*). The proportion of crude oil converted to gasoline after distillation and fractionation, naturally quantified on $(0, 1)$, is defined as the dependent variable of the regression model.

The FBR model for the proportion of converted crude oil, $Y_i \sim FB(\mu_i, \phi, p, w)$ for $i = 1, \ldots, n$, is defined as:

$$logit(\mu_i) = \beta_0 + \beta_1 X_{i1}, \quad i = 1, \ldots, n,$$

where X_1 is a quantitative covariate about the standardized temperature (originally in degrees Fahrenheit) at which all gasoline has vaporized and β_0, β_1 are unknown regression parameters. The response variances, being functions of the corresponding means (see [3.3]), will also vary with the covariates, thus inducing a preliminary form of heteroscedasticity.

However, to further improve the fit of the FBR model one can let also the parameter ϕ depend directly on the same covariate rather than being constant for all observations

$$log(\phi_i) = \delta_0 + \delta_1 X_{i1}, \quad i = 1, \ldots, n$$

where δ_0, δ_1 are unknown regression parameters. To compare the FBR model with the beta and with the BRR ones, we simulated MCMCs of length 10,000, discarded the

first half values and used a thinning interval set equal to 3 for the beta model, 1 for the BRR model and 10 for FBR model. These values satisfy the various diagnostic tests mentioned in section 3.4.

The results are shown in Tables 3.1 and 3.2, and in Figure 3.2.

Parameter	β_0	β_1	δ_0	δ_1	p	w	α
FBR	−1.4848	0.5300	−3.4124	0.1626	0.5866	0.2659	
BRR	−1.4132	0.5140	−3.2077	0.1604			0.0984
Beta	−1.4853	0.5318	−3.1811	0.1755			

Table 3.1. *Posterior means of the parameters under the three models*

Model	**DIC**	**EAIC**	**EBIC**
FBR	−86.9641	−71.59964	−62.80523
BRR	−71.93797	−65.75476	−58.42608
Beta	−72.39978	−68.42512	−62.56218

Table 3.2. *Model comparison criteria for the FBR, the BRR and the beta regression models*

We may observe that we get similar estimates for the four regression parameters under the competing models. Indeed, in Figure 3.2 we plotted the three fitted regression curves into three different plots to avoid overlapping. The FBR exhibits a better fit than the other two models, especially with respect to DIC, whereas the BR displays the worst one.

To better grasp the behavior of the models, the left panel in Figure 3.2 reports the group regression means of the FBR. The FBR and the beta regression models display a nearly identical behavior in terms of regression curves for the mean of the response variable. In fact, the better fit of the FBR is due to its ability to locate and accurately describe two groups with a different rate curve. Thus, the FBR efficiently exploits its greater flexibility. On the contrary, the flexibility of the BRR model does not seem to provide a better fit since the uniform component, equal to 1/2, lies completely outside the observed scatter plot.

3.6. Conclusion

The FBR model proves to be a good compromise between tractability and flexibility when modeling continuous responses bounded to the unit interval. Although in this chapter we consider only response variables restricted to the interval $(0, 1)$, the model we propose can be easily extended via an obvious linear transformation of the response to deal with variables taking values on a generic bounded interval.

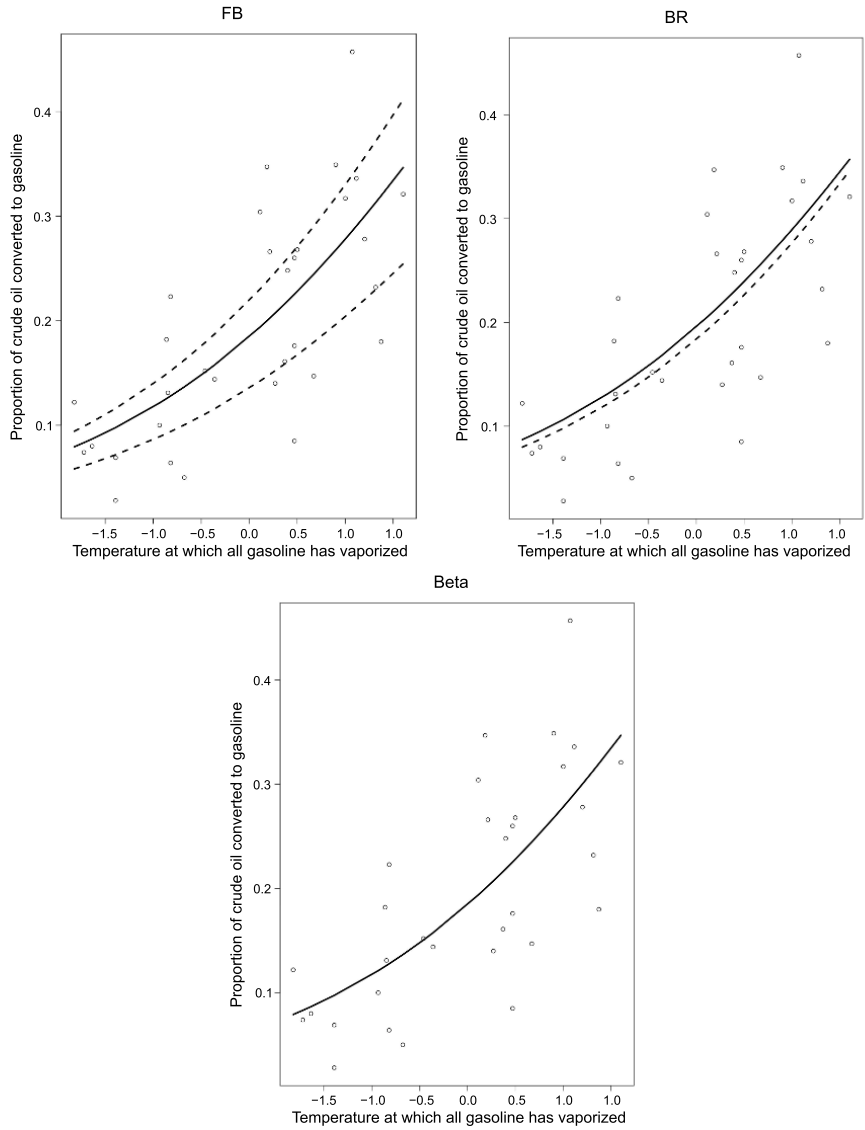


Figure 3.2. *From the left: fitted regression curves for the mean of the response variable (solid) and for the components λ_{1i} and λ_{2i} (dashed) under the FBR model; fitted regression curves for the mean of the response variable (solid) and for the group means (dashed, the uniform component 1/2 lies out of the plot) under the BRR model; fitted regression curve for the mean of the response variable under the beta regression model (solid)*

Our preliminary results show that the FB greatly expands the modeling potential of the beta, without demanding the theoretical and computational intricacy of a general beta mixture. In particular, the special mixture structure defining the FB ensures good theoretical properties that lead to computational tractability in terms of posterior computation.

At the same time, the FBR model displays easiness of interpretation. More specifically, the FBR directly models the overall regression mean μ as an arbitrary suitably chosen function of the covariates. This mean can be interpreted as weighted average of the two group regression means (one of which lies above the other for any value of the covariates), whose difference is represented by a further parameter w. The mixing weight p and a precision parameter ϕ complete its description. Interestingly, the model incorporates a form of heteroscedasticity natural for this type of data, since the response variance not only depends on ϕ, but also on μ.

From an applicative viewpoint, the illustrative example shows that the FBR model outperforms the beta and BRR models even when no clear evidence of bimodality is present. Though further analysis is required, this seems to be an indication of a broad applicability of the model.

Finally, it seems also worthwhile to stress the possibility that some of the parameters ϕ, p and w of the FB model may depend on covariates too. This greatly further expands its flexibility, enabling it to model fairly complex data patterns.

A deeper theoretical as well as applicative insight into the model is now available in the study of Migliorati *et al*. (2018).

3.7. References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, Springer, 199–213.

Albert, J. (2009). *Bayesian computation with R*. Springer Science & Business Media.

Bayes, C.L., Bazàn, J.L., Garcìa, C. (2012). A new robust regression model for proportions. *Bayesian Analysis*, 7(4), 841–866.

Branscum, A.J., Johnson, W.O., Thurmond, M.C. (2007). Bayesian beta regression: Applications to household expenditure data and genetic distance between foot-and-mouth disease viruses. *Australian & New Zealand Journal of Statistics*, 49(3), 287–301.

Brooks, S. (2002). Discussion on the paper by Spiegelhalter, Best, Carlin and Van der Linde.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1–38.

Ferrari, S., Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7), 799–815.

Ferrari, S.L., Espinheira, P.L., Cribari-Neto, F. (2011). Diagnostic tools in beta regression with varying dispersion, *Statistica Neerlandica*, 65(3), 337–351.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*, Springer Science & Business Media.

García, C., Pérez, J.G., van Dorp, J.R. (2011). Modeling heavy-tailed, skewed and peaked uncertainty phenomena with bounded support, *Statistical Methods & Applications*, 20(4), 463–486.

Gelfand, A.E., Smith, A.F. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, 85(410), 398–409.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2014). *Bayesian data analysis*, 3rd edition, Taylor & Francis, CRC Press.

Hahn, E.D. (2008), Mixture densities for project management activity times: A robust approach to {PERT}. *European Journal of Operational Research*, 188(2), 450–459.

Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D. (2000). Winbugs–A bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10(4), 325–337.

Markatou, M. (2000). Mixture models, robustness, and the weighted likelihood methodology. *Biometrics*, 56, 483–486.

McCullagh, P., Nelder, J.A. (1989). *Generalized linear models*, 37, CRC press.

Mengersen, K.L., Robert, C.P., Guihenneuc-Jouyaux, C. (1999). MCMC convergence diagnostics: A review. *Bayesian Statistics*, 6, 415–440.

Migliorati, S., Di Brisco, A.M., Ongaro, A. (2018). A new regression model for bounded responses, *Bayesian Anal.*, 13(3), 845–872, https://doi.org/10.1214/17-BA1079

Migliorati, S., Ongaro, A.,Monti, G.S. (2017). A structured dirichlet mixture model for compositional data: inferential and applicative issues. *Statistics and Computing*, 27(4), 963–983, http://dx.doi.org/10.1007/s11222-016-9665-y

Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS*. John Wiley & Sons, New York.

Ongaro, A., Migliorati, S. (2013). A generalization of the dirichlet distribution. *Journal of Multivariate Analysis*, 114, 412–426.

Paolino, P. (2001). Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis*, 9(4), 325–346.

Prater, N. (1956). Estimate gasoline yields from crudes. *Petroleum Refiner*, 35(5), 236–238.

R Core Team. (2016). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.

Smithson, M., Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables, *Psychological Methods*, 11(1), 54.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, 64(4), 583–639.

Tanner, M.A., Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398), 528–540.

Thomas, A. (1994). Bugs: A statistical modelling package. *RTA/BCS Modular Languages Newsletter*, 2, 36–38.

4

S-weighted Instrumental Variables

This chapter deals with two problems – with the situation when the orthogonality condition is broken and with the problem when an atypical data set contains a significant amount of information in a group of good leverage points but includes also a "troublesome" group of outliers.

Several robust methods were recently modified in order to overcome problem with the broken orthogonality condition, employing typically the idea of instrumental variables. In an analogous way, modified S-weighted estimator is also able to cope with broken orthogonality condition. We prove its consistency and we offer a small pattern of results of simulations.

It is believed that the bad leverage points are a more challenging problem in identification of underlying regression model than outliers. We show that sometimes outliers can also represent an intricate task.

4.1. Summarizing the previous relevant results

The median is the only classical statistic that is able to cope with high contamination, even 50%, and to give reasonable information about the location parameter of a data set. When Peter Bickel (Bickel 1975) opened the problem of possibility to construct an analogy of median in the framework of regression model, that is, an estimator of regression coefficients with 50% breakdown point, nobody had an idea how long and painful way to the solution we would have to go.

It seemed several times that we had achieved solution but finally always a bitter disappointment arrived. For instance, as the median is in fact the 50% quantile, we hoped that Koenker and Bassett's *regression quantiles* were the solution (Koenker and Bassett 1978). However, results by Maronna and Yohai (1981), establishing the maximal value of breakdown point of M-estimators, ruined our dreams.

Chapter written by Jan Ámos VÍŠEK.

Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining,
First Edition. Edited by Christos H. Skiadas and James R. Bozeman.

By proposing the *repeated median,* Siegel (Siegel 1982) has broken this long years lasting nightmare. But only proposals of the *least median of squares* (LMS) and the *least trimmed squares* (LTS) by Rousseeuw (Rousseeuw 1983, 1984) and (Hampel *et al.* 1986) brought feasible methods. In fact, he "rediscovered" the power of such statistical notion as the *order statistics of (squared) residuals* (see Hájek and Šidák 1967). Unfortunately, at those days we have not at hand a proper tool for studying the asymptotic properties of these estimators (the proof of consistency of LTS arrived after 20 years from its proposal (see Víšek 2006) and this technical problem was (except of others) an impulse for proposing S-estimator (Rousseeuw and Yohai 1984) with an immediately available proof of consistency and the simultaneous preservation of high breakdown point.

The algorithms for all these estimators were also successfully found. For LTS, it was based on repeated application of algorithm for the *ordinary least squares* and it was so simple that it was not published (as such, see Víšek 1990) until the moment when an improvement for large data set became inevitable (Číek and Víšek 2000, Hawkins 1994, Hawkins and Olive 1999). The algorithm for the S-estimator was a bit more complicated but feasible (see Campbell and Lopuhaa 1998).

Nevertheless, results by Hettmansperger and Sheater (1992), although they were wrong (due to the bad algorithm they used for LMS – for efficient algorithm, see Boček and Lachout 1993), they warned us that the situation need not be so simple as we had assumed. It led to a return to the order statistics of squared residuals and to the proposal of the *least weighted squares* (LWS) in Vsek (2000). It profited from extremely simple algorithm, basically the same as the algorithm for LTS (see Víšek 2006b), however, the study of its properties was tiresome and clumsy (see Víšek 2002). A significant simplification came with generalization of Kolmogorov–Smirnov result for the regression scheme (see Víšek 2011a), together with the fact that the rank of given order statistic is given by the value of empirical distribution function of these order statistics at given order statistic (see Víšek 2011b). It opened a way for defining an estimator covering all above-mentioned estimators as special cases – *S-weighted estimator* – and to describe its asymptotics (see Víšek 2015, 2016).

Due to the character of data in the social sciences, we can expect that the orthogonality condition is frequently broken. That was the reason why there are several attempts to modify the robust methods to be able to cope with the broken orthogonality condition, similarly as the *ordinary least squares* were "transformed" into the *instrumental variables* (e.g. Carroll and Stefanski 1994, Cohen-Freue *et al.* 2013, Desborges and Verardi 2012, Heckman *et al.* 2006, Víšek 1998, 2004, 2006a,b, 2017, or Wagenvoort and Waldmann 2002). This chapter offers a similar modification of the *S-weighted estimator,* which is able to cope with the broken orthogonality condition (*S-weighted instrumental variables*).

At the end of chapter, we answer to the problem whether the leverage points represent always more complicated problem than outliers. And the answer is bit surprising.

4.2. The notations, framework, conditions and main tool

Let $\mathcal{N}$ denote the set of all positive integers, R the real line and R^p the p-dimensional Euclidean space. All random variables are assumed to be defined on a basic probability space $(\Omega, \mathcal{A}, P)$. (We will not write – as the mathematical rigor would ask it – the random variable as $X(\omega)$ (say) but sometimes by including (ω) we emphasize the exact state of things.) For a sequence of $(p+1)$-dimensional random variables (r.v.'s) $\{(X_i', e_i)'\}_{i=1}^{\infty}$, for any $n \in \mathcal{N}$ and a fixed $\beta^0 \in R^p$ the linear regression model given as

$$Y_i = X_i'\beta^0 + e_i = \sum_{j=1}^{p} X_{ij}\beta_j^0 + e_i, \quad i = 1, 2, ..., n \qquad [4.1]$$

will be considered. (It is clear that the results of paper can be applied for the panel data – the model [4.1] will be used to keep the explanation as simple as possible.) We will need some conditions on the explanatory variables and the disturbances.

CONDITION $\mathcal{C}1$.– *The sequence* $\left\{(X_i', e_i)'\right\}_{i=1}^{\infty}$ *is sequence of independent* $p+1$*-dimensional random variables (r.v.'s) distributed according to distribution functions (d.f.)* $F_{X,e_i}(x,r) = F_{X,e}(x, r\sigma_i^{-1})$ *where* $F_{X,e}(x,r)$ *is a parent d.f. and* $\sigma_i^2 = \mathrm{var}\,(e_i)$. *Further,* $\mathbb{E}e_i = 0$ *and*

$$0 < \liminf_{i\to\infty} \sigma_i \leq \limsup_{i\to\infty} \sigma_i < \infty.$$

We denote $F_{e|X}(r|X_1 = x)$ *as the conditional d.f. corresponding to the parent d.f.* $F_{X,e}(x,r)$. *Then, for all* $x \in R^p$ $F_{e|X}(r|X_1 = x)$ *is absolutely continuous with density* $f_{e|X}(r|X_1 = x)$ *bounded by* U_e *(which does not depend on* x*).*

In what follows, $F_X(x)$ and $F_e(r)$ will denote the corresponding marginal d.f.s of the *parent d.f.* $F_{X,e}(x,r)$. Then, assuming that e is a *"parent" r.v.* distributed according to *parent d.f.* $F_e(r)$, we have $F_{e_i}(r) = P(e_i < r) = P(\sigma_i \cdot e < r) = P(e < \sigma_i^{-1} \cdot r) = F_e(\sigma_i^{-1} \cdot r)$, etc. Condition $\mathcal{C}1$ implies that the marginal d.f. $F_X(x)$ does not depend on i, that is, the sequence $\{X_i\}_{i=1}^{\infty}$ is sequence of independent and identically distributed (i.i.d.) r.v.'s.

Let, for any $\beta \in R^p$, $a_i = |Y_i - X_i'\beta|$ be absolute values of the ith residual and $F_{i,\beta}(v)$ its d.f., i.e. $F_{i,\beta}(v) = P(a_i(\beta) < v)$. Then put

$$\overline{F}_{\beta}^{(n)}(v) = \frac{1}{n}\sum_{i=1}^{n} F_{i,\beta}(v). \qquad [4.2]$$

Further, let $F_{\beta}^{(n)}(v)$ be the empirical distribution function (e.d.f.) of the absolute values of residuals, that is,

$$F_{\beta}^{(n)}(v) = \frac{1}{n}\sum_{i=1}^{n} I\left\{a_i(\beta) < v\right\}. \qquad [4.3]$$

It seems strange to consider the e.d.f. of a_i's, as they are heteroscedastic, but lemma 4.1 shows that it makes sense. Finally, let $a_{(1)} \leq a_{(2)} \leq ... \leq a_{(n)}$ denote the order statistics of absolute values of residuals and $\tilde{F}_{\beta}^{(n)}(v)$ be a continuous and strictly increasing modification of $F_{\beta}^{(n)}(v)$ defined as follows. Let $\tilde{F}_{\beta}^{(n)}(v)$ coincide with $F_{\beta}^{(n)}(v)$ at $a_i(\beta),\ i = 1, 2, ..., n$ and let it be continuous and strictly monotone between any pair of $a_{(i)}(\beta)$ and $a_{(i+1)}(\beta)$. Then it holds as follows.

LEMMA 4.1.– *Let condition $\mathcal{C}1$ hold. Then for any $\varepsilon > 0$, there is a constant K_ε and $n_\varepsilon \in \mathcal{N}$ so that for all $n > n_\varepsilon$*

$$P\left(\left\{\omega \in \Omega : \sup_{r \in R^+}\sup_{\beta \in R^p} \sqrt{n}\left|\tilde{F}_{\beta}^{(n)}(v) - \overline{F}_{\beta}^{(n)}(v)\right| < K_\varepsilon\right\}\right) > 1 - \varepsilon. \qquad [4.4]$$

The proof that employs Skorohod's embedding into Wiener process (see Breiman 1968) is a slight generalization of lemma 1 of Víšek (2011a) and it is based on the fact that $R^p \times R^+$ is separable space and that $\tilde{F}_{\beta}^{(n)}(v)$ is monotone.

Condition $\mathcal{C}2$ specifies the character of objective and weight functions.

CONDITION $\mathcal{C}2$.–

– $w : [0,1] \to [0,1]$ *is a continuous, non-increasing weight function with* $w(0) = 1$. *Moreover, w is Lipschitz in absolute value, i.e. there is L such that for any pair $u_1, u_2 \in [0,1]$ we have* $|w(u_1) - w(u_2)| \leq L \times \times |u_1 - u_2|$.

– $\rho : (0,\infty) \to (0,\infty)$, $\rho(0) = 0$, *non-decreasing on* $(0,\infty)$ *and differentiable (denote the derivative of ρ by ψ).*

– $\psi(v)/v$ *is non-increasing for* $v \geq 0$ *with* $\lim_{v \to 0_+} \frac{\psi(v)}{v} = 1$.

4.3. S-weighted estimator and its consistency

DEFINITION 4.1.– Let $w : [0,1] \to [0,1]$ and $\rho : [0,\infty] \to [0,\infty]$ be a weight function and an objective function, respectively. Then

$$\hat{\beta}^{(SW,n,w,\rho)} = \underset{\beta \in R^p}{\arg\min} \left\{ \sigma(\beta) \in R^+ : \frac{1}{n} \sum_{i=1}^{n} w\left(\frac{i-1}{n}\right) \rho\left(\frac{a_{(i)}(\beta)}{\sigma}\right) = b \right\} \quad [4.5]$$

where $b = I\!E_{F_e}\left\{w(F_{\beta^0}(|e|))\rho(e)\right\}$ is called the *S-weighted estimator* (see Víšek 2015).

REMARK 4.1.– Note that we cannot write [4.5] simply $\rho\left(\frac{a_i(\beta)}{\sigma}\right)$ because we would assign the weight $w\left(\frac{i-1}{n}\right)$ to other residual. (Let us recall that $\mathrm{var}_{F_e}(e) = 1$, so that the scale of e need not appear in the definition of b.)

Employing a slightly modified argument of Rousseeuw and Yohai (1984), we can show that $\hat{\beta}^{(SW,n,w,\rho)}$ has the solution

$$\sum_{i=1}^{n} w\left(\frac{i-1}{n}\right) X_{j_i} \psi\left(\frac{|Y_{j_i} - X'_{j_i}\beta|}{\hat{\sigma}_n}\right) \cdot \mathrm{sign}(Y_{j_i} - X'_{j_i}\beta) = 0$$

where j_i is the index of observation corresponding to $a_{(i)}$ and $\hat{\sigma}_n$ fulfills the constraint

$$\frac{1}{n} \sum_{i=1}^{n} w\left(\frac{i-1}{n}\right) \rho\left(\frac{a_{(i)}(\beta)}{\hat{\sigma}_n}\right) = b. \quad [4.6]$$

Then by following Hájek and Šidák (1967) and putting

$$\pi(\beta, j) = i \in \{1, 2, ..., n\} \quad \Leftrightarrow \quad a_j(\beta) = a_{(i)}(\beta), \quad [4.7]$$

we arrive at

$$\sum_{i=1}^{n} w\left(\frac{\pi(\beta,i)-1}{n}\right) X_i \psi\left(\frac{|Y_i - X'_i\beta|}{\hat{\sigma}_n}\right) \cdot \mathrm{sign}(Y_i - X'_i\beta) = 0 \quad [4.8]$$

and utilizing the equality $n^{-1}\left(\pi(\beta,i) - 1\right) = F_\beta^{(n)}(a_i(\beta))$ (see Víšek 2011b), we finally obtain

$$\sum_{i=1}^{n} w\left(F_\beta^{(n)}(a_i(\beta))\right) X_i \psi\left(\frac{|Y_i - X'_i\beta|}{\hat{\sigma}_n}\right) \cdot \mathrm{sign}(Y_i - X'_i\beta) = 0. \quad [4.9]$$

Then the fact that $\psi(0) = 0$ allows us to write the normal equation [4.8] as

$$\sum_{\{i\,:\,r_i(\beta)\neq 0\}} w\left(F_\beta^{(n)}(a_i(\beta))\right)\left[\psi\left(\frac{a_i(\beta)}{\hat{\sigma}_n}\right)\cdot\frac{\hat{\sigma}_n}{a_i(\beta)}\right]X_i\left(Y_i - X_i'\beta\right)$$

$$= \sum_{i=1}^{n} \tilde{w}\left(\tilde{F}_\beta^{(n)}(a_i(\beta)), \hat{\sigma}_n\right) X_i\left(Y_i - X_i'\beta\right) = 0. \qquad [4.10]$$

Note that if w and ρ fulfill condition $\mathcal{C}2$, then $\tilde{w}$ is well defined and it also fulfills $\mathcal{C}2$ for any fixed $\sigma > 0$. Note also that [4.10] coincides with the *normal equations* of the LWS only if $\rho(v) = v^2$ compared with (Víšek 2011b). Otherwise, first, $\tilde{w}(v)$ is implicitly modified by $\psi(v)$ and second, $\tilde{w}(v)$ depends also on $\hat{\sigma}_n$. As the S-weighted estimator controls the influence of residuals by the weight and objective functions, the Euclidean metrics is substituted by a Riemannian one and the consequence is that – contrary to the *ordinary least squares* – we need an identification condition.

CONDITION $\mathcal{C}3$.– *There is the only solution of the equation*

$$\beta' I\!E\left[\sum_{i=1}^{n} w\left(\overline{F}_\beta^{(n)}(|e_i|)\right)\cdot X_i\cdot\psi(e_i - X_i'\beta)\right] = 0, \qquad [4.11]$$

(for $\tilde{w}$, see [4.9]) at $\beta = \beta^0$.

REMARK 4.2.– Note that [4.11] is for the classical *ordinary least squares* fulfilled because $\tilde{w} \equiv 1$. Similarly, it can be shown that when w is zero-one function and ρ is quadratic function (as for the LTS) that [4.11] also holds but in that case it is technically rather complicated (see (Víšek 2006)).

THEOREM 4.1.– *Let conditions $\mathcal{C}1$, $\mathcal{C}2$ and $\mathcal{C}3$ be fulfilled and $\hat{\sigma}_n$ be a weakly consistent estimator of $\mathrm{var}_{F_e}(e)$ fulfilling the constraint [4.6]. Then any sequence $\left\{\hat{\beta}^{(SW,w,\rho,\hat{\sigma}_n,n)}\right\}_{n=1}^{\infty}$ of the solutions of sequence of normal equations [4.9] for $n = 1, 2, ...,$ is weakly consistent.*

The proof is a slight generalization of the proof of theorem 1 from Víšek (2015).

4.4. S-weighted instrumental variables and their consistency

Due to Euclidean geometry, the solution of the extremal problem that defines the *ordinary least squares*, namely

$$\hat{\beta}^{(OLS,n)} = \underset{\beta \in R^p}{\arg\min} \sum_{i=1}^{n} \left(Y_i - X_i'\beta\right)^2, \qquad [4.12]$$

is given as the solution of *normal equations*

$$\sum_{i=1}^{n} X_i \left(Y_i - X_i'\beta\right) = 0. \qquad [4.13]$$

Having performed a straightforward algebra and the substitution from [4.1], we arrive at

$$\hat{\beta}^{(OLS,n)} = \beta^0 + \left(\frac{1}{n}\sum_{i=1}^{n} [X_i \cdot X_i']\right)^{-1} \cdot \frac{1}{n}\sum_{i=1}^{n} [X_i e_i] \qquad [4.14]$$

which indicates that if the *orthogonality condition* is broken, i.e. $\mathbb{E}[X_1 \cdot e] \neq 0$ (for e, see condition $\mathcal{C}1$), $\hat{\beta}^{(OLS,n)}$ is biased and inconsistent. Then we look for some instrumental variables $\{Z_i\}_{i=1}^{\infty}$, usually i.i.d. r.v.'s, such that $\mathbb{E}Z_1 = 0$, $\mathbb{E}[Z_1 \cdot Z_1']$ positive definite matrix, $\mathbb{E}[Z_1 \cdot e] = 0$ and define the estimator by *means of the instrumental variables* (IV) as the solution of the *normal equations*

$$\sum_{i=1}^{n} Z_i \left(Y_i - X_i'\beta\right) = 0. \qquad [4.15]$$

(An alternative way how to cope with the broken *orthogonality condition* is to utilize the *orthogonal regression* – sometimes called the *total least squares*, e.g. Paige and Strako 2002). There are several alternative ways to define the *instrumental variables* – see (Víšek 2017) and references given there – but all of them are practically equivalent to [4.15]; for the discussion which summarizes also geometric background of the *instrumental variables*, see again Víšek (2017). To prove the unbiasedness and consistency of classical instrumental variables, we do not need (nearly) any additional assumptions except of those which are given several lines above [4.15].

DEFINITION 4.2.– Let $\{Z_i\}_{i=1}^{\infty}$ be a sequence of i.i.d. r.v.'s, such that $I\!E Z_1 = 0$, $I\!E[Z_1 \cdot Z_1']$ positive definite matrix, $I\!E[Z_1 \cdot e] = 0$. The solution of the *normal equation*

$$I\!N E_{Y,X,Z}^{(\tilde{w},\rho,\hat{\sigma}_n,n)}(\beta) = \sum_{i=1}^{n} \tilde{w}\left(\tilde{F}_{\beta}^{(n)}(a_i(\beta)), \hat{\sigma}_n\right) Z_i\left(Y_i - X_i'\beta\right) = 0 \qquad [4.16]$$

will be called the estimator by *means of the S-weighted instrumental variables* (briefly, the *S-weighted instrumental variables*) and denoted by $\hat{\beta}^{(SWIV,w,\rho,\hat{\sigma}_n,n)}$.

To be able to prove the consistency of $\hat{\beta}^{(SWIV,w,\rho,\hat{\sigma}_n,n)}$, we will need some additional assumptions and an identification condition, similar to condition $\mathcal{C}3$. We will start with an enlargement of notations.

Let for any $\beta \in R^p$ and $u \in R$ $F_{\beta' Z X' \beta}(u) = P(\beta' Z_1 X_1' \beta < u)$ and $F_{\beta' Z X' \beta}^{(n)}(u) = \frac{1}{n}\sum_{i=1}^{n} I\{\beta' Z_i(\omega) X_i'(\omega)\beta < u\}$ be the d.f. of $\beta' Z_1 X_1' \beta$ and e.d.f. of $\{\beta' Z_i(\omega) X_i'(\omega)\beta\}_{i=1}^{n}$, respectively. Further, for any $\lambda \in R^+$ and any $a \in R$ put

$$\begin{aligned} \gamma_{\lambda,a} &= \sup_{\|\beta\|=\lambda} F_{\beta' Z X' \beta}(a) \text{ and } \tau_{\lambda} \\ &= -\inf_{\|\beta\|\leq\lambda} \beta' I\!E\left[Z_1 X_1' \cdot I\{\beta' Z_1 X_1' \beta < 0\}\right]\beta. \end{aligned} \qquad [4.17]$$

CONDITION $\mathcal{C}4$.– *The instrumental variables $\{Z_i\}_{i=1}^{\infty}$ are independent and identically distributed with distribution function $F_Z(z)$. Further, the joint distribution function $F_{X,Z}(x,z)$ is absolutely continuous with a density $f_{X,Z}(x,z)$ bounded by $U_{ZX} < \infty$. Further for any $n \in \mathcal{N}$, we have $I\!E\sum_{i=1}^{n}\{w(F_{\beta^0}(|e_i|))\times \ \times\psi(e_i)\cdot Z_i\} = 0$ and the matrices $I\!E Z_1 Z_1'$ as well as $I\!E\sum_{i=1}^{n}\{w(F_{\beta^0}(|e_i|))\times \ \times\psi(e_i)\cdot Z_i X_i'\}$ are positive definite. Moreover, there is $q > 1$ so that $I\!E\{\|Z_1\| \times \ \times \|X_1\|\}^q < \infty$. Finally, there is $a > 0$, $b \in (0,1)$ and $\lambda > 0$ so that*

$$a \cdot (b - \gamma_{\lambda,a}) \cdot \tilde{w}(b) > \tau_{\lambda} \qquad [4.18]$$

for $\gamma_{\lambda,a}$ and τ_{λ} given by [4.17].

LEMMA 4.2.– *Let conditions $\mathcal{C}1$, $\mathcal{C}2$, $\mathcal{C}4$ be fulfilled and $\hat{\sigma}_n$ be a weakly consistent estimator of $\mathrm{var}_{F_e}(e)$ fulfilling the constraint [4.6]. Then for any $\varepsilon > 0$, there is $\zeta > 0$ and $\delta > 0$ such that*

$$P\left(\left\{\omega \in \Omega : \inf_{\|\beta\|\geq\zeta} -\frac{1}{n}\beta' I\!N E_{Y,X,Z}^{(\tilde{w},\rho,\hat{\sigma}_n,n)}(\beta) > \delta\right\}\right) > 1 - \varepsilon.$$

In other words, any sequence $\left\{\hat{\beta}^{(SWIV,\tilde{w},\rho,\hat{\sigma}_n,n)}\right\}_{n=1}^{\infty}$ *of the solutions of the sequence of normal equations [4.16]* $I\!N E_{Y,X,Z}^{(\tilde{w},\rho,\hat{\sigma}_n,n)}(\beta) = 0$ *is bounded in probability.*

The proof is formally nearly the same as the proof of lemma 1 in Vísek (2009). The allowance for the heteroscedasticity of disturbances requires some *formally* straightforward modifications. The fact that the modifications are relatively simple and straightforward is due to the fact that the complicated steps were made in (Víšek 2011b) but the background of proof is different from the proof in (Víšek 2009). The approximation of empirical d.f. is not by the underlying d.f. as the limit of the empirical d.f.'s but we employ the knowledge about convergence of the difference of the empirical d.f.'s and the arithmetic mean of the d.f.'s of individual disturbances (see lemma 4.1) □

LEMMA 4.3.– *Let conditions* $\mathcal{C}1$, $\mathcal{C}2$ *and* $\mathcal{C}4$ *be fulfilled and* $\hat{\sigma}_n$ *be a weakly consistent estimator of* $\mathrm{var}_{F_e}(e)$ *fulfilling the constraint [4.6]. Then for any* $\varepsilon > 0$, $\delta \in (0,1)$ *and* $\zeta > 0$ *there is* $n_{\varepsilon,\delta,\zeta} \in \mathcal{N}$ *so that for any* $n > n_{\varepsilon,\delta,\zeta}$ *we have*

$$P\left(\left\{\omega\in\Omega\ :\ \sup_{\|\beta\|\le\zeta}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\tilde{w}(\tilde{F}_{\beta}^{(n)}(a_i(\beta)),\hat{\sigma}_n)\beta' Z_i\,(e_i - X_i'\beta)\right.\right.\right.$$

$$\left.\left.\left.-\beta' I\!E\left[\tilde{w}(\overline{F}_{\beta}^{(n)}(a_i(\beta)),\hat{\sigma}_n)Z_i\,(e_i - X_i'\beta)\right]\right\}\right| < \delta\right\}\right) > 1-\varepsilon$$

(for $a_i(\beta)$*, see a line above [4.2] and for* $\overline{F}_{\beta}^{(n)}(v)$*, see [4.2]).*

The proof has formally similar structure as the proof of lemma 2 in Vísek (2009). It is a bit more complicated because instead of employing a limiting distribution, we need to estimate differences of empirical d.f. of $a_i(\beta)$'s from a sequence of the arithmetic means of underlying d.f.'s $\left\{\overline{F}_{\beta}^{(n)}(v)\right\}_{n=1}^{\infty}$ (see [4.2]). □

LEMMA 4.4.– *Let conditions* $\mathcal{C}1$, $\mathcal{C}2$ *and* $\mathcal{C}3$ *hold and* $\hat{\sigma}_n$ *be a weakly consistent estimator of* $\mathrm{var}_{F_e}(e)$ *fulfilling the constraint [4.6]. Then for any positive* ζ

$$\beta' I\!E\left[\tilde{w}\left(F_{\beta}(a_i(\beta)),\hat{\sigma}_n\right) Z_i\,(e_i - X_i'\beta)\right] \qquad [4.19]$$

(for $\tilde{w}$*, see [4.10]) is uniformly in* $i \in \mathcal{N}$*, uniformly continuous in* β *on* $\mathcal{B} = \{\beta\in R^p : \|\beta\| \le \zeta\}$*, i.e. for any* $\varepsilon > 0$ *there is* $\delta > 0$ *so that for any pair of vectors* $\beta^{(1)}, \beta^{(2)} \in R^p$, $\left\|\beta^{(1)} - \beta^{(2)}\right\| < \delta$ *we have*

$$\sup_{i\in\mathcal{N}}\left|\left[\beta^{(1)}\right]' I\!E\left[\tilde{w}\left(F_{\beta^{(1)}}(a_i(\beta^{(1)})),\hat{\sigma}_n\right) Z_i\left(e_i - X_i'\beta^{(1)}\right)\right]\right.$$

$$- \left[\beta^{(2)}\right]' I\!E \left[\tilde{w}\left(F_{\beta^{(2)}}(a_i(\beta^{(2)})), \hat{\sigma}_n\right) Z_i \left(e_i - X_i'\beta^{(2)}\right)\right] \Bigg| < \varepsilon.$$

The proof is a chain of approximations utilizing simple estimates of upper bounds of differences of the values of [4.19] for close pair of points in R^p. □

Similarly as for the *S-weighted estimator,* we need for the *S-weighted instrumenal variables* the identification condition.

CONDITION $\mathcal{C}4$.– *For any* $n \in \mathcal{N}$, *the equation*

$$\beta' \sum_{i=1}^{n} I\!E \left[w\left(\overline{F}_{\beta}^{(n)}(|e_i|)\right) \cdot Z_i \cdot \psi(e_i - X_i'\beta)\right] = 0 \qquad [4.20]$$

in the variable $\beta \in R^p$ *has a unique solution at* $\beta = \beta^0$.

THEOREM 4.2.– *Let conditions* $\mathcal{C}1$, $\mathcal{C}2$, $\mathcal{C}3$ *and* $\mathcal{C}4$ *be fulfilled and* $\hat{\sigma}_n$ *be a weakly consistent estimator of* $\mathrm{var}_{F_e}(e)$ *fulfilling the constraint [4.6]. Then any sequence* $\left\{\hat{\beta}^{(SWIV,w,\rho,\hat{\sigma}_n,n)}\right\}_{n=1}^{\infty}$ *of the solutions of normal equations [4.16]* $I\!N E_{Y,X,Z}^{(\tilde{w},\rho,\hat{\sigma}_n,n)}(\beta) = 0$ *is weakly consistent.*

PROOF.– Without loss of generality assume that $\beta^0 = 0$ (as $\hat{\sigma}^{(SWIV,w,\rho,\hat{\sigma}_n,n)}$ is scale and regression equivariant). To prove the consistency, we have to show that for any $\varepsilon > 0$ and $\delta > 0$, there is $n_{\varepsilon,\delta} \in \mathcal{N}$ such that for all $n > n_{\varepsilon,\delta}$

$$P\left(\left\{\omega \in \Omega : \left\|\hat{\beta}^{(SWIV,\tilde{w},\rho,\hat{\sigma}_n,n)} - \beta^0\right\| < \delta\right\}\right) > 1 - \varepsilon. \qquad [4.21]$$

So fix $\varepsilon_1 > 0$ and $\delta_1 > 0$. According to lemma 4.2, $\delta_1 > 0$ and $\theta_1 > 0$, so that for ε_1 there is $n_{\delta_1,\varepsilon_1} \in \mathcal{N}$; for any $n > n_{\delta_1,\varepsilon_1}$

$$P\left(\left\{\omega \in \Omega : \inf_{\|\beta\| \geq \theta_1} -\frac{1}{n}\beta' I\!N E_{Y,X,Z}^{(\tilde{w},\rho,\hat{\sigma}_n,n)}(\beta) > \delta_1\right\}\right) > 1 - \frac{\varepsilon_1}{2}$$

(denote the corresponding set by B_n). It means that for all $n > n_{\delta_1,\varepsilon_1}$, all solutions of the normal equations [4.16] $I\!N E_{Y,X,Z}^{(\tilde{w},\rho,\hat{\sigma}_n,n)}(\beta) = 0$ are inside the ball $\mathcal{B}(0,\theta_1)$ with probability at least $1 - \frac{\varepsilon_1}{2}$. If $\theta_1 \leq \delta$, we have finished the proof. Generally, of course, we can have $\theta_1 > \delta$.

Then, using lemma 4.3 we may find for ε_1, $\delta = \min\{\frac{\delta_1}{2}, \delta_1\}$ and θ_1 such $n_{\varepsilon_1,\delta,\theta_1} \in \mathcal{N}$, $n_{\varepsilon_1,\delta,\theta_1} \geq n_{\delta_1,\varepsilon_1}$, so that for any $n > n_{\varepsilon_1,\delta,\theta_1}$, there is a set C_n (with $P(C_n) > 1 - \frac{\varepsilon}{2}$) such that for any $\omega \in C_n$

$$\sup_{\|\beta\| \leq \theta_1} \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \tilde{w}\left(\tilde{F}_{\beta}^{(n)}(a_i(\beta)), \hat{\sigma}_n\right) \beta' Z_i \left(e_i - X_i'\beta\right) \right.\right.$$

$$\left.\left. -\beta' I\!E\left[\tilde{w}\left(\tilde{F}_{\beta}(a_i(\beta)), \hat{\sigma}_n\right) Z_i \left(e_i - X_i'\beta\right)\right] \right\} \right| < \delta.$$

But it means that

$$\inf_{\|\beta\|=\theta_1} \left\{ -\beta' \frac{1}{n} \sum_{i=1}^{n} I\!E\left[\tilde{w}\left(\tilde{F}_{\beta}(a_i(\beta)), \hat{\sigma}_n\right) Z_i \left(e_i - X_i'\beta\right)\right] \right\} > \frac{\delta_1}{2} > 0. \qquad [4.22]$$

Further consider the compact set $C = \{\beta \in R^p : \delta_1 \leq \|\beta\| \leq \theta_1\}$ and find

$$\tau_C = \inf_{\beta \in C} \left\{ -\beta' \frac{1}{n} \sum_{i=1}^{n} I\!E\left[\tilde{w}\left(\tilde{F}_{\beta}(a_i(\beta)), \hat{\sigma}_n\right) Z_i \left(e_i - X_i'\beta\right)\right] \right\}. \qquad [4.23]$$

Then there is a $\{\beta_k\}_{k=1}^{\infty}$ such that

$$\lim_{k\to\infty} \beta_k' \frac{1}{n} \sum_{i=1}^{n} I\!E\left[\tilde{w}\left(\tilde{F}_{\beta_k}(a_i(\beta_k)), \hat{\sigma}_n\right) Z_i \left(e_i - X_i'\beta_k\right)\right] = -\tau_C.$$

On the other hand, due to compactness of C, there is a β^* and a subsequence $\left\{\beta_{k_j}\right\}_{j=1}^{\infty}$ such that

$$\lim_{j\to\infty} \beta_{k_j} = \beta^*$$

and due to the uniform continuity (uniform in $i \in \mathcal{N}$ as well as in $\beta \in C$) of $\beta' I\!E\left[\tilde{w}\left(\tilde{F}_{\beta}(a_i(\beta)), \hat{\sigma}_n\right) Z_i \left(e_i - X_i'\beta\right)\right]$ (see lemma 4.4), we have

$$-\left[\beta^*\right]' \frac{1}{n} \sum_{i=1}^{n} I\!E\left[\tilde{w}\left(\tilde{F}_{\beta^*}(a_i(\beta^*)), \hat{\sigma}_n\right) Z_i \left(e_i - X_i'\beta^*\right)\right] = \tau_C. \qquad [4.24]$$

Employing once again the uniform continuity (uniform in $i \in \mathcal{N}$ and $\beta \in C$) of $\beta' I\!E\left[\tilde{w}\left(F_\beta(a_i(\beta)), \hat{\sigma}_n\right) Z_i\left(e_i - X_1'\beta\right)\right]$ together with condition $\mathcal{C}4$ and [4.22] we find that $\tau_C > 0$, otherwise there has to be a solution of [4.20] inside the compact C, which does not contain $\beta = 0$.

Now, using lemma 4.3 once again we may find for ε_1, δ_1, θ_1 and τ_C $n_{\varepsilon_1,\delta_1,\theta_1,\tau_C} \in \mathcal{N}, n_{\varepsilon_1,\delta_1,\theta_1,\tau_C} \geq n_{\varepsilon_1,\delta,\theta_1}$, so that for any $n > n_{\varepsilon_1,\delta_1,\theta_1,\tau_C}$ there is a set D_n (with $P(D_n) > 1 - \frac{\varepsilon}{2}$) such that for any $\omega \in D_n$

$$\sup_{\|\beta\| \leq \theta_1} \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \tilde{w}\left(\tilde{F}_\beta^{(n)}(a_i(\beta)), \hat{\sigma}_n\right) \beta' Z_i\left(e_i - X_i'\beta\right) \right.\right.$$

$$\left.\left. -\beta' I\!E\left[\tilde{w}\left(\tilde{F}_\beta(a_i(\beta)), \hat{\sigma}_n\right) Z_i\left(e_i - X_i'\beta\right)\right] \right\} \right| < \frac{\tau_C}{2}. \qquad [4.25]$$

But [4.23] and [4.25] imply that for any $n > n_{\varepsilon_1,\delta_1,\theta_1,\tau_C}$ and any $\omega \in B_n \cap D_n$ we have

$$\inf_{\|\beta\| > \delta_1} -\frac{1}{n} \beta' I\!N E_{Y,X,Z}^{(\tilde{w},\rho,\hat{\sigma}_n,n)}(\beta) > \frac{\tau_C}{2}. \qquad [4.26]$$

Of course, $P\left(B_n \cap D_n\right) > 1 - \varepsilon_1$. But it means that all solutions of normal equations [4.16] are inside the ball of radius δ_1 with probability at least $1 - \varepsilon_1$, i.e. in other words, $\hat{\beta}^{(SWIV,w,\rho,\hat{\sigma}_n,n)}$ is weakly consistent. □

4.5. Patterns of results of simulations

In the simulations, we compared *S-weighted instrumental variables* with classical *instrumental variables* (which is not robust) and with three other robust versions of *instrumental variables*, namely *instrumental weighted variables* (see Víšek 2017), *S-instrumental variables* and *W-instrumental variables* (see Cohen-Freue *et al.* 2013, Desborges and Verardi 2012; unfortunately the description of these estimators would require rather large space, so we only refer to original papers). The best results from these three alternative estimators were achieved by the *S-instrumental variables* and *instrumental weighted variables*, and we decided to report, in Tables 4.1, 4.2 and 4.3, *S-instrumental variables* (the lack of space has not allowed us to present more).

4.5.1. *Generating the data*

The data were generated for $i = 1, 2, .., n,\ t = 1, 2, ..., T$ according to the model

$$Y_{it} = 1 - 2 \cdot X_{it1} + 3 \cdot X_{it2} - 4 \cdot X_{it3} + 5 \cdot X_{it4} + \sigma_{it} \cdot e_{it},$$

with $X_{it+1} = 0.9 \cdot X_{it} + 0.1 \cdot v_{it} + 0.5 \cdot e_{it}$ where the initial value $\{X_{i1}\}_{i=1}^{n}$, the innovations $\{v_{it}\}_{i=1,t=1}^{n,T}$ and the disturbances $\{e_{it}\}_{i=1,t=1}^{n,T}$ were i.id. four dimensional normal vectors with the zero means and the unit covariance matrix. Sequence $\{\sigma_{it}\}_{i=1,t=1}^{n,T}$ is i.i.d., distributed uniformly over $[0.5, 5.5]$. In the role of the objective function, we have employed Tukey's ρ given for some $c > 0$ as

$$\begin{aligned} \rho_c(x) &= \frac{x^2}{2} - \frac{x^4}{2 \cdot c^2} + \frac{x^6}{6 \cdot c^4} \quad \text{for } abs(x) \leq c, \\ &= \frac{c^2}{6} \qquad\qquad\qquad\qquad \text{otherwise.} \end{aligned}$$

For $0 < h < g < 1$, the weight function $w(r) : [0, 1] \to [0, 1]$ is equal to 1 for $0 \leq r \leq h$, it is equal to 0 for $g \leq r \leq 1$ and it decreases from 1 to 0 for $h \leq r \leq g$, i.e. putting $c = g - h$ and $y = g - r$, we compute

$$w(r) = 3\frac{y^2}{c^2} - 3\frac{y^4}{c^4} + \frac{y^6}{c^6}, \qquad [4.27]$$

i.e. between h and g the weight function borrowed the shape from Tukey's ρ.

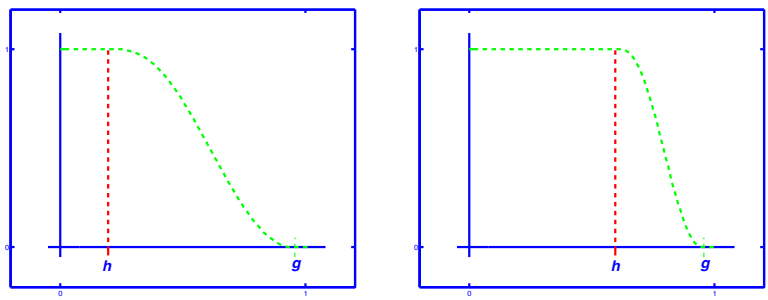


Figure 4.1. *The examples of possible shapes of weight function. For a color version of this figure, see www.iste.co.uk/skiadas/data1.zip*

The data were contaminated so that we selected randomly one block (i.e. one $t \in \{1, 2, ..., T\}$) and either the bad leverage points were created as $X^{(new)} = 5 \cdot X^{(original)}$ and $Y^{(wrong)} = -Y^{(correct)}$ or the outliers were created as $Y^{(wrong)} = -3 \cdot Y^{(correct)}$. The data contained the same number of good leverage points $X^{(new)} = 20 \cdot X^{(original)}$ (with the response Y calculated correctly) as bad leverage points.

4.5.2. *Reporting the results*

We have generated 500 sets, each containing $n \cdot T$ observations (it is specified in heads of tables) and then we calculated the estimates

$$\left\{\hat{\beta}^{(index,k)} = (\hat{\beta}_1^{(index,k)}, \hat{\beta}_2^{(index,k)}, \hat{\beta}_3^{(index,k)}, \hat{\beta}_4^{(index,k)}, \hat{\beta}_5^{(index,k)})'\right\}_{k=1}^{500} \quad [4.28]$$

where the abbreviations IV, SIV and $SWIV$ at the position of "*index*" indicate the method employed for the computation, namely IV – for the *instrumental variables*, SIV – for *S-instrumental variables estimator* and finally $SWIV$ – for *S-weighted instrumental variables estimator*. The empirical means and the empirical mean squared errors (MSE) of estimates of coefficients (over these 500 repetitions) were computed, i.e. we report values (for $j = 1, 2, 3, 4$ and 5)

$$\begin{aligned} \hat{\beta}_j^{(index)} &= \frac{1}{500}\sum_{k=1}^{500} \hat{\beta}_j^{(index,k)} \text{ and } \widehat{\text{MSE}}\left(\hat{\beta}_j^{(index)}\right) \\ &= \frac{1}{500}\sum_{k=1}^{500}\left[\hat{\beta}_j^{(index,k)} - \beta_j^0\right]^2 \end{aligned} \quad [4.29]$$

where $\beta^0 = [1, -2, 3, -4, 5]'$ and the index have the same role as above. The results are given in tables in the form as follows: the first cell of each row indicates the method, e.g. $\hat{\beta}^{(IV)}$, the next five cells contain then just $\hat{\beta}^{(IV)}$ $\left(\widehat{MSE}(\hat{\beta}^{(IV)})\right)$ for the first, the second up to the fifth coordinate.

As discussed previously, it is believed that the leverage points are more complicated problem than outliers. Table 4.3 offers results indicating that the "classical" estimators as the LMS, the LTS or the *S-estimator* can exhibit a problem when data contain a group of good leverage points (far away from the main bulk of data) and some outliers (not very far from the bulk of data). As the mean squared errors of the *S-estimates* below indicate that the *S-estimator* have used the information in data less efficiently than *S-weighted estimator* (see [4.5]). (Due to the lack of space we present only the results for the *S-estimator* – which were the best among the "classical" estimators (LMS, LTS, LWS and *S-estimator*). The reason for large MSE of the *S-estimates* is the depression of the information brought by good leverage points. It happened due to the implicit estimation of variance of disturbances.

$\boldsymbol{T=1,\quad n\cdot T=100,\quad h=0.98,\quad g=0.99}$					
$\hat{\beta}^{(IV)}_{(MSE)}$	$0.970_{(0.372)}$	$-1.924_{(0.375)}$	$2.835_{(0.419)}$	$-3.781_{(0.429)}$	$4.706_{(0.479)}$
$\hat{\beta}^{(SIV)}_{(MSE)}$	$0.993_{(0.105)}$	$-1.986_{(0.133)}$	$2.979_{(0.141)}$	$-4.021_{(0.151)}$	$4.987_{(0.142)}$
$\hat{\beta}^{(SWIV)}_{(MSE)}$	$0.992_{(0.106)}$	$-1.990_{(0.105)}$	$3.002_{(0.122)}$	$-4.000_{(0.120)}$	$4.992_{(0.105)}$

$\boldsymbol{T=2,\quad n\cdot T=200,\quad h=0.98,\quad g=0.99}$					
$\hat{\beta}^{(IV)}_{(MSE)}$	$0.966_{(0.319)}$	$-1.873_{(0.404)}$	$2.814_{(0.378)}$	$-3.808_{(0.369)}$	$4.690_{(0.605)}$
$\hat{\beta}^{(SIV)}_{(MSE)}$	$0.993_{(0.056)}$	$-2.004_{(0.082)}$	$3.007_{(0.077)}$	$-4.017_{(0.071)}$	$4.984_{(0.084)}$
$\hat{\beta}^{(SWIV)}_{(MSE)}$	$0.993_{(0.059)}$	$-1.997_{(0.069)}$	$3.009_{(0.061)}$	$-4.002_{(0.058)}$	$4.992_{(0.068)}$

$\boldsymbol{T=3,\quad n\cdot T=300,\quad h=0.98,\quad g=0.99}$					
$\hat{\beta}^{(IV)}_{(MSE)}$	$0.982_{(0.259)}$	$-1.879_{(0.323)}$	$2.795_{(0.363)}$	$-3.734_{(0.453)}$	$4.678_{(0.532)}$
$\hat{\beta}^{(SIV)}_{(MSE)}$	$1.002_{(0.037)}$	$-2.017_{(0.050)}$	$2.995_{(0.057)}$	$-4.009_{(0.057)}$	$4.989_{(0.058)}$
$\hat{\beta}^{(SWIV)}_{(MSE)}$	$0.999_{(0.039)}$	$-2.006_{(0.041)}$	$2.990_{(0.050)}$	$-3.989_{(0.046)}$	$4.995_{(0.047)}$

$\boldsymbol{T=4,\quad n\cdot T=400,\quad h=0.98,\quad g=0.99}$					
$\hat{\beta}^{(IV)}_{(MSE)}$	$0.961_{(0.213)}$	$-1.887_{(0.280)}$	$2.863_{(0.290)}$	$-3.764_{(0.403)}$	$4.743_{(0.380)}$
$\hat{\beta}^{(SIV)}_{(MSE)}$	$0.995_{(0.027)}$	$-2.022_{(0.046)}$	$2.986_{(0.052)}$	$-4.017_{(0.047)}$	$4.981_{(0.049)}$
$\hat{\beta}^{(SWIV)}_{(MSE)}$	$0.994_{(0.029)}$	$-2.013_{(0.038)}$	$2.992_{(0.042)}$	$-4.014_{(0.038)}$	$4.986_{(0.036)}$

$\boldsymbol{T=5,\quad n\cdot T=500,\quad h=0.98,\quad g=0.99}$					
$\hat{\beta}^{(IV)}_{(MSE)}$	$0.964_{(0.194)}$	$-1.859_{(0.360)}$	$2.806_{(0.393)}$	$-3.781_{(0.334)}$	$4.717_{(0.407)}$
$\hat{\beta}^{(SIV)}_{(MSE)}$	$1.003_{(0.025)}$	$-2.007_{(0.042)}$	$2.995_{(0.041)}$	$-4.006_{(0.042)}$	$4.997_{(0.045)}$
$\hat{\beta}^{(SWIV)}_{(MSE)}$	$1.002_{(0.025)}$	$-2.006_{(0.032)}$	$2.991_{(0.033)}$	$-4.000_{(0.033)}$	$5.004_{(0.036)}$

Table 4.1. *The contamination by leverage points on the level of 1%, $n = 100$. The values of variance of the disturbances randomly selected from $[0.5, 5.5]$*

$\boldsymbol{T=5,\quad n\cdot T=100, h=0.940, g=0.948}$					
$\hat{\beta}^{(IV)}_{(MSE)}$	$0.879_{(4.420)}$	$-1.505_{(6.863)}$	$2.335_{(6.609)}$	$-3.096_{(7.212)}$	$3.730_{(7.995)}$
$\hat{\beta}^{(SIV)}_{(MSE)}$	$0.992_{(0.662)}$	$-1.953_{(0.946)}$	$2.824_{(1.243)}$	$-3.920_{(1.017)}$	$4.672_{(2.158)}$
$\hat{\beta}^{(SWIV)}_{(MSE)}$	$0.982_{(0.178)}$	$-1.982_{(0.336)}$	$2.981_{(0.349)}$	$-4.018_{(0.296)}$	$4.954_{(0.362)}$

$\boldsymbol{T=10,\quad n\cdot T=200,\quad h=0.940,\quad g=0.948}$					
$\hat{\beta}^{(IV)}_{(MSE)}$	$0.862_{(2.967)}$	$-1.604_{(3.871)}$	$2.548_{(3.839)}$	$-3.227_{(4.665)}$	$4.011_{(5.258)}$
$\hat{\beta}^{(SIV)}_{(MSE)}$	$0.990_{(0.138)}$	$-2.001_{(0.349)}$	$2.971_{(0.350)}$	$-3.997_{(0.273)}$	$4.933_{(0.389)}$
$\hat{\beta}^{(SWIV)}_{(MSE)}$	$0.990_{(0.082)}$	$-1.993_{(0.140)}$	$3.010_{(0.138)}$	$-3.992_{(0.133)}$	$5.010_{(0.154)}$

$\boldsymbol{T=15,\quad n\cdot T=300,\quad h=0.940,\quad g=0.948}$					
$\hat{\beta}^{(IV)}_{(MSE)}$	$0.755_{(1.912)}$	$-1.479_{(3.644)}$	$2.431_{(3.242)}$	$-3.324_{(4.295)}$	$3.980_{(4.988)}$
$\hat{\beta}^{(SIV)}_{(MSE)}$	$0.984_{(0.053)}$	$-2.020_{(0.219)}$	$2.934_{(0.233)}$	$-4.020_{(0.236)}$	$4.897_{(0.218)}$
$\hat{\beta}^{(SWIV)}_{(MSE)}$	$0.985_{(0.048)}$	$-2.008_{(0.107)}$	$2.995_{(0.121)}$	$-4.017_{(0.104)}$	$4.975_{(0.112)}$

$\boldsymbol{T=20,\quad n\cdot T=400,\quad h=0.940,\quad g=0.948}$					
$\hat{\beta}^{(IV)}_{(MSE)}$	$0.774_{(1.463)}$	$-1.562_{(2.618)}$	$2.577_{(2.490)}$	$-3.374_{(2.845)}$	$4.220_{(2.826)}$
$\hat{\beta}^{(SIV)}_{(MSE)}$	$0.992_{(0.036)}$	$-1.988_{(0.199)}$	$2.934_{(0.191)}$	$-3.974_{(0.176)}$	$4.948_{(0.169)}$
$\hat{\beta}^{(SWIV)}_{(MSE)}$	$0.994_{(0.033)}$	$-1.986_{(0.076)}$	$3.006_{(0.078)}$	$-3.981_{(0.077)}$	$5.017_{(0.072)}$

$\boldsymbol{T=25,\quad n\cdot T=500,\quad h=0.940,\quad g=0.948}$					
$\hat{\beta}^{(IV)}_{(MSE)}$	$0.794_{(1.074)}$	$-1.629_{(1.644)}$	$2.494_{(1.923)}$	$-3.551_{(1.722)}$	$4.314_{(2.187)}$
$\hat{\beta}^{(SIV)}_{(MSE)}$	$0.990_{(0.034)}$	$-1.983_{(0.168)}$	$2.930_{(0.172)}$	$-3.991_{(0.151)}$	$4.944_{(0.209)}$
$\hat{\beta}^{(SWIV)}_{(MSE)}$	$0.993_{(0.028)}$	$-1.985_{(0.062)}$	$2.984_{(0.069)}$	$-3.995_{(0.060)}$	$4.996_{(0.065)}$

Table 4.2. *The contamination by leverage points on the level of 5%, $n = 100$. The values of variance of the disturbances randomly selected from $[0.5, 5.5]$.*

Number of observations in each data set = 500					
Contamination level = 1%, $h = 0.973$, $g = 0.989$					
$\hat{\beta}^{(S)}_{(MSE)}$	$1.010_{(0.024)}$	$2.003_{(0.031)}$	$-3.021_{(0.032)}$	$3.975_{(0.035)}$	$-4.974_{(0.023)}$
$\hat{\beta}^{(SW)}_{(MSE)}$	$1.002_{(0.022)}$	$2.001_{(0.013)}$	$-3.012_{(0.011)}$	$3.986_{(0.010)}$	$-4.990_{(0.011)}$
Contamination level = 2%, $h = 0.963$, $g = 0.978$					
$\hat{\beta}^{(S)}_{(MSE)}$	$0.993_{(0.027)}$	$2.014_{(0.032)}$	$-2.973_{(0.027)}$	$3.985_{(0.028)}$	$-4.996_{(0.023)}$
$\hat{\beta}^{(SW)}_{(MSE)}$	$0.992_{(0.030)}$	$2.008_{(0.005)}$	$-3.000_{(0.004)}$	$4.000_{(0.005)}$	$-5.003_{(0.004)}$
Contamination level = 5%, $h = 0.921$, $g = 0.942$					
$\hat{\beta}^{(S)}_{(MSE)}$	$0.985_{(0.028)}$	$1.948_{(0.040)}$	$-2.967_{(0.034)}$	$3.919_{(0.038)}$	$-4.955_{(0.030)}$
$\hat{\beta}^{(SW)}_{(MSE)}$	$1.014_{(0.027)}$	$2.002_{(0.002)}$	$-3.006_{(0.002)}$	$3.998_{(0.002)}$	$-5.003_{(0.001)}$

Table 4.3. *Contamination by outliers: For randomly selected observations, we put $Y_i = 5 * Y_i^{original}$ and data contained also good leverage points $X_i = 10 * X_i^{original}$ and responses Y_i's were computed correctly*

Generally, the implicit estimation of variance of the disturbances (e.g. by LMS, LTS or LWS) is the significant advantage (from the computational point of view) because the estimators do not need any studentization – contrary to M-estimators – see (Bickel 1975). Sometimes, it can betray us.

4.6. Acknowledgment

This study was performed with the support of the Czech Science Foundation project P402/12/G097'DYME – Dynamic Models in Economics.

4.7. References

Bickel, P.J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc*. 70, 428–433.

Boček, P., Lachout, P. (1993). Linear programming approach to LMS-estimation. *Memorial volume of Comput. Statist. & Data Analysis*, 19(1995), 129–134.

Breiman, L. (1968). *Probability*. Addison-Wesley Publishing Company, London.

Campbell, N.A., Lopuhaa, H.P., Rousseeuw, P.J. (1998). On calculation of a robust S-estimator if a covariance matrix. *Statistics in medcine*, 17, 2685–2695.

Carroll, R.J., Stefanski, L.A. (1994). Measurement error, instrumental variables and correction for attenuation with applications to meta-analyses. *Statistics in Medicine*, 13, 1265–1282.

Číek, P., Víšek, J.Á. (2000). The least trimmed squares. *User Guide of Explore*, Humboldt University, Berlin.

Cohen-Freue, G.V., Ortiz-Molina, H., Zamar, R.H. (2013). Natural robustification of the ordinary instrumental variables estimator. *Biometrics*, 69, 641–650.

Desborges, R., Verardi, V. (2012). A robust instrumental-variable estimator. *The Stata Journal*, 12, 169–181.

Hájek, J., Šidák, Z. (1967). *Theory of Rank Test*. Academic Press, New York.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A. (1986). *Robust Statistics – The Approach Based on Influence Functions*. John Wiley & Sons, New York.

Heckman, J., Urza, S., Vytlacil, E.J. (2006). Understanding instrumental variables in models with essential heteroscedasticity. Working paper 12574, National Bureau of Economic Research, 2006.

Hawkins, D.M. (1994). The feasible solution algorithm for least trimmed squares regression. *Computational Statistics and Data Analysis* 17, 185–196.

Hawkins, D.M., Olive, D.J. (1999). Improved feasible solution algorithms for breakdown estimation. *Computational Statistics & Data Analysis* 30, 1–12.

Hettmansperger, T.P., Sheather, S.J. (1992). A cautionary note on the method of least median squares. *The American Statistician* 46, 79–83.

Koenker, R., Bassett, G. (1978). Regression quantiles. *Econometrica*, 46, 33–50.

Maronna, R.A., Yohai, V. (1981). Asymptotic behavior of general M-estimates for regression and scale with random carriers. *Z. Wahrscheinlichkeitstheorie verw*, Gebiete 58, 7–20.

Paige, C.C., Strako, Z. (2002). Scaled total least squares fundamentals. *Numerische Mathematik*, 91, 117–146.

Rousseeuw, P.J. (1983). Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications B*, Grossmann, W., Pflug, G., Vincze, I., Wertz, W. (eds), Reidel, Dordrecht, 283–297.

Rousseeuw, P.J. (1984). Least median of square regression. *Journal of Amer. Statist. Association*, 79, 871–880.

Rousseeuw, P.J., Yohai, V. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis: Lecture Notes in Statistics No. 26*, Franke, J., Härdle, W.H, Martin, R.D. (eds), Springer Verlag, New York, 256–272.

Siegel, A.F. (1982). Robust regression using repeated medians. *Biometrica*, 69, 242–244.

Víšek, J.Á. (1990). Empirical study of estimators of coefficients of linear regression model. *Technical report of Institute of Information Theory and Automation*, Czechoslovak Academy of Sciences, 1699.

Víšek, J.Á. (1998). Robust instruments. In *Robust'98*, Antoch, J., Dohnal, G. (eds), Union of Czech Mathematicians and Physicists, Matfyzpress, Prague, 195–224.

Víšek, J.Á. (2000). Regression with high breakdown point. In *Robust 2000*, Antoch, J., Dohnal, G. (eds), The Union of the Czech Mathematicians and Physicists and the Czech Statistical Society 2001, Matfyzpress, Prague, 324–356.

Víšek, J.Á. (2002). The least weighted squares II. Consistency and asymptotic normality. *Bulletin of the Czech Econometric Society*, 9, 1–28.

Víšek, J.Á. (2004). Robustifying instrumental variables. In *Proceedings of COMPSTAT' 2004*, Antoch, J. (ed.), Physica-Verlag/Springer, 1947–1954.

Víšek, J.Á. (2006). The least trimmed squares. Part I – Consistency. Part II – $\sqrt{n}$-consistency. Part III – Asymptotic normality and Bahadur representation. *Kybernetika*, 42, 1–36, 181–202, 203–224.

Víšek, J.Á. (2006). Instrumental weighted variables. *Austrian Journal of Statistics*, 35, 379–387.

Víšek, J.Á. (2006). Instrumental weighted variables – algorithm. In *Proceedings of COMPSTAT 2006*, Rizzi, A., Vichi, M. (eds), Physica-Verlag, Springer Company, Heidelberg, 777–786.

Víšek, J.Á. (2009). Consistency of the instrumental weighted variables. *Annals of the Institute of Statistical Mathematics*, 61, 543–578.

Víšek, J.Á. (2011). Empirical distribution function under heteroscedasticity. *Statistics*, 45, 497–508.

Víšek, J.Á. (2011). Consistency of the least weighted squares under heteroscedasticity. *Kybernetika*, 47, 179–206.

Víšek, J.Á. (2015). S-weighted estimators. In *Proceedings of the 16th Conference on the Applied Stochastic Models, Data Analysis and Demographics*, Skiadas, C.H. (ed.), 1031-1042 or *Stochastic and Data Analysis Methods and Applications in Statistics and Demography*, Bozeman, J.R., Oliveira, T., Skiadas, C.H. (eds), 437–448.

Víšek, J.Á. (2016). Representation of *SW*-estimators. In *Proceedings of the 4th Stochastic Modeling Techniques and Data Analysis International Conference with Demographics Workshop, SMTDA 2016*, Skiadas, C.H. (ed.), 425–438.

Víšek, J.Á. (2017). Instrumental weighted variables under heteroscedasticity. Part I. Consistency. Part II. Numerical study. *Kybernetika*, 53(2017), 1-25, 26–58.

Wagenvoort, R., Waldmann, R. (2002). On B-robust instrumental variable estimation of the linear model with panel data. *Journal of Econometrics*, 106, 297–324.

PART 2

Models and Modeling

5

Grouping Property and Decomposition of Explained Variance in Linear Regression

The quantification of the relative importance of predictors on a response variable has been an active subject of research for many years. Regression analysis may be used for that purpose but estimating importance of predictors via (standardized) regression coefficients is not adequate in the presence of correlations between these variables. Therefore, alternative methods have been considered. Grouping property is respected when estimators of importance tend to equate for highly correlated predictors. We will analyze the respect of grouping property for several methods used to quantify the relative importance of predictors through decomposition of the explained variance in linear regression. After being criticized by several authors, Correlation-Adjusted marginal coRrelation (CAR) scores have been recommended again as estimators of importance of predictors and presented with respect to the grouping property. We will show that CAR scores actually do not respect this property. We will explain, in turn, why some other variance decomposition methods do respect grouping property and we will formulate recommendations for quantifying the relative importance of predictors.

5.1. Introduction

The quantification of relative importance of predictors on a response variable has been a subject of research in biostatistics, psychology, economics or market research. Many methods have been investigated, sometimes reinvented by researchers across different fields (see Grömping 2007, 2009, 2015 for an overview). Some approaches relate to game theory (*lmg, pmvd, owen*). Others are based on regularization techniques (lasso, elasticnet). Random forests have also been proposed both for estimation of relative importance and for variable selection (see Grömping 2009, Genuer 2012, Genuer *et al.* 2015). Some methods consist of decomposing the variance explained by the linear regression among the different predictors. When the

Chapter written by Henri WALLARD.

Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining,
First Edition. Edited by Christos H. Skiadas and James R. Bozeman.

predictors are mutually decorrelated, the R^2 can be naturally decomposed by allocating to each predictor the square of the correlation coefficient between this predictor and the variable to predict, but with collinearity between predictors there is no longer a unique and natural way to decompose the explained variance. Multicollinearity designates situations in which two or more explanatory variables in a multiple regression model are highly correlated.

Several methods used to decompose the explained variance in linear regression can be presented in a unified perspective using the singular value decomposition of the matrix of (standardized) observations. One of these methods was proposed in 1962 and designated as the Gibson method (Gibson 1962), but has later been criticized (see Johnson 2000, Grömping 2015). These criticisms lead several authors to recommend other approaches, for instance (Genizi 1993) and (Johnson 2000). However, in 2011, (Zuber and Strimmer 2011) have applied again the Gibson method under the name of CAR scores for regression analysis and variable selection and suggested that CAR scores would benefit from the grouping property as introduced by Zou and Hastie (Zou and Hastie 2005), in the sense that if two predictors tend to correlate their CAR scores or the square of their CAR scores would tend to equate. As a result, this method would have interesting properties for the quantification of the importance of predictors and their selection. We will demonstrate that, actually, CAR scores do not respect the grouping property as defined by Zou and Hastie (2005), but the Genizi and Johnson method does respect such grouping property. Our analysis leads us to recommend against the usage of Gibson CAR scores.

5.2. CAR scores

5.2.1. *Definition and estimators*

5.2.1.1. *Presentation of CAR scores*

Let Y be a random variable and $\boldsymbol{X}$=$(X_1,...,X_p)^T$ be a random vector of dimension p (of finite variance). The covariance matrix is var($\boldsymbol{X}$)=$\boldsymbol{\Sigma}_{XX}$. Let us note also $\boldsymbol{P}_{XX}$ the correlation matrix. We can decompose the covariance matrix as:

$$\boldsymbol{\Sigma}_{XY} = \boldsymbol{V}^{1/2}\boldsymbol{P}_{XX}\boldsymbol{V}^{1/2} \qquad [5.1]$$

where $\boldsymbol{V}$ is the diagonal matrix containing the variances of the X_i. If we note $\boldsymbol{P}_{XY}$=$(\rho_{X_1y},...,\rho_{X_py})^T$ the vector or marginal correlations between Y and $\boldsymbol{X}$, the vector of CAR scores is noted as $\boldsymbol{\omega}$=$(\omega_1,...,\omega_p)^T$ and given by:

$$\boldsymbol{\omega} = \boldsymbol{P}_{XX}^{-1/2}\boldsymbol{P}_{XY} \qquad [5.2]$$

CAR stands for *"Correlation-Adjusted marginal coRrelation"*.

As in Zuber and Strimmer (2011), we will also introduce the *best linear predictor* of Y, the linear combination of the explanatory variables as follows:

$$Y^* = a + \boldsymbol{b}^T \boldsymbol{X}$$

that minimizes the mean squared prediction error E[$(Y\text{-}Y^*)^2$]. In the approach used by Zuber and Strimmer (2011), the coefficients a and $\boldsymbol{b}$=$(b_1,...,b_d)^T$ are considered as *constant* for the interpretation of the grouping property, but this restriction leads to wrong conclusions regarding the interest of the method, as shown later.

It results from the definition of CAR scores above that the sum of the square of the CAR scores adds up to the R^2 of the linear regression (see Johnson 2000, Zuber and Strimmer 2011). This is shown in the following equation:

$$\boldsymbol{\omega^T \omega} = \boldsymbol{P'}_{XY} \boldsymbol{P}_{XX}^{-1} \boldsymbol{P}_{XY} = R^2 \qquad [5.3]$$

It is possible to use the squared CAR scores to quantify the relative importance of each predictor as follows. Formally:

$$\Phi^{CAR}(X_j) = \omega_j^2 \qquad [5.4]$$

5.2.1.2. *CAR scores with two predictors*

In the case with two predictors X_1 and X_2 and if we note $(\beta std)_1$ and $(\beta std)_2$, the standardized regression coefficients of Y on X_1 and X_2, and note cor(X_1,X_2)=ρ_{12}, we can write the following result:

$$\omega_1^2 - \omega_2^2 = ((\beta std_1)^2 - (\beta std_2)^2)\sqrt{1 - \rho_{12}^2} \qquad [5.5]$$

This equation is used by Zuber and Strimmer (2011) with the restriction that the coefficients βstd_i are kept constant when ρ_{12} tends toward 1 so both side of equation 5.5 tend toward 0. The consequences of this restriction are discussed later in this chapter.

5.2.1.3. *Estimators of CAR scores using singular value decomposition*

When applied to linear regression on a set of observations, CAR scores as well as some other explained variance decomposition methods can be presented using matrix calculus.

Let us consider a linear model of the form below with p predictors:

$$Y_i = \beta_0 + X_{i1}\beta_1 + ... + X_{ip}\beta_p + \varepsilon_i$$

with independent errors terms ε_i of expectation 0 and constant positive variance σ^2.

Let us note as X the matrix of n observations of the p predictors ($i = 1, \ldots, n$ and $j = 1, \ldots, p$), X is a (n, p) matrix deemed to be of rank p, and let us note Y the column vector of the n observations of Y.

$$\boldsymbol{X} = (x_{ij})$$

$$\boldsymbol{Y} = (y_i)$$

We will note Y^*as the values predicted by the linear model:

$$Y^* = X\ \beta^* = (X'X)^{-1}X'Y$$

We will assume in the rest of this chapter that Y and the p columns of X, designated as X_j, are standardized: mean of all variables equal to 0 and variance equal to 1.

We will refer to the singular value decomposition (SVD) of the matrix of observations using the notations from Johnson (2000) as reminded in the following. The SVD of X is:

$$\boldsymbol{X = P\Delta Q'} \qquad [5.6]$$

We will define a particular matrix Z as:

$$\boldsymbol{Z = PQ'} \qquad [5.7]$$

Z is of rank p and is the Mahalanobis transform of X. Z_j is the j_{th} column of Z. Johnson (2000) pointed out that the columns of Z are also characterized as the best-fitting approximations to the columns of X in the sense that they minimize the sum of squares of the differences between the original variables and the orthogonal variables.

Estimators of the CAR scores can be calculated using the SVD of the matrix X. The vector of the estimators of the CAR scores is:

$$\begin{aligned}\hat{\boldsymbol{\Omega}} &= \frac{1}{n-1}\boldsymbol{(X'X)^{-1/2}X'Y} \\ &= \frac{1}{n-1}\boldsymbol{Q\Delta^{-1}Q'Q\Delta P'Y} = \frac{1}{n-1}\boldsymbol{QP'Y}\end{aligned} \qquad [5.8]$$

We have:

$$\hat{\boldsymbol{\Omega}} = \frac{1}{n-1}\boldsymbol{Q}\boldsymbol{P}'\boldsymbol{Y} = \frac{1}{n-1}\boldsymbol{Z}'\boldsymbol{Y} \quad [5.9]$$

The estimator of the CAR score for predictor j is equal to cor(Y,Z_j).

As Z is a unitary matrix as a result of the properties of SVD, the column vectors of Z are all of norm 1 and are all orthogonal and the sum of the squares of the CAR scores adds up to the R^2 of the model. We will refer to the squares of the CAR scores, called squared CAR scores, which represent the proportion of R^2 allocated to a given predictor with the CAR scores method.

5.2.2. *Historical criticism of the CAR scores*

Zuber and Strimmer (2011) proposed the CAR scores as a way to decompose the explained variance. This method was initially proposed in 1962 by Gibson and explained by Grömping (2007, 2015). The CAR scores were at that time called Gibson scores. Zuber and Strimmer point out that the CAR scores are computed in regressing the variable to predict on the Mahalanobis-decorrelated predictors and that these decorrelated predictors are the "*nearest*" to the original standardized predictors. This proximity had been identified and demonstrated by Johnson in 1966 as referenced by Johnson (2000) and had actually been viewed by several authors as a reason not to use the CAR scores. For instance, Johnson (2000) explained that these decorrelated variables Z are only approximation of the original variables and may not be close representations of the original variables if two or more of the predictor variables are highly correlated. Similarly, Grömping (2015) underlined that in case of relevant correlations among the X variables, the Z variables can be far from being good representatives for the corresponding X variables. Grömping reminds us that Green *et al.* (1978) proposed the modification in the values from Gibson by relating the Z variables back to the X variables leading to the Green method. Genizi (1993) and Johnson (2000) also proposed another way to decompose the explained variance. So the literature on variance decomposition methods had criticized Gibson CAR scores, and this is precisely why other methods such as Green *et al.* (1978), Fabbris (1980), Genizi (1993), or Johnson (2000) were proposed (see Grömping 2015, 2016). The usage of SVD will enable the analysis of the respective properties of these various methods.

5.3. Variance decomposition methods and SVD

We can formalize several methods used to decompose the explained variance in linear regression as a succession of two steps, first to allocate the explained variance

in p positive terms and then further reallocate these p terms back to the original predictors. We can formalize this considering a vector A of p positive terms adding up to the R^2 and a matrix of weights Π with all positive terms, the only additional condition required being that each column of Π adds up to 1:

$$\boldsymbol{A} = (a_j)$$

$$\boldsymbol{\Pi} = (\pi_{ij})$$

The relative importance estimators for the predictors can then be computed as the column vector $\boldsymbol{R_w}$:

$$\boldsymbol{R_w} = \boldsymbol{\Pi A}$$

Table 5.1 summarizes the empirical estimators of relative importance for the methods proposed by Fabbris, Genizi–Johnson (these two being identical) and Gibson CAR scores. We are using the notation $\boldsymbol{M^{\cdot 2}}$ for the Hadamard squared matrix obtained when each term of the matrix $\boldsymbol{M}$ is elevated to its square.

Methods	Matrix Π	Vector A
Fabbris	$\boldsymbol{Q^{\cdot 2}}$	$(\frac{1}{n-1})^2(\boldsymbol{P'Y})^{\cdot 2}$
Genizi–Johnson	$(\frac{1}{n-1})^2(\boldsymbol{Z'X})^{\cdot 2}$	$(\frac{1}{n-1})^2(\boldsymbol{Z'Y})^{\cdot 2}$
CAR scores	$\boldsymbol{I}$	$(\frac{1}{n-1})^2(\boldsymbol{Z'Y})^{\cdot 2}$

Table 5.1. *Allocation of explained variance*

Line 2 of Table 5.1 presents that any orthogonal matrix $\boldsymbol{O}$ can be used instead of $\boldsymbol{Z}$ to generate an allocation of the explained variance, as $\boldsymbol{\pi_{ij}}$ is the square of cor($\boldsymbol{O_i}$,$\boldsymbol{X_j}$), so the columns of $\boldsymbol{\Pi}$ all add up to $\boldsymbol{1}$. The relative importance measures computed using the Genizi–Johnson method are designated as relative weights as in Johnson (2000), $\boldsymbol{RW_i}$ for predictor $\boldsymbol{i}$.

5.4. Grouping property of variance decomposition methods

In the perspective of regularization and variable selection via the elastic net, Zou and Hastie (2005) introduced the general concept of grouping property for any given type of regression method. Qualitatively speaking a regression method will be said to exhibit the grouping property if the coefficients of a group of highly correlated variables tend to be equal (up to a change in sign if negatively correlated) and also that in the extreme situation when variables would be identical, the regression-based method should assign to each of them identical coefficients. The grouping property

appears useful for a consistent estimation of variable importance in case of multicollinearity.

Zuber and Strimmer (2011) claimed that variable importance derived from CAR scores respects the grouping property. We will demonstrate that this method actually does not respect the grouping property as originally defined by Zou and Hastie (2005).

The grouping property as presented by Zou and Hastie (2005) applies to the usual linear regression model: given p predictors $x_1,...,x_p$, the response y is predicted by:

$$\hat{y} = \hat{\beta}_0 + x_1\hat{\beta}_1 + ... + x_p\hat{\beta}_p$$

and where a model fitting procedure produces the vector of coefficients $\hat{\beta}=(\hat{\beta}_1,...,\hat{\beta}_p)$.

By comparison, Zuber and Strimmer consider situations where the coefficients in $b=(b_1,...,b_d)^T$ are kept *constants*. This is the particular way to analyze the respect of the grouping property when using CAR scores does not align with the full generality of the original definition proposed by Zou and Hastie (2005). In reality, keeping the vector b as constant when the correlation of the predictor variables vary results in modifying Y^*. With this approach, it is impossible to analyze the performance of models with variable selection to model a given and fixed variable Y against various subsets of predictors.

If we follow the original approach of Zou and Hastie (2005), we can in turn analyze models where several predictors will be increasingly correlated while the response Y is kept constant, and the vector of estimated coefficients $\hat{\beta}$ will vary when we use various subsets of predictors or when we change the structure of the correlations between predictors.

This will lead to different conclusions from Zuber and Strimmer (2011), and this is extremely important in practice when real data are analyzed.

5.4.1. *Analysis of grouping property for CAR scores*

Based on the formula 5.5 for the case with two predictors, Zuber and Strimmer (2011) state that when ρ_{12} tends toward 1 the product of the two terms on the right side of equation 5.5 above tends toward 0 and conclude that the two squared CAR scores for X_1 and X_2 become identical with absolute value of the correlation between the two predictors.

But quite contrary when the variable to predict Y is fixed (and not collinear to X_1+ X_2 which is a particular case) while the correlation between X_1 and X_2 tends toward 1, the quantity $((\beta std_1)^2 - (\beta std_2)^2)$ is not capped and the product of the two quantities on the right side of equation 5.5 does not tend toward 0. This will be illustrated in the case with two predictors.

5.4.2. *Demonstration with two predictors*

Let us now consider two standardized and decorrelated variables $\boldsymbol{E_1}$ and $\boldsymbol{E_2}$. Let us also chose two real values ϕ and ψ.

$$\boldsymbol{X_1} = \boldsymbol{cos}(\phi)\boldsymbol{E_1} - \boldsymbol{sin}(\phi)\boldsymbol{E_2}$$
$$\boldsymbol{X_2} = \boldsymbol{cos}(\phi)\boldsymbol{E_1} + \boldsymbol{sin}(\phi)\boldsymbol{E_2}$$
$$\boldsymbol{Y} = \boldsymbol{cos}(\psi)\boldsymbol{E_1} + \boldsymbol{sin}(\psi)\boldsymbol{E_2}$$

We have (with $\phi \neq \mathbf{0}$):

$$\beta\boldsymbol{std_1} = \frac{\boldsymbol{sin}(\phi - \psi)}{\boldsymbol{sin}(\mathbf{2}\phi)}$$
$$\beta\boldsymbol{std_2} = \frac{\boldsymbol{sin}(\phi + \psi)}{\boldsymbol{sin}(\mathbf{2}\phi)}$$

Let us note:

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{cos}(\phi) & \boldsymbol{cos}(\phi) \\ -\boldsymbol{sin}(\phi) & \boldsymbol{sin}(\phi) \end{pmatrix}$$

The matrix composing the SVD of the matrix $\boldsymbol{X}$ are given as (see formulas [5.6] and [5.7]):

$$\boldsymbol{P} = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{pmatrix}$$

$$\boldsymbol{\Delta} = \begin{pmatrix} \sqrt{\mathbf{1} + \boldsymbol{\rho_{12}}} & \mathbf{0} \\ \mathbf{0} & \sqrt{\mathbf{1} - \boldsymbol{\rho_{12}}} \end{pmatrix}$$

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{cos}(\frac{\pi}{4}) & -\boldsymbol{sin}(\frac{\pi}{4}) \\ \boldsymbol{sin}(\frac{\pi}{4}) & \boldsymbol{cos}(\frac{\pi}{4}) \end{pmatrix}$$

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{cos}(\frac{\pi}{4}) & \boldsymbol{sin}(\frac{\pi}{4}) \\ -\boldsymbol{sin}(\frac{\pi}{4}) & \boldsymbol{cos}(\frac{\pi}{4}) \end{pmatrix}$$

We also have:

$$\boldsymbol{\rho_{12}} = \boldsymbol{cos}(\mathbf{2}\phi)$$

If we come back to equation [5.5], we can now write:

$$((\beta std_1)^2 - (\beta std_2)^2) = -\frac{sin(2\psi)}{sin(2\phi)} \quad [5.10]$$

This confirms the conclusion that the first term of the product on the right of equation [5.10] is not capped and tends toward the infinite in the general case ($\psi \neq 0$) if ϕ tends toward 0.

Using again formula [5.5], we can also express the CAR scores using directly ψ and ϕ.

$$\begin{aligned} \omega_1 &= cos(\psi + \frac{\pi}{4}) \\ \omega_2 &= cos(\psi - \frac{\pi}{4}) \\ \omega_1^2 - \omega_2^2 &= -sin(2\psi) \end{aligned} \quad [5.11]$$

The difference of squared CAR scores for the two predictors depends only on the correlation between Y and E_1. However, it does not depend on ρ_{12}, which is also cos(2ϕ), correlation between X_1 and X_2. In the case of two predictors, Z is actually strictly constant (see also Thomas *et al.* 2014). If we look at the representation of the variables in a plane, Z_1 and Z_2 are symmetric around E_1 and form with E_1 an angle of $\pm\frac{\pi}{4}$.

5.4.3. *Analysis of grouping property using SVD*

An important property of the CAR scores is that they do not depend on the matrix Δ but on PQ' as shown in equations [5.7] and [5.9]. The matrix P and Q' are known to include the eigenvectors of $X'X$ and XX', but they do not depend on the eigenvalues of $X'X$ that are the terms of the diagonal matrix Δ^2. The repartition of the eigenvalues of Δ^2 describes collinearity between predictors. The fact that the CAR scores do not depend on Δ is the underlying reason why the CAR scores will not tend to equate when the correlation between predictors increases: high or low eigenvalues of $X'X$ can be achieved with identical matrix P and Q'.

In turn, if we look (Table 5.1) at the matrix $Z'X$ involved in the variance decomposition for Genizi–Johnson, this property will be respected as we will demonstrate below. We can express the matrix $Z'X$ as follows, which when elevated to the Hadamard squarer is the matrix Π for Genizi–Johnson as per Table 5.1:

$$\boldsymbol{Z'X = QP'P\Delta Q' = Q\Delta Q'} \quad [5.12]$$

Equation 5.12 shows that the matrix $\boldsymbol{Z'X}$ is symmetric: $\boldsymbol{Z'X = X'Z}$, so we can write the matrix $\boldsymbol{\Pi}$ as $\boldsymbol{X'Z}$, and each relative weight as follows:

$$\boldsymbol{RW_i = \sum_{k=1}^{p} cor^2(X_i, Z_k) cor^2(Y, Z_k)} \quad [5.13]$$

The difference of relative weights between two predictors is:

$$\boldsymbol{RW_i - RW_j = \sum_{k=1}^{p} ((cor^2(X_i, Z_k)) - cor^2(X_j, Z_k)) cor^2(Y, Z_k)} \quad [5.14]$$

We have:

$$\begin{aligned} \boldsymbol{cor^2(X_i, Z_k) - cor^2(X_j, Z_k)} \;=\; & \boldsymbol{(cor(X_i, Z_k) - cor(X_j, Z_k))} \\ & \boldsymbol{\times (cor(X_i, Z_k) + cor(X_j, Z_k))} \end{aligned}$$

As a result:

$$\boldsymbol{cor^2(X_i, Z_k) - cor^2(X_j, Z_k) \leq 2\,|cor(X_i, Z_k) - cor(X_j, Z_k)|} \quad [5.15]$$

As all variables are standardized, the correlations equate to the covariance and we can use the additive property of covariance to write:

$$\boldsymbol{cor^2(X_i, Z_k) - cor^2(X_j, Z_k) \leq 2\,|cov(X_i, Z_k) - cov(X_j, Z_k)|} \quad [5.16]$$

$$\boldsymbol{cor^2(X_i, Z_k) - cor^2(X_j, Z_k) \leq 2\,|(cov(X_i - X_j, Z_k)|} \quad [5.17]$$

We also have:

$$\boldsymbol{|(cov(X_i - X_j, Z_k)| \leq \|X_i - X_j\|} \quad [5.18]$$

If we note that:

$$\boldsymbol{\|X_i - X_j\| = \sqrt{2(1 - \rho_{ij})}} \quad [5.19]$$

As the sum of the $cor^2(Y, Z_k)$ adds up to the R^2 in the end:

$$|RW_i - RW_j| \leq 2\sqrt{2(1-\rho_{ij})}R^2 \qquad [5.20]$$

This proves that relative weights (Johnson–Genizi measures) respect the grouping property in the general sense as defined by Zou and Hastie (2005), but as shown before CAR scores do not respect that grouping property. In addition, we can see that starting from equation [5.13] using $Y = Xb$ with b fixed does not change the rest of the subsequent demonstration. This means that like the CAR scores, the relative weights from Genizi–Johnson measures will also tend to equate if the coefficients b are kept constant as considered by Zuber and Strimmer (2011) for highly correlated predictors. So the *restricted* property as used by Zuber and Strimmer is equally respected by relative weights and CAR scores. But only relative weights do respect the *full* original grouping property as defined by Zou and Hastie (2005). These results show why relative weights are to be preferred to CAR scores.

There is also an interesting result regarding the variance decomposition via an orthogonal matrix in the case of two predictors. Using X as defined above, let us consider an orthonormal matrix O as defined in the following:

$$O = \begin{pmatrix} cos(\omega) & -sin(\omega) \\ sin(\omega) & cos(\omega) \end{pmatrix}$$

and Y similarly as in the case above with two predictors $Y = (cos(\psi), sin(\psi))^T$.

If we substitute the matrix O to the matrix Z in the computation of the relative weights (Genizi–Johnson measures), we can easily compute the difference between the two relative weights for variables 1 and 2 denoted as RW_1^O and RW_2^O when using the matrix O for decomposition:

$$\begin{aligned} RW_1^O - RW_2^O \; = \; & cos(2(\phi+\omega)cos^2(\psi-\omega) \\ & + cos(2(\phi-\omega)sin^2(\psi-\omega) \end{aligned} \qquad [5.21]$$

When ϕ tends toward 0, $RW_1^O - RW_2^O$ tends toward $\cos(2\omega)$ as per equation [5.21]. So the two relative weights will tend to equate in case the correlation between the two predictors tends toward 1 if and only if $\omega = \pm\pi/4$ or $\omega = \pm 3\pi/4$, which means if $O = \pm Z$ or $O = \pm Z'$. This result shows that in the case of two predictors the only variance decomposition via orthogonalization that respect grouping property is Genizi–Johnson measures or associated decomposition related to Z.

In conclusion, we have identified an important difference between CAR scores decomposition and Genizi–Johnson decomposition. Given a variable to predict $\boldsymbol{Y}$, kept constant, if the correlation between two of the predictors tends toward $\mathbf{1}$, or if in the data set there are highly correlated predictors, their Genizi–Johnson importance (called relative weights by Johnson) will always tend to equate while this is not the case with CAR scores. This has important consequences as we show in the following using the numeric examples.

5.4.4. *Application to the diabetes data set*

In the documentation associated with the R package care, Zuber and Strimmer (2011) use the diabetes data from Efron *et al.* (2004). The diabetes data have 10 variables ("age", "sex", "bmi", "bp", "s1", "s2", "s3", "s4", "s5", "s6") and 442 measurements. The data are standardized. We will add a new predictor variable using a parameter to adjust the correlation of this new variable with one of the original predictors. We considered one of the original variables, $\boldsymbol{bp}$, and generated a random set of 442 observations called $\boldsymbol{\epsilon}$, with a mean of $\mathbf{0}$ and a standard deviation of $\mathbf{0.2}$. We then generated a variable $\boldsymbol{bp'}$ using original variable from the diabetes data, the variable $\boldsymbol{bmi}$ and adding a terms as follows:

$$\mathbf{bp'} = \mathbf{bp}\boldsymbol{cos}(\phi) + (\mathbf{bmi} + \boldsymbol{\epsilon})\boldsymbol{sin}(\phi) \qquad [5.22]$$

By adjusting the value of ϕ, we can vary the correlation between $\boldsymbol{bp}$ and $\boldsymbol{bp'}$. The results of the squared CAR scores as a function of the correlation between $\boldsymbol{bp}$ and $\boldsymbol{bp'}$ are shown in Figure 5.1. There is no convergence between the squared CAR scores of $\boldsymbol{bp}$ and $\boldsymbol{bp'}$ when the correlation between these two variables tends toward 1.

We made another analysis this time by introducing a variable $\boldsymbol{bp}^*$ that is now computed as:

$$\mathbf{bp}^* = \mathbf{bp} + \boldsymbol{\epsilon} \qquad [5.23]$$

In that case, the correlation between $\boldsymbol{bp}$ and $\boldsymbol{bp}^*$ is 0.981. We compared the ranking of the predictors by decreasing order of their importance measure according to three different methods: Random Forest, "lmg" (see Grömping 2007) and CAR scores. The ranks are presented in Table 5.2. We can observe that the two highly correlated predictors $\boldsymbol{bp}$ and $\boldsymbol{bp}^*$ remain close in the ranking, and even ranked one after the other when random forest or lmg-Shapley is used, but this is not the case with squared CAR scores.

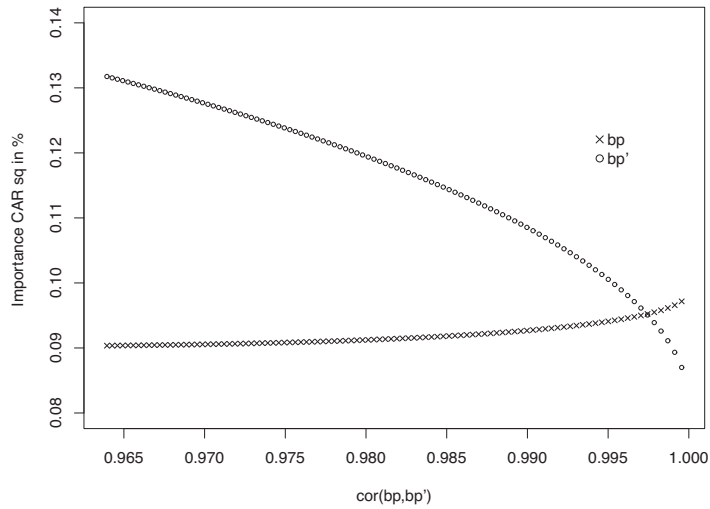


Figure 5.1. *Squared CAR scores. Diabetes data. Predictor bp' added correlated to bp*

Method	Random forest	Img-Shapley	Genizi–Johnson	CAR sores
s5	1	2	2	2
bmi	2	1	1	1
bp*	3	4	4	6
bp	4	3	3	3
s4	5	6	6	5
s3	6	5	5	4
s2	7	10	10	10
s1	8	8	8	11
s6	9	7	7	7
sex	10	9	11	8
age	11	11	9	9

Table 5.2. *Ranking predictors by decreasing importance: random forest, Img-Shapley, CAR scores and Genizi–Johnson*

5.5. Conclusions

CAR scores have been considered for many years but were criticized by several authors because of the fact that the decorrelated predictors used to decompose the explained variance of the variable to predict can be poor representatives of the original predictors. CAR scores have been recommended again recently for the quantification of relative importance of predictors and been credited with the respect of grouping

property. Using theoretical demonstrations and simulations based on data sets, we have proven that actually CAR scores do not respect the grouping property. Even when the correlation between predictors is very high and tends toward 1, the CAR scores do not tend to equate. Highly correlated predictors can be allocated very different squared CAR scores and the quantification of their importance will differ. As a result, with CAR scores highly correlated predictor can suffer from wrong interpretation of their relative importance and inconsistent selection in a model. In turn, the relative weights introduced by Genizi (1993) and Johnson (2000) do fully respect the definition of grouping property.

In the absence of proper new justification of the CAR scores, we do not see any reason to overcome the past criticisms this method had faced over a long period or to recommend again their usage. Alternative methods should be preferred to estimate the relative importance of predictors.

For classical variance decomposition, relative weights can be easily implemented (see Genizi 1993, Johnson 2000, Grömping 2015).

Regularization and variable selection can also be implemented with elastic net including in the case of high dimensional data knowing that elastic net unlike lasso respects grouping property (see Zou and Hastie 2005).

Random forests take into account nonlinearities and interactions without modeling them (see Grömping 2009). This method can be recommended to quantify the relative importance of predictors and also for variable selection including the case of high dimensional data (see Genuer 2012, Genuer *et al.* 2015). (Gregorutti *et al.* 2015) has also shown that they tend to respect grouping property.

Lastly, it is important to note that variance decomposition should not be seen as a substitute for linear regression models, path analytical models and models based on theory-driven explanations. However, when a model based on theory is not available variance decomposition, elastic net or random forests can help identify and select important variables.

5.6. References

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004). Least angle regression (with discussion). *Ann. Statist*, 32, 407–499.

Fabbris, L. (1980). Measures of predictor variable importance in multiple regression. *Quality and quantity*, Elsevier Scientific Publishing Company, 4, 787–792.

Genizi, A. (1993). Decomposition of R^2 in multiple regression with correlated regressors. *Statistica Sinica*, 3, 407–420.

Genuer, R., Poggi, J-M., Tuleau-Malot, C. (2012). Variable selection using Random Forests. *Pattern Recognition Letters*, 31(14), 2225–2236, https://hal.archives-ouvertes.fr/hal-00755489.

Genuer, R., Poggi, J-M., Tuleau-Malot, C. (2015). VSURF: an R package for variable selection using random forests. *The R Journal*, 7/2, ISSN 2073-4589.

Gibson, W.A. (1962). Orthogonal predictors: A possible solution for the Hoffman-Ward controversy. *Psychlogical Reports*, 11, 32–34.

Green, P.E., Carroll, J.D., DeSarbo, W.S. (1978). A new measure of regressor importance in multiple regression. *J. Marketing Res.*, 15, 356–360.

Gregorutti, B., Michel, B., Saint-Pierre, P. (2015). Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics and Data Analysis*, 90, 15–35.

Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, 61(2), 139, 308–319, https://doi.org/10.1198/000313007X188252.

Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, 63(4), 308–319, https://doi.org/10.1198/tast.2009.08199.

Grömping, U. (2015). Variable importance in regression models. *Wired Comput. Stats.*, 7, 137–152, https://doi.org/10.1002/wics.

Grömping, U. (2016). Variable importance in regression models, corrigenda. *Wired Comput. Stats.*, 8, 154–157, https://doi.org/10.1002/wics.

Johnson, J.W. (2000). A heuristic method for estimating the relative weights of predictor variables in multiple regression. *Multivariate Behavioral Research*, 35, 1–19.

Thomas, R., Zumbo, B., Kwan, E., Schweitzer, L. (2014). On Johnson's (2000) relative weights method for assessing variable importance: a reanalysis. *Multivariate Behavioral Research*, 49, 329–338.

Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67, Part 2, 301–320.

Zuber, V., Strimmer, K. (2011). High-dimensional regression and variable selection using CAR scores. *Statistical Applications in Genetics and Molecular Biology*, 10(1). 10:34, https://doi.org/10.2202/1544-6115.1730.

6

On GARCH Models with Temporary Structural Changes

When an economic shock like the Lehman Crisis occurred, it is expected to investigate its influence based on economic time series. The intervention analysis by Box and Tiao is a method for such purpose. Most of the intervention analyses are based on ARIMA models, but some are based on GARCH models. The GARCH models have been developed for analyzing time series of stock returns. Usually, the expected value function of a GARCH model is assumed to be constant. However, this assumption is not appropriate when a time series includes a varying trend. Our first purpose is to propose a trend model, which can be easily taken in intervention analysis. Based on this trend model, we generalize a GARCH model for an intervention analysis for both trend and volatility. An identification method is also provided and evaluated by simulation studies. Usability of the proposed model is demonstrated by applying to real stock returns.

6.1. Introduction

The events such as the Lehman crash and the enforcement of large-scale monetary policy have a strong influence on society, and are strongly reflected in the economic time series. It is important to analyze the magnitude of the influence of such an event using changes that occurred in time series. The intervention analysis by Box and Tiao (1975) is a method for such a purpose. Most of the intervention analyses are based on ARIMA models. In recent years, a model based on the GARCH model was also proposed. The GARCH model is often used to analyze the volatility of the return on average stock price or stock price index. Ho and Wan (2002) proposed a test to detect structural breaks using a model with a special GARCH structure applying intervention analysis. Watanabe and Nagashima (2016) proposed a GARCH model with an intervention term in the constant term of volatility. These models are considered for intervention with volatility. However,

Chapter written by Norio WATANABE and Fumiaki OKIHARA.

Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining,
First Edition. Edited by Christos H. Skiadas and James R. Bozeman.

these models are based on the GARCH model with a constant expected value, and trends are not considered.

Usually, the expected value function of GARCH models is assumed to be constant. However, this assumption is not appropriate when a time series includes a varying trend. As models which are applicable to the expected value of the GARCH model, there are the polynomial regression trend and fuzzy trend models proposed by Kuwahara and Watanabe (2008). However, polynomial trends imply poor fitness to data, and fuzzy trend models are not easy to apply intervention analysis. In this paper, we first propose a parametric trend model based on a GARCH model, which can be applied as usual stock returns. Furthermore, we propose a GARCH model incorporating intervention analysis for both trend and volatility.

We propose an identification method of the model and do simulation studies for evaluation. Usability of the proposed model is demonstrated by applying to practical stock returns.

6.2. The model

6.2.1. *Trend model*

We consider expressing expected value function of stock returns using the smoothed step function. Let $\{X_t : t = 1, 2, ..., T\}$ be a series of stock returns and μ_t be expected value function of X_t. An interval $0 < t \leq T$ is divided into small intervals of length L_1. Let L_1 be a natural number. The midpoint of the first interval is 0. Let K be the number of intervals which is given by

$$K = \left[\frac{T + L_1}{L_1}\right] + 1. \qquad [6.1]$$

where $[x]$ means the smallest integer not exceeding x. The midpoint A_k of each interval satisfies

$$A_1 = 0 \qquad [6.2]$$

$$A_k = A_{k-1} + L_1 \qquad (2 \leq k \leq K). \qquad [6.3]$$

The expected value function μ_t is defined as follows: if $A_k - \frac{L_1}{2} \leq t \leq A_k - \frac{d}{2}$

$$\mu_t = \frac{m_{k-1} - m_k}{2}(-\cos(\frac{t - (A_k - \frac{d}{2})}{L_1 - d})\pi + 1) + m_k \qquad [6.4]$$

if $A_k - \frac{d}{2} \leq t \leq A_k + \frac{d}{2}$

$$\mu_t = m_k \qquad [6.5]$$

if $A_k + \frac{d}{2} \leq t \leq A_k + \frac{L_1}{2}$

$$\mu_t = \frac{m_{k+1} - m_k}{2}(-\cos(\frac{t - (A_k + \frac{d}{2})}{L_1 - d})\pi + 1) + m_k \qquad [6.6]$$

We define, $\mu_t = m_1$ when $A_1 < t \leq A_1 + \frac{d}{2}$, $\mu_t = m_K$ when $A_K - \frac{d}{2} \leq t \leq T$. $m_k (k = 1, ..., K)$ is a latent variable that determines the level of the trend. d is a width that takes a constant value within each interval and when $d = L_1$, it corresponds to a step function. An example of the trend is shown in Figure 6.1.

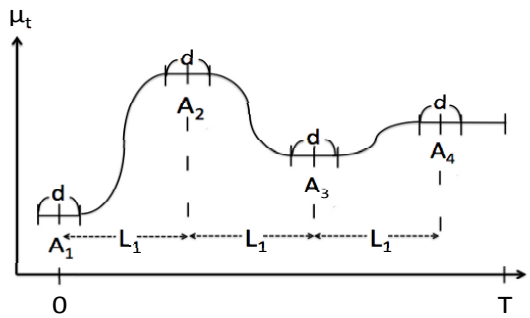


Figure 6.1. *An example of the trend*

6.2.2. *Intervention GARCH model*

Now we consider a GARCH model incorporating intervention analysis for both trend and volatility. In order to represent the intervention effects on the volatility,

the intervention term to the constant term of volatility proposed by Watanabe and Nagashima (2016) is incorporated into our model. Our model is defined as follows:

$$X_t = \mu_t + \varepsilon_t \quad [6.7]$$

$$\varepsilon_t = h_t z_t \qquad z_t \sim NID(0,1) \quad [6.8]$$

$$h_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1}^2 + I_t \quad [6.9]$$

$$I_t = \begin{cases} \gamma^{(v)} & (\text{if} \quad t = T_v) \\ 0 & (\text{otherwise}) \end{cases} \quad [6.10]$$

where we assumed that

$$0 < \alpha_0 \ , \ 0 < \alpha_0 + \gamma^{(v)} \ , \ 0 \leq \alpha_1 \ , \ 0 \leq \beta_1 \ , \ \alpha_1 + \beta_1 < 1. \quad [6.11]$$

The terms I_t represent the intervention effects on the volatility. In this chapter, we consider only the simplest type for I_t. The conditional standard deviation h_t is called volatility usually.

We propose intervention trend model for μ_t. Intervention trend model is a model incorporating a temporary structural change in the model in section 6.2.1. Let $T_m (1 \leq T_m \leq T)$ be the time at which an event begins to affect the expected value. Let L_2 be the length of time the effect continues, assumed to be a natural number. For the interval $1 \leq t < T_m$, the interval division in section 6.2.1 is performed in reverse from time T_m. Let K_1 be the number of interval $1 \leq t < T_m$ which is given by

$$K_1 = \left[\frac{T_m + L_1 - 1}{L_1} \right] + 1. \quad [6.12]$$

The middle point of the interval $A_k (k = -K-1, -K-1+1, ..., -1)$ satisfies

$$A_{-1} = T_m \quad [6.13]$$

$$A_k = A_{k+1} - L_1 \quad (-2 \geq k \geq -K_1). \quad [6.14]$$

In each interval, the expected value function is defined in the same way as in section 6.2.1. However, $\mu_t = m_{-1}$ when $A_{-1} - \frac{d}{2} \leq t < A_{-1}$, $\mu_t = m_{-K_1}$ when $1 \leq t \leq A_{-K_1} + \frac{d}{2}$.

The expected value function in the interval, which is from T_m to $T_m + L_2$, is defined as follows:

$$\mu_t = m_1. \qquad [6.15]$$

After the effect of shock, which is $T_m + L_2 < t\ leT$, starting from the point of time $T_m + L_2$ as in the trend model in section 2.1 interval division is performed.

For the interval $T_m + L_2 < t \leq T$, the interval division is performed in the same way as in section 6.2.1 from $T_m + L_2$ and an mean value function is defined. Let K_2 be the number of interval $T_m + L_2 < t \leq T$ which is given by

$$K_2 = \left[\frac{T - (T_m + L_2) + L_1}{L_1}\right] + 1. \qquad [6.16]$$

The middle point of the interval A_k $(k = 1, 2, ..., K_2)$ satisfies

$$A_1 = T_m + L_2 \qquad [6.17]$$

$$A_k = A_{k-1} + L_1 \quad (2 \leq k \leq K_2). \qquad [6.18]$$

In each interval, the expected value function are defined in the same way as in section 6.2.1. However, when $A_1 < t \leq A_1 + \frac{d}{2}$, $\mu_t = m_1$, and when $A_{K_2} - \frac{d}{2} \leq t \leq T$, $\mu_t = m_{K_2}$. An example of the intervention trend is shown in Figure 6.2.

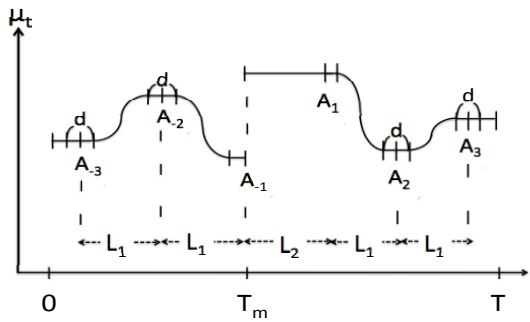


Figure 6.2. *An example of Intervention on the trend*

6.3. Identification

The parameters in the model can be estimated by the maximum likelihood method, if L_1, L_2 and d are given. In this chapter, d is also assumed to be a natural number. The log likelihood function can be written as

$$\ln L = -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\ln(h_t^2) - \frac{1}{2}\sum_{t=1}^{T}(\frac{(X_t-\mu_t)^2}{h_t^2}) \qquad [6.19]$$

where T is the length of series and the initial value h_0 is given.

In this chapter, we identify the trend by the information criterion AIC. For identification of the trend, we prepare the set of candidate value of L_1, L_2 and d and select among them by AIC. In this chapter, $T^{(\mu)}$ and $T^{(v)}$ are assumed to be known, but it is also possible to select by AIC if they are unknown.

In the trend model, when the length of the last interval is too short, the latent variable estimation becomes unstable. Therefore, in such a case, the last two intervals should be combined.

6.4. Simulation

The purpose of this simulation study is to verify the validity of the trend identification method by AIC shown in section 6.3. In this section, we discuss a simulation study focused only on estimating expected value for the proposed model. Also the details of the simulation result are omitted, and only some results are shown.

6.4.1. *Simulation on trend model*

We define the index on trend fluctuations in the series as follows:

$$P_t = \frac{\max(m_k) - \min(m_k)}{\sigma} \qquad [6.20]$$

where m_k is the latent variable of the trend model and σ^2 is the standard deviation of the series. Large P_t means that the expected value function fluctuates wide. In this section, σ^2 is given as 0.5. We generate 100 series for some values of P_t in each model in section 6.2.1. The best model for the generated series is selected by AIC from three models: (1) constant expected value with GARCH, (2) trend with GARCH and (3) constant expected value with Gaussian white noise.

When a model containing trends is selected by AIC, we need to know how close the estimated trend is close to the true trend. Thus, we define the index as follows:

$$D_t = \frac{\Sigma(\mu_t - \bar{x_t})^2 - \Sigma(\mu_t - \hat{\mu}_t)^2}{\max(m_k) - \min(m_k)}, \qquad [6.21]$$

where $\bar{x_t}$ is the sample mean of the series and is an estimate of the expected value function by assuming constant expected value. $\hat{\mu}_t$ is the estimate of the trend derived from the estimate of m_k.

In this example, $T = 300$, $L_1 = 50$ and d is fixed to 1. The value of P_t, D_t and the rate at which the trend model was selected are shown in Table 6.1. A scatter plot of P_t and D_t is shown in Figure 6.3. From Table 6.1 and Figure 6.3, it can be seen that the model is well selected when P_t is larger.

No.	P_t	Rate	D_t	No.	P_t	Rate	D_t	No.	P_t	Rate	D_t
1	0.837	1.00	0.0080	11	0.671	0.96	0.0058	21	0.500	0.94	-0.0024
2	0.807	1.00	0.0615	12	0.658	1.00	0.0258	22	0.488	0.86	-0.0075
3	0.794	1.00	0.0478	13	0.636	1.00	0.0317	23	0.475	0.77	-0.0098
4	0.781	1.00	0.0415	14	0.623	0.98	0.0131	24	0.457	0.91	0.0016
5	0.763	1.00	0.0324	15	0.610	0.98	0.0117	25	0.439	0.97	0.0045
6	0.750	0.98	0.0114	16	0.592	0.99	0.0137	26	0.421	0.99	0.0154
7	0.746	0.97	0.0137	17	0.579	1.00	0.0228	27	0.409	0.88	-0.0076
8	0.728	1.00	0.0211	18	0.567	0.95	0.0143	28	0.396	0.77	-0.0120
9	0.715	0.98	0.0156	19	0.544	0.99	0.0188	29	0.365	0.77	-0.0115
10	0.702	1.00	0.0207	20	0.531	0.99	0.0154	30	0.352	0.97	0.0068

Table 6.1. *Result of simulation ($T = 300$, $L_1 = 50$)*

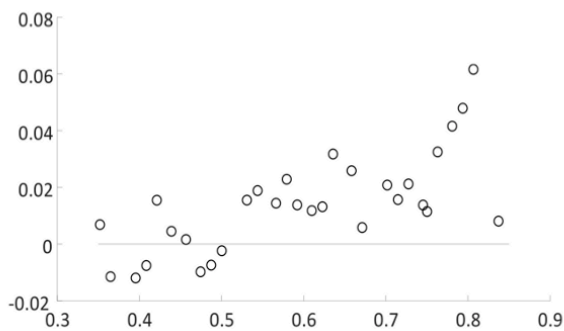


Figure 6.3. *A scatter plot of P_t and D_t*

6.4.2. *Simulation on intervention trend model*

Let the magnitude of the effect on the trend be $\gamma^{(m)}$. Then the latent variable m_1 in the intervention trend model is given by

$$m_1 = \gamma^{(m)} + m_{-1}. \qquad [6.22]$$

We generate 100 series for some values of P_t and the magnitude of effect $\gamma^{(m)}$ in the intervention trend model in section 6.2.3. The best model for generated series is selected by AIC from models by adding the intervention trend with GARCH to three models in the previous section. In this example, $T = 400$, $T_m = 200$, $L_1 = 30$, $L_2 = 60$, and d is fixed to 1. The value of P_t, D_t,$\gamma^{(m)}$ and the rate at which the intervention trend model is selected, are shown in Table 6.2. From Table 6.2, it can be seen that the model is well selected when $\gamma^{(m)}$ is large. The histogram of the estimated values of $\gamma^{(m)}$ in No. 1 in Table 6.2 is shown in Figure 6.4.

No.	$\gamma^{(m)}$	P_t	D_t	Rate	No.	$\gamma^{(m)}$	P_t	D_t	Rate
1	1.00	0.803	0.1162	0.96	7	0.50	0.596	0.0508	0.84
2	0.75	0.791	0.1083	0.88	8	0.25	0.614	0.0299	0.69
3	0.50	0.812	0.0517	0.82	9	1.00	0.410	0.1029	0.96
4	0.25	0.809	0.0468	0.61	10	0.75	0.391	0.1028	0.88
5	1.00	0.609	0.1357	0.98	11	0.50	0.413	0.0149	0.85
6	0.75	0.624	0.1039	0.95	12	0.25	0.405	0.0020	0.65

Table 6.2. *Result of simulation ($T = 400$, $T_m = 200$, $L_1 = 30$, $L_2 = 60$)*

The vertical line in Figure 6.4 is the value of true $\gamma^{(m)}$, and Figure 6.4 shows that the estimation procedure works well.

6.5. Application

In this section, we apply the proposed model to practical data. The series of daily stock index to be used is shown in Figure 6.5. Figure 6.6 shows the series of its returns. This is series of TOPIX from January 27, 2012, to November 20, 2013. From Figure 6.5, it seems that the expected value of the series of returns include a varying trend. Figure 6.5 shows also that the rising effect appears from the time of the vertical line, which is the day the Japanese House of Representatives dissolved. Furthermore, a little after the effect of the shock appears in the expected value, the effect of shock appears also in the volatility. Therefore, it is necessary to use different time points T_m and T_v.

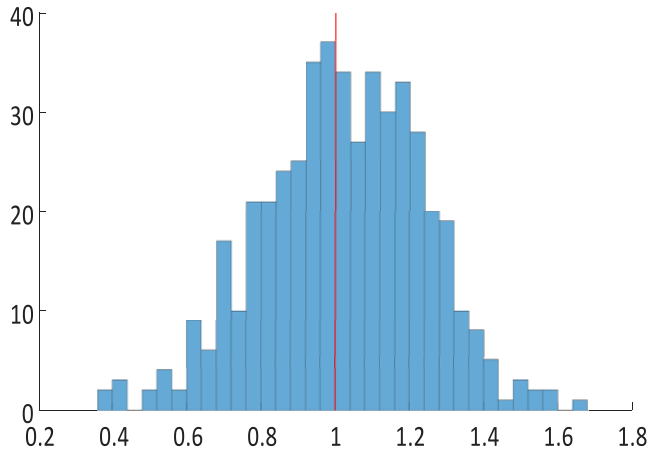


Figure 6.4. *The histogram of estimated value of* $\gamma^{(m)}$

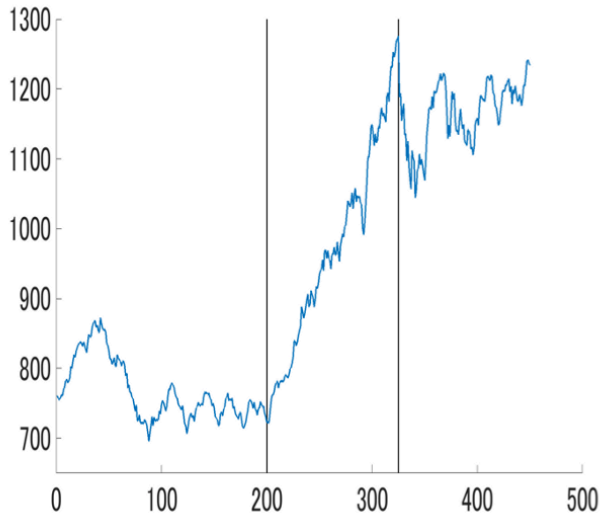


Figure 6.5. *TOPIX (Tokyo Price Index)*

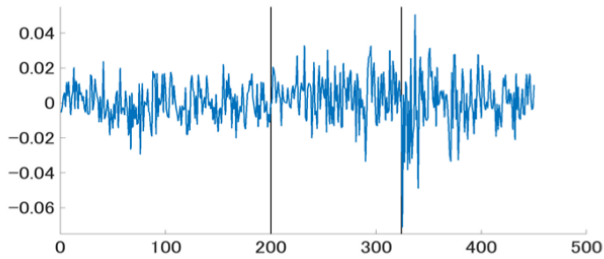


Figure 6.6. *Returns of TOPIX*

We apply a constant expected value, trend model proposed in section 6.2.1 and intervention trend model proposed in section 6.2.2. In both models, we assume the intervention GARCH (1,1) model for volatility, which is shown in section 6.2.2. The length of series of stock returns is $T = 450$. The time point T_m is the day when the Japanese House of Representatives dissolved 200 and T_v is 325. In this chapter, we set $d = 1$. L_1, L_2 and K selected by AIC for trend model and intervention trend model are shown in Table 6.3. The AIC is shown in Table 6.4. In Table 6.4, it can be considered that a model having a trend is appropriate for these data. Furthermore, Table 6.4 shows that the model considering effect on shock is more suitable. Table 6.5 shows the estimated parameters in the intervention trend model. The whole estimated expected value function is shown in Figure 6.7, and the series $t = 1, 2, ..., 250$ is shown in Figure 6.8.

$\boldsymbol{\mu}_t$	**Length of interval**	**Number of interval**
Trend	$L_1 = 77$	$K = 6$
Intervention trend	$L_1 = 66, L_2 = 119$	$K_1 = 4, K_2 = 2$

Table 6.3. *Slected L_1, L_2 and K*

$\boldsymbol{\mu}_t$	AIC
Constant	-2688.81
Trend	-2695.26
Intervention trend	-2700.46

Table 6.4. *AIC of models*

$\boldsymbol{a}$	**b**	$\boldsymbol{a}$	**b**	$\boldsymbol{a}$	**b**	$\boldsymbol{a}$	**b**
m_{-4}	0.004786	m_{-1}	-0.001372	α_0	0.000009	$\gamma^{(v)}$	0.0016311
m_{-3}	-0.003408	m_1	0.004501	α_1	0.052395		
m_{-2}	0.000752	m_2	0.000920	β_1	0.873430		

Table 6.5. *The estimated parameters*

The solid line is the estimated value by the intervention trend, the dotted line is by the trend model and the gray line is by assuming the expected value constant. The whole estimated values of volatility are shown in Figure 6.9, and the series of $t = 1, 2, ..., 250$ are shown in Figure 6.10. The type of line is the same as shown in Figures 6.7 and 6.8.

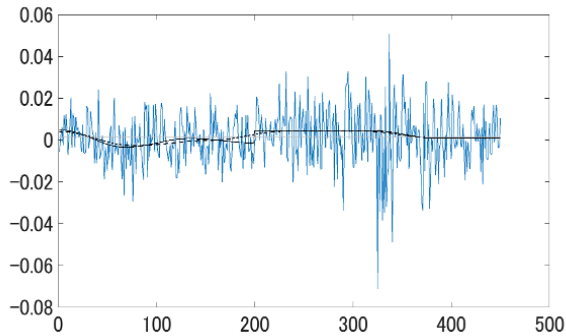


Figure 6.7. *Estimated trend (all)*

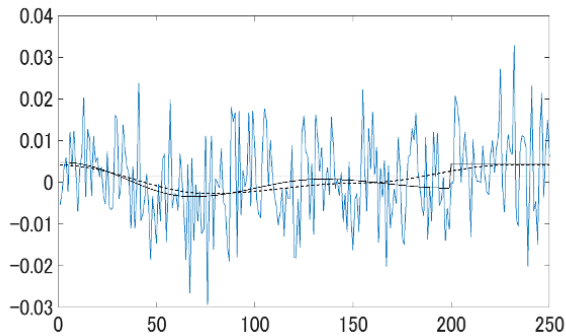


Figure 6.8. *Estimated trend (t=1,2,...,250)*

As shown in Figures 6.9 and 6.10, the volatility in the trend model is lower compared to the constant expected value model. This means that volatility is overestimated if a varying trend is not considered. Moreover, from the time point T_m, the estimate of the volatility by the model without intervention term becomes higher than the model with the intervention term. Therefore, it is necessary to apply a model reflecting the characteristics of data as proposed in this chapter.

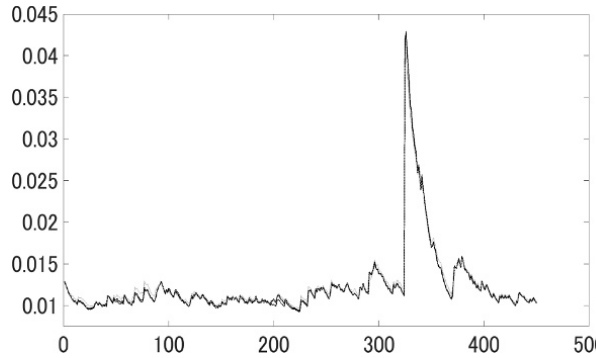


Figure 6.9. *Estimated volatality (all)*

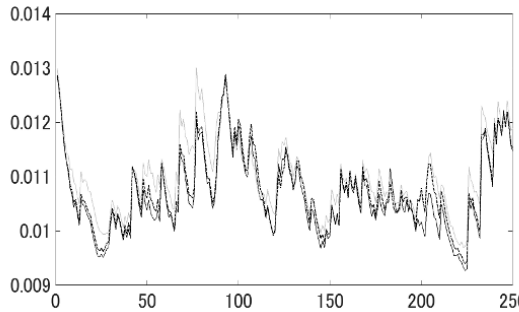


Figure 6.10. *Estimated volatality: (t=1,2,...,250)*

6.6. Concluding remarks

In this chapter, we proposed a new trend model and the intervention GARCH model based on it. Through application to practical data, it was shown that there is a possibility of overestimating volatility when the usual GARCH model is applied. This means that it is necessary to consider GARCH models with trend for adequate estimation of volatility.

It should be noted that the calculation for estimation requires the numerical optimization under the constraint. Some software might fail in such calculation. In this study, we use the software MATLAB R2016a for parameter estimation.

In this chapter, the pattern of intervention for our proposed trend model is limited. As future work, other intervention patterns should be considered for various applications.

6.7. References

Box, G.E.P., Tiao, G.C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the Statistical Association*, 70(349), 70–79.

Ho, A.K.F., Wan, A.T.K. (2002). Testing for covariance stationary of stock returns in the presence of structural breaks: An intervention analysis. *Applied Economics Letters*, 9, 441–447.

Kuwahara, Y., Watanabe, N. (2008). Analysis of financial time series data by fuzzy trend model (Japanese). *Intelligence and Information (Journal of Japan Society for Fuzzy Theory and Intelligent Informatics)*, 20(2), 92–102.

Watanabe, N., Nagashima, M. (2016). An intervention analysis based on the GARCH model. *SMTDA 2016 Proceedings*, 515–525.

7

A Note on the Linear Approximation of TAR Models

The linear approximation of nonlinear time series models is not an easy task. In this chapter, we give a definition of linear process and we distinguish between linear approximation and linear representation of nonlinear models, briefly giving some examples that better clarify this distinction. The attention is here focused on the threshold autoregressive models whose linear approximation is discussed starting from a motivating example and some theoretical issues.

7.1. Introduction

The complexity of most nonlinear models often leads to evaluate if a linear representation or a linear approximation can be admitted for this class of models. In the presence of linear representation, the aim can be ascribed to the need to take advantage (under proper assumptions) of the large and strengthened literature developed in the linear domain (to cite the main references, Box and Jenkins 1976, Brockwell and Davies 1991) whereas linear approximations can be seen as a tool for model selection (or more generally to select candidate models for the data under analysis) to "filter" the dynamic relationship among variables such that the "purely" nonlinear component, obtained in output, can be properly examined.

Before showing the main advantages obtained from the linearization, it is useful to clarify when a stochastic process $\{X_t\}$, with $t \in \mathbb{Z}$, is said to be linear.

Chapter written by Francesco GIORDANO, Marcella NIGLIO and Cosimo Damiano VITALE.

Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining,
First Edition. Edited by Christos H. Skiadas and James R. Bozeman.

Let $\{X_t\}$ be a mean zero stationary process and let $\{e_t\}$ be a sequence of white noise, with $E[e_t] = 0$ and $E[e_t^2] = \sigma^2 < \infty$. From the Wold decomposition, X_t can be expressed as:

$$X_t = \sum_{i=0}^{\infty} \psi_i e_{t-i} + D_t \qquad [7.1]$$

with $\psi_0 = 1$, $\sum_{i=1}^{\infty} \psi_i^2 < \infty$, $E[e_t D_s] = 0$, for all $s, t \in \mathbb{Z}$, and D_t a deterministic component.

Starting from the decomposition [7.1], a zero mean stationary process X_t is said to be linear if it can be given as:

$$X_t = \sum_{i=0}^{\infty} \psi_i e_{t-i} \qquad [7.2]$$

with $\{e_t\} \sim$IID$(0, \sigma^2)$[1].[1]

It can be easily shown that the ARMA(p, q) model belongs to the linear class (some authors identify the linear class with ARMA models) and its widely known structure is given by:

$$X_t - \sum_{i=1}^{p} \phi_i X_{t-i} = e_t - \sum_{j=1}^{q} \theta_j e_{t-j} \qquad [7.3]$$

where well-defined assumptions are given on the parameters ϕ_i and θ_j to guarantee the stationarity and invertibility of the model (Box and Jenkins 1976).

Starting from this definition of linear process, the aim of the present paper is to show how to obtain the "best linear approximation" (in terms of $\mathbb{L}^2$ norm) of a nonlinear process. In particular, in section 7.2 we further clarify the difference between linear representation and linear approximation of nonlinear models, and then in section 7.3, we provide new results on the linear approximation of the threshold autoregressive (TAR) model (Tong 1990). Some examples with simulated data give evidence of the advantages that can be obtained from the use of the theoretical issues proposed.

1 Note that in some cases, $\{e_t\}$ is assumed to be a sequence of uncorrelated Gaussian random variables and so the independence is guaranteed as well.

7.2. Linear representations and linear approximations of nonlinear models

The linearization of nonlinear processes has been differently intended in the literature. Ozaki (1992) proposes a local linearization of a nonlinear continuous dynamical system using (under proper requirements) a discrete time autoregressive approximation over a sufficiently small time interval Δt; Francq and Zakoïan (1998) investigate on the properties of the estimators of the parameters of the so-called *weak* ARMA models when some assumptions, usually given on the innovations e_t (Box and Jenkins 1976), do not hold. The estimation procedure is based on the minimization of the squared deviations about the linear conditional expectation and for this reason the estimated model is seen as weak linear representation of nonlinear models.

If we want to face organically the linearization problem of the nonlinear process Y_t, we can consider two main approaches: the first considers the *linear representation* of the nonlinear model (where the nonlinear structure is rewritten in alternative linear form, after the introduction of proper assumptions); the second approach makes a distinction between the linear (X_t) and the "purely" nonlinear (V_t) component of the process Y_t, such that it can be decomposed as:

$$Y_t = X_t + V_t. \qquad [7.4]$$

This decomposition is usually made through linear approximations, often obtained from proper expansions of Y_t.

Examples of the first and the second approach have been differently proposed in the literature.

Consider a GARCH(p, q) model (Bollerslev 1986) for the conditional variance of Y_t, it can be shown that this model admits a linear representation. Let $Y_t \sim$GARCH(p, q):

$$\begin{aligned} Y_t &= h_t \epsilon_t \\ h_t &= c + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j h_{t-j}, \end{aligned} \qquad [7.5]$$

with ϵ_t and i.i.d. sequence with $E[\epsilon_t] = 0$ and $E[\epsilon_t^2] = 1$. If we fix $u_t = \epsilon_t^2 - h_t$, model [7.5] becomes:

$$\epsilon_t^2 = c + \sum_{i=1}^{\max\{p,q\}} (\alpha_i + \beta_i)\epsilon_{t-i}^2 + u_t - \sum_{j=1}^{p} \beta_j u_{t-j}$$

with $\alpha_i = 0$, for $i > q$ and $\beta_i = 0$, for $i > p$. In other words $\epsilon_t^2 \sim$ARMA$(\max\{p,q\},p)$ model.

The distinction between the linear and the "purely" nonlinear component, introduced with the second approach, is traditionally based on the Volterra series expansion of Y_t (among the others, Priestley 1988, Tong 1990). In more detail, let $f(Y_t, Y_{t-1}, Y_{t-2}, \ldots) = e_t$, with $f(\cdot)$ an invertible function, then $Y_t = g(e_t, e_{t-1,\ldots})$ where $g(\cdot)$ is a well-behaved nonlinear function that can be expanded, near the origin $\mathbf{0} = (0, 0, \ldots)$, in Taylor series. Under these conditions, Y_t can be given as:

$$Y_t = k_0 + \sum_{i=0}^{\infty} k_i e_{t-i} + \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} k_{ij} e_{t-i} e_{t-j} + \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{w=0}^{\infty} k_{ijw} e_{t-i} e_{t-j} e_{t-w} + \ldots \quad [7.6]$$

where $k_0 = g(\mathbf{0})$, $k_i = \frac{\partial g}{\partial e_{t-i}} \mid_{\mathbf{0}}$, $k_{ij} = \frac{\partial g}{\partial e_{t-i}\partial e_{t-j}} \mid_{\mathbf{0}}$ and so on.

It is clear that when $k_{ij} = k_{ijw} = \ldots = 0$, the linear approximation of Y_t is obtained.

An example of linear approximation of a nonlinear process can be easily shown if we consider the bilinear model (Subba Rao 1981):

$$Y_t + \sum_{j=1}^{p} a_j Y_{t-j} = \sum_{j=0}^{r} c_j \epsilon_{t-j} + \sum_{i=1}^{m}\sum_{i'=1}^{k} b_{ii'} X_{t-i} \epsilon_{t-i'} \quad [7.7]$$

where the "purely" nonlinear component is given by the last term on the right of equation [7.7]. If $b_{ii'} = 0$, for $i < i'$, the "purely" nonlinear component (even called superdiagonal bilinear model) is such that its terms are (at least) uncorrelated and it makes the derivation of the linear approximation easier. In other cases, this approximation is more difficult and for this reason it has been investigated using different expansions (Guegan 1987).

The bilinear model [7.7] is often seen as first example of generalization of the linear ARMA model in nonlinear domain.

Another example is given by the threshold autoregressive model (Tong and Lim 1980):

$$Y_t = \sum_{j=1}^{k} \left(\phi_0^{(j)} + \sum_{i=1}^{p} \phi_i^{(j)} Y_{t-i} \right) \mathbb{I}(Y_{t-d} \in \mathcal{R}_j) + \epsilon_t \quad [7.8]$$

where k is the number of autoregressive regimes, p is the autoregressive order, Y_{t-d} is the threshold variable, d is the threshold delay, $\mathcal{R}_j = [r_{j-1}, r_j)$, for $j = 1, \ldots, k$, such that $\mathcal{R} = \bigcup_{j=1}^{k} \mathcal{R}_j$ and $-\infty = r_0 < r_1 < \ldots < r_{k-1} < r_k = +\infty$.

When $k = 1$, the threshold model [7.8] degenerates to a linear autoregressive model, whereas when $k > 1$, the linear approximation is not so immediate.

In section 7.3, we present some new results of the linear approximation of model [7.8]. It is based on the use of an alternative representation of the threshold model as discussed in the following.

7.3. Linear approximation of the TAR model

The investigation of threshold models has been differently faced. (Petruccelli 1992) has evaluated the ability of threshold models to approximate exponential autoregressive and invertible bilinear processes. Here, the attention is focused on a different aspect related to the ability of the linear ARMA models to approximate the SETAR structures.

Let Y_t be a threshold model [7.8] that, for easy of exposition, is assumed to have $k = 2$ regimes and null intercepts ($\phi_0^{(j)} = 0$, for $j = 1, 2, \ldots, k$):

$$Y_t = \sum_{i=1}^{p} \phi_i^{(1)} Y_{t-i} \mathbb{I}(Y_{t-d} \leq r_1) + \sum_{i=1}^{p} \phi_i^{(2)} Y_{t-i} [1 - \mathbb{I}(Y_{t-d} \leq r_1)] + \epsilon_t. \quad [7.9]$$

Model [7.9] can be alternatively written as:

$$\begin{aligned} \mathbf{Y}_t &= \mathbf{\Phi}_1 \mathbf{Y}_{t-1} \mathbb{I}(Y_{t-d} \leq r_1) + \mathbf{\Phi}_2 \mathbf{Y}_{t-1} [1 - \mathbb{I}(Y_{t-d} \leq r_1)] + \boldsymbol{\epsilon}_t, \\ &= \mathbf{\Phi}_2 \mathbf{Y}_{t-1} + \boldsymbol{\epsilon}_t + (\mathbf{\Phi}_1 - \mathbf{\Phi}_2) \mathbf{Y}_{t-1} \mathbb{I}(Y_{t-d} \leq r_1) \end{aligned} \quad [7.10]$$

where

$$\mathbf{Y}_t = \begin{bmatrix} Y_t \\ \ldots \\ Y_{t-p+1} \end{bmatrix}_{(p \times 1)}, \quad \mathbf{\Phi}_j = \left[\begin{array}{ccc|c} \phi_1^{(j)} & \ldots & \phi_{p-1}^{(j)} & \phi_p^{(j)} \\ \hline & \mathbf{I}_{p-1} & & \mathbf{0} \end{array}\right]_{(p \times p)},$$

$$\boldsymbol{\epsilon}_t = \begin{bmatrix} \epsilon_t \\ \mathbf{0} \end{bmatrix}_{(p \times 1)}$$

for $j = 1, 2$, with $\mathbf{I}$ the identity matrix and $\mathbf{0}$ the null vector.

From equation [7.10], it seems easy to discriminate the linear and the nonlinear components of the threshold model: in fact, if we use the same approach considered for model [7.7], the last term of [7.10] could represent the "purely" nonlinear component of the model.

Note that the bilinear and the threshold models have a not negligible difference: as remarked before, if we consider the "purely" nonlinear component of model [7.7], it can be shown that, under proper conditions on the values of i and i', its terms are uncorrelated (Granger and Andersen 1978), whereas similar results do not hold for the "purely" nonlinear component of the threshold model [7.10]. It implies that the linear approximation cannot be limited to the first two terms of [7.10] but it needs a more detailed investigation.

In this regard, consider the following example.

EXAMPLE 7.1.– Let Y_t be a stationary and ergodic threshold autoregressive model (for the stationarity and ergodicity conditions, see Petruccelli and Woolford 1984) with autoregressive order $p = 1$:

$$Y_t = \begin{cases} -1.44Y_{t-1} + \epsilon_t & Y_{t-1} \leq 0 \\ 0.18Y_{t-1} + \epsilon_t & Y_{t-1} > 0 \end{cases} \qquad [7.11]$$

with $\epsilon_t \sim N(0, 1)$. If we generate $T = 1,000$ artificial data (with burn-in 500) from model [7.11], the plots of the autocorrelation function (ACF) and the partial autocorrelation function (PACF) are shown in Figure 7.1.

Following the decomposition in equation [7.10], it seems that the linear component should be an AR(1) structure that, if fitted to the artificial data, does not catch all the linearity of the generating process, as can be clearly evaluated from the correlograms of the residuals in Figure 7.2. □

A first and naïve linear approximation of the nonlinear process Y_t can be given defining, for the parameters of the linear model, a set of values obtained as weighted mean of the parameters of the threshold model, as illustrated in the following example.

EXAMPLE 7.2.– Let Y_t be a stationary and ergodic threshold autoregressive model:

$$Y_t = \begin{cases} 0.12Y_{t-1} + \epsilon_t & Y_{t-1} \leq 0 \\ 0.36Y_{t-1} + \epsilon_t & Y_{t-1} > 0 \end{cases} \qquad [7.12]$$

with $\epsilon_t \sim N(0, 1)$. We generate $T = 1,000$ artificial data (with burn-in 500) whose corresponding correlograms are shown in Figure 7.3.

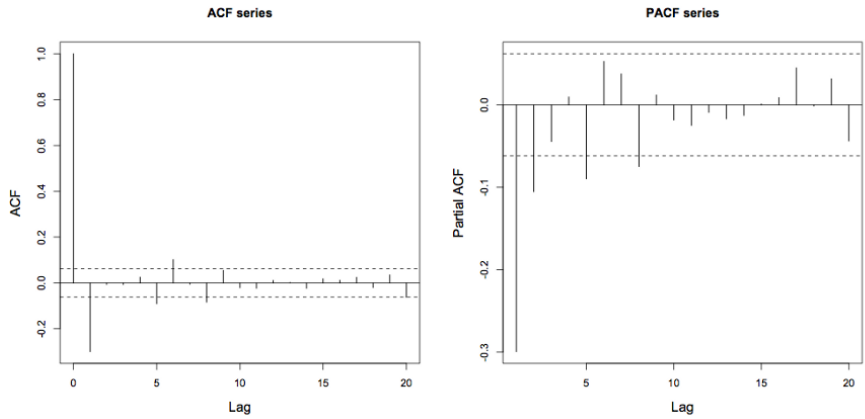


Figure 7.1. *ACF and PACF of the artificial data generated from model [7.11]*

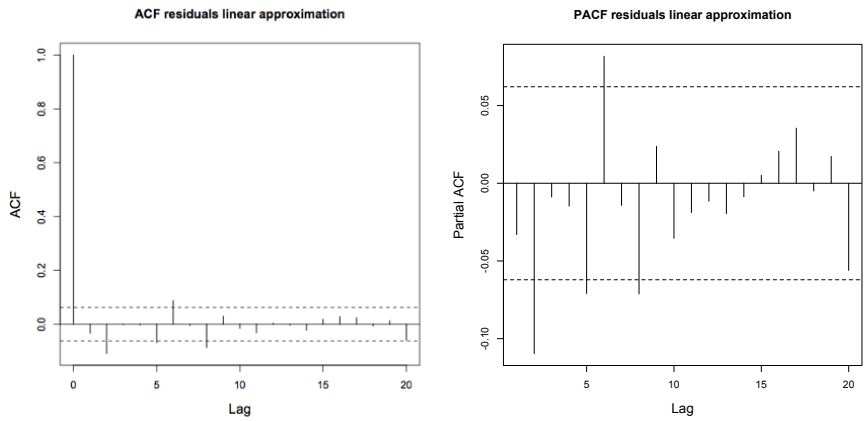


Figure 7.2. *ACF and PACF of the residuals obtained after fitting an AR(1) model to the artificial data generated from model [7.11]*

If we further generate $T = 1,000$ artificial data from an autoregressive model $X_t = \phi X_{t-1} + \epsilon_t$ (with the same innovations of model [7.12]), where ϕ is a weighted mean of the parameters used in model [7.12], such that $\phi = 0.12 * \lambda + 0.36 * (1 - \lambda)$, with $\lambda = P[Y_{t-1} \leq 0]$, the correlograms of the simulated data are shown in Figure 7.4.

From the comparison of Figures 7.3 and 7.4, it can be noted that the ACF and PACF of both series are similar: it gives empirical evidence of the ability of the autoregressive approximation to catch the linear component of the series Y_t such that the "purely" nonlinear component (V_t) can be properly investigated. □

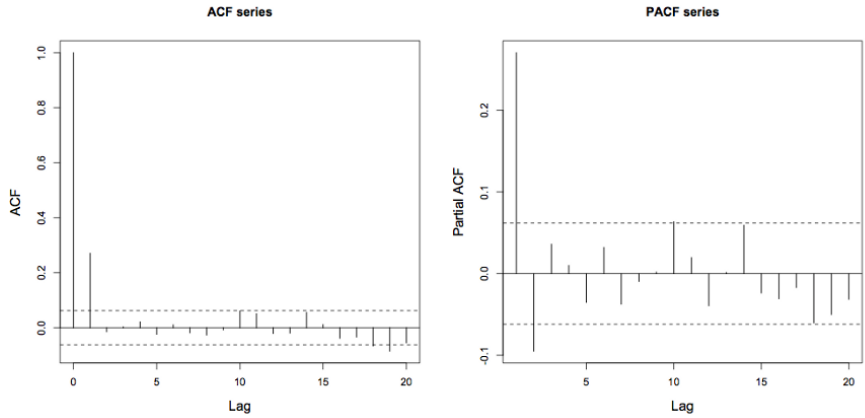


Figure 7.3. *ACF and PACF of the artificial data generated from model [7.12]*

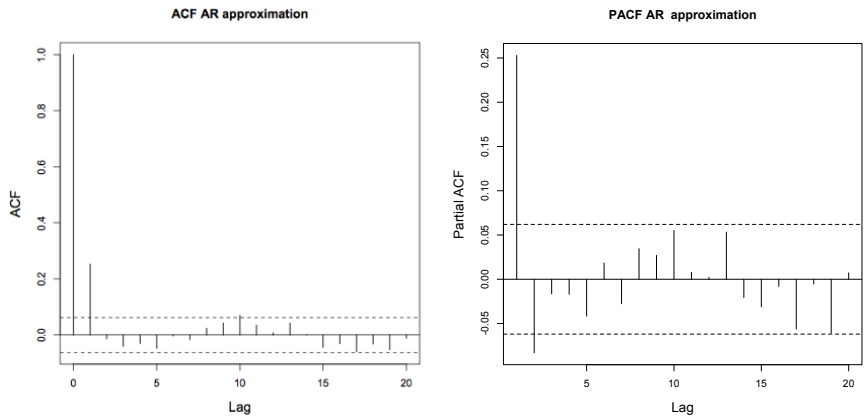


Figure 7.4. *ACF and PACF of the artificial data generated from the linear approximation of model [7.12]*

The empirical evidence of examples 7.1 and 7.2 introduces what we state in the following proposition (whose proof is omitted for brevity):

PROPOSITION 7.1.– *Let Y_t be a stationary and ergodic threshold process [7.9] with $E[Y_t^2] < \infty$. The best linear approximation, in $\mathbb{L}^2$ norm, of Y_t is given by $X_t \sim$ARMA(2p; 2p) model.*

Proposition 7.1 allows us to further investigate and to revise the results of example 7.1.

Example 1 (cont.). Given the artificial data generated in example 7.1 and following the results of proposition 7.1, the best linear approximation of Y_t is given by an ARMA(2,2) model whose ACF of the residuals (that represent the "purely" nonlinear component) and of the squared residuals are presented in Figure 7.5. It can be clearly noted that, differently from the results in example 7.1, the linear approximation completely catches the linear component (X_t) of the generating process, whereas the squared residuals show the existence of a nonlinear component (V_t) that can be evaluated. □

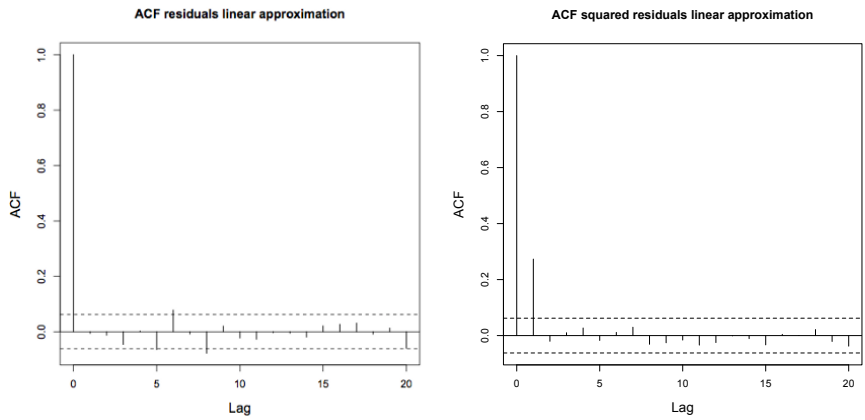


Figure 7.5. *On the left, ACF of the residuals of the ARMA approximation of the data generated from model [7.11]; on the right, ACF of the squared residuals*

Even the results of example 7.2 can be further discussed.

Example 2 (cont.). It can be noted that the linear AR(1) structure considered can be seen as the dominant part of the ARMA(2,2) model that, from proposition 7.1, represents the best linear approximation of the generating process Y_t.

It can be empirically evaluated in Figure 7.6 where the ACF and PACF of the residuals of the ARMA(2,2) model fitted to Y_t are represented.

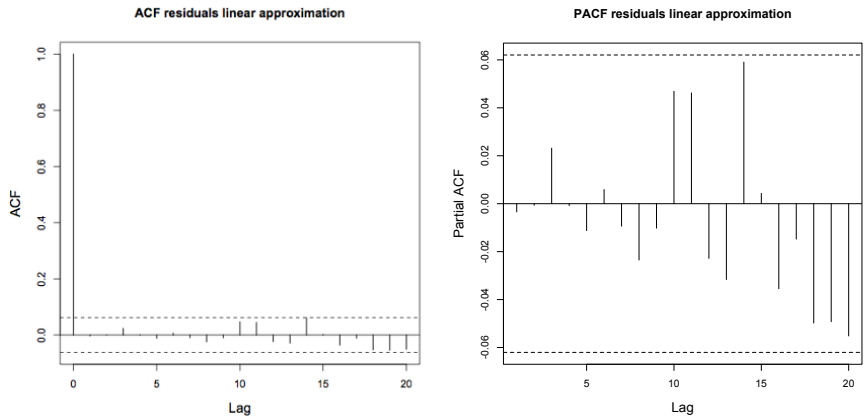


Figure 7.6. *ACF (left panel) and PACF (right panel) of the residuals of the ARMA(2,2) approximation of the data generated from the SETAR model in example 7.2*

Further note that when in decomposition [7.4] $X_t \equiv 0$, the process Y_t becomes "purely" nonlinear and so the autocorrelations of the series cannot be significantly different from zero, as stated in corollary 7.1.

COROLLARY 7.1.– *Given model [7.9], under the assumptions of proposition 7.1, there exists a threshold process where the linear component is identically null.*

It can be empirically illustrated showing that proper combinations of the parameters of the autoregressive regimes can lead to $X_t \equiv 0$ in [7.4].

EXAMPLE 7.3.– Let Y_t be a stationary and ergodic threshold autoregressive model:

$$Y_t = \begin{cases} 0.50Y_{t-1} + \epsilon_t & Y_{t-1} \leq 0 \\ -0.90Y_{t-1} + \epsilon_t & Y_{t-1} > 0 \end{cases} \qquad [7.13]$$

with $\epsilon_t \sim N(0,1)$. In Figure 7.7(a), the ACF does not show a significant linear dependence among the data that on the contrary becomes evident if we consider the ACF of Y_t^2 (b). In fact, if we compute the parameter ϕ as in example 7.2, its value is very near to zero and so the nonlinear structure of data prevails (it can be clearly appreciated from the correlogram of Y_t^2). □

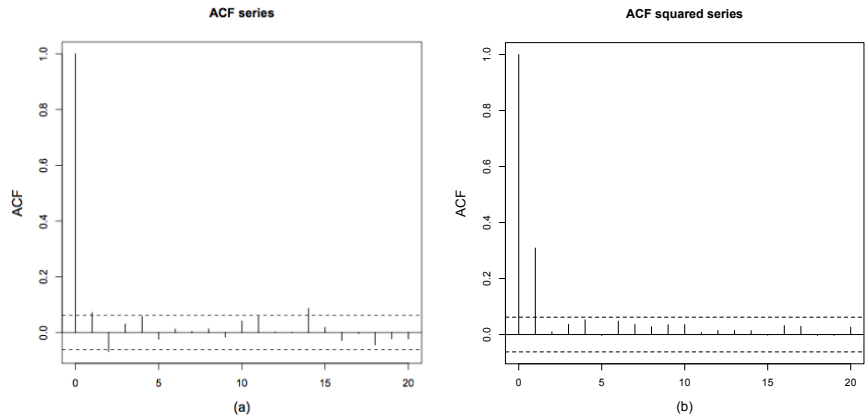


Figure 7.7. *(a) ACF of the artificial data (Y_t) generated from model [7.13]; (b) ACF of Y_t^2*

REMARK 7.1.– What stated in corollary 7.1, and empirically shown in example 7.3, has a main remarkable consequence: the linear approximation of the threshold model can be seen as a proper reparameterization of the process Y_t. In fact, when the parameters of the autoregressive regimes assume well-defined values of the parametric space (such that the linear component X_t becomes identically null), the process Y_t is "purely" nonlinear.

To conclude, it is interesting to note what distinguishes our results from those given in (Francq and Zakoïan 1998). As said before, they consider the estimation of a linear ARMA model [7.3] under "weak" assumptions on the innovations e_t and their aim is to show that, under proper conditions, the strong consistency and asymptotic normality of the estimators still hold.

These results are not negligible: in fact, if applied in nonlinear domain, they allow us to state that if well-defined assumptions are verified on the generating process, the estimated ARMA model is a "weak" linearization of the nonlinear model. The results can be even applied to the general nonlinear threshold generating process [7.9] under proper assumptions related to the existence of moments and the geometric ergodicity (An and Huang 1996), theorem 3.2.

A problem that has not been addressed in Francq and Zakoïan (1998) is the identification of the linear approximation that on the contrary has been introduced in this study where, given a threshold autoregressive model, a relation between the order of the autoregressive regimes and the order of the ARMA model is stated. Further, we have highlighted that the ARMA approximation is obtained, for this class of models,

reparameterizing the threshold process that, under proper conditions on the values assumed by its autoregressive coefficients, becomes a "purely" nonlinear process.

7.4. References

An, H.Z., Huang, F.C. (1996). The geometrical ergodicity of nonlinear autoregressive models. *Statistica Sinica*, 6, 943–956.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31, 307–327.

Box, G.E.P., Jenkins, G.M. (1976). *Time series analysis: Forecasting and control*. Holden-Day, New York.

Brockwell, P.J., Davies, R.A. (1991). *Time series: Theory and methods*, Springer, New York.

Francq, C., Zakoïan, J.M. (1998). Estimating linear representation of nonlinear processes. *Journal of Statistical Planning and Inference*, 68, 145–165.

Granger, C.W.J., Andersen, A.P. (1978). *An introduction to bilinear time series*. Vandenhoeck & Ruprecht, Göttingen.

Guegan, D. (1987). Different representations for bilinear models. *Journal of the Time Series Analysis*, 8, 389–408.

Ozaki, T. (1992). A bridge between nonlinear time series models and nonlinear stochastic dynamical systems: A local linearization approach. *Statistica Sinica*, 2, 113–135.

Petruccelli, J.D. (1992). On the approximation of time series by Threshold Autoregressive models. *Sankhya (B)*, 54, 106–113.

Petruccelli, J.D., Woolford, S.W. (1984). A threshold AR(1) model. *Journal of Applied Probabability*, 21, 270–286.

Priestley, M.B. (1988). *Non-linear non-stationary time series analysis*. Academic Press, London.

Subba Rao, T. (1981). On the Theory of Bilinear Time Series Models. *Journal of the Royal Statistical Society (B)*, 43, 244–255.

Tong, H. (1990). *Non-linear Time Series: A Dynamical System Approach*. Oxford University Press, New York.

Tong, H., Lim, K.S. (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society (B)*, 42, 245–292.

8

An Approximation of Social Well-Being Evaluation Using Structural Equation Modeling

Periurban livestock has proved to be important since it caters to food cities, transforms waste from local restaurants and markets, and provides economic, cultural and social incentives that provide well-being to families and the entire community. The objective was established by indicators of social welfare in periurban communities through categorical principal components analysis and structural equation modeling to understand why farming families continue to survive in spite of high production costs, pressure for land use and environmental pollution.

8.1. Introduction

Urban spaces have gradually gained ground with respect to rural areas. Migration of people from the countryside to cities is becoming more frequent as they seek better employment opportunities, wages and living conditions. This causes communities on the periphery to grow and provide adequate public services.

Health is one of the important needs that a human has, and according to World Health Organization (WHO) (OMS 2015), this concept involves not only the absence of diseases or pathologies, but also includes social and psychological aspects.

Chapter written by Leonel SANTOS-BARRIOS, Monica RUIZ-TORRES, William GÓMEZ-DEMETRIO, Ernesto SÁNCHEZ-VERA, Ana LORGA DA SILVA and Francisco MARTÍNEZ-CASTAÑEDA.

Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining,
First Edition. Edited by Christos H. Skiadas and James R. Bozeman.

Welfare is another aspect that is closely related to health; Graaff (1967) mentions that it is the state in which one is well, without diseases, with health, energy, etc. Therefore, it can be compared with happiness, joy, satisfaction, etc.

FAO (2014) has given importance to small-scale family farming activities carried out in rural, urban and periurban communities because it estimates that 20% of total consumption in cities around the world is produced by small-scale systems. In Mexico, for example, according to SAGARPA (2012), about 25% of the pork meat is produced by family farm. This kind of farm has a main characteristic that it is carried out in confined spaces, generally in the backyards of houses, and uses organic waste of the house.

Despite the studies about the importance of small-scale family farming, the role in health and social welfare of this kind of farm is not clear, so the objective of this study was to establish indicators of social welfare in periurban communities through structural equation models (SEMs) to understand why farming families continue to survive despite high production costs, pressure for land use and environmental pollution.

8.2. Wellness

It is the state in which one is well, without diseases, with health, energy, etc. Therefore, it can be brought by happiness, joy, satisfaction, etc. This concept includes material and immaterial factors; while material factors are measurable, immaterial factors are difficult to measure as they represent states of mind and psychological factors of the human being (Graaff 1967).

The concept of well-being is mainly concerned with qualitative factors, which are generated through the integration of growth and development for the satisfaction of both material and immaterial needs of individuals in a society (Graaff 1967).

The well-being of a society is not only the absence of illnesses or weaknesses, but it also can be expressed by the strengths of the community, such as facing stress in complex situations (Rodríguez *et al.* 2008).

8.3. Social welfare

According to Putnam (2000), social capital plays a significant role in every aspect of our personal and community life, so it is affected by physical health (Cabañero *et al*. 2004) and especially by mental health (U.S. Department of Health and Human Services 2001). Social welfare depends on the role of the people who

work in a society and what we do about certain circumstances (Keyes 1998). It is composed of the following dimensions:

1) Social integration (SI): It is defined as "the assessment of the quality of the relationships we maintain with society and community" (Keyes 1998). This dimension considers that "healthy people feel part of a society", and establish social networks, which includes family, friends, neighbors, etc.

2) Social acceptance (SA): It is imperative for mental health (Keyes 1998) to be and to belong in a community, and this feeling must have three qualities: trust, acceptance and positive attitudes toward others (attribution of honesty, kindness, kindness, capacity), and acceptance of the positive and negative aspects of our own lives.

3) Social contribution (SC): The feeling of being useful is important for self-confidence because if you "are a vital member of society, you have something useful to offer the world" (Keyes 1998), and according to Bandura (1997), to be useful is a synonymous of utility, profit and efficiency.

4) Social update (SU): This dimension considers that society and institutions are dynamic entities. They move in a certain direction in order to achieve benefits (trust in progress and social change). In addition, social updating implies the belief that society controls its destiny and drawing roads for the future. People who are healthier from the mental point of view trust in the future of society, their potential for growth and development and their ability to produce well-being (Keyes 1998).

5) Social coherence (SCo): The cohesion refers to the ability of understanding social dynamics, such as the way of organizing the social world and the concern to know what is happening in the world. Healthy people not only care about their world, but also they have the feeling that they are able to understand what is happening around them (Keyes 1998). According to Blanco and Diaz (2005), people find a logic in the events that surround them.

8.4. Methodology

The data were obtained through semistructured interviews, thus monitoring 70 farms by applying a questionnaire to measure social well-being, which was adapted from Keyes (1998). Questions were related to social integration, social acceptance; social contribution, social update and cohesion of producers of bovine, porcine, ovine and poultry from the community of San Miguel Coatlinchan, Texcoco, State of Mexico, Mexico.

Only sixty-nine answers were registered. The questionnaire has several items, which are related to each dimension, as discussed in section 8.5.

In this work, we construct each of the described dimensions based on categorical principal component (CATPCA), after which the well-being indicator, based on these components, is constructed by an SEM. It allows us to analyze how each component contributes to well-being as it will be described in the following section.

8.5. Results

On the basis of obtained components by CATPCA, a model was formulated by SEM in order to verify the impact of each of these components on welfare.

The obtained model is shown in Figure 8.1.

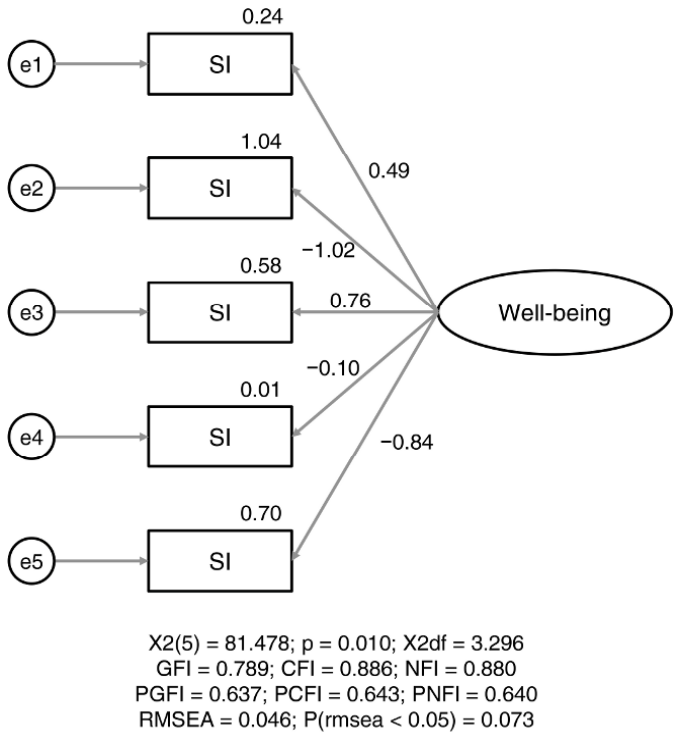


Figure 8.1. *The structural equation model based on the constructed components*

The fundamental principle on which the model is presented is based on the covariance/correlation between the manifest variables, which is the result of the existence of a common factor between them.

As pointed out by Kenny *et al*. (2011), McIntosh (2007) and Mulaik *et al*. (1998), there is no real consensus regarding the quality evaluation of SEMs. Nevertheless, some indicators considered relevant were analyzed.

Using the absolute indices, the quality of the model (X2/df) and the covariance ratio, observed among the manifested variables, explained by the adjusted model (Goodness of Fit Index) were analyzed.

Using the relative indices, it was possible to analyze the percentage increase in the quality of the adjustment of the adjusted model in relation with the model of total independence (normed fit index).

Likewise, the comparative fit index was used in order to correct the underestimation that usually occurs when normal fit index (NFI) is used with small samples.

The use of the parsimony indexes (parsimony goodness of fit index (PGFI), parsimony comparative fit index (PCFI) and parsimony normed fit index (PNFI)) aimed to compensate the "artificial" improvement in the model.

Finally, the root mean square error of approximation (RMSEA) statistic was used in order to compare the adjustment of the obtained model with the sample moments in relation to what would have been obtained with the population moments.

The results obtained through the mentioned indicators allow us to affirm that the model obtained presents a quality of adjustment, meeting the minimum acceptable conditions to proceed with the analysis.

Well-being factor scores will be obtained by the following formula:

$$\text{Well-being} = -0.017 \times SI - 0.612 \times SA - 0.047 \times SC + 0.003 \times SU + 0.072 \times SCo$$

where:

– SI is the social integration;

– SA is the social acceptance;

– SC is the social contribution;

– SU is the social update;

– SCo is the social coherence.

Table 8.1 illustrates how the issues were grouped in order to construct the mentioned factors/variables.

Variables	Items	CATPCA	
		Cronbach's alpha	Percentage variance
Social integration	I feel like I am an important part of society.	0.796	55.015
	I believe that people value me.		
	If I have something to say, everyone hears.		
	I am close to people.		
	If I had something to say, I do not think people take it seriously.		
Social acceptance	I think people are not trustworthy.	0.668	42.936
	I believe that people only think about themselves.		
	I think we should not trust people.		
	People are becoming more selfish.		
	People are becoming more dishonest.		
	People do not care about others.		
Social contribution	I can contribute to the world.	0.722	47.375
	I have nothing to contribute from the world.		
	My activities do not contribute to society.		
	I do not have the time or energy to contribute to society.		
	I think what I do is important to society.		
Social update	For me there is no social progress.	0.768	51.921
	Society offers no incentives for people like me.		
	I see that society is constantly evolving.		
	I do not believe that institutions such as justice and government improve my life.		
	Society shows no progress.		
Social coherence	I do not understand what's happening in the world.	0.760	51.025
	The world is too complicated for me.		
	The world is not worth it.		
	Many cultures seem strange and I do not understand them.		

Table 8.1. *Components obtained by CATPCA*

8.6. Discussion

Blanco and Diaz (2005) mentioned that Keyes assessment is not strong enough and they recommended to include other variables such as social action, recent social action and satisfaction. Martínez *et al*. (2016) only used three components of Keyes theory: SI, SC and SU. In the present study, SI, SA and SC were negative and SU and SCo were positive; all components had Cronbach's alpha of greater than 0.66.

8.7. Conclusions

On the basis of this particular study applied to 69 farms, we constructed a well-being indicator using an SEM, where the components obtained by principal components analysis, SI, SA and SC contributed negatively to well-being, whereas SU and SCo contributed positively.

8.8. References

Bandura, A. (1997). *Self-Efficacy. The Exercise of Control*. W H Freeman/Times Books/Henry Holt & Co., New York.

Blanco, A. and Díaz, D. (2004). Bienestar social y trauma psicosocial: una visión alternativa al trastorno de estrés postraumático. *Clínica y Salud*, 15, 227–252.

Blanco, A., Diaz D. (2005). El bienestar social: concepto y medición. *Psicothema*, 17.

Cabañero, M.J., Richard, M., Cabrero, J., Orts, M.I., Reig, A., Tosal, B. (2004). Fiabilidad y validez de una Escala de Satisfacción con la Vida de Diener en una muestra de mujeres embarazadas y puérperas. *Psicothema*, 16.

FAO. (2014). Año de la agricultura familiar. http://www.fao.org/family-farming-2014/es.

Graaff, J. de V. (1967). *Teoría de la Economía del Bienestar*. Amorrurto Editores, Buenos Aires.

Henson, R.K., Roberts, J.K. (2006). Use of exploratory factor analysis in published research. *Educ. Psychol. Measure*., 66, 393–416.

Kenny, D.A., McCoach, D.B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Struct. Equ. Model*., 10, 333–351.

Keyes, C. (1998). Social well-being. *Soc. Psychol. Q.*, 61, 121–140.

Martínez Servin, L.G., River Heredia, M.E., Martínez Beiza, I., Val Arreola Manuel, D., Tena Martínez, J. (2016). Sistemas familiares de producción lechera en el estado de Michoacán: un análisis psicosocial de sus redes y capital social. In *Ganadería, Sociedad y Recursos Naturales*, Cavallotti Vázquez, B.A., Ramírez Valverde, B., Cesín Vargas, J.A. (eds). Univercidad Autónoma Chapingo.

McIntosh, C.N. (2007). Rethinking fit assessment in structural equation modelling: A commentary and elaboration on Barrett. *Pers. Individ. Diff.*, 42(5), 859–867.

Mulaik, S.A., James, L.R., Van Alstine, J., Bennett, N., Lind, S., Stilwell, C.D. (1998). Evaluation of goodness-of-fit indices for structural equation models. *Psychol. Bull.*, 105(3), 430–445.

OMS. (2015). Cómo define la OMS la salud? http://www.who.int/suggestions/ faq/es/.

Putnam, R. D. (2000). *Bowling Alone: The Collapse and Revival of American Community*. Simon and Schuster, New York.

Rodríguez Jorge J., Robert, K., Sergio, A.G. (2008). *Epidemiología de los transtornos mentales en America Latina y el Caribe*, Scientific and Technical Publication No. 632, Organización Panamericana de la Salud, Washington DC.

SAGARPA. (2012). La granja ecológica integral. http://www.sagarpa.gob.mx/desarrolloRural/Documents/fichasaapt/La%20granja%20ecol%C3%B3gica%20integral.pdf.

U.S. Department of Health and Human Services (2001). Mental health: culture, race, and ethnicity. A supplement to mental health: a report of the Surgeon General. Office of the Surgeon General, Center for Mental Health Services, and National Institute of Mental Health, Rockwell, USA. Available: https://www.ncbi.nlm.nih.gov/books/NBK44243.

9

An SEM Approach to Modeling Housing Values

Although hedonic regression remains a popular technique for estimating property values, structural equation modeling (SEM) is increasingly seen as a realistic analytical alternative. This chapter presents an SEM analysis of a historical data set for a large Canadian realtor. An iterative approach was adopted for the modeling. The first phase focused on internal relationships between houses' structural characteristics and the second on housing values and their determinants. In the final phase, advertised list prices and location details were the priority. A comprehensive evaluation of the resulting holistic model revealed a wealth of significant structural relationships – particularly between house style, structure and attributes.

9.1. Introduction

Housing is a durable, highly differentiated asset, characterized by being fixed in location (Kinnard 1968). Effective price estimation (Karanka *et al*. 2013) is crucial to its successful acquisition. With Hedonic Price Theory (Rosen 1974), a house is considered a "basket" of attributes – z_1, z_2, …, z_n – against which the house price, $P = f(z_1, z_2, \ldots, z_n)$, can be derived. Here, attributes are typically related to structural, locational, neighborhood and environmental (e.g. noise and pollution) characteristics (Anas and Eum 1984). Depending on the functional form f adopted for computing P, various hedonic regression formulations are available for estimation purposes (Palmquist 1984). Some of the most common in use include the linear, semi-log and log linear but a mixture of these together with transformations such as the Box-Cox

Chapter written by Jim FREEMAN and Xin ZHAO.

Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining,
First Edition. Edited by Christos H. Skiadas and James R. Bozeman.

is becoming increasingly prevalent (see, for instance, Linneman and Voith 1991, Fletcher *et al*. 2004).

In terms of mass property appraisal modeling, the "physical–neighborhood–location" framework has long been established. But as neighborhood is strongly linked to location, physical (*structural*) and *locational* characteristics tend to be the focus in practice.

– *Structural characteristics*

Structural characteristics of a house usually include the *age* of the house; the square meters of *living area*; the number, size and type of *rooms* in the house; and the attachment or lack of *garage spaces* or *chattels*. These characteristics contain both quantitative and qualitative dimensions, for example *living area*, *lot size* and number of *rooms* are quantitative characteristics, whereas *chattels* can be viewed as qualitative.

– *Locational characteristics*

Because houses are "locationally immobile" and spatially no two properties are the same, location can have a huge impact on house values (Des Rosiers *et al*. 2000). Location is often assessed in relation to accessibility – for example proximity to a central business district – also to neighborhood in terms of access to educational and entertainment facilities, etc. (McCluskey *et al*. 2000).

Although hedonic prices models are widely employed in residential property value studies, they are frequently beset by a range of technical problems including multicollinearity, heteroscedasticity and autocorrelation (Bowen *et al*. 2001) – all of which are actually straightforwardly handled by SEM.

This is rationale for the modeling that follows. Using the AMOS 16 package, the SEM analysis approach is illustrated for a Canadian house sales application, which is detailed in section 9.2. An overview of the work is presented in section 9.3 and conclusions in section 9.4.

9.2. Data

The data were taken from Freeman and Janssen (1993) and relate to houses located in 10 selected neighborhoods of Edmonton, Canada. Each home was listed and sold individually through the realtor's multiple listing system and related to the 18-month period after January 1, 1988.

Of the 240 houses that made up the sample, 90% (216) were bungalows and the remainder were two-story homes. The mean list price for all houses was $140,324.63, whereas that for bungalows was $137,129.22 – compared to $169,083.33 for two-story homes. Ages of houses ranged from 5 to 32 years with a mean of 22 years; correspondingly the sizes of properties ranged from 90 to 267 m^2 (the mean value was approximately 125 m^2).

Except for list price, variables for the study were identified by house and lot attributes as presented in Table 9.1.

Variable	**Definition**
House attributes	
STYLE	1 if bungalow, 2 if two story
R	Number of rooms
B	Number of bathrooms
BR	Number of bedrooms
S	Living area (square meters)
A	Age (years)
BAS	Basement (from 1 (open) to 3 (finished))
G	Number of garage spaces
ATT	Dummy variable, 1 if attached, 0 detached
F	Number of fireplaces (wood-burning)
C	Number of chattels (appliances)
Lot attributes	
LOTS	Lot size (square meters)
CO	1 if corner lot, 0 otherwise
CUL	1 if cul-de-sac, 0 otherwise
LA	1 if lane behind, 0 otherwise
E	Exposure of yard (N, NE, E = 1, otherwise 0)
Zj	Dummy variable represents zone j (j = 0, 1, …, 9)
TIME	Driving time

Table 9.1. *Definition of variables*

9.3. Analysis

Prior to the SEM modeling, exploratory factor analysis (EFA) was conducted to help determine the underlying structure of the data and gain insight into the possible

latent constructs that might exist. This provided the basis of the initial measurement model, which was then tested using confirmatory factor analysis. Following this, a progression of structural relationships between latent and manifest variables was introduced and validated (Byrne 2009). Three phases were involved in the SEM modeling overall.

Exploratory factor analysis (EFA)

Factors of interest were first identified using principal components analysis with varimax rotation (Field 2009).

This was then followed up by a principal axis factoring/oblimin rotation analysis to help establish the latent structure of the model.

For the sake of practicality, it was decided not to introduce the locational variables Z0-Z9 until phase two of the modeling (Zhao 2012). These were, therefore, omitted from the EFA, which generated factors as follows:

– *Factor 1* ("House Style") linked to the STYLE (bungalow or two-story) and ATT (attached or detached) variables;

– *Factor 2* ("House Structure") included the R (number of rooms), B (number of bedrooms) and BR (number of bathrooms) variables;

– *Factor 3* ("House Attributes") represented the F (number of fireplaces), G (number of garage spaces) and C (number of chattels) variables.

Based on the latter, the following three research hypotheses were adopted for the study:

– H1: house style positively impacts house structure;

– H2: house style positively impacts house attributes;

– H3: house structure positively impacts house attributes.

Phase 1

In the first phase of the SEM modeling, internal relationships between the structural characteristics of a house were investigated, without consideration of the value of the house – see Figure 9.1 which shows the formal path diagram of the theoretical (measurement) model to be evaluated.

AMOS provides a large number of standard diagnostics for assessing the effectiveness of a particular model choice. Selected values of these for model 1a are summarized in Table 9.2.

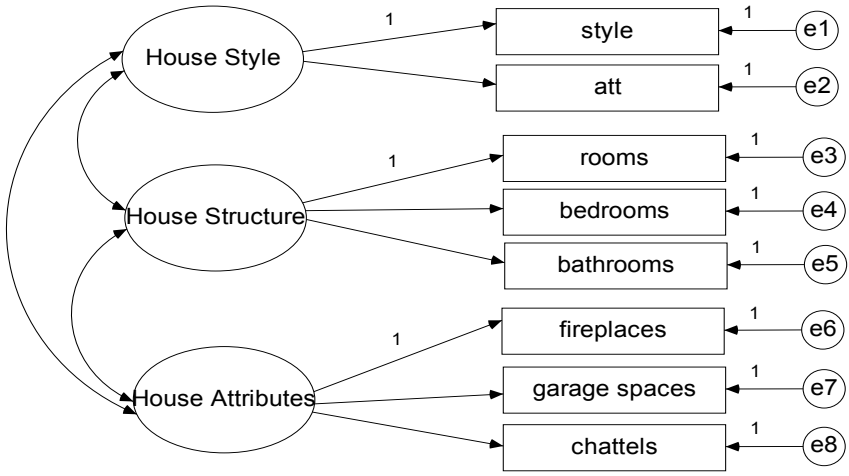


Figure 9.1. *Measurement model 1a*

Indices	Value	Threshold	Acceptability
χ^2	53.879 (p<.001)	Significant at 0.01 level	
df	17		
CMIN/DF	3.169	<2-3	Unacceptable
CFI	0.911	>.9	Acceptable
GFI	0.944	>.9	Acceptable
AGFI	0.881	>.9	Unacceptable
RMSEA	0.095	<.08	Unacceptable

Table 9.2. *Goodness of fit summary: model 1a*

Although the latter model appears satisfactory according to some of the indices shown here, there is scope for improvement with others. As a refinement, it was decided to introduce a covariance link between the error terms e2 and e4. Note that the chattels item (which turned out to have a zero regression weight) was

simultaneously dropped. The resultant estimated model 1b is shown in Figure 9.2 with associated diagnostics in Table 9.3.

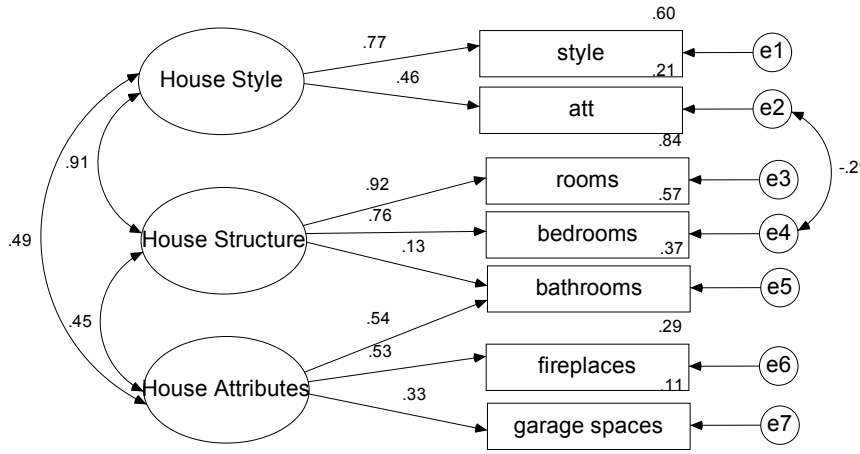


Figure 9.2. *Estimated model 1b*

Indices	Value	Threshold	Acceptability
χ^2	11.665 (p=.233)	Significant at 0.01 level	
df	9		
CMIN/DF	1.296	<2-3	Acceptable
CFI	0.994	>0.9	Acceptable
GFI	0.987	>0.9	Acceptable
AGFI	0.958	>0.9	Acceptable
RMSEA	0.035	<0.08	Acceptable

Table 9.3. *Goodness of fit summary: model 1b*

Obviously, this is a much better fit. No less importantly, all loadings here (except for bathrooms) were found to be statistically significant ($P < 0.05$) and all relationships in the expected direction.

Phase 2

In the second phase, the model was modified to investigate the relationships between house values and structural and location characteristics.

Therefore, a second-order factor "House Value" was introduced into the model. Furthermore, the latent variable "Location" was included as an underlying indicator of accessibility and neighborhood. The resultant estimated model is shown in Figure 9.3. Unfortunately, the accompanying diagnostics in Table 9.4 can be seen to be far from satisfactory.

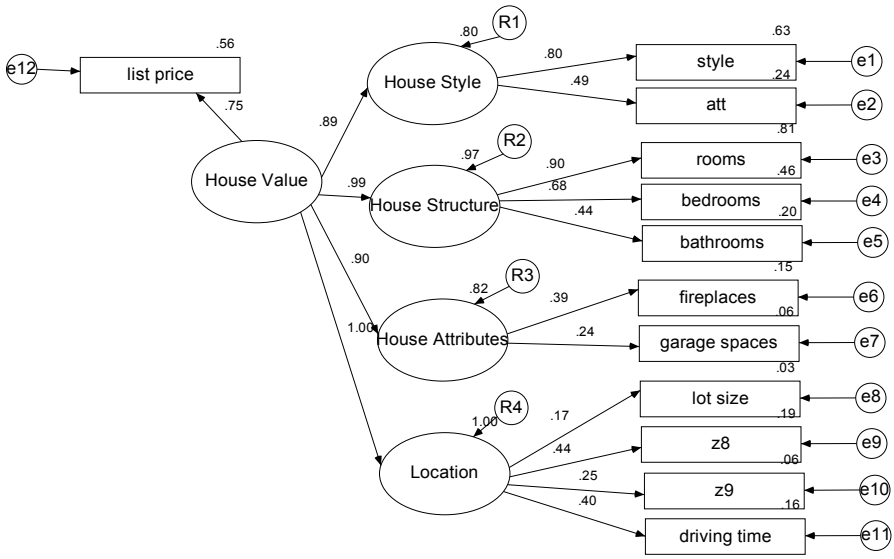


Figure 9.3. *Estimated model 2a*

Indices	Value	Threshold	Acceptability
χ^2	269.893 (p<.000)	Significant at 0.01 level	
df	53		
CMIN/DF	5.092	<2-3	Unacceptable
CFI	0.737	>.9	Unacceptable
GFI	0.819	>.9	Unacceptable
AGFI	0.734	>.9	Unacceptable
RMSEA	0.131	<.08	Unacceptable

Table 9.4. *Goodness of fit summary: model 2a*

To rectify the situation, it was decided to allow for selected error terms to be correlated. This was done iteratively in line with successive AMOS modification index outputs. Details of the revised estimated model 2b are presented in Figure 9.4.

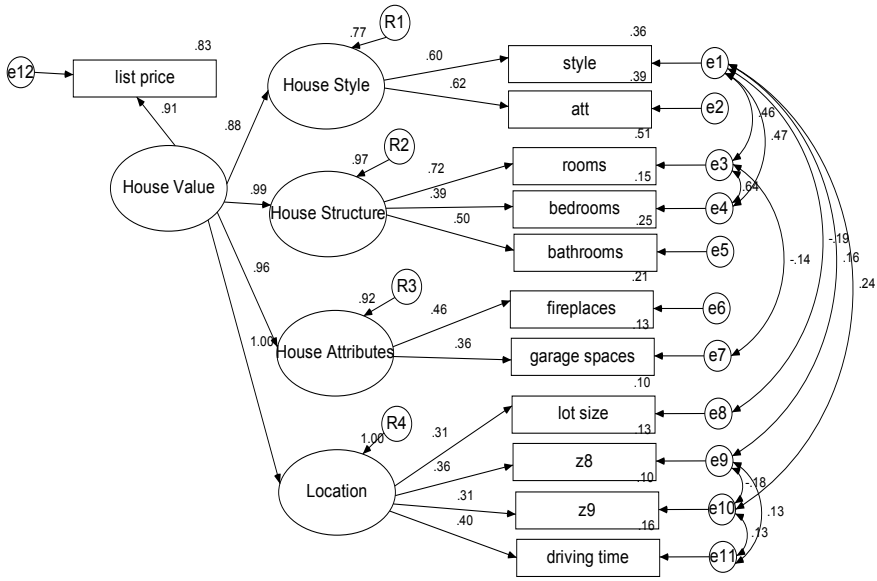


Figure 9.4. *Estimated model 2b*

Indices	Value	Threshold	Acceptability
χ^2	72.436 (p=.003)	Significant at 0.01 level	
df	43		
CMIN/DF	1.685	<2-3	Acceptable
CFI	0.964	>.9	Acceptable
GFI	0.955	>.9	Acceptable
AGFI	0.919	>.9	Acceptable
RMSEA	0.054	<.08	Acceptable

Table 9.5. *Goodness of fit summary: model 2b*

Apart from the χ^2 statistic, here the fit results all look satisfactory. However, because large sample sizes – as in this study – are notoriously linked to significant χ^2 results, the CMIN/DF criterion (following convention) was used as a proxy. On this basis, the model would, therefore, be judged to have satisfactory fit characteristics throughout. In addition, all indicator estimates are statistically significant at the 5% level with the arrow directions too, in line with expectations.

Phase 3

In the third phase, the latent variable "House Price" was introduced into the model, as estimated in Figure 9.5 with corresponding diagnostics in Table 9.6. (The logic behind this adaptation is that list price is literally more a reflection of House Price than House Value.)

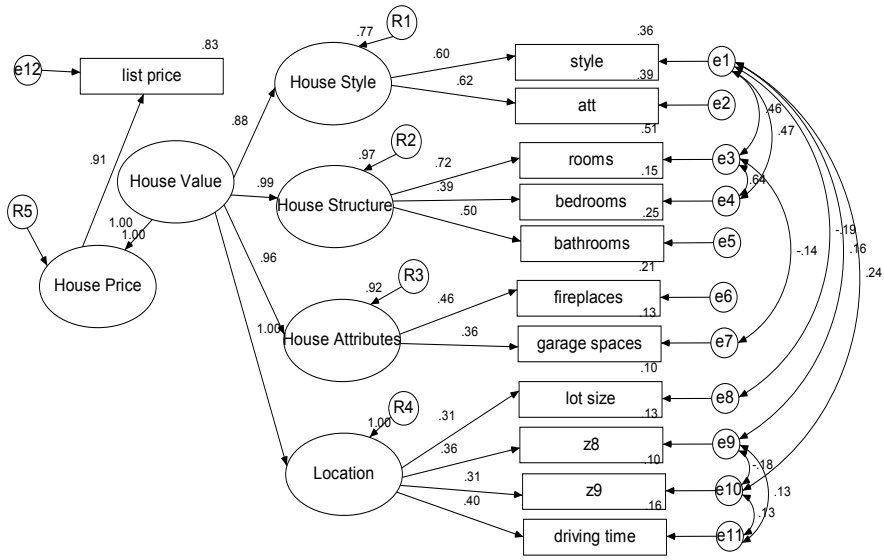


Figure 9.5. *Estimated model 3*

Indices	Value	Threshold	Acceptability
χ^2	72.436 (p=.003)	Significant at 0.01 level	
df	43		
CMIN/DF	1.685	<2-3	Acceptable
CFI	0.964	>.9	Acceptable
GFI	0.955	>.9	Acceptable
AGFI	0.919	>.9	Acceptable
RMSEA	0.054	<.08	Acceptable

Table 9.6. *Model 3: goodness of fit summary*

As with Table 9.5, the latter diagnostics confirm that the model satisfactorily fits the data.

Similarly, the standardized estimates of all model loadings shown in Figure 9.5 are highly significant (at the 0.1% level).

House Value is also found to be positively related to the four latent variables: House Style, House Structure, House Attribute and Location.

In addition, supporting evidence for research hypotheses H1, H2 and H3 is provided by supplementary correlation output as follows:

	Sample correlation	Inference
H1: House Style positively impacts House Structure	0.866*	Accepted
H2: House Style positively impacts House Attributes	0.432*	Accepted
H3: House Structure positively impacts House Attributes	0.409*	Accepted

*$P < 0.05$.

Finally, the squared multiple correlations in Figure 9.5 reveal that 77% of the variance of House Style, 97% of the variance of House Structure, 92% of the variance of House Attributes and 100% of the variance of Location are individually accounted for by House Value. Of interest the percentages here – as well as model loadings in Figure 9.5 are identical to those for model 2b (Figure 9.4) – suggesting that the distinction between the House Price and House Value latent constructs is probably more semantic than real.

9.4. Conclusions

Despite the dearth of published research on SEM modeling of house values, it is very encouraging to discover how strikingly the above findings are in agreement with those of Liu and Wu (2009). For their analysis of residential data from Dalian, China, two measures, house price and satisfaction degree, were chosen to represent house value (not just list price as in our own case). Furthermore, a four-factor (location, residential district, structure and neighborhood environment) framework was adopted for their modeling. Notwithstanding these differences, however, it is salutary to note that Liu and Wu too concluded location and structure to be, far and away, the most important determinants of house value.

9.5. References

Anas, A. and Eum, S.J. (1984). Hedonic analysis of a housing market in disequilibrium. *Journal of Urban Economics*, 15(1), 87–106.

Bowen, W.M., Mikelbank, B.A. and Prestegaard, D.M. (2001). Theoretical and empirical considerations regarding space in hedonic house price estimation. *Growth and Change*, 32(4), 466–490.

Byrne, B.M. (2009). *Structural Equation Modelling with Amos: Basic Concepts, Applications, and Programming*, 2nd edition Abingdon, Oxfordshire:Routledge Academic.

Des Rosiers, F., Theriault M. and Villeneuve, P. (2000). Sorting out access and neighbourhood factors in hedonic price modelling. *Journal of Property Investment & Finance*, 18(3), 291–315.

Field, A. (2009). Discovering statistics using SPSS, 3rd edition. London: SAGE.

Fletcher, M., Mangan, J. and E. Raeburn. (2004). Comparing hedonic models for estimating and forecasting house price. *Property Management*, 22(3), 189–200.

Freeman, J. and Janssen, C. (1993). Analysis of the effect of Realtor Competition on Housing Prices. Research Papers in Management Science, Faculty of Business University of Alberta.

Karanka, J., O'Neill, R., Weadon, N., Sanderson R. and Jenkins C. (2013). *Official House Price Statistics Explained*. Newport, UK: Office for National Statistics.

Kinnard, W.N.K. (1968). Reducing uncertainty in real estate decisions. *The Real Estate Appraiser*, 34(7), 10–16.

Linneman, P. and Voith, R. (1991). Housing price functions and ownership capitalization rates. *Journal of Urban Economics*, 30(1), 100–111.

Liu, Y. and Wu, Y.X. (2009). Analysis of residential product's value based on structural equation model and hedonic price theory. *2009 International Conference on Management Science & Engineering*, September 16, Moscow, Russia, 1950–1956.

McCluskey, W.J., Deddis, W.G., Lamont, I.G. and Borst, R.A. (2000). The application of surface generate interpolation models for the prediction of residential property values. *Journal of Property Investment & Finance*, 18(2), 162–176.

Palmquist, R.B. (1984). Estimating the demand for the characteristics of housing. *The Review of Economics and Statistics*, 66(3) 394–404.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy*, 82(1), 34–55.

Zhao, X. (2012). Structural equation modelling analysis of Canadian house sales Data. Unpublished MSc dissertation. University of Manchester, Manchester, UK.

10

Evaluation of Stopping Criteria for Ranks in Solving Linear Systems

Linear systems of algebraic equations arising from mathematical formulation of natural phenomena or technological processes are common. Many of these systems of equations are large, the matrices derived are mainly sparse and need to be solved iteratively. Moreover, interpretation is crucial in making decision. Bioinformatics, internet search engines (web pages) and social networks are some of the examples with large and high sparsity matrices. For some of these systems, only the actual ranks of the solution vector is interesting rather than the vector itself. In this case, it is desirable that the stopping criterion reflects the error in ranks rather than the residual vector that might have a lower convergence. In this chapter, we evaluated stopping criteria on Jacobi, successive over relaxation and power series iterative schemes. Numerical experiments were performed and results show that Kendall's correlation coefficient τ gives good stopping criterion of ranks for linear system of equations.

10.1. Introduction

Sparse and large linear systems of equations are common in many physical applications. In particular, the Internet search engine is one area where such systems are encountered and have been studied intensively. The concepts have been deployed in areas such as social networks, bioinformatics and infectious disease management (Wills and Ipsen 2009). With growing technology, the size of data from these fields is reaching billions, and numerical computations are becoming more demanding (Boldi *et al.* 2007). Moreover, for the case of search engines, only a few relevant pages for query are provided, which is literally termed as ranking of web pages as explained by Kohlscütter *et al.* (2009). Indeed, ranking requires efficient algorithms, understanding parameter influence on convergence and stability or error tolerance as

Chapter written by Benard ABOLA, Pitos BIGANDA, Christopher ENGSTRÖM and Sergei SILVESTROV.

Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining,
First Edition. Edited by Christos H. Skiadas and James R. Bozeman.

suggested by Engström and Silvestrov (2015). Recently, PageRank, connecting a line of nodes with a complete graph, was studied, and explicit formula to determine ranks was proposed, which was mainly aimed at minimizing errors in estimating ranks as pointed out in Engström and Silvestrov (2016). Much as these attempts have been made to improve ranking processes by increasing convergence rate and formulating formulas, the issue on the quality of ranks obtained still remains unresolved. As a matter of fact, paying much attention to the accurate solutions of linear system of equations underscore the practical significance for ranking in areas where they are applied.

In this chapter, we focus on the termination criterion as means of obtaining good ranks. We first outline some studies carried out on stopping criteria in solving the linear system.

Bennani and Braconnier (1993) studied convergence detection of iterative solvers. They pointed out that such criterion may allow excessive number of iterations to be performed. Geometric distance such as absolute error and normalized residual for linear systems have been preferred as terminating criteria when using Jacobi iterative schemes (Gliech *et al.* 2005). Another study was done by Qiu and Cho (2006) who applied ranking distance as termination criterion to approximate PageRank vector. They noted that there are instances in which the relative error may be large but the ranks are pretty good.

An essential goal in PageRank computing is to rank webpages (Gong *et al.* 2014), while accuracy of the PageRank vectors is secondary. In the view of Haveliwala (1999), if the PageRank vector is to be used for establishing the importance of pages, the convergence should be measured based on how the ordering changes as the number of iterations increases. Berkhin (2005b) questioned the sense to iterate beyond the accuracy that establishes the order of the pages in search engines analysis. Boldi *et al.* (2007) said, "it is the relative order of pages with respect to PageRank that is actually interesting for search engines". Rank ordering as an algorithm provides a few numbers of iterations sufficient for the purpose of identifying important webpages, and moreover, time complexity of algorithm can be drastically reduced (Bidoki *et al.* 2007). In fact, with the advancement in technology and adaptation of PageRank algorithm in areas like object tracking or target filtering, the importance of stopping criterion for ranks should not be underscored. This motivates the need to evaluate stopping criteria in relation to ranks. This chapter aims to answer the following questions: (1) Which stopping criterion out performs the others? (2) Which criteria are similar? and (3) Is there any link between the stopping criterion and the iterative methods in ranking problems?

10.2. Methods

This section describes some notions used in the methods to an easy understanding. Preliminaries of concepts such as key definitions and theorems are outlined. Also, iterative methods and stopping criteria are briefly described.

10.2.1. *Preliminaries*

Consider a large sparse system of equation

$$A\mathbf{x} = \mathbf{b}, \qquad [10.1]$$

where $A \in \mathbb{R}^{n\times n}$ is a non-singular matrix and $\mathbf{x}$, $\mathbf{b} \in \mathbb{R}^n$ are vectors. To avoid confusion, all the matrices and vectors are real numbers unless otherwise stated.

DEFINITION 10.1.– (Hadjidimos 2000) If a matrix A satisfies $A \geq 0$, then it is said to be non-negative. The matrix A is said to be an L-matrix if and only if $a_{i,i} > 0$, $i = 1, \cdots, n$ and $a_{i,j} \leq 0$, $i \neq j$. A matrix A is said to be an M-matrix if it is both an L-matrix and invertible.

THEOREM 10.1.– *(Perron–Frobenius Theorem, Meyer 2000) Let A be an $n \times n$ positive matrix with spectral radius ρ. Then the following statement holds:*

1) there is a positive real eigenvalue $\lambda_1 = \rho = \max\{|\lambda| : \lambda \in \mathbb{C}\}$;

2) there is an eigenvector $\boldsymbol{\varsigma} > 0$ such that $A\boldsymbol{\varsigma} = \lambda_1\boldsymbol{\varsigma}$;

3) the eigenvalue λ_1 has multiplicity 1;

4) apart from $\boldsymbol{\varsigma}$, there are no positive eigenvector of A other than positive scalar multiples of $\boldsymbol{\varsigma}$.

Having stated some essential definitions and a theorem that will be referred to in this work, we next present the formulation of two iterative schemes, that is the Jacobi and Successive overrelaxation methods. Solving equation [10.1] iteratively involves splitting matrix A as

$$A = M - N, \qquad [10.2]$$

where M is non-singular. It is also a convergent splitting of A if the spectral radius of iterative matrix, $M^{-1}N$, is less than 1 (Li and Wu 2014). Following representation [10.2], the iterative solution of equation [10.1] becomes

$$\begin{aligned} \mathbf{x}^{(m+1)} &= M^{-1}N\mathbf{x}^{(m)} + M^{-1}\mathbf{b}, \\ &= T\mathbf{x}^{(m)} + \mathbf{g}, \end{aligned} \qquad [10.3]$$

where $T = M^{-1}N$ is called iterative matrix and $\mathbf{g} = M^{-1}\mathbf{b}$ is a vector.

10.2.2. *Iterative methods*

This section presents iterative techniques applied in solving large sparse linear systems, namely Jacobi method, successive overrelaxation (SOR) method and power series method. Note that the performance of iterative solver(s) depends mostly on the structure of iterative matrix, which we have taken into account without much details described in this paper (see Young 1971).

10.2.2.1. *Jacobi Iterative method*

To derive the Jacobi iterative formula, the matrix A is split as $A = D - (L + U)$. Using representation [10.2], $M = D$, where $D = \text{diag}(A)$, and $N = L + U$, where L and U are, respectively, strictly lower and upper triangular $n \times n$ matrices, whose entries are the negatives of the entries of A, respectively, below and above the leading diagonal of A. It follows from equation [10.3] that

$$\mathbf{x}^{(m+1)} = D^{-1}(L + U)\mathbf{x}^{(m)} + D^{-1}\mathbf{b}, \quad m \geq 0, \qquad [10.4]$$

where $\mathbf{x}^{(0)}$ is the initial estimate of the unique solution $\mathbf{x}$ of [10.1]. Equation [10.4] is called the point Jacobi iterative method (Varga 1962) and the matrix

$$J = D^{-1}(L + U) \qquad [10.5]$$

is called the point Jacobi iterative matrix associated with the matrix A. The method requires a simultaneous storage of all the components of the vector $\mathbf{x}^{(m)}$ while computing the components of the vector $\mathbf{x}^{(m+1)}$.

10.2.2.2. *SOR iterative method*

In a similar way as the Jacobi method, we split matrix A as $A = M - N$, but $M - N \equiv D - L - U = (\frac{D}{\omega} - L) - ((\frac{1}{\omega} - 1)D + U)$, where the quantity ω is called the relaxation factor (Young 1971). From [10.3], we have

$$(D - \omega L)\mathbf{x}^{m+1} = [(1 - \omega)D + \omega U]\,\mathbf{x}^{m} + \omega\mathbf{b}. \qquad [10.6]$$

Since $D - \omega L$ is a non-singular, [10.6] is equivalent to

$$\mathbf{x}^{(m+1)} = (D - \omega L)^{-1} \left[(1 - \omega)D + \omega U\right] \mathbf{x}^{(m)} + \omega (D - \omega L)^{-1} \mathbf{b}, \qquad [10.7]$$

which is called the point SOR iterative method. The matrix J_{SOR} is given by

$$J_{SOR} = (D - \omega L)^{-1} \left[(1 - \omega)D + \omega U\right],$$

is called the point SOR matrix. To find an optimal value of ω, it is assumed that the decomposed matrix has Property "A" (Hadjidimos 2000), and the optimal choice of the parameter can be obtained from Theorem 10.2.

THEOREM 10.2.– *(Optimal Successive Overrelaxation parameter, Young 1971) Let J and J_{SOR} be Jacobi and SOR iterative matrices, respectively, and derived from a matrix A. If $\mu(J)$ is the eigenvalues of J and the spectral radius, $\rho(J) < 1$, then the optimal SOR parameter, $\omega_o = \frac{2}{1+\sqrt{1-\rho^2(J)}}$.*

Before we have a look at how to approximate ω_o, it is important to note that explicit formula that compute the optimal parameter for SOR method in general does not exist. There are some special forms of matrices, such as tri-diagonal, property "A" and weakly cyclic, whose relaxation parameters can be derived precisely. This implies that determining $\rho(J)$ is practically challenging so one need to approximate ω_o, which was the case in this problem. Now, using equations [10.5] and [10.13], the matrix, $J = L + U = cP$, taking 1-norm of J, we have $\rho(J) \leq c$, and using this bound, one gets an estimate of the parameter as $\omega_o = \frac{2}{1+\sqrt{1-c^2}}$. Since the damping factor $c = 0.85$, an approximate value of $\omega_o = 1.3099$. This value works only if the matrix has property "A".

Alternatively, if J is weakly cyclic then for each eigenvalue of J, $\mu(J)$ with $\frac{1}{2} < \mu(J) < 1$, set $\omega_o = \frac{2}{1+\sqrt{2c-1}}$, where $c = \mu(J)$ as earlier mentioned. Hence, a better choice of ω_o is 1.0889. Exploring connectivity of the graph that generated an iterative matrix J is important to avoid extreme initial guess of ω_o.

10.2.2.3. *Power series method*

The formulation of power series iterative method arise from equation [10.12], that is multiplying both sides by $(I - cP)^{-1}$, we obtain

$$\boldsymbol{\pi} = (1 - c)(I - cP)^{-1}\mathbf{v}. \qquad [10.8]$$

Expressing the term $(I - cP)^{-1}$ as geometric series yields

$$(I - cP)^{-1} = I + cP + c^2P^2 + \cdots = \sum_{j=0}^{\infty} (cP)^j.$$

Substituting in equation [10.8] gives

$$\boldsymbol{\pi} = (1-c)[\sum_{j=0}^{\infty}(cP)^j]\mathbf{v}. \qquad [10.9]$$

10.3. Formulation of linear systems

Linear systems of equations considered in this chapter are those arising from PageRank problems (Langville and Meyer 2006). We briefly describe how one can formulate the system from a web link graph $\mathcal{G}$ with n vertices. That is, we let P be weighted adjacency (stochastic) $n \times n$ matrix derived from outgoing vertices of the graph. If $\mathcal{G}$ has no outgoing links in some vertices, then P is a substochastic matrix. We remedy this by adding to P rows corresponding to dangling vertices (vertices without outgoing links) a positive probability distribution $\mathbf{v}$ over all vertices. A stochastic matrix $\mathcal{P}$ obtained after adjustment for dangling vertices is defined as

$$\mathcal{P} = P + \mathbf{d}\mathbf{v}^T,$$

where $\mathbf{d}$ is a column vector such that

$$d_i = \begin{cases} 1, & \text{if vertex } i \text{ is dangling,} \\ 0, & \text{otherwise,} \end{cases} \qquad [10.10]$$

and $\mathbf{v} = \frac{\mathbf{e}}{n}$, n is the dimension of $\mathcal{P}$ and $\mathbf{e} = (1, 1, \cdots, 1)$.

Suppose $c \in (0, 1)$ is the probability that a web surfer follows the link as described by $\mathcal{P}$ and jump to any vertex in the graph according to $\mathbf{e}\mathbf{v}^T$ with probability $(1-c)$, then the matrix defined in PageRank problem becomes

$$\mathcal{T} = c(P + \mathbf{d}\mathbf{v}^T) + (1-c)\mathbf{e}\mathbf{v}^T. \qquad [10.11]$$

Using Theorem 10.1, we can formulate the problem to determine a stationary distribution of the web link graph by multiplying [10.11] by an eigenvector $\boldsymbol{\pi}$ corresponding to eigenvalue 1. Rearranging, we get

$$(I - cP)\boldsymbol{\pi} = (1-c)\mathbf{v}, \qquad [10.12]$$

where c is damping factor, P is a stochastic matrix and $(I - cP)$ is M-matrix. This is the version of linear system defined in Engström and Silvestrov (2015) as the

eigenvalue problem. However, the version of the system considered is the one in which $\|\boldsymbol{\pi}\|_1 \neq 1$ and is given as

$$(I - cP)\boldsymbol{\pi} = n_g \mathbf{v}, \qquad [10.13]$$

where n_g is the size of the one vector $\mathbf{e}$ in case $\mathbf{v}$ is uniform.

REMARK 10.1.– From equation [10.13], and comparing it with equation [10.1], then $A \equiv (I - cP)$, $\mathbf{x} = \boldsymbol{\pi}$ and $\mathbf{b} = n_g \mathbf{v}$.

10.4. Stopping criteria

Whenever one attempts to solve large linear system of equations, the solution is always approximated because of round off errors. Therefore, some degree of accuracy needs to be adopted that will depend on stopping criteria among others. This section briefly outlines five stopping criteria used in large linear system solvers. We intend to use them based on their popularity in PageRank problem, which is a sister problem to linear system of equations [10.12].

1) Assume that the residual vector at the $(m+1)\text{th}$ iteration is $r^{(m)}$, then the norm of the residual is

$$\|\mathbf{r}^{(m)}\|_1 = max_i |\mathbf{x}^{(m)} - A\mathbf{x}^{(m-1)}|, \qquad [10.14]$$

where $\mathbf{x}^{(m)}$ is the approximate solution vector at $m\text{th}$ iteration.

2) Componentwise backward error: The criterion allowed for determination of finite bound when the matrix A is sparsed (Arioli *et al.* 1992) is

$$max_i \frac{|\mathbf{r}^m|_i}{(|A|.|\mathbf{x}^m| + |\mathbf{b}|)_i}. \qquad [10.15]$$

At this point, the use of 1-norm appears in many literature but infinity norm and 2-norm may also work. However, one must be conscious on technical reasons to use a norm or combination of norms. We highlight this in the next criterion.

3) Normwise backward stopping criterion: The idea to include this bound arises from the fact that in most iterative methods their convergence solely depends on eigensystem of iteration matrix where the backward error is also unknown. Moreover, such iteration involves successive matrix-vector computation; this result to dense matrix. Hence, a solution obtained may be quite near machine precision, which in

turn results in stopping iteration too early. To remedy this, one would need to choose some larger threshold to determine termination, i.e.

$$\frac{\|\mathbf{r}^m\|_\infty}{\|A\|_\infty.\|\mathbf{x}^m\|_1 + \|\mathbf{b}\|_\infty}. \qquad [10.16]$$

4) The ratio of residual to infinity norm of vector $\mathbf{b}$ neglecting the effect of matrix A:

$$\frac{\|\mathbf{r}^m\|_\infty}{\|\mathbf{b}\|_\infty}. \qquad [10.17]$$

This criterion has been discussed in Arioli *et al.* (1992) in detail.

5) Before defining the criterion, let us consider the following lemma:

LEMMA 10.1.– *If $J \in \mathbb{R}^{n\times n}$ and $\|J\| < 1$, then $I-J$ is non-singular and $(I-J)^{-1} = \sum_{k=0}^{\infty} J^k$ with $\|(I-J)^{-1}\| \leq \frac{1}{1-\|J\|}$.*

PROOF.– To prove non-singularity, we use proof by contradiction. Let $I - J$ be singular, then for some vector $\mathbf{x} \in \mathbb{R}^n$ we have $(I - J)\mathbf{x} = 0$ but $\|\mathbf{x}\| = \|J\mathbf{x}\|$, hence $\|J\| \geq 1$. Thus, $I - J$ is non-singular.

To prove the second part of the lemma, consider the identity $\sum_{k=0}^{N} J^k(I - J) = I - J^{N+1}$, since $\|J\| < 1$, it follows that $J^k \to 0$ as $k \to \infty$ because $\|J^k\| \leq \|J\|^k$ for some k. Therefore, $\lim_{N\to\infty} \sum_{k=0}^{N} J^k(I - J) = I$. This is equivalent to $(I - J)^{-1} = \lim_{N\to\infty} \sum_{k=0}^{N} J^k$.

Taking matrix norm on both sides, we get $\|(I - J)^{-1}\| \leq \frac{1}{1-\|J\|}$.

In the second lemma, we derived the bound of successive residual, which will turn out to be stopping criterion.

LEMMA 10.2.– *Suppose a stationary iterative scheme is defined as $\mathbf{x}^{(m+1)} = J\mathbf{x}^{(m)} + \mathbf{d}$, where $\|J\| < 1$, $\mathbf{d}$ is a constant and $m = 1, 2, \cdots,$. Then the estimate $\frac{\|J\|}{1-\|J\|}\|\mathbf{x}^{(m)} - \mathbf{x}^{(m-1)}\| < \epsilon$.*

PROOF.– Let $\mathbf{x}$ be the exact solution obtained when using the iterative scheme, then at the mth iteration, we can write the residual as

$$\begin{aligned} \mathbf{x} - \mathbf{x}^{(m)} &= J\mathbf{x} - J\mathbf{x}^{(m-1)} \\ &= J\mathbf{x} - J\mathbf{x}^{(m)} + J\mathbf{x}^{(m)} - J\mathbf{x}^{(m-1)}. \end{aligned}$$

Collecting similar terms and taking vector norm on both sides yields

$$\begin{aligned} \|(\mathbf{x}-\mathbf{x}^m)-J(\mathbf{x}-\mathbf{x}^m)\| &= \|J(\mathbf{x}^m-\mathbf{x}^{(m-1)})\| \\ \|(\mathbf{x}-\mathbf{x}^m)(I-J)\| &\leq \|J\|\|(\mathbf{x}^m-\mathbf{x}^{(m-1)})\|, \\ \|(\mathbf{x}-\mathbf{x}^m)\| &\leq \|J\|\|(I-J)^{-1}\|\|(\mathbf{x}^m-\mathbf{x}^{(m-1)})\|, \end{aligned}$$

using lemma 10.1, we get $\frac{\|J\|}{1-\|J\|}\|\mathbf{x}^{(m)}-\mathbf{x}^{(m-1)}\|$, which is less or equal to some tolerance (ϵ). Now, we write criterion V as $\frac{c}{1-c}\|\mathbf{x}^{(m)}-\mathbf{x}^{(m-1)}\|$, where c (damping factor) is approximated by $\|J\|$.

6) Kendall's τ rank correlation: This is one of the many correlation indices for comparing orders. It is a non-parametric correlation index and widely used for ranking aggregation in the web community (Dwork *et al.* 2001). It also helps to determine how fast the computation of PageRanks converges (Kamvar *et al.* 2003). Kendall's τ is defined as follows:

DEFINITION 10.2.– Let $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^n$ be two vectors of rank values. Given a pair (x_i, y_i) and (x_j, y_j), $1 \leq i, j \leq n$, then the pair is said to be

– concordant iff $x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i < y_j$;

– discordant iff $x_i > x_j$ and $y_i < y_j$ or $x_i < x_j$ and $y_i > y_j$;

– neither concordant nor discordant iff $x_i = x_j$ (x- tie) or $y_i = y_j$ (y- tie) or $x_i = x_j = y_i = y_j$ (joint tie).

Let n_c and n_d be the number of concordant pairs and discordant pairs, respectively. Kendall's τ is calculated as

$$\tau = \begin{cases} \frac{n_c - n_d}{n(n-1)/2}, & \text{if no tie} \\ \frac{n_c - n_d}{\sqrt{(n_0-n_1)(n_0-n_2)}}, & \text{otherwise,} \end{cases} \qquad [10.18]$$

where $n_0 = n(n-1)/2$, $n_1 = \sum_i t_i(t_i-1)/2$, $n_1 = \sum_j u_j(u_j-1)/2$, t_i is the number of ties in the ith group of ties for $\mathbf{x}$ and u_j is the number of ties in the jth group of ties for $\mathbf{y}$.

Kendall's correlation ranges between 1 and -1. If $\tau = 1$, there are no non-joint ties and the two total orders induced by the vectors are the same. The converse is true for $\tau = -1$, i.e. no non-joint ties and the two total orders are of opposite signs. When $\tau = 0$, the pairs are not correlated.

Top k lists: This technique was introduced by Fagin *et al.* (2003). It is understood as a ranking metric in which the first k elements are considered in the list of n, where $k \leq n$. The top k lists can be ranked in many ways (Fagin *et al.* 2003). However, due to its simplicity, Kendall's metric is mostly used.

Let $N = \{1, \ldots, n\}$ be a set of size n. Suppose that r_1 and r_2 are two top k rankings on N. According to Rolland (2007) and Fagin *et al.* (2003), Kendall's distance metric is defined as $d_K(r_1, r_2) = \sum_{\{i,j\} \in N} K_{i,j}(r_1, r_2)$, where

$$K_{i,j}(r_1, r_2) = \begin{cases} 0, & \text{if } i \text{ and } j \text{ appear in the same order in } r_1 \text{and } r_2, \\ 1, & \text{if } i \text{ and } j \text{ appear in the opposite order,} \\ 0, & \text{if } i \text{ is ahead of } j \text{ in } r_1. \end{cases} \quad [10.19]$$

Alternatively, $d_K(r_1, r_2)$ is the total sum of pairwise discordances between two k lists.

To determine the stopping criterion of the iterative methods, one assumes that $r_1^{(m)}$ and $r_2^{(m+1)}$ are the ranks at mth and $(m+1)$th iterations, respectively. Then, the Kendall's distance for top k lists is expressed as

$$d_K\left(r_1^{(m)}, r_2^{(m)}\right) = \sum_{\{i,j\} \in N} K_{i,j}\left(r_1^{(m)}, r_2^{(m+1)}\right).$$

We normalized d_K, hence we obtain Kendall's τ correlation.

10.5. Numerical experimentation of stopping criteria

In this section, we evaluate the stopping criteria for ranks of linear system of equations. We feel that it is essential to base the stopping criteria on intended purpose, that is induced ranks rather than solution of the equation. To that effect, good evaluation framework should be carried out to ensure sound stopping criteria for ranks. We set up the evaluations as follows:

– convergence of five stopping criteria;

– quantize and ranks at different iterations;

– Kendall's coefficient τ against number of iterations for different stopping criteria by iterative method;

– top k list (100 and 300) against number of iterations by iterative method.

10.5.1. *Convergence of stopping criterion*

In this section, we present evaluation of five criteria mostly used in iterative schemes. We performed 20–100 iterations and results are presented in Table 10.1 and Figure 10.1. The findings revealed that criterion I (1-norm of residual $\|\mathbf{x}^{(m)} - A\mathbf{x}^{(m-1)}\|$) and criterion IV ($\frac{\|\mathbf{x}^m - A\mathbf{x}^{m+1}\|_\infty}{\|\mathbf{b}\|_\infty}$) seem to suit convergence of ranks of linear system, since their error tolerance was within the range as suggested in (Engström and Silvestrov 2015) and (Stewart 1994). Also, criteria V ($\frac{c}{1-c\|}\|\mathbf{x}^{(m)} - \mathbf{x}^{(m-1)}\|$) is good alternative if one wish to choose any other, more specifically for slow convergence scheme. While criterion II and III performed badly, they rather have faster convergence than expected.

Method	**Error by criterion**	**No. of iterations**			
		20 ($\times 10^{-3}$)	40 ($\times 10^{-4}$)	60 ($\times 10^{-6}$)	100 ($\times 10^{-9}$)
Power series and Jacobi	I	38.8	15.0	58.0	86.8
	II	6.0	2.25	8.70	1.30
	III	15.6	3.26	7.75	5.51
	IV	38.8	15.0	58.0	86.8
	V	219.6	85	328.0	492.0
SOR	I	37.1	14.0	48.9	64.68
	II	5.7	2.07	7.34	9.68
	III	14.7	2.87	6.36	3.96
	IV	37.1	14.0	48.9	64.68
	V	210.5	77.0	277.6	366.0

Table 10.1. *Error tolerance and stopping criteria by iterative methods*

The behavior of error around 10^{-15} could probably be due to machine precision being similar. In MATLAB software, double precision in IEEE format can store up to 16 digits only.

10.5.2. *Quantiles*

The quantiles (10%, 25%, 50%, 75% and 90%) of solutions for the iterative methods on four sets of iteration were performed and results are presented in Table 10.2. Criterion I was used as a stopping requirement because it is simple and found to be better as noted earlier. We found that no significant difference in solutions at

different quantiles points. Further, lower ranks seem to converge much faster than higher ranks (see Table 10.2).

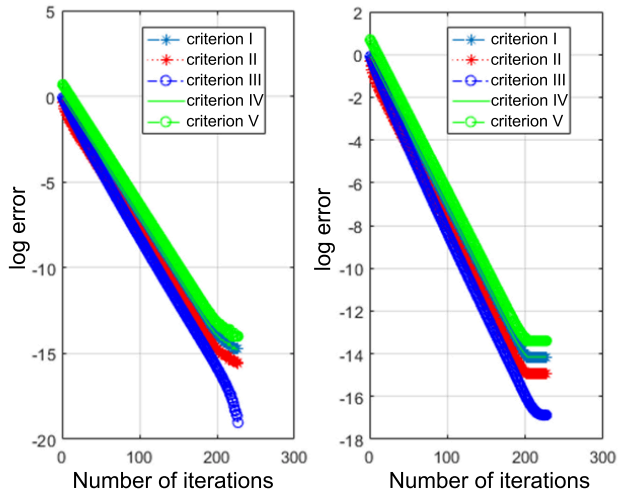


Figure 10.1. *Convergence of five criteria using power series/Jacobi iterative method (left) and SOR (right). For a color version of this figure, see www.iste.co.uk/skiadas/data1.zip*

When the numbers of iteration was increased to 60, we noted that ranks seem to have stabilized. To get a better picture, we present a plot in Figure 10.1. It can be clearly seen that in the early states of iterations, ranks were unstable; however, after at least 50 iterations, most ranks achieved their limiting values. Moreover, the iterative scheme had converged already in at most 30 iterations. Hence, the scheme had converged before the ranks.

10.5.3. *Kendall correlation coefficient as stopping criterion*

Further, we explored the use of Kendall correlation coefficient τ as a stopping criterion. We first assumed that the exact solution vector is known, then we determined its correlation at different iteration points. To avoid heavy computation, the top 100 and 300 lists were considered for analysis of correlation. Taking an interval of 10 and considering the 10th iterations as the starting point and ending at 220th iterations. The followings were observed: the first time convergence of rank was at about 60th iterations, as shown in Figure 10.3. This has been revealed by both iterative methods and this could be the stopping point, which match with the finding using criterion I and II.

		No. of Iterations	
Method	**Quartiles**	**40**	**60**
		solution	solution
Power series and Jacobi	10%	3.5630	3.5637
	25%	4.6043	4.6051
	50%	5.4261	5.4273
	75%	5.9494	5.9515
	90%	6.2534	6.2565
SOR	10%	3.5630	3.5637
	25%	4.6045	4.6051
	50%	5.4263	5.4273
	75%	5.9497	5.9516
	90%	6.2539	6.2565

Table 10.2. *Quartiles, number of iterations and solutions using stopping criterion I and II*

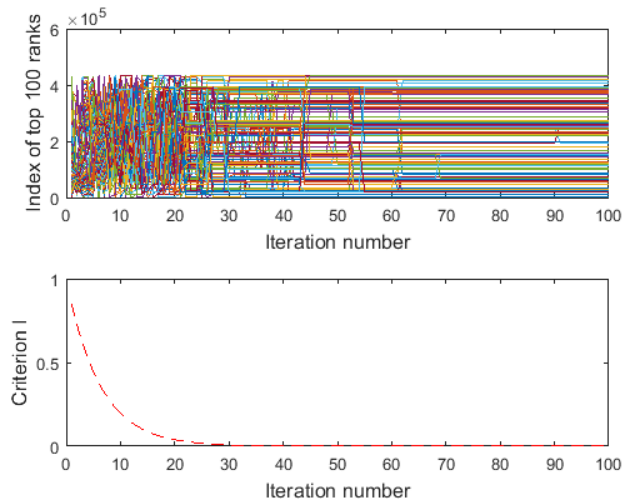


Figure 10.2. *Index of first 100 top ranks and error (criterion I) against iteration number using power series method. For a color version of this figure, see www.iste.co.uk/skiadas/data1.zip*

Second, we observed that a proper rank can be achieved at a particular iteration and lost as iteration progresses. This seems to be pronounced in SOR as compared to Jacobi or power series methods, particularly with top 100 ranks. However, for top 300 ranks it seems to achieve their convergence once and this can be seen in Figure 10.3. Based on these results, determining convergence rate of ranks using Kendall τ coefficient seems to be an effective technique as compared to the others, in particular, when two top k lists are compared and where they coupled is the best stopping point. We evaluated this argument using top 100 and top 300, and in all cases the results were promising. The number of iterations required were 85 and 69 for Jacobi and SOR methods, respectively, as presented in Figure 10.3.

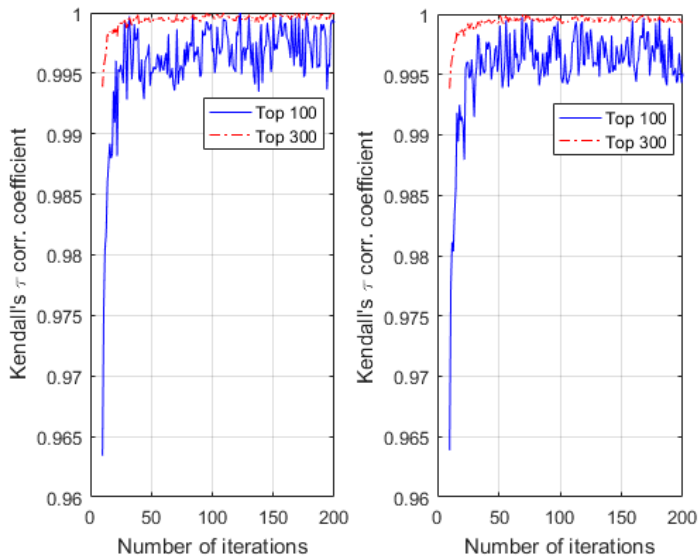


Figure 10.3. *Convergence of top 100 and top 300 ranks using Kendall's τ correlation coefficient with Jacobi (left) and SOR (right) iterative schemes. For a color version of this figure, see www.iste.co.uk/skiadas/data1.zip*

10.6. Conclusions

Stopping criterion of ranks for linear system of equations is of practical significance, particularly if ranking is to be meaningful and applicable in areas such as search engines, financial networks, bioinformatics and so on. In this chapter, we evaluated several criteria and the findings revealed that Kendall's τ method seems to

be an effective stopping criterion for ranks of linear systems of equations as compared to the others. It was found that this technique, which is based on the correlation coefficient between successive iterates of solution vectors together with two top k lists (for example top 100 and 300) has good convergence, and we named it Kendall's τ 2 top k list method. Comparing the method with other stopping criteria, it was observed that the stopping point can easily be identified as shown in Figure 10.3.

Further, criteria I and IV were found to be competing with Kendall's τ 2 top k list method, whereas criterion V seems to be relevant for slow convergence linear system problem. We conclude that this study should play a complementary role when the purpose is to determine ranks in linear system of equations.

10.7. Acknowledgments

This research was supported by the Swedish International Development Cooperation Agency (Sida), International Science Programme (ISP) in Mathematical Sciences (IPMS) and Sida Bilateral Research Program (Makerere University and University of Dar-es-Salaam). We are also grateful to the research environment Mathematics and Applied Mathematics (MAM), Division of Applied Mathematics, Mälardålen University for providing an excellent and inspiring environment for research education and research.

10.8. References

Arioli, M., Duff, I., Ruiz, D. (1992). Stopping criteria for iterative solvers. *Journal SIAM Journal on Matrix Analysis and Applications*, 13(1), 138–144.

Bennani, M., Braconnier, T. (1993). Stopping criteria for eigensolver, *Submitteto IMA Num. Annal*.

Berkhin, P. (2005). A survey on PageRank computing. *Internet Mathematics*, 2(1), 73-120.

Bidoki, Z., Mohammand, A., Yazdani, N. (2007). DistanceRank: an intelligent ranking algorithm for web pages. *Information Processing and management*, 18(1), 134–160.

Boldi, P., Santini, M., Vigna, S. (2005). Paradoxical effects in PageRank incremental computations. *Internet mathematics*, 2(3), 387–404.

Dwork, C., Kumar, R., Naor, M., Sivakamar, D. (2001). Rank aggregation methods for the web. *World Wide Web,* 10, 613–622.

Engström, C., Silvestrov, S. (2015). A componentwise PageRank algorithm. *16th ASMDA Conference Preceedings*, ISAST.

Engström, C., Silvestrov, S. (2016). PageRank, connecting a line of nodes with a complete graph. In: *Engineering Mathematics II. Springer Proceedings in Mathematics & Statistics*, Silvestrov S., Rančić M. (eds.), 179, Springer, Cham.

Fagin, R., Kumar, R., Sivakumar, D. (2003). Comparing top lists. *J. Discrete Mathematics*, 18(1), 134–160.

Gliech, D., Zhukov, L., Berhim, P. (2005). Fast parallel PageRank: A Linear System Approach, *WWW2005,* Chiba and Japan.

Gong, C., Fu, K., Loza, A., Wu, Q., Liu, J., Yang, J. (2014). PageRank tracker: From ranking to tracking, *IEEE Transactions on Cybernetics*, 44(6), 882–893, 2014.

Hadjidimos, A. (2000). Successive overrelaxation (SOR) and related methods. *Journal of Computational and Applied Mathematics*, 123, 177-199.

Haveliwala, T.H. (1999). Efficient computation of PageRank. *Tech. Rep, 31*, Stanford University.

Kamvar, S.D., Haveliwala, T.H., Manning, C.D., Golub, G.H. (2003). Extrapolation methods for accelarating PageRank computations, *World Wide Web,* 12, 261-270.

Kohlscütter, C., Chirita, P., Nejdl, W. (2009). Efficient Parallel Computation of PageRank, http://www.l3s.de/ chirita/publications/kohlschuetter06efficient.pdf.

Langville, A.N., Meyer, C.D. (2006). A reoredering for PageRank problem. *SIAM J. Sci. Comput.*, 6(27), 2112–2120.

Li, C., Wu, S. (2014). Some new comparison theorems for double splitting of matrices. *Applied Mathematics and Information Sciences*, 8(5), 2523–2526.

Meyer, C.D. (2000). Matrix analysis and applied linear algebra. *Society for Industrial and Applied Mathematics*, Philadelphia, PA, USA.

Mielou, J.C., Spiteri, P., El Baz, D. (2008). A new stopping criterion for linear pertubed asyncronous iterations, *Journal of Computational and Applied Mathematics*, 219, 471–483.

Qiu, F., Cho, J. (2006). Automatic identification of user interest for personalised search. *WWW '06: Proc. 15th International Conference on World Wide Web*, ACM Press, New York, 727-736.

Rolland, A. (2007). *A Note on Top-K Lists: Aveage Distance between Two Top-K Lists*. University Lyon 2.

Stewart, W.J. (1994). *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press.

Varga, R.S. (1962). *Matrix iterative Analysis,* Prentice-Hall, Englewood Cliffs, NJ.

Wills, R.S. (2009). When ranks trumps precion: Using the power method to compute Google's Pagerank. http://www.lib.ncsu.edu/theses/available/etd-061220007-173712/.

Wills, R.S., Ipsen, I.C.F. (2009). Ordinal ranking for google PageRank. *SIAM J. Matrix Anal. Appl.*, 30(4), 1677-1696.

Young, D.M. (1971). *Iterative solutions for Linear Systems*. Academic Press, New York.

11

Estimation of a Two-Variable Second-Degree Polynomial via Sampling

In various fields of environmental and agriculture sciences, the estimation of two-variable second-degree polynomial coefficients via sampling is of major importance, as it gives very useful information. In this chapter, we propose a very simple and very low budget systematic sampling plan for the estimation of the coefficients A, B, C, D, E and H of the polynomial $f(x,y) = (Ax^2 + By^2 + Cxy + Dx + Ey + H)^{-1}$, which is sometimes found to be a probability density function. The above polynomial is defined on a domain $D = [a, b] \times [c, d]$, which can be represented by the domain $D = [0, 1] \times [0,1]$ for convenience. Numerical methods, such as Simpson's rule, are applied. The comparison between means of both estimated and theoretic functions is used to confirm the accuracy of the results. The stability of the numerical methods allows us to get results with very good accuracy for small sample sizes. Illustrative examples are given in the following.

11.1. Introduction

All across the world, people are facing a wealth of environmental problems every day. Point-source pollutants have a major impact on environmental concentrations on a local scale and also contribute to the concentrations on a larger regional scale (van Leeuwen 2010). In accordance with the U.S. Environmental Protection Agency (EPA), point-source pollution is defined as "any single identifiable source of pollution from which pollutants are discharged, such as a pipe, ditch, ship or factory smokestack" (Hill 1997). Environmental authorities are concerned with locating and punishing violations of environmental protection regulations, but even if laws are followed, these types of practices occurred in the past before the laws were enacted and the pollutants are still around (Osborne 2017).

Chapter written by Ioanna PAPATSOUMA, Nikolaos FARMAKIS and Eleni KETZAKI.

Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining,
First Edition. Edited by Christos H. Skiadas and James R. Bozeman.

From the mathematical point of view, these sources are called point sources because, in mathematical modeling, they can be approximated as a mathematical point to simplify analysis.

Let us consider the bivariate function:

$$f(x,y) = \left(Ax^2 + By^2 + Cxy + Dx + Ey + H\right)^{-1} \quad [11.1]$$

defined on a domain $D = [a, b] \times [c, d]$. The coefficient H is equal to the inverse value of the function f when both x and y are 0, $H^{-1} = f(0,0)$, and the rest coefficients are estimated as described in section 11.2.

11.2. Proposed method

We will describe a method to estimate the coefficients A, B, C, D and E of the function given in [11.1], where the studied area is $D = [0, 1] \times [0, 1]$. The proposed method consists of the following steps:

Step 1. We conduct systematic sampling (Farmakis 2016) by taking samples of linear subspaces of R^2.

Step 2. In each sampling space (Step 1), we integrate the function given in [11.1] with respect to x or y, for x or y between 0 and 1.

Step 3. We apply the Simpson's rule of integration (Davis and Rabinowitz 2007) to three points, which are equally spaced in the interval $[0, 1]$, $0, 0.5$ and 1.

Step 4. We equate each integration result (Step 2) with its approximation (Step 3).

Step 5. We solve the system derived from Step 4.

11.2.1. *First restriction*

Let us consider that $y = 0$:

$$f_1(x) = f(x,0) = \left(Ax^2 + Dx + H\right)^{-1} \quad [11.2]$$

or

$$Ax^2 + Dx + H = 1/f_1(x) \quad [11.3]$$

The sampling is done on the x-axis of the given field D (Figure 11.1a). We integrate equation [11.3] with respect to x, x between 0 and 1, and we get:

$$I_1 = (2A + 3D + 6H)/6 \qquad [11.4]$$

By applying Simpson's rule of integration to the three points, which are equally spaced in the interval [0,1] $x = 0, 0.5$ and 1, we get:

$$I_1 = \left(1/f_1(0) + 4/f_1(0.5) + 1/f_1(1)\right)/6 = 1/6\left(\sum 1/f_1\right) \qquad [11.5]$$

If we equate [11.4] and [11.5], we get the so-called $\tilde{g}_1$ estimator:

$$6I_1 = 2A + 3D + 6H = \tilde{g}_1 = 1/f_1(0) + 4/f_1(0.5) + 1/f_1(1) = \sum 1/f_1 \qquad [11.6]$$

11.2.2. *Second restriction*

Let us consider that $x = 0$:

$$f_2(y) = f(0, y) = \left(By^2 + Ey + H\right)^{-1} \qquad [11.7]$$

or

$$By^2 + Ey + H = 1/f_2(y) \qquad [11.8]$$

The sampling is done on the y-axis of the given field D (Figure 11.1b). We integrate equation [11.8] with respect to y, y between 0 and 1, and we get:

$$I_2 = (2B + 3E + 6H)/6 \qquad [11.9]$$

By applying Simpson's rule of integration to the three points, which are equally spaced in the interval [0,1], we get:

$$I_2 = \left(1/f_2(0) + 4/f_2(0.5) + 1/f_2(1)\right)/6 = 1/6\left(\sum 1/f_2\right) \qquad [11.10]$$

If we equate [11.9] and [11.10], we get the so-called $\tilde{g}_2$ estimator:

$$6I_2 = 2B + 3E + 6H = \tilde{g}_2 = 1/f_2(0) + 4/f_2(0.5) + 1/f_2(1) = \sum 1/f_2 \qquad [11.11]$$

11.2.3. *Third restriction*

Let us consider that $x = y$:

$$f_3(x) = f(x,x) = \left((A+B+C)x^2 + (D+E)x + H\right)^{-1} \quad [11.12]$$

or

$$(A+B+C)x^2 + (D+E)x + H = 1/f_3(x) \quad [11.13]$$

The sampling is done on the diagonal line $y = x$ of the given field D (Figure 11.1c). We integrate equation [11.13] with respect to x, x between 0 and 1, and we get:

$$I_3 = \left(2(A+B+C) + 3(D+E) + 6H\right)/6 \quad [11.14]$$

By applying Simpson's rule of integration to the three points, which are equally spaced in the interval $[0,1]$, we get:

$$I_3 = \left(1/f_3(0) + 4/f_3(0.5) + 1/f_3(1)\right)/6 = 1/6\left(\sum 1/f_3\right) \quad [11.15]$$

If we equate [11.14] and [11.15], we get the so-called $\tilde{g}_3$ estimator:

$$6I_3 = 2(A+B+C) + 3(D+E) + 6H = \tilde{g}_3 = 1/f_3(0) + 4/f_3(0.5) + 1/f_3(1) = \sum 1/f_3 \quad [11.16]$$

11.2.4. *Fourth restriction*

Let us consider that $y = x/2$:

$$f_4(x) = f(x,x/2) = \left((A+B/4+C/2)x^2 + (D+E/2)x + H\right)^{-1} \quad [11.17]$$

or

$$(A+B/4+C/2)x^2 + (D+E/2)x + H = 1/f_4(x) \quad [11.18]$$

The sampling is done on the diagonal line $y = x/2$ of the given field D (Figure 11.1d). We integrate equation [11.18] with respect to x, x between 0 and 1, and we get:

$$I_4 = (4A + B + 2C + 6D + 3E + 12H)/12 \quad [11.19]$$

By applying Simpson's rule of integration to the three points, which are equally spaced in the interval [0,1], we get:

$$I_4 = (1/f_4(0) + 4/f_4(0.5) + 1/f_4(1))/6 = 1/6(\sum 1/f_4) \quad [11.20]$$

If we equate [11.19] and [11.20], we get the so-called $\tilde{g}_4$ estimator multiplied by 2:

$$12I_4 = 4A + B + 2C + 6D + 3E + 12H = 2\tilde{g}_4 = 2/f_4(0) + 8/f_4(0.5) + 2/f_4(1) = \sum 1/f_4 \quad [11.21]$$

11.2.5. *Fifth restriction*

Let us consider that $x = y/2$:

$$f_5(y) = f(y/2, y) = \left((A/4 + B + C/2)y^2 + (D/2 + E)y + H\right)^{-1} \quad [11.22]$$

or

$$(A/4 + B + C/2)y^2 + (D/2 + E)y + H = 1/f_5(y) \quad [11.23]$$

The sampling is done on the line $y = 2x$ of the given field D (Figure 11.1e). We integrate equation [11.23] with respect to y, y between 0 and 1, and we get:

$$I_5 = (A + 4B + 2C + 3D + 6E + 12H)/12 \quad [11.24]$$

By applying Simpson's rule of integration to the three points, which are equally spaced in the interval [0,1], we get:

$$I_5 = (1/f_5(0) + 4/f_5(0.5) + 1/f_5(1))/6 = 1/6(\sum 1/f_5) \quad [11.25]$$

If we equate [11.24] and [11.25], we get the so-called $\tilde{g}_5$ estimator multiplied by 2:

$$12I_5 = A + 4B + 2C + 3D + 6E + 12H = 2\tilde{g}_5 = 2/f_5(0) + 8/f_5(0.5) + 2/f_5(1) = \sum 1/f_5 \quad [11.26]$$

11.2.6. *Coefficient estimates*

Figure 11.1 illustrates the restrictions on the use of the proposed method and summarizes the sampling spaces (bold line) using systematic sampling.

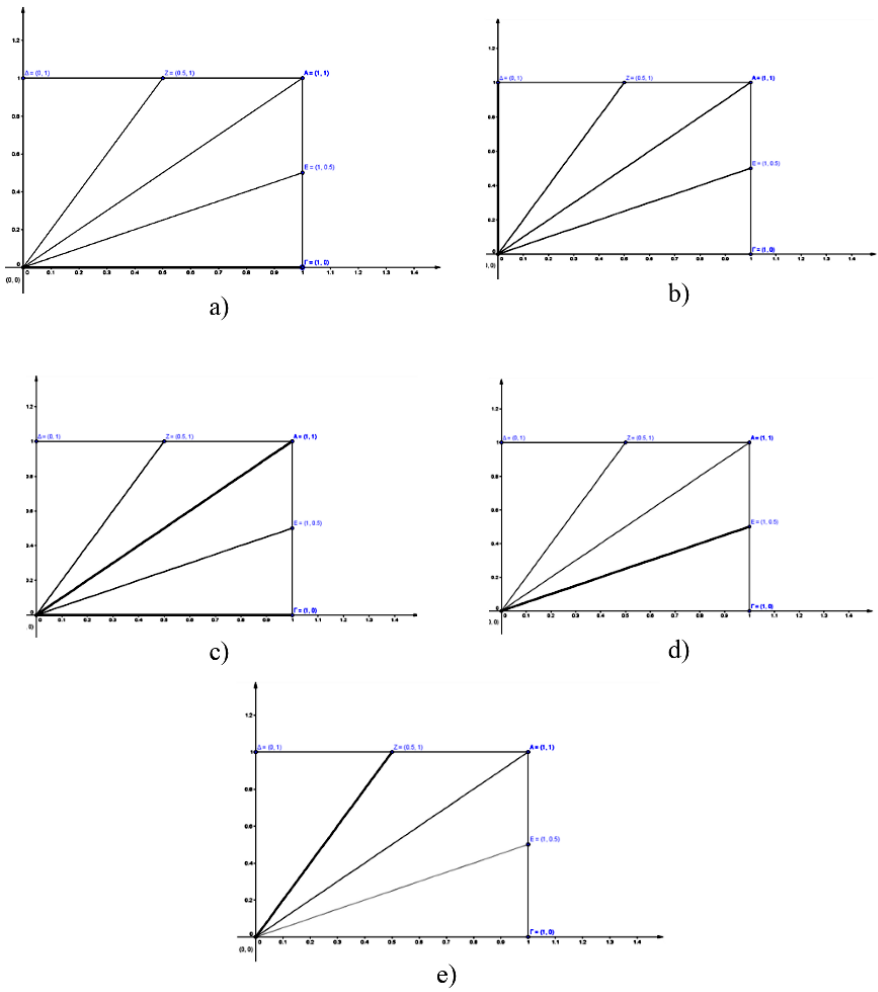


Figure 11.1. *Restrictions on the use of the proposed method*

By solving the following system derived from [11.6], [11.11], [11.16], [11.21] and [11.26]:

$$\begin{bmatrix} 2 & 0 & 0 & 3 & 0 & 6 \\ 0 & 2 & 0 & 0 & 3 & 6 \\ 2 & 2 & 2 & 3 & 3 & 6 \\ 4 & 1 & 2 & 6 & 3 & 12 \\ 1 & 4 & 2 & 3 & 6 & 12 \end{bmatrix} \cdot \begin{bmatrix} A \\ B \\ C \\ D \\ E \\ H \end{bmatrix} = \begin{bmatrix} \tilde{g}_1 \\ \tilde{g}_2 \\ \tilde{g}_3 \\ 2\tilde{g}_4 \\ 2\tilde{g}_5 \end{bmatrix} \quad [11.27]$$

we estimate the coefficients of the bivariate function as follows:

$$\begin{aligned} A &= \tilde{g}_2 + \tilde{g}_3 - 2\tilde{g}_5 \\ B &= \tilde{g}_1 + \tilde{g}_3 - 2\tilde{g}_4 \\ C &= 3H + (\tilde{g}_3 - \tilde{g}_2 - \tilde{g}_1)/2 \\ D &= -2H + (\tilde{g}_1 - 2\tilde{g}_2 - 2\tilde{g}_3 + 4\tilde{g}_5)/3 \\ E &= -2H - (2\tilde{g}_1 - \tilde{g}_2 + 2\tilde{g}_3 - 4\tilde{g}_4)/3 \end{aligned} \quad [11.28]$$

where $H = f^{-1}(0,0)$, and the formula for the $\tilde{g}_i$ estimators is given by:

$$\tilde{g}_i = \begin{cases} \sum 1/f_i, & if\ i = 1,2,3 \\ (\sum 1/f_i)/2, & if\ i = 4,5 \end{cases} \quad [11.29]$$

11.3. Experimental approaches

11.3.1. *Experiment A*

Five different three-point measurements of pollutant concentration over a field have been recorded. The inverse values of a function measuring pollutant concentration obeying the five restrictions described are presented in Table 11.1.

The last column includes the $\tilde{g}_i$ estimators, $i = 1,2,3,4,5$, derived from the experimental measurements.

Points			$\tilde{g}$
0	**0.5**	**1**	
18	16.75	17	102
18	18.25	22	113
18	17.5	23	111
18	16.6875	18.25	103
18	17.6875	21.75	110.5

Table 11.1. *Pollutant concentration*

It can be easily observed that the coefficient H is equal to 18. The rest of the coefficients are estimated as follows:

$$A = \tilde{g}_2 + \tilde{g}_3 - 2\tilde{g}_5 = 113 + 111 - 221 = 3$$

$$B = \tilde{g}_1 + \tilde{g}_3 - 2\tilde{g}_4 = 102 + 111 - 206 = 7$$

$$C = 3H + \left(\tilde{g}_3 - \tilde{g}_2 - \tilde{g}_1\right)/2 = 54 + \left(111 - 113 - 102\right)/2 = 2 \qquad [11.30]$$

$$D = -2H + \left(\tilde{g}_1 - 2\tilde{g}_2 - 2\tilde{g}_3 + 4\tilde{g}_5\right)/3 = -36 + \left(102 - 226 - 222 + 442\right)/3 = -4$$

$$E = -2H - \left(2\tilde{g}_1 - \tilde{g}_2 + 2\tilde{g}_3 - 4\tilde{g}_4\right)/3 = -36 - \left(204 - 113 + 222 - 412\right)/3 = -3$$

The requested function is given as

$$f\left(x,y\right) = \left(3x^2 + 7y^2 + 2xy - 4x - 3y + 18\right)^{-1} \qquad [11.31]$$

We use statistical software to investigate the above function's local extrema. The denominator of the function has a local minimum at point $(5/8, 1/8)$, which equals

to $265/16 = 16.5625$. This implies that the pollutant concentration function reaches its maximum value at the same point, which equals to $f(5/8, 1/8) = 16/265 = 0.0604$ (accurate to four decimal places). Figure 11.2 confirms our analysis.

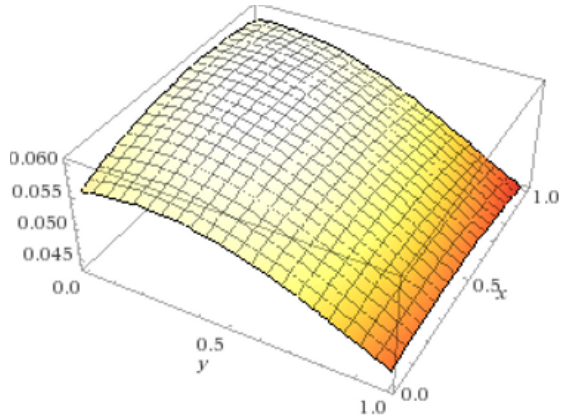


Figure 11.2. *Point-source pollutant. For a color version of this figure, see www.iste.co.uk/skiadas/data1.zip*

11.3.2. *Experiment B*

Five different three-point measurements of intensity of radiation at a height of 3 m over a field have been recorded. The inverse values of a function measuring the intensity of radiation obeying the five restrictions described are presented in Table 11.2.

The last column includes the $\tilde{g}_i$ estimators, $i = 1, 2, 3, 4, 5$, derived from the experimental measurements.

Points			$\tilde{g}$
0	**0.5**	**1**	
11	10	11	62
11	10.75	12	66
11	9	9	56
11	9.3125	9.25	57.5
11	9.625	9.5	59

Table 11.2. *Intensity of radiation*

It can be easily observed that the coefficient H is equal to 11. The rest coefficients are estimated as follows:

$$A = \tilde{g}_2 + \tilde{g}_3 - 2\tilde{g}_5 = 113 + 111 - 221 = 3$$

$$B = \tilde{g}_1 + \tilde{g}_3 - 2\tilde{g}_4 = 102 + 111 - 206 = 7$$

$$C = 3H + \left(\tilde{g}_3 - \tilde{g}_2 - \tilde{g}_1\right)/2 = 33 + \left(111 - 113 - 102\right)/2 = 2 \qquad [11.32]$$

$$D = -2H + \left(\tilde{g}_1 - 2\tilde{g}_2 - 2\tilde{g}_3 + 4\tilde{g}_5\right)/3 = -36 + \left(102 - 226 - 222 + 442\right)/3 = -4$$

$$E = -2H - \left(2\tilde{g}_1 - \tilde{g}_2 + 2\tilde{g}_3 - 4\tilde{g}_4\right)/3 = -36 - \left(204 - 113 + 222 - 412\right)/3 = -3$$

The requested function is given as

$$f(x,y) = \left(3x^2 + 7y^2 + 2xy - 4x - 3y + 11\right)^{-1} \qquad [11.33]$$

We use statistical software to investigate the above function's local extrema. The denominator of the function has a local minimum at point $(30/39, 28/39)$, which equals to $341/39 = 8.7436$. In other words, the pollutant concentration function, given in [11.33], reaches its maximum value at the same point, which equals to $f(30/39, 28/39) = 39/341 = 0.1144$ (accurate to four decimal places). Figure 11.3 confirms our analysis.

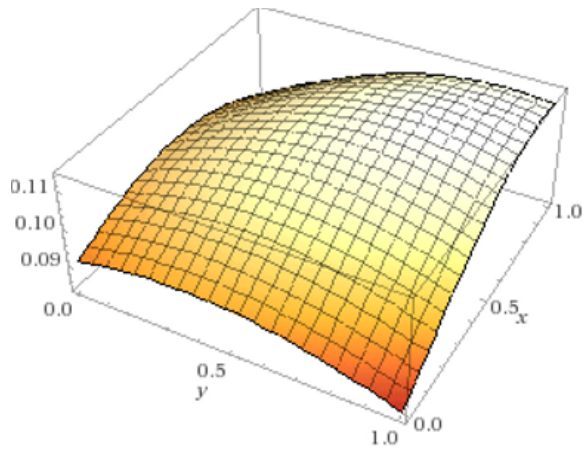


Figure 11.3. *Point-source of radiation. For a color version of this figure, see www.iste.co.uk/skiadas/data1.zip*

11.4. Conclusions

In this chapter, we have proposed a simple and easy-to-remember method to estimate the coefficients of a two-variable second-degree polynomial via sampling. We have used the systematic sampling to define the areas of sampling and the Simpson's rule of integration to approximate definite integrals on the domain $D = [0,1] \times [0,1]$. The proposed method can be characterized as a low-budget sampling because of the small sample size. It is advantageous because it can be applied to detect point-source pollutants and point-source radiation, and thus determine risk assessment in public health. Finally, the stability of the numerical methods allows us to get accurate results.

11.5. References

Davis, P.J. and Rabinowitz, P. (2007). *Methods of Numerical Integration*, 2nd edition. Dover Publications, Mineola, NY.

Farmakis, N. (2016). *Introduction to Sampling*. Kyriakidis Bros – Publications SA (in Greek), Thessaloniki, Greece.

Hill, M.S. (1997). *Understanding Environmental Pollution*. Cambridge University Press, Cambridge, UK.

Osborne, F.H. (2017). *Kean University Continuing Education. Implementing the Science Standards K-4. Point Source and Non-Point-Source Pollution.* http://www.kean.edu/~fosborne/resources/ex12j.htm (Accessed July 20, 2017).

van Leeuwen, C.J. (2010). *Risk Assessment of Chemicals: An Introduction*, 2nd edition. Springer, Dordrecht, The Netherlands.

PART 3

Estimators, Forecasting and Data Mining

12

Displaying Empirical Distributions of Conditional Quantile Estimates: An Application of Symbolic Data Analysis to the Cost Allocation Problem in Agriculture

This chapter uses the symbolic data analysis tools in order to display and analyze the conditional quantile estimates, with an application to the cost allocation problem in agriculture. After recalling the conceptual framework of the estimation of agricultural production costs, the first part presents the empirical data model, the quantile regression approach and the interval data techniques used as symbolic data analysis tools. The second part presents the comparative analysis of the econometric results for wheat between 12 European member states, using principal component analysis and hierarchic clustering of estimates and range of estimation intervals, discussing the relevance of the displays obtained for intercountry comparisons based on specific productivity.

12.1. Conceptual framework and methodological aspects of cost allocation

Successive reforms of the common agricultural policy, as well as integration of the new member state agricultures resulting from enlargement of the European Union (EU) have raised recurring needs for estimating costs of production of major agricultural products both in the context of competitive markets as in markets subject to regulation. The analysis of agricultural production costs, whether

Chapter written by Dominique DESBOIS.

Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining,
First Edition. Edited by Christos H. Skiadas and James R. Bozeman.

retrospective or prospective, is also a tool for analyzing margins for farmers. It allows us to assess the price competitiveness of farmers, one of the major elements for development and sustainability of food chains in the European regions. Thus, the estimation of production costs provides partial but certainly needed insight into the issues raised by the adaptation of European agriculture in the context of agricultural markets, whether they are national, European or international.

Given these issues, in contexts of either ex ante or ex post scenarios, for evaluation of eventual options of public agricultural policy, we must be able to provide information on the entire distribution of production costs and to meet the needs of simulations and impact assessment in the various common market organizations. In this perspective, based on the observation of asymmetry and heteroskedasticity within the empirical distribution of agricultural inputs, we propose a methodology adapted to the problem of estimating the specific costs of production for the main agricultural products in a European context where agricultural holdings remain mainly oriented toward multiple productions, despite a preponderance of farms specializing in some of the more integrated agricultural production sectors.

To this end, we propose a methodology to obtain and analyze estimates of these quantile-specific costs that are conditioned by the product mix of farmers. To demonstrate the relevance of this approach, we will apply this methodology to estimate the specific costs of wheat, agricultural commodity most commonly produced in the EU28, to a set of 12 European countries (EU12) where this production is significant in 2006.

We first present the empirical estimation of specific production cost model derived from an econometric approach to cost allocation using a microeconomic model to build an input–output matrix (Divay and Meunier 1980). Second, we introduce the methodology for estimating conditional quantiles as proposed by Koenker and Bassett (1978). Third, we present the symbolic data analysis tools used in our procedures, mainly based on the concepts and methods described in the study of Bock and Diday (2000), using symbolic principal component and divisive clustering analyses. Fourth, the various displays provided by these symbolic tools are commented and discussed based on the conditional quantile estimates obtained for wheat production at the European level. Eventually, we conclude with the effectiveness of this approach for the wheat production with a proposal to extend it to other main agricultural products at the regional level of analysis.

12.2. The empirical model of specific production cost estimates

In the EU-Farm Accounting Data Network (FADN) survey, the input recording occurs in aggregated form at the farm level and does not provide direct estimates of

production costs incurred by each agricultural commodity produced. In contrast, the EU-FADN survey provides at the farm level, on one hand, the amount of gross product generated by each of the various commodities and, on the other hand, the sum of expenses for all the specific agricultural inputs recorded. So, it becomes possible to estimate, with an econometric model regressing the specific inputs on the gross products, the allocation coefficients of agricultural inputs to the main agricultural products, denoted as "specific coefficients" of production.

Following Desbois *et al*. (2015), the econometric model decomposes linearly x the sum of specific inputs for each farm holding i according to the gross product Y of each agricultural commodity j, expressed by the following stochastic equation:

$$x_i = \sum_{j=1}^{P} \gamma_j Y_i^j + \varepsilon_i \text{ with } \varepsilon_i \text{ iid} \quad [12.1]$$

12.3. The conditional quantile estimation

To take into account the intrinsic heterogeneity of the distribution of the specific costs, we estimate the specific production coefficients of equation [12.1] accordingly with the methodology of quantile regression (Koenker and Bassett 1978), the solution being expressed in terms of conditional quantile of order q:

$$\hat{\mu}_q(i) = \sum_{j=1}^{P} \hat{\gamma}_j^{(q)} Y_i^j \quad [12.2]$$

According to Cameron and Trivedi (2005), we assume that the data generator process is a linear model with multiplicative heteroscedasticity characterized in matrix form by:

$$x = Y'\beta + u \text{ with } u = Y'\alpha \times \varepsilon \text{ and } Y'\alpha > 0 \quad [12.3]$$

where $\varepsilon \sim \text{iid}[0, \sigma]$ is a random-vector identically and independently distributed with zero mean and constant variance σ^2. Under this assumption

$$\mu_q(x|Y, \beta, \alpha),$$

the qth conditional quantile of the production cost x, the production cost, conditioned by Y and the α and β parameters, is derived analytically as follows:

$$\mu_q(x|Y, \beta, \alpha) = Y'[\beta + \alpha \times F_e^{-1}(q)] \quad [12.4]$$

where F_ε is the cumulative distribution function of the errors.

Two kinds of models can be distinguished:

1) $x = Y'\beta + u$ with $u = K\epsilon$, homoscedastic errors $V(\epsilon|Y) = \sigma^2$, denoted as the *location-shift* model, that is, the linear model of conditional quantile with homogeneous slopes; while $Y'\alpha = K$ is constant, the conditional quantiles $\mu_q(x|Y, \beta, \alpha) = Y'\beta + KF_e^{-1}(q)$ get all the same slope, but differ only by a constant gap, growing as q, the quantile order, increases;

2) $x = Y'\beta + (Y'\alpha) \times \varepsilon$ and $Y'\alpha > 0$ with heteroscedastic residuals, referred as the *location-scale shift* model, i.e. the linear model of heterogeneous conditional quantile slopes.

Weighted conditional quantiles were proposed as weighted L-estimates in heteroscedastic linear models (Koenker and Zhao 1994) defined by the weighting $\{\omega_i; i = 1, \cdots, n\}$. Those weights correspond to the inverse of the FADN sampling frequency in order to ensure the country representativeness of the estimates.

The observation weighting leads to a quantile regression scheme solving the minimization problem [12.5]:

$$\hat{\beta}_\omega(q) = \underset{\beta \in R^p}{Argmin} \left\{ \sum_{i; y_i \geq x_i'\beta} [\omega_i q |y_i - x_i'\beta|] + \sum_{i; y_i < x_i'\beta} [\omega_i (1-q) |y_i - x_i'\beta|] \right\}$$

Given the FADN sample size and its non-random selection, we opt for the method of resampling-based procedure based on the Markov Chain Marginal Bootstrap (MCMB) because no assumption on distributions of hazards is needed; this method gives robust empirical confidence intervals in a reasonable computation time (He and Hu 2002).

12.4. Symbolic analyses of the empirical distributions of specific costs

The symbolic data analysis tools used in this chapter are mainly based on the vertices principal component analysis (V-PCA; Cazes *et al*. 1997) and the PCA of the range transformation of interval data (RT-PCA; Lauro and Palumbo 2000), with the divisive clustering method (Chavent and Bock 2000) chosen to ensure the best mathematical coherency for the display and analysis of the conditional quantile estimates.

Let us denote by $\Delta = \{\delta_1, \cdots, \delta_l, \cdots \delta_L\}$, the empirical distributions of specific costs, as symbolic objects (SOs) representing the L countries, each one described by a set of M modalities of the conditional quantile estimator distribution for the jth product and the lth country:

$$\hat{\beta}_l^j = \left\{\hat{\beta}_l{}^j(q_1), \ldots, \hat{\beta}_l{}^j(q_m), \ldots, \hat{\beta}_l{}^j(q_M)\right\}.$$

Let us consider the estimation intervals of the *m*th MCMB conditional quantile estimates for the *l*th country and the *j*th product, which are denoted by:

$$\delta_l^m(j) = \left[Inf_{estim}\left(\hat{\beta}_l{}^j(q_m)\right); Sup_{estim}\left(\hat{\beta}_l{}^j(q_m)\right)\right] = \left[\underline{\delta_l^m(j)}; \overline{\delta_l^m(j)}\right].$$

These estimation intervals can be represented by the pair $\left(c_l^m(j); z_l^m(j)\right)$, where $c_l^m(j)$ is the conditional central estimate and $z_l^m(j) = \overline{\delta_l^m(j)} - \underline{\delta_l^m(j)}$ is the interval range transformation.

Because the MCMB intervals are not symmetric, we consider the hyper-volume, associated with the description of the SOs,

$$\delta_l, 1 \leqslant l \leq L,$$

computed as the Cartesian product $z_l^1 \times \cdots \times z_l^j \times \cdots \times z_l^p$ of the p associated quantile descriptors.

The description potential (DP) measurement is defined by:

$$\pi(\delta_l) = \prod_{j=1}^{p} \mu\left(z_l^j / s_j\right) = \prod_{j=1}^{p} \mu\left(\bar{z}_l^j\right), \qquad [12.6]$$

where is the normalized range with respect to the domain s_j. However, if the measurement of one of the descriptors tends to zero, then the DP tends to zero.

To overcome this problem, the linear description potential (LDP), we used a_l of SO δ_l, defined by De Carvalho (1997) as:

$$\sigma(a_l) = \sum_{j=1}^{p} \mu\left(\bar{z}_l^j\right) \qquad [12.7]$$

The range transformation principal component analysis (RT-PCA) is defined by the factorial decomposition of the total LDP:

$$LDP_{tot} = \sum_{l=1}^{L} \sigma(a_l) \qquad [12.8],$$

allowing a geometric representation of hypervolumes in which the *inf* vertices are translated to the origin. With regard to the orthogonality between couples of sides of each hypercube, the search of the optimal subspace to visualize the size and the shape of each quantile distribution as a SO can be implemented by a non-centered

PCA with respect to the *sup* vertices. The non-centered PCA performed on the matrix $\sqrt{\bar{z}_l^j}$ is decomposing the LDP_{tot} criterion.

Hence, our mixed strategy of factorial analysis of the distribution of specific costs combines the V-PCA and the RT-PCA in the three steps approach of the symbolic PCA (S-PCA) defined by Lauro and Palumbo (2000) to take into account the differences in scale and shape between empirical distributions of specific costs.

Also based on the estimation intervals of conditional quantiles, we use DIV[1], a divisive hierarchical clustering procedure (Chavent and Bock 2000), to obtain criteria for classifying countries according to their specific wheat costs.

The DIV evaluation criterion of the partition P_K with K clusters

$$P_K = (C_1, \cdots, C_k, \cdots, C_K)$$

is the sum of the homogeneity indices, such as:

$$W(P_K) = \sum_{C_k \in P_K} I(C_k) \qquad [12.9]$$

with the homogeneity index $I(C_k) = \frac{1}{2n_k} \sum_{\omega_l \in C_k} \sum_{\omega_{l'} \in C_k} D^2{}_E(\delta_l, \delta_{l'})$,

and the Euclidean distributional distance between two SO

$$D_E(\delta_l, \delta_{l'}) = \left(\sum_{j=1}^{p} d_j^2\left(\delta_l^j, \delta_l^j\right)\right)^{1/2} \qquad [12.10]$$

where $d_j\left(\delta_l^j, \delta_{l'}^j\right) = \left(\sum_{m=1}^{M} \left\{d^2\left(\underline{\gamma_l^m[j]}, \underline{\gamma_{l'}^m[j]}\right) + d^2\left(\overline{\gamma_l^m[j]}, \overline{\gamma_{l'}^m[j]}\right)\right\}\right)^{1/2}$

defining the interval Euclidean distance, while d is the Euclidean distance and $\gamma_l^m[j] = \delta_l^m[j]/s^m[j]$, the estimates normalized with respect to the $s^m[j]$ standard deviation.

12.5. The visualization and the analysis of econometric results

In 2010, according to Eurostat estimates, the EU27 accounts for 21% of world wheat production. The EU12 countries are among the largest producers in terms of amounts collected, which are arranged in the descending order as follows: France (FRA, 27.9%), Germany (DEU, 17.6%), United Kingdom (UKI, 10.9%), Poland

1 DIV is the divisive hierarchical clustering procedure of the Sodas 2.5 software.

(POL, 6.9%), Italy (ITA, 5.0%), Spain (ESP, 4.3%), Denmark (DNK, 3.7%), Hungary (HUN, 2.7%), Sweden (SVE, 1.6%), Belgium (BEL, 1.4%) and 1.1% for Austria (OST) and The Netherlands (NLD), 84.3% of the European production.

Table 12.1 presents estimates of conditional quantiles (the first decile D1, the lower quartile Q1, the median Q2, the upper quartile Q3, the ninth decile D9), with the ordinary least squares estimate (OLS) for wheat production costs based on Surry *et al*. (2012). They are issued from the quantile regression of specific agricultural production costs (the SE281 accounting aggregate in EU-FADN) based on the decomposition of the gross product among 15 speculations for a subset of 12 European countries (EU12) in 2006.

WHEAT	D1 [Inf ; Sup]	Q1 [Inf ; Sup]	Q2 [Inf ; Sup]	Q3 [Inf ; Sup]	D9 [Inf ; Sup]	OLS [Inf ; Sup]
Austria	[125.4 ; 256.2]	[153.8 ; 260.4]	[227.7 ; 346.3]	[262.6 ; 398.8]	[279.0 ; 466.0]	[174.2 ; 279.0]
Belgium	[294.0 ; 412.6]	[348.9 ; 465.7]	[347.6 ; 467.0]	[386.1 ; 581.9]	[510.6 ; 872.0]	[408.0 ; 604.9]
Denmark	[195.5 ; 240.3]	[246.1 ; 286.3]	[331.5 ; 401.3]	[423.0 ; 502.0]	[598.2 ; 769.4]	[406.6 ; 478.1]
Germany	[256.3 ; 318.7]	[292.8 ; 342.0]	[343.4 ; 388.0]	[366.0 ; 435.0]	[427.5 ; 515.9]	[319.6 ; 354.0]
France	[300.2 ; 353.8]	[353.8 ; 397.6]	[426.3 ; 464.7]	[496.3 ; 539.5]	[546.2 ; 622.6]	[419.2 ; 454.1]
Spain	[176.6 ; 258.4]	[156.6 ; 242.6]	[139.0 ; 306.4]	[225.7 ; 391.5]	[442.7 ; 679.1]	[246.6 ; 374.9]
Hungary	[201.1 ; 254.3]	[223.0 ; 278.0]	[289.3 ; 378.7]	[364.6 ; 427.8]	[316.6 ; 479.0]	[352.3 ; 445.6]
Italy	[131.2 ; 200.8]	[201.0 ; 265.6]	[288.3 ; 354.1]	[338.6 ; 419.2]	[394.0 ; 556.4]	[90.7 ; 309.0]
Netherlands	[31.6 ; 431.4]	[34.4 ; 296.2]	[3.3 ; 249.5]	[316.4 ; 595.2]	[455.8 ; 979.8]	[-13.2 ; 640.6]
Poland	[221.5 ; 260.7]	[268.5 ; 298.7]	[334.8 ; 371.8]	[409.8 ; 458.8]	[482.2 ; 557.6]	[380.8 ; 407.8]
Sweden	[250.6 ; 444.8]	[201.3 ; 356.1]	[255.9 ; 407.9]	[319.0 ; 511.2]	[438.9 ; 718.5]	[209.5 ; 336.1]
United Kingdom	[282.0 ; 309.8]	[283.0 ; 318.0]	[322.4 ; 364.2]	[356.2 ; 417.2]	[392.6 ; 481.2]	[283.5 ; 331.6]

Table 12.1. *The specific cost estimates (€) for 1,000 € of wheat gross product, EU12-FADN 2006 (source: FADN-based author's computations)*

The visualization of the relative position of countries is provided by the first factorial plan (Figure 12.1) from the V-PCA of point estimates, accounting for 90.7% of total inertia. Accounting for 64.4% of total inertia, the first principal component F1 is a size axis positively correlated (>0.77) to all conditional estimators, especially to Q2 (0.89) and Q1 (0.90). The second principal component F2, accounting for 26.3% of total inertia, is positively correlated with D9 (0.81) and Q3 (0.52) and negatively with D1 (–0.45) and Q1 (–0.37).

The hierarchical clustering procedure[2] gives two interesting partitions at low level of semipartial R squared, P4 (8%) and P5 (4%), projected on first factorial plan (Figure 12.2): {NLD} with highest D9 estimate, {BEL, FRA} with highest Q2 estimates, {OST, DEU} the lowest D9 estimates, {ESP, ITA, HUN} with Q2 lower values than {DNK, POL, SVE, UKI} characterized by more central estimates.

2 The "proc CLUSTER" procedure of the SAS software.

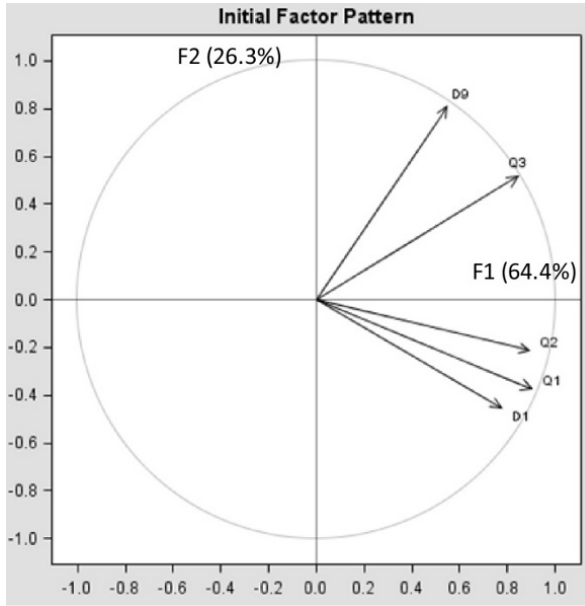


Figure 12.1. *The first factorial plan (F1xF2) of the V-PCA, with the conditional quantile projections (D1, first decile; Q1, lower quartile; Q2, median; Q3, upper quartile; D9, ninth decile) (source: FADN-based author's computations)*

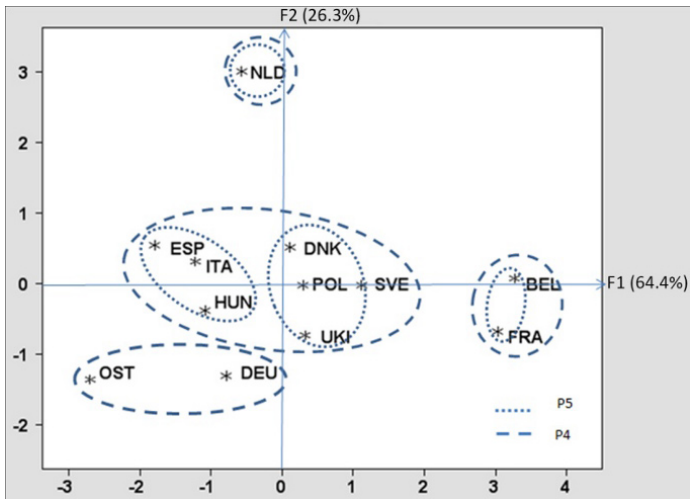


Figure 12.2. *The first factorial plan (F1xF2) of the V-PCA, with the empirical distributions of EU-12 countries (source: FADN-based author's computations)*

However, S-PCA of the interval ranges gives complementary information (see Figure 12.3).

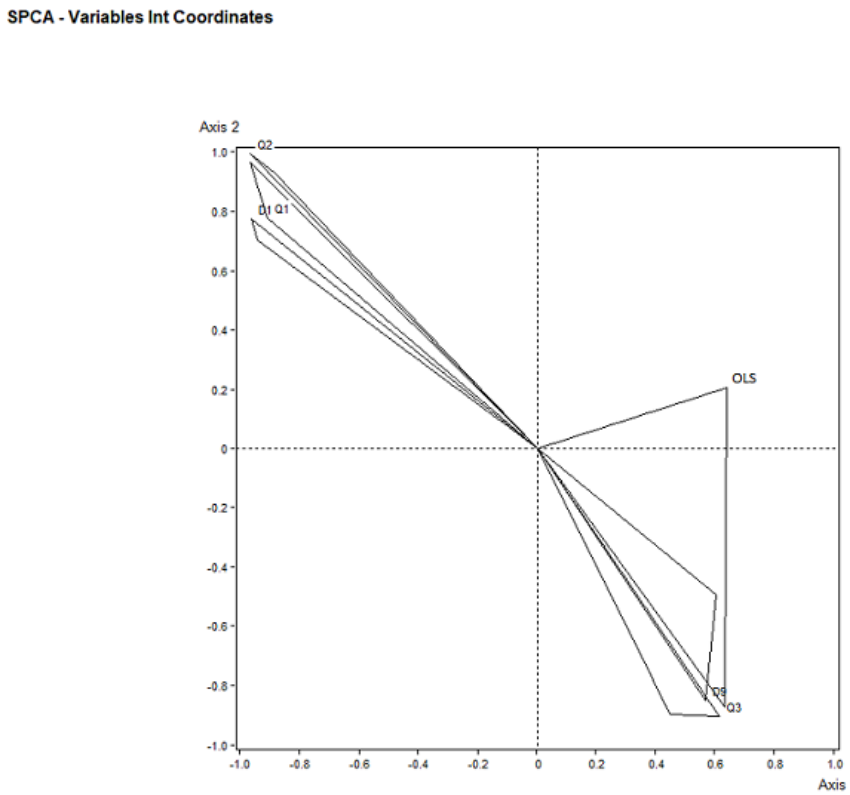


Figure 12.3. *S-PCA of empirical distributions of specific costs, with the projected interval coordinates of quantile estimates (source: FADN-based author's computations)*

For Axis 1 component, only the correlation interval of D9 is strictly positive ([0.260, 0.990]); all other correlations intervals contain origin. Meanwhile, the radii of correlation intervals of D1 and Q1 are the highest ones (0.87 and 0.88, respectively). Hence, examination of Axis 1 correlation projections identifies D9 as an estimator indicating heterogeneity for conditional distributions. Along Axis 2, we have observed very strong correlations between estimate intervals of D1 and Q1 as indicators of skewness by lower estimates, as well as strong correlation between estimate intervals of Q3 and D9 as indicators of skewness by higher estimates. The positive correlation between the OLS vector and the D9 and Q3 vectors along Axis 1 is an indicator of the asymmetry to the right of the distribution due to the greater

dispersion by highest values. The larger width of the OLS projection indicates a greater dispersion of the OLS estimator relative to the conditional quantile estimators.

On the first factorial plane of SOs (Figure 12.4), the rectangle of France (FRA), consisting of the projections of the edges of estimate hyperrectangle parallel to the first two principal components, differs from that of Germany (DEU), which means specific differences in costs both in terms of scale (along Axis 1, the first principal component, whose central dispersion rather reflects the differences between median estimates) and in the shape of specific cost distributions (along Axis 2, the second principal component, opposing the estimates of lower quantiles to those of the upper quantiles).

The distributional ranges may partially overlap on the lowest quantiles as indicated by the positions of the projected hyperrectangles of Austria (OST) and France (FRA). Note also the projection of the Netherlands (NED) hyperrectangle that includes all other projected hyperrectangles, indicating the most heterogeneous distribution, followed by Sweden (SVE)- and Belgium (BEL)-projected rectangles whose lengths according to Axis 1 are among the largest. Hence, S-PCA can be used as a procedure to characterize distributions of specific costs: the Netherlands, Sweden and Belgium are associated to the location-scale shift model, whereas other countries rather belong to the location-shift (homogeneous) model.

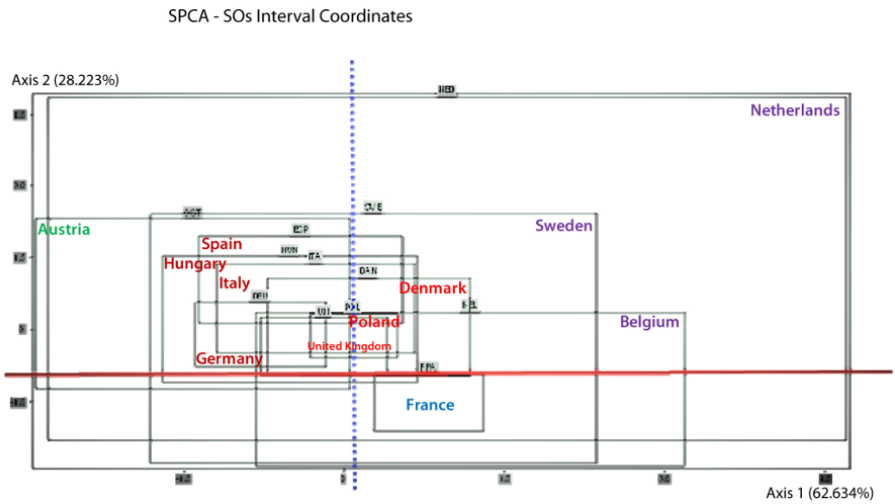


Figure 12.4. *S-PCA of empirical distributions of specific costs, with the projected hyper-rectangles of EU12 countries (source: FADN-based author's computations). For a color version of this figure, see www.iste.co.uk/skiadas/data1.zip*

The divisive clustering procedure (Chavent and Bock 2000), applied to the confidence intervals of the quantile estimates, confirms these differences and similarities indicating the cost structure to the classes of countries (Figure 12.5).

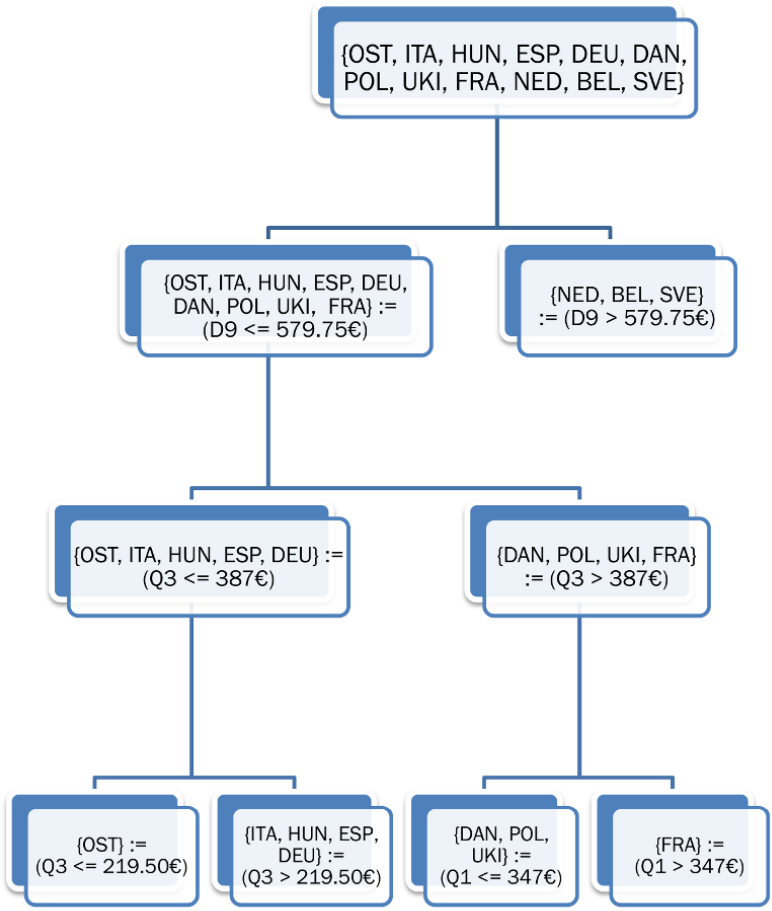


Figure 12.5. *The interval divisive hierarchical clustering of empirical distributions of specific cost, with the EU-12 countries, 2006 (source: FADN-based author's computations). For a color version of this figure, see www.iste.co.uk/skiadas/data1.zip*

First, at the top of the divisive tree, we discuss about "global" ranking between distributional scales: on the one hand, the countries {the Netherlands, Belgium, Sweden} associated with the local-scale model where the conditional quantiles of the highest decile D9 of specific costs exceed €579.75 and where wheat is a

marginal output for most of the farmers; on the other hand, the countries where D9 estimates are less or equal to that value associated with the local-shift model.

Second, those countries associated with the local-shift model, where wheat is a substantial output for most of the farmers, are ordinated between two subgroups divided by the level of the highest quartile (Q3) conditional estimates: {Austria, Italia, Hungary, Spain, Germany} where Q3 estimates are lower or equal to €387; {Denmark, Poland, United Kingdom, France} where estimates are higher than €387, which we denote as the "intermediate" ranking.

Eventually, at the bottom of the tree, the divisive clustering algorithm highlights a partial order authorizing some "local" rankings based on conditional estimates of Q3 and the lowest quartile (Q1): {Austria} with Q3 estimates lower than €219.50, {Italy, Hungary, Spain, Germany} with Q3 estimates higher than €219.50; a second split between {Denmark, Poland, United Kingdom} with Q1 estimates lower than €347 and {France} with Q1 estimates greater than €347.

These dividing values, and the *global*, *intermediate* and *local* rankings they define, can be used in intercountry comparison for the purpose of specific productivity assessment in the context of an increasingly competitive wheat market.

12.6. Conclusion

In this chapter, we proposed a symbolic data analysis approach in order to display the distributions of quantile conditional estimates. We apply this approach to the specific costs of production for wheat in 12 European countries. This case study demonstrates the relevance of our approach in allowing the identification of significant differences particularly between France and Germany, two of the main European producers of wheat.

With regard to location and shape of the conditional quantile estimated distribution, the national differences are mainly based on the highest conditional decile and quartile, hence displaying the asymmetric heterogeneity by the greatest values of the empirical distribution of conditional quantile estimates.

The skewness of conditional quantile distributions for production cost being inherited from the heterogeneity of conditions and techniques in agricultural production among European producers, we propose to extend our symbolic approach to other main agricultural products such as cow's milk and pork on the basis of regional estimates in order to analyze more precisely spatial sources of heterogeneity.

12.7. Acknowledgments

This study has received funding from the 7th Framework Program of the European Community (FP7 / 2007-2013) under the authorization no. 212292.

12.8. References

Bock, H., Diday, E. (2000). *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*. Springer-Verlag, Berlin.

Cameron, A.C., Trivedi, P.K. (2005). *Microeconometrics. Methods and Applications*. Cambridge University Press, Cambridge.

Cazes, P., Chouakria, A., Diday, E., Schektman, Y. (1997). Extension de l'analyse en composantes principales à des données de type intervalle. *Revue de statistique appliquée*, 45(3), 5–24.

Chavent, M., Bock, H.H. (2000). Clustering methods for symbolic objects. In *Analysis of Symbolic Data*, Bock, HH., Diday, E. (eds), 294–341. Springer, Berlin.

De Carvalho, F.A.T. (1997). Clustering of constrained symbolic objects based on dissimilarity functions. *Indo–French Workshop on Symbolic Data Analysis and its Applications*, University of Paris IX.

Desbois, D., Butault, J.-P., Surry, Y. (2015). Distribution des coûts spécifiques de production dans l'agriculture de l'Union européenne : une approche reposant sur la méthode de régression quantile, *9es Journées de Recherches en Sciences Sociales*, December 10th-11th, 2015.

Divay, J.F., Meunier, F. (1980). Deux méthodes de confection du tableau entrées-sorties. *Annales de l'INSEE*, 37, 59–109.

He, X., Hu. F. (2002). Markov chain marginal bootstrap, *Journal of the American Statistical Association*, 97, 783–795.

Koenker, R., Bassett, G. (1978). Regression quantiles. *Econometrica*, 46, 33–50.

Koenker, R., Zhao, Q. (1994). L-estimation for linear heteroscedastic models. *Journal of Nonparametric Statistics*, 3, 223–235.

Lauro, C.N., Palumbo, F. (2000). Principal component analysis of interval data: a symbolic data analysis approach, *Computational Statistics*, 15(1), 73–87.

Surry, Y., Desbois, D., Butault, J.-P. (2012). Quantile Estimation of Specific Costs of Production. FACEPA, D8.2.

13

Frost Prediction in Apple Orchards Based upon Time Series Models

The scope of this work was to evaluate the autoregressive integrated moving average (ARIMA) model as a frost forecast model for South Tyrol in Italy using weather data of the past 20 years that were recorded by 150 weather stations located in this region. Accurate frost forecasting should provide growers with the opportunity to prepare for frost events in order to avoid frost damage. The radiation frost in South Tyrol occurs during the so-called frost period, i.e. in the months of March, April and May during calm nights between sunset and sunrise. In case of a frost event, the farmers should immediately switch on water sprinklers. The ice cover that builds on the trees protects the buds and blossoms from damage. Based on the analysis of time series data, the linear regression (LR) and ARIMA models were compared and evaluated. The best result was achieved by the ARIMA model, with the optimal value of 1.0 for recall in case of forecast of 95% confidence intervals. This means that all frost cases could be correctly predicted. Despite the encouraging results for recall, the rate of false positives with a sensitivity of 21% is too high, such that further investigations are desirable (e.g. testing VARIMA models, which are a multivariate extension of ARIMA models). The graphical illustration of the 95% confidence intervals of the ARIMA model forecast and the linear models forecast should be helpful in frost prediction and could be integrated in the electronic monitoring system that permits forecasting of frost weather phenomena.

13.1. Introduction

Accurate frost forecasting should provide growers in South Tyrol with the opportunity to prepare for frost events in order to avoid frost damage. The higher the level of forecast accuracy, the lower the risk of frost damage. Damage to apple orchards brought by freezing night temperatures can cause high crop yield losses to

Chapter written by Monika A. TOMKOWICZ and Armin O. SCHMITT.

Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining,
First Edition. Edited by Christos H. Skiadas and James R. Bozeman.

the growers. The critical periods for frost damage in apple orchards are the months of spring: March, April and May. The radiation frost occurs during clear nights with little or no wind after sunset and lasts until after sunrise.

Overplant conventional sprinklers are widely used in South Tyrol as effective frost protection method for apple orchards. The ice cover prevents the temperature of the protected plant from falling below the freezing point. Sprinkling must start with the onset of the critical temperature and be maintained until the temperature rises above 0° C. This work describes frost prediction in apple orchards based upon a non-seasonal autoregressive integrated moving average (ARIMA) model and three different linear (LR) models.

The general autoregressive moving average (ARMA) model was described first in 1951 by Peter Whittle in his thesis "Hypothesis testing in time series analysis". The ARIMA model is a generalization of the ARMA model. Nowadays, it is widely used in time series analysis. However, there exists only little literature about frost forecasting with ARIMA.

In Castellanos *et al.* (2009), the authors used the ARIMA model to forecast the minimum monthly absolute temperature and the average monthly minimum temperature following the Box and Jenkins methodology.

Another interesting research project about frost forecasting of minimum temperature in the Alpine area is described by Eccel *et al.* (2008). In this work, a simple LR model, a random forest (RF) model and a neural network (NN) model were compared and evaluated. The results achieved by RF were slightly superior to those of other methods. The LR model for frost forecasting was introduced and implemented by Snyder *et al.* (2005).

13.2. Weather database

The weather database holds data from about 150 weather stations, which have been operating since 1993. The weather stations are distributed in apple orchards at an elevation between 200 and 1,100 m.a.s.l.

Currently, the database continues to receive data every 5 min via radio waves or GPRS. The measurements include atmospheric conditions, air and soil temperature, relative air humidity, soil humidity at a depth of 10, 30 and 50 cm, wind speed and direction, precipitation amounts and the relative humidity at leaf surfaces. Moreover, the database contains information of the geographic coordinates of each station (latitude, longitude and altitude).

The historical climate patterns of the past 20 years stored in the weather database can serve as indicator of the climate for future time points. Based on the measurements of the past years, we calculated the forecast of frost weather phenomena and compared the prediction against the observed temperature in order to evaluate the results.

Measurement	Unit
Wet bulb temp (60 cm)	°C
Dry bulb temp (60 cm)	°C
Rel. air humidity	%
Air temp (2 m)	°C
Wind speed	m/sec
Wind direction	N/S/E/W
Leaf surface humidity	%
Precipitation	mm
Irrigation	ON/OFF
Irrigation	mm
Soil temp (–25 cm)	°C
Min interval air temp (2 m)	°C
Max interval air temp (2 m)	°C
Min interval rel. air humidity	%
Max interval rel. air humidity	%
Max interval wind speed	m/sec
Soil humidity (–10 cm)	%
Soil humidity (–30 cm)	%
Soil humidity (–50 cm)	%

Table 13.1. *Variables that were recorded by the weather stations*

13.3. ARIMA forecast model

ARIMA models are a widely used approach to time series forecasting based on autocorrelations in the data.

13.3.1. *Stationarity and differencing*

ARIMA models as described by Hyndman and Athanasopoulos (2012) require that the time series to which they are applied be stationary. A stationary time series is one whose properties like the mean, variance and autocorrelation do not depend on the time point at which the series is observed. A stationary time series has no predictable pattern in the long term. The time plots show a horizontal pattern with constant variance. A non-stationary time series can be transformed into a stationary one by computing the differences between consecutive observations. This transformation is known as differencing. The first-order differenced series can be written as

$$y'_t = y_t - y_{t-1} \qquad [13.1]$$

If the result of the first-order differencing is still a non-stationary time series, second-order differencing can be applied to obtain a stationary time series (Hyndman and Athanasopoulos 2012):

$$y''_t = y'_t - y'_{t-1} \qquad [13.2]$$

One approach to identify non-stationarity is an autocorrelation function (ACF) plot. The ACF plot shows the autocorrelations, which measure the relationship between y_t and y_{t-k} for k (k = 1, 2, 3….) lags. In case of non-stationarity, the ACF will slowly decrease.

Furthermore, widely used tests are the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test and the augmented Dickey–Fuller (ADF) test.

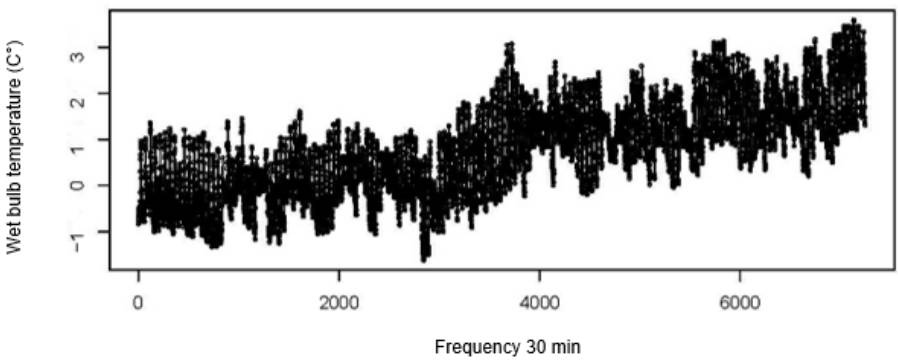


Figure 13.1. *Non-stationary time series*

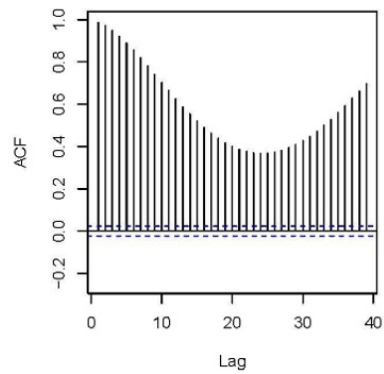


Figure 13.2. *ACF before differencing*

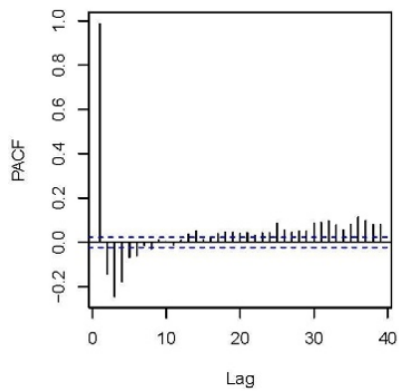


Figure 13.3. *PACF before differencing*

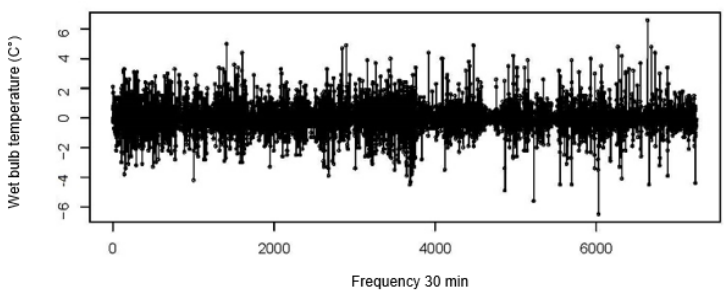


Figure 13.4. *Differenced time series*

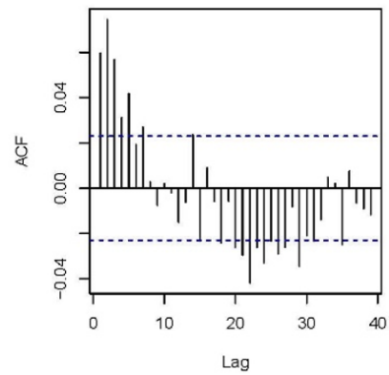


Figure 13.5. *ACF after differencing*

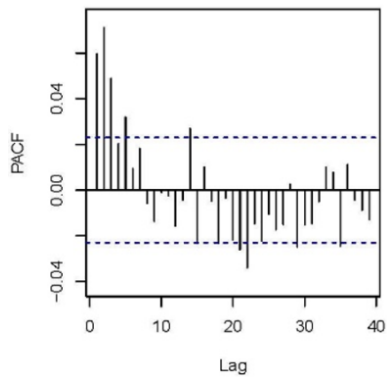


Figure 13.6. *PACF after differencing*

13.3.2. *Non-seasonal ARIMA models*

The non-seasonal ARIMA model described by Hyndman and Athanasopoulos (2012) is a combination of an AR(p) model, differencing and an MA(q) model, and can be written as:

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \cdots + \theta_1 e_{t-1} + \theta_q e_{t-q} + e_t \qquad [13.3]$$

where:

– y'_t is the differenced series;

– e_t is white noise;

– c is a constant;

– p and q are the order of the AR and the MA model, respectively.

A non-seasonal ARIMA model is written as

$$ARIMA(p, d, q) \quad [13.4]$$

where:

– p is the order of the autoregressive part $\mathrm{AR}(p)$;

– d is the degree of first differencing involved;

– q stands for order of the moving average part $\mathrm{MA}(q)$.

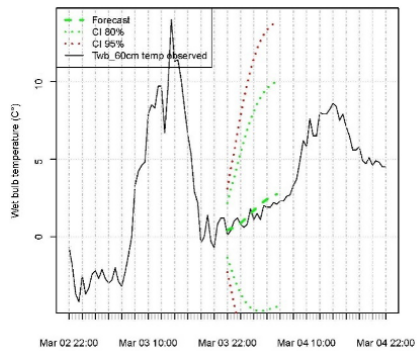


Figure 13.7. *ARIMA(4,1,4) forecast from 22:00 until sunrise. The observed temperature is shown as full line*

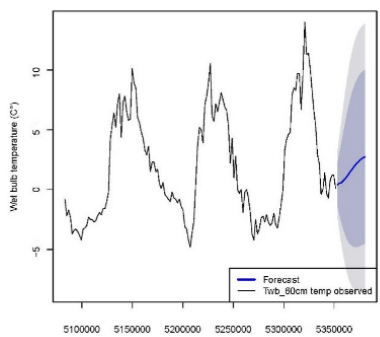


Figure 13.8. *ARIMA(4,1,4) forecast shown with the temperature of the previous three days*

In Figure 13.7, a forecast plot for ARIMA(4,1,4) built on the wet bulb temperature time series is shown (March 4, 2006) for the station 14 in Terlano as a green dashed line together with the observed data and the confidence intervals. In this example, the point forecast suits well the real data. Figure 13.8 shows the forecast plot (blue line) with blue shadowed confidence intervals together with the observed time series data of the previous 3 days.

13.4. Model building

As preliminary step for the analysis, we tested ARIMA models for the wet bulb and dry bulb temperature time series. We ran numerous trials in order to optimize the length of time series to be included in the model and the time frequencies. The results were best when the frequency was observed at 30 min and the length was approximately equal to the length of the frost period. We also compared the results of the manually created ARIMA models following the steps of the procedure described by Hyndman and Athanasopoulos (2012) with those created by the automatic ARIMA function *auto.arima()* from the R package "*forecast*" for "Forecasting Functions for Time Series and Linear Models". Altogether, the results obtained manually and automatically were quite comparable.

For the calculation of the sunrise and sunset time, we used the geographic coordinates of the station from the database.

13.4.1. *ARIMA and LR models*

The scope of the first analysis was to compare automatically modeled ARIMA [13.4] for the dry bulb temperature time series and three LR forecast models, which are variations of the model described by Snyder *et al.* (2005):

$$LRmodel1: TdbSunrise \sim TdbSunset + TdpSunset \quad [13.5]$$

$$LRmodel2: TdbSunrise \sim TdbSunset + RHSunset \quad [13.6]$$

$$LRmodel3: TdbSunrise \sim TdbSunset + TwbSunset \quad [13.7]$$

where:

– $TdbSunrise$ is the dry bulb temperature at 60 cm above ground at sunrise;

– $TdbSunset$ is the dry bulb temperature at 60 cm above ground at sunset;

– $TdpSunset$ stands for the dew point temperature at sunset;

– $TwbSunset$ is the wet bulb temperature at 60 cm above ground at sunset;

– $RHSunset$ is relative humidity at sunset.

For the test, 100 forecasts for each model were calculated and tested on a randomly chosen data set. The point forecast, lower bound of the 80% and the 95% confidence interval were calculated.

13.4.2. *Binary classification of the frost data*

The following test conditions were defined:

– *frost*: positive condition, when the predicted variable dry bulb temperature at 60 cm above ground falls below 0°C at sunrise;

– *no frost*: negative condition, when the predicted variable dry bulb temperature at 60 cm above ground does not fall below 0°C at sunrise.

13.4.3. *Training and test set*

The data set for the forecast was selected in the following manner:

1) choose randomly one station from all stations;

2) choose randomly 1 year from the range of years;

3) choose randomly 1 day for the forecast from the relevant frost period (from March until May).

We randomly selected 47 frost cases and 53 days without frost, altogether 100 days. As a training set, a 30-day time period before the previously randomly selected day of the forecast was chosen. As a test set served the day of the forecast itself. Next, the following two-step procedure was conducted:

1) calculate the ARIMA non-seasonal model on the training set;

2) test the model on the test set.

13.5. Evaluation

In order to assess the quality of a forecast, we considered the following quantities: accuracy, recall and specificity. Accuracy is defined as the ratio of all correctly recognized cases to the total number of test cases. The recall is defined as ratio of true positives to all frost cases. The specificity is the ratio of true negatives to all no frost cases.

On the basis of the recall value for the point forecast, the LR model 3 with recall value equal to 70% could be identified as the best model. The two other LR models and the ARIMA model reached a recall value of about 60%. The specificity values for all models were between 96% and 100%. The accuracy for the point forecast was between 79% (ARIMA) and 85% (LR model 3). The test results for the 80% and 95% confidence intervals for the LR models were quite similar. Their recall values ranged from 68% to 83%, the specificity from 85% to 91% and the accuracy from 79% and 85%. The model 2 was the best in this group for both confidence intervals. For the test level of the 80% and 95% confidence interval lower bounds, the best model was ARIMA, which reached higher values for recall than the LR models. In case of the 95% CI lower bound, the optimal value of 1.0 for the recall was achieved. Unfortunately, the payoff of the good results for recall was a low value for specificity of 20% only, which resulted in a low accuracy of 58%.

	Point forecast					
Model	Precision	NPV	Specificity	Recall	Accuracy	F_2
db_sr ~ db_st + dp	1.00	0.75	1.00	0.62	0.82	0.67
db_sr ~ db_st + RH	1.00	0.75	1.00	0.62	0.82	0.67
db_sr ~ db_st + wb	0.97	0.79	0.98	0.70	0.85	0.74
ARIMA	0.93	0.73	0.96	0.60	0.79	0.64
	80 % CI					
Model	Precision	NPV	Specificity	Recall	Accuracy	F_2
db_sr ~ db_st + dp	0.97	0.76	0.89	0.68	0.79	0.72
db_sr ~ db_st + RH	1.00	0.80	0.91	0.74	0.83	0.78
db_sr ~ db_st + wb	0.95	0.79	0.87	0.74	0.81	0.78
ARIMA	0.66	0.94	0.57	**0.96**	0.75	0.88
	95 % CI					
Model	Precision	NPV	Specificity	Recall	Accuracy	F_2
db_sr ~ db_st + dp	0.95	0.81	0.87	0.77	0.82	0.80
db_sr ~ db_st + RH	0.95	0.85	0.87	0.83	0.85	0.85
db_sr ~ db_st + wb	0.93	0.83	0.85	0.81	0.83	0.83
ARIMA	0.53	1.00	0.21	1.00	0.58	0.85

Table 13.2. *Evaluation results of ARIMA and linear regression models*

13.6. ARIMA model selection

The scope of the second analysis was to study which ARIMA model parameters were selected for the forecast by the automatic model building function *auto.arima*() from the R package "forecast". The forecast was made for 500 selected days. We

chose randomly one station from the range of all stations, 1 year, 1 day from the frost period and calculated the model for the previous 30 days before the randomly selected day.

The results of the trial showed that there was a wide range of possible model parameter sets. The distribution of p and q values is shown in Figures 13.9 and 13.10, respectively. At the same time, it is notable that the p and q values are correlated. The higher is the p, the higher the q value. The correlation between p and q is shown in Figure 13.11. On the other hand, the p and q values are not correlated when d = 0, which confirmed the test for association between paired samples using Pearson's product moment correlation coefficient. The *P*-value for the statistical significance was above the conventional threshold of 0.05, so the correlation is not statistically significant.

	if d=0	if d=1
p	Count	
0	0	2
1	9	199
2	60	44
3	39	87
4	9	29
5	4	18
∑	121	379

Figure 13.9. *Count of p*

	if d=0	if d=1
q	Count	
0	0	11
1	24	71
2	32	182
3	40	61
4	18	46
5	7	8
∑	121	379

Figure 13.10. *Count of q*

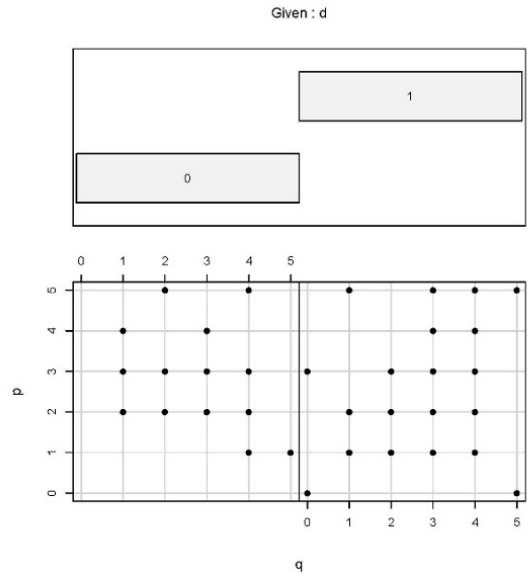


Figure 13.11. *Conditional plot of p versus q at given d*

13.7. Conclusions

This work described frost prediction in apple orchards based upon time series models: a non-seasonal ARIMA model and three different LR models. The model should help in the design of an electronic monitoring system that permits forecasting of frost weather phenomena. Based on analysis of time series data and numerous trials, the proposed models could be compared and evaluated. The following observations regarding temperature forecast for up to 12 h after sunset were made:

– for the test level of the lower bound of the 80% and 95% confidence intervals, the ARIMA model reached higher values for recall (the ratio of correctly recognized frost cases to all frost cases) than the LR models. In case of ARIMA models and the lower bound of the 95% CI, the optimal value of 1.0 for the recall was achieved, which means that all frost cases could be correctly forecast;

– unfortunately, the payoff of the good results for recall is the low value for the specificity of only 20%. This means risk of frequent false alarms.

Despite the high risk of false alarms, the ARIMA model offers encouraging results worth further investigations. LR models can be further improved as well. Here is a list of several complementary analysis steps, which could be tried out toward more accurate forecasting:

– vector ARIMA models (VARIMA), a multivariate extension of ARIMA models, should be tested. The vector of predictors variable in LR models could be extended by wind speed and soil temperature, which would likely lead to more precise forecast;

– the length of the training data can be still optimized in order to find the optimal fit;

– the orders p and q of the ARIMA model should be studied in order to find out potential correlations with temperature. This would allow to exclude some of the models;

– a similarity study of the forecast coming from different stations should be made. Such similarity information could turn out to be helpful in the ARIMA model selection.

13.8. Acknowledgments

The authors would like to thank Professor Johann Gamper of the Free University of Bozen-Bolzano, Faculty of Computer Science, who supported this research project.

13.9. References

Beratungsring. (2012). Leitfaden 2012.

Bootsma, A., Murray, D. (1985). *Freeze Protection Methods for Crops*. Factsheet. Ministry of Agriculture, Food and Rural Affairs, Guelph, ON.

Bowermann, B.L., O'Connell, R.T., Koehler, A.B. *Forecasting, Time Series, and Regression*. Thomson, Brooks/Cole, Pacific Grove, CA.

Castellanos, M.T., Tarquis, A.M., Morató, M.C. and Saa, A. (2009). Forecast of frost days based on monthly temperatures. *Spanish Journal of Agricultural Research*, 7(3), 513–524.

Eccel, E., Ghielmi, L., Granitto, P., Barbiero, R., Grazzini, F., Cesari, D. (2008). Techniche di post-elaborazione di previsione di temperatura minima a confronto per un' area alpina. *Italian Journal of Agrometeorology* 3, 38–44.

Hyndman, R.J., Athanasopoulos, G. (2012). *Forecasting: Principles and Practice*. OTexts, Melbourne.

Oberhofer, H. (1969). Erfahrungen aus den Spätfrösten 1969. Obstbau Weinbau Mitteilungen des Südtiroler Beratungsringes.

Oberhofer, H. (1986). Eine Frostnacht – eine Lehre? Obstbau Weinbau Mitteilungen des Südtiroler Beratungsringes.

Snyder, R.L., de Melo-Abreu, J.P., Matulich, S. (2005). *Frost Protection: Fundamentals, Practice, and Economics*, volume 1. FAO, Rome.

Waldner, W. (1993). Was bewirken die Spätfröste Ende März? Obstbau Weinbau Mitteilungen des Südtiroler Beratungsringes.

14

Efficiency Evaluation of Multiple-Choice Questions and Exams

Multiple-choice questions are common in Israeli institutions of higher education. They can be checked and graded automatically using artificial intelligence methods so that the answer sheets are aligned and segmented automatically into the relevant regions, and then the answers marked by the students are read. In the next step, the grades can be easily calculated by comparing the marked data with the correct answers. To evaluate the efficiency of the exam in addition to the basic statistical analysis of the grades, we propose efficiency measures for each question as well as for the whole exam. These efficiency measures attempt to answer the following questions: how many of the "strong" students have answered a particular question correctly and how many of the "weak" students have failed in a particular question. A question is considered efficient if most "strong" students succeed in it, whereas most "weak" ones fail. In a similar fashion, an exam questionnaire is considered efficient if the majority of its questions are efficient. Our measures can be used both for multiple-choice and numeric answers. We have performed the proposed statistical analysis on the grades of a number of real-life examinations, and our conclusion is that the proposed analysis and efficiency measures are beneficial for the purpose of estimating the quality of the exam and discovering the inefficient questions, that is, the ones that fail to separate the "strong" and the "weak" students.

14.1. Introduction

Multiple-choice questions are a well-known method of examination often used in academic institutions of higher education (see Wood 1977). They are easy to check and can even be graded automatically using scanners or camera-based systems that utilize image processing and computer vision techniques (Gershikov and Kosolapov, 2014a, 2014b; Kosolapov *et al*. 2014). Most of the automatic systems use specially

Chapter written by Evgeny GERSHIKOV and Samuel KOSOLAPOV.

Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining,
First Edition. Edited by Christos H. Skiadas and James R. Bozeman.

tailored optical mark recognition techniques, which are much faster and more reliable than general purpose optical character recognition techniques (Bergeron 1998). The use of machine vision methods for applications, where visual information has to be translated to quantitative data, has accelerated in recent years due to technological advances in the areas of mobile devices and digital photography. Automatic checking of exams also has the advantage of easy statistical analysis of the students' performance in the exam at the global level as well as at the individual question level. This is because during the grading process, all the necessary data for such analysis has already been collected.

Once the grades have been derived, the grades statistics can be analyzed by a number of well-known statistical methods: classical test theory, factor analysis, cluster analysis, item response theory and model analysis (Ding and Beichner 2009). Additionally, after the grades have been derived and analyzed, it is important to compare the performance of the group of students in this particular test or quiz to other groups of students or past examinations and determine the level of knowledge of the students versus the level of knowledge required by the exam. Clearly, a high difficulty level of the questions or a low knowledge level of the examined students may result in the same low performance in the exam. The opposite case is also true: high performance of the examinees due to an easy exam or excellent knowledge of the exam subject demonstrated by the students. To identify these cases, we suggest a different kind of mathematical analysis in addition to the regular statistical analysis of the grades by calculating the average, the standard deviation, the median, the histogram of the grades, the passing/failing percent of students and other similar values.

Our idea is to use efficiency measures for each question. One of these efficiency measures attempts to answer the following question: how many of the "strong" students answered a particular question correctly. Another measure attempts to evaluate the performance of the "weak" students: how many of them failed in a particular question. A question is considered efficient if most "strong" students succeed in it while most "weak" ones fail. In a similar fashion, an exam questionnaire is considered efficient if the majority of its questions are efficient. In the following section, we present our efficiency measures. For best performance, we believe these measures should be calculated iteratively.

14.2. Exam efficiency evaluation

14.2.1. *Efficiency measures and efficiency weighted grades*

Assume that the exams were checked and graded using the regular method of point allocation to the different questions without any other weighting. We first

define $Eff33Gd_{i,0}$, the initial "good" efficiency of an exam question number i, as the ratio between the number of "strong" students that answered this question correctly N_{Gd}^{i} and the number of "strong" students in the whole exam N_{Gd}. The "strong" students are defined as those in the top 33% of the final grades. Thus,

$$Eff33Gd_{i,0} = \frac{N_{Gd}^{i}}{N_{Gd}}. \qquad [14.1]$$

In a similar fashion, we define $Eff33Bd_{i,0}$, the initial "bad" efficiency of an exam question number i, as the ratio between the number of "weak" students that answered this question incorrectly N_{Bd}^{i} and the number of "weak" students in the whole exam N_{Bd}. The "weak" students are defined as those in the bottom 33% of the final grades. Thus,

$$Eff33Bd_{i,0} = \frac{N_{Bd}^{i}}{N_{Bd}}. \qquad [14.2]$$

We can now define the efficiency of question i as either the minimum of the two efficiencies:

$$Eff33_{i,0} = min\left\{Eff33Gd_{i,0}, Eff33Bd_{i,0}\right\} \qquad [14.3]$$

or as the average of the two efficiencies:

$$Eff33_{i,0} = \frac{Eff33Gd_{i,0} + Eff33Bd_{i,0}}{2}. \qquad [14.4]$$

We prefer the second choice as the first option is much more demanding for a question to be considered efficient. This means that the minimum option is more suitable for readers who prefer more of a challenge.

It is reasonable to average the individual question grades by the efficiencies of equations [14.3] or [14.4] so that an efficient question contributes more to the resulting grade than an inefficient one. We now define the efficiency weighted grade of each student as:

$$WGrades_0 = \sum_{i=1}^{N_Q} Eff33_{i,0} \cdot Grade_i. \qquad [14.5]$$

Here, N_Q is the number of questions in the exam and $Grade_i$ is the grade of an individual examinee of question number *i*.

14.2.2. *Iterative execution*

To improve the performance of our efficiency measures, we suggest repeating the process described above, using the following steps in iteration k $(k = 1, 2, 3, ...)$:

1) divide the students into three classes: the "strong" students, the "weak" students and the average ones, where the "strong" ones are the students in the top 33% of the weighted grades $WGrades_k$ and the "weak" ones are the students in the bottom 33% of the weighted grades $WGrades_k$;

2) calculate the efficiencies $Eff33Gd_{i,k}$ and $Eff33Bd_{i,k}$ for question number i using the classes derived in the previous stage and equations [14.1] and [14.2]. The number of students in the "strong" and "weak" classes are substituted for N^i_{Gd} and N^i_{Bd}, respectively;

3) calculate the question efficiencies $Eff33_{i,k}$ using equations [14.3] or [14.4];

4) calculate the weighted grades $WGrades_k$ using equation [14.5].

Repeat the process until the maximal number of iterations is reached (e.g. 50) or the efficiencies converge subject to a certain stop criterion, for example,

$$\frac{1}{N_Q}\sum_{i=1}^{N_Q} |Eff33Gd_{i,k} - Eff33Gd_{i,k-1}| < TH \quad [14.6]$$

and

$$\frac{1}{N_Q}\sum_{i=1}^{N_Q} |Eff33Bd_{i,k} - Eff33Bd_{i,k-1}| < TH. \quad [14.7]$$

TH here is a small threshold, e.g. 0.01.

Finally, the exam efficiency score is calculated as

$$Eff33 = \frac{1}{N_Q}\sum_{i=1}^{N_Q} Eff33_{i,klast}, \quad [14.8]$$

where k_{last} is the last iteration of the algorithm.

14.2.3. *Postprocessing*

The weighted grades $WGrades_{klast}$ can be adjusted to have a distribution on the same scale as the one of the non-weighted grades, given by $Grades = \sum_{i=1}^{N_Q} Grade_i$. There are several options for such an adjustment. Only one of following options should be used:

1) Adjust the weighted grades to have the desired median value $GradesMedian$ that can be chosen to be the same as for the non- weighted grades or any other value, e.g. 50 for grades on the scale of 0–100. This step is done simply by dividing $WGrades_{klast}$ by their current median, multiplying it by $GradesMedian$, and then rounding:

$$WGrades_{klast}^{new} = round\left(\frac{WGrades_{klast} \cdot GradesMedian}{median\left(WGrades_{klast}\right)}\right). \qquad [14.9]$$

An additional step of setting grades above the maximal possible grade (e.g. 100) to that maximal value should be added.

2) Adjust the weighted grades to have the desired maximal value $GradesMax$, which can be chosen to be the same as for non-weighted grades or another value, e.g. 100. This step is done in a similar fashion to the one above. No additional steps are required here if $GradesMax$ is chosen reasonably.

3) Adjust the weighted grades to be in the same scale as the non-weighted grades, e.g. 0–100, by scaling them using the final question efficiencies $Eff33_{i,klast}$. The maximal possible weighted grade that the student can achieve is

$$WGradesMax = \sum_{i=1}^{N_Q} Eff33_{i,klast} \cdot Points_i, \qquad [14.10]$$

where $Points_i$ is the number of points allocated to question number i. Thus, the scaling is done simply by the following formula:

$$WGrades_{klast}^{new} = round\left(\frac{WGrades_{klast} \cdot GradesMax}{WGradesMax}\right). \qquad [14.11]$$

Here, $GradesMax$ is usually chosen to be the maximal value of the regular non-weighted grades, e.g. 100.

In this work, we prefer the third option for postprocessing the grades.

14.3. Real-life experiments and results

We applied the algorithms described in the previous section to a number of real-life examinations performed in an electrical engineering course in an academic college. The results for one of the exams are shown in Figure 14.1. The exam consisted of 10 multiple-choice questions with five possible answers for each. We label this exam as "Exam1". When analyzing the results, the efficiencies were calculated using the average of the "good" and "bad" efficiency values, as given in equation [14.4]. As can be seen in Figure 14.1, the efficiencies range from as low as 0.08 to as high as 0.94 before averaging, and the range is 0.5–0.76 after averaging. Also, there is no significant correlation between the success rate in a certain question and its efficiency score except for the radical cases of a very high success rate (close to 100%) or a very low one (close to 0%). We consider the success rate to be a poor criterion to measure efficiency since it can be the same for questions where the "strong" students succeeded and the "weak" ones failed and for questions with the opposite results.

Based on both good and bad efficiencies, we can classify the questions into categories as shown in Table 14.1. There are 16 categories, similar to the four shown in the table. Another four categories of inefficient questions, which we find interesting, are given in Table 14.2 and allow the definition of very easy and very hard questions differently than just based on the success rate. The proposed efficiencies allow the comparison of two questions with exactly the same success rate, but different performance of "strong" and "weak" students.

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
%Success									
57.97	43.48	56.52	8.70	15.94	94.20	69.57	43.48	28.99	81.16
Eff33%Gd									
0.63	0.54	0.71	0.08	0.17	1.00	0.92	0.54	0.58	0.88
Eff33%Bd									
0.76	0.53	0.53	0.94	0.82	0.24	0.59	0.71	0.82	0.41
Eff33%									
0.70	0.54	0.62	0.51	0.50	0.62	0.76	0.63	0.70	0.65

low efficiency high efficiency

Figure 14.1. *Results for Exam1: success rates and efficiencies for 10 multiple-choice questions. Exam efficiency is 0.62. For a color version of this figure, see www.iste.co.uk/skiadas/data1.zip*

Criterion	"Strong" students category	"Weak" students category	Efficiency category
$0.25 < Eff33Gd_i < 0.5$ and $0.25 < Eff33Bd_i < 0.5$	Inefficient question	Inefficient question	Inefficient
$Eff33Gd_i < 0.25$ and $Eff33Bd_i < 0.25$	Highly inefficient question	Highly inefficient question	Highly inefficient
$0.5 < Eff33Gd_i < 0.75$ and $0.5 < Eff33Bd_i < 0.75$	Normal efficiency question	Normal efficiency question	Efficient
$0.75 < Eff33Gd_i$ and $0.75 < Eff33Bd_i$	Highly inefficient question	Highly inefficient question	Highly efficient

Table 14.1. *Question categories based on good and bad efficiencies*

We can now define an efficient exam as follows:

– the majority of questions were efficient;

– there were no questions with average efficiency below 0.5.

Criterion	Category
$0.75 < Eff33Gd_i$ and $Eff33Bd_i < 0.5$	Easy question
$Eff33Gd_i < 0.5$ and $0.75 < Eff33Bd_i$	Hard question
$0.75 < Eff33Gd_i$ and $Eff33Bd_i < 0.25$	Primitive (very easy) question
$Eff33Gd_i < 0.25$ and $0.75 < Eff33Bd_i$	Challenging (very hard) question

Table 14.2. *Special question categories based on good and bad efficiencies*

The results for another exam, labeled Exam2, are shown in Figure 14.2. Here, the efficiencies range from 0.15 to 1 before averaging and from 0.57 to 0.87 after averaging for the 12 available multiple-choice questions. The exam scored higher on the total efficiency scale, with 0.71 compared to just 0.62 for Exam1. Based on the categories in Tables 14.1 and 14.2, we can classify questions 2, 9 and 10 as highly efficient, questions 3, 6, 8 and 12 as efficient, questions 1, 5 and 11 as easy (which is also supported by the success rate), question 4 as hard (despite the 51% success rate) and question 7 as challenging. There were no inefficient questions in Exam2, as defined in Table 14.1.

There were no primitive questions as well. This is supported also by the overall, relatively high, average efficiency. In Exam1, on the other hand, there was one primitive question (number 6), two challenging questions (number 4 and 5), an easy question (number 10) and one question close to being inefficient (number 2). Checking our criterion for efficiency of the whole exam, we can see that Exam2 is efficient, whereas Exam1 is marginally efficient.

Graphs of the non-weighted grades in ascending order and the corresponding adjusted weighted grades are shown in Figure 14.3. The weighted grades were post-processed using the third method in section 14.2.2. The grades change when using weighting to a small extent, but enough to see different weighted grades for students who answered the same number of questions correctly.

	% success		Eff3%Gd	Eff3%Bd	Eff3	
Q1	80.49		0.92	0.33	0.625	x
Q2	41.46		0.85	0.89	0.87	v
Q3	53.66		0.69	0.78	0.735	
Q4	51.22		0.38	0.89	0.635	x
Q5	78.05		1	0.33	0.665	x
Q6	36.59		0.54	0.78	0.66	
Q7	9.76		0.15	1	0.575	x
Q8	29.27		0.62	0.89	0.755	
Q9	53.66		0.77	0.78	0.755	v
Q10	48.78		0.85	0.78	0.815	v
Q11	78.05		0.85	0.33	0.59	x
Q12	56.1		0.85	0.67	0.76	

Figure 14.2. *Results for Exam2: success rates and efficiencies for 10 multiple-choice questions. Exam efficiency is 0.71*

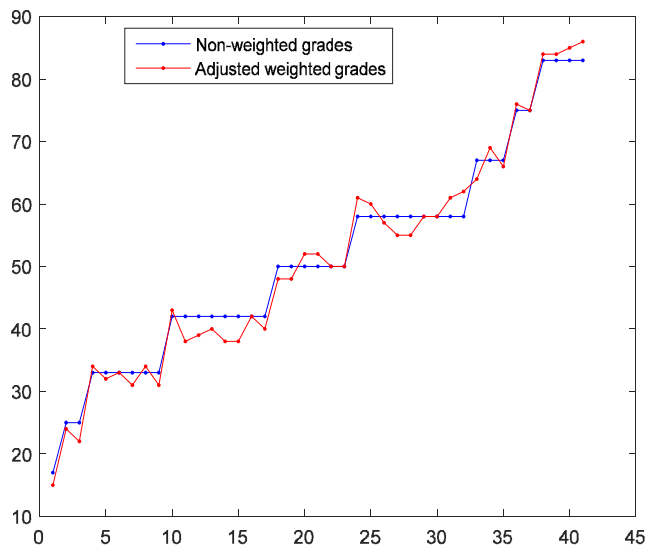


Figure 14.3. *Comparison of the non-weighted grades (plotted in ascending order) and the corresponding adjusted weighted grades for Exam2. For a color version of this figure, see www.iste.co.uk/skiadas/data1.zip*

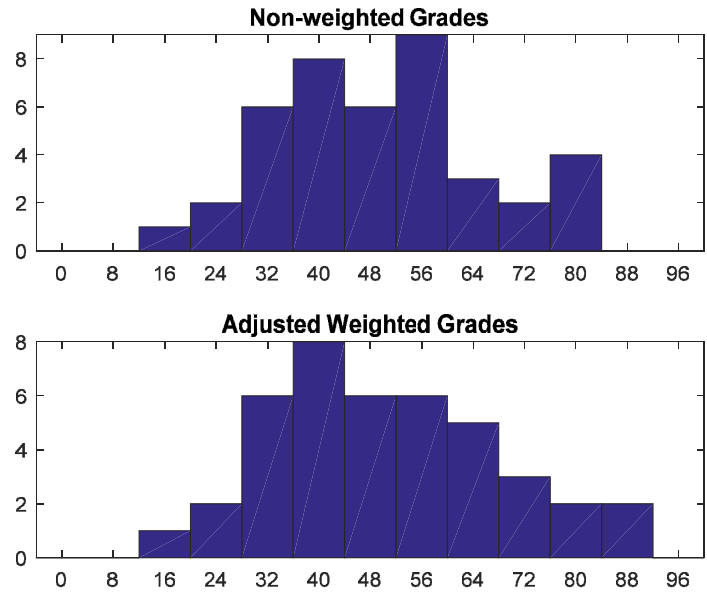


Figure 14.4. *Histograms of the non-weighted and adjusted weighted grades for Exam2. The average was 50.98 and the standard deviation was 18 for the weighted grades, whereas the average was 51.37 and the standard deviation was 17 for the non-weighted grades. The median was 50 in both cases*

A comparison of the histograms of the weighted and non-weighted grades, adjusted for efficiency, is shown in Figure 14.4 for Exam2. The average and median of the grades remain practically the same in both cases: with an average of around 51 and a median of 50 for grades on the scale of 0–100. The weighted grades become less discrete, meaning the differences between adjacent grades on the grade scale are smaller, allowing differentiation between two examinees that replied the same number of questions, but one of whom succeeded in questions that were more difficult for the group of examined students. This student will get a higher-weighted grade than the other one by a few points, allowing the examiner to tell them apart (whether it will affect their actual final grades or not).

14.4. Conclusions

In this chapter, we use statistical analysis of exam grades to evaluate their efficiency. The efficiency measures are proposed at both the individual question level and the exam level. One of these efficiency measures attempts to answer the

following question: how many of the "strong" students have answered a particular question correctly? Another measure attempts to evaluate the performance of the "weak" students: how many of them have failed in a particular question? A question is considered efficient if most "strong" students succeed in it, whereas most "weak" ones fail. In a similar fashion, an exam questionnaire is considered efficient if the majority of its questions are efficient. Our measures can be used both for multiple-choice and numeric answers (where points are granted if the student writes the expected numeric value or one close to it).

We also propose a different method to grade the exams using weighted averaging using the question efficiencies as weight coefficients. This method has the benefit of differentiating between the students that successfully solve more difficult questions and those that solve the easier ones even when the non-weighted grade is the same.

We performed the proposed statistical analysis on the grades for a number of real-life examinations and have presented and discussed the results. Our conclusion is that the proposed analysis and efficiency measures are beneficial for the purpose of estimating the quality of the exam and locating its weakest links: the questions that fail to differentiate between the "strong" and the "weak" students.

14.5. References

Bergeron, B.P. (1998). Optical mark recognition. Tallying information from filled-in "bubbles". *Postgraduate Medicine*, 104(2), 23–25.

Ding, L., Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics - Physics Education Research*, 5, 1–17.

Gershikov, E., Kosolapov, S. (2014a). On image based fast feedback systems for academic evaluation. *International Journal of Signal Processing Systems*, 3(1), 19–24.

Gershikov, E., Kosolapov, S. (2014b). Camera-based instant feedback systems. *International Journal of Advanced Computing*, 47(1), 1463–1473.

Kosolapov, S., Sabag, N., Gershikov, E. (2014). Evaluation of two camera-based approaches for collecting student feedback. *Educational Alternatives*, 12, 1173–1182.

Wood, R. (1977). Multiple choice: A state of the art report. *Evaluation in Education, International Progress*, 1(3), 191–280.

15

Methods of Modeling and Estimation in Mortality

The scope of this chapter is to present several important methods of modeling and estimation of mortality and health. Both subjects have gained significant attention from demographers and other scholars since the 17th Century, when the first life tables were invented. In the 19th and of course 20th Centuries, there were significant advances in the scientific progress in both fields. The Gompertz law (1825) is one example as well as the developments of scientific research concerning population health. A brief discussion of these developments will be carried out in this chapter.

15.1. Introduction

The study of the mortality of a population is not an easy venture. It dates back to the scientific revolution of the 17th Century and is still going on today with new approaches and the refinement of the old ones. Soon after the invention of the first life tables by Graunt (1662) and Halley (1693), two main approaches were developed: the first approach deals with the further improvement of life tables as a scientific instrument and the second one deals with the effort to formulate a general law for mortality in the human species. Thus, there are numerous publications for these efforts.

In this chapter, only the most important of those concerning mortality laws will be included. Thus, in the first part of this chapter a short historical valuation of the life tables will be presented. Afterward, a brief discussion of the most spread methods for the evaluation of mortality laws will be briefly discussed. Among them

Chapter written by Christos H. SKIADAS and Konstantinos N. ZAFEIRIS.

Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining,
First Edition. Edited by Christos H. Skiadas and James R. Bozeman.

are the famous Gompertz model, the Gompertz–Makeham model, the Heligman–Pollard model and others.

In the second part of the chapter, the term health and its relationship with the analysis of mortality will be discussed. As matter of fact, this term became part of the scientific literature in the second half of the 20th Century, when health was defined as "a state of complete physical, mental and social well-being" in the World Health Organization's (WHO) publications (WHO 1948). Afterward, several efforts expanded the work with more refined analytic instruments, including among them stochastic theories. The more important of these approaches will be discussed here.

15.2. The appearance of life tables

It was the middle of the 17th Century when Black Death, the notorious epidemic, hit London once again. During the outbreaks of the plague, thousands of people were killed annually. Therefore, the Bills of Mortality were created, being the weekly mortality statistics for the capital of Great Britain (see Adams 1971). By the year 1662, John Graunt, a draper from Hampshire, "addressed to the Right Honourable Baron Roberts, Lord Privie-Seal, his Natural and Political Observations, based upon Bills of Mortality as they were called" (see Jones 1941). In this book (available at http://www.edste-phan.org/Graunt/bills.html), several attempts were made to answer questions about the numbers of deaths, survival health, sex ratio, family, population age structure and growth, etc. Even from that time Graunt had delineated the framework for the development of modern demography and mortality statistics.

Meanwhile, Edmond Halley was born in London in 1656. He started to study astronomy at a young age and he became famous for the identification of the homonymous Comet, namely Halley's Comet. However, his contribution to demography and mortality analysis is also of great importance. He studied the number of deaths and births of a city known in those days as Breslau of the Hamburg empire, which nowadays is called Wroclaw and belongs to Poland. He published the results of his research in 1693 and he further developed the well-known and universally used life table. His work was greatly accepted during his life and afterward by setting the foundations for a more detailed analysis of mortality in the years after.

In subsequent years, many improvements have been made to the life tables in order for a more effective use of them to be accomplished. A review of these methods can be found in Preston *et al*. (2001, pp. 38–91) and Namboodiri (1990).

15.3. On the law of mortality

Soon after the invention of life tables, many scholars started to work on the mathematization of mortality patterns and the formulation of a law of mortality. Abraham De Moivre in his publication of 1725 assumed a linear change in the survival function. Thus, the force of mortality μ is related to age x using the following formula:

$$\mu_x = \frac{1}{\omega - x}$$

and the probability of survival:

$$xp_0 = \left(1 - \frac{x}{\omega}\right)$$

ω is a maximum age a person can reach in a population; in this case, it was considered to be 86 years.

The study of mortality was a widespread effort over the next few centuries. Babbage (1823), for example, proposed a quadratic formula for the probabilities of survival but the most famous approach was of Gompertz (1825). In his paper, he says about causes of death:

> …the one, chance, without previous disposition to death or deterioration; the other, a deterioration, or an increased inability to withstand destruction. If, for instance, there be a number of diseases to which the young and old were equally liable, and likewise which should be equally destructive whether the patient be young or old, it is evident that the deaths among the young and old by such diseases would be exactly in proportion of the number of young to the old ; provided those numbers were sufficiently great for chance to have its play ; and the intensity of mortality might then be said to be constant ; and were there 130 other diseases but such as those, life of all ages that would be of equal value, and the number of living and dying from a certain number living at a given earlier age, would decrease *in geometrical progression*, as the age increased by equal intervals of time….

Thus, the geometric increase in the force of mortality can be described as:

$$\mu_x = Bc^x,$$

where x is the age and B is the constant. This mathematical formula has been described in different ways as well as the survival curve of a population. According to Preston *et al*. (2001, pp. 192–200), the number of survivors in each age interval can be estimated as:

$$l(x) = Ca^{b^x},$$

where, if y is the age (the original signs have been kept) and n is the width of the age class:

$$b = \left(\frac{ln\left(\frac{l(y+2n)}{l(y+n)}\right)}{ln\left(\frac{l(y+n)}{l(y)}\right)}\right)^{\frac{1}{n}}, a = \left(\frac{ln\left(\frac{l(y+n)}{l(y)}\right)}{b^y(b^n-1)}\right) \text{ and } c = l(y)\exp(-b^y lna).$$

Gompertz himself proposed a second formula (Gompertz 1860), which was further modified into a third one (Gompertz 1862). In any case, Gompertz's model remains a very famous one still in use today. Of the other scholars working in this field, Makeham's approach (Makeham 1890) is also very much in use. Actually, this is considered to be a modification of the Gompertz first law by supposing that besides the geometric increase in the force of mortality, mortality also consists of a constant element, thus

$$\mu_x = A + Bc^x.$$

Later on, Makeham (1890) added another element that increases arithmetically throughout life. Thus, the force of mortality can be given as:

$$\mu_x = A + Hx + Bc^x.$$

Another important contribution to the field is of Weibull (1951), in which the force of mortality is given by:

$$\mu_x = Ax^B,$$

where x is the age and A and B are parameters.

Several other scientists proposed different mortality laws (e.g. Beard 1971) until Heligman and Pollard proposed a very effective model in 1980.

According to the Heligman–Pollard formula (1980, Figure 15.1), the q_x distribution of a life table can be modeled as:

$$\frac{q_x}{p_x} = A^{(x+B)^C} + De^{-E(lnx-lnF)^2} + GH^x \qquad [15.1]$$

where x is the age and A, B, C, D, E, F, G and H are parameters that should be estimated. In fact, the odds of mortality at age x is described by the summation of three components (Figure 15.1). The first component, which includes the parameters A, B and C, represents the fall in mortality during early childhood as the child adapts to its new environment and gains immunity from diseases from the outside world. Of the three parameters, A measures the level of mortality and C measures the rate of mortality decline in childhood. The higher the value of C, the faster is the decline in mortality with increasing age. The parameter B represents the location of infant mortality within the range (q_1, ½). In practice, it is close to zero in modern times.

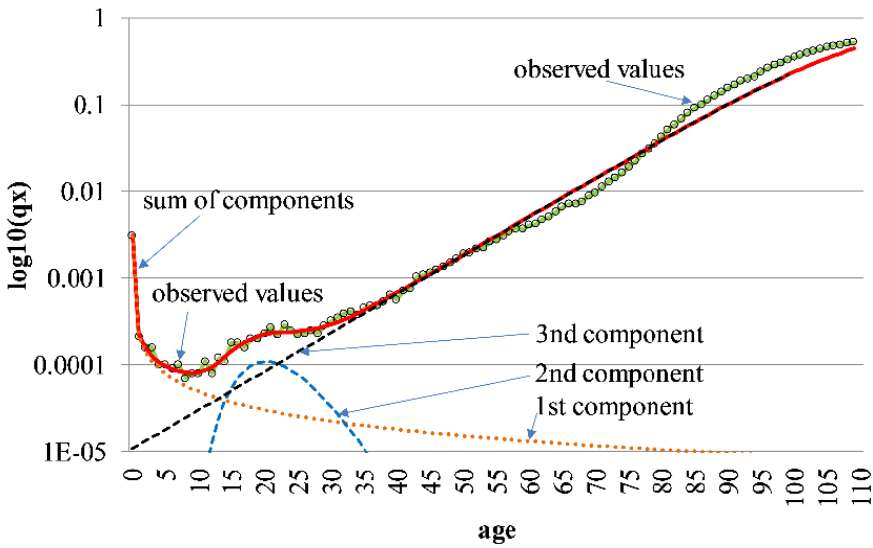


Figure 15.1. *The probabilities of death [log10(qx)] and the Heligman–Pollard formula (Greece, females, 2010–2013) (own calculations; data source: Human mortality database; www.mortality.org). For a color version of this figure, see www.iste.co.uk/skiadas/data1.zip*

The second component includes the parameters D, E and F. It describes the accident hump between ages 10 and 40, which appears either as a distinct hump in the mortality curve or at least as a flattening out of the mortality rates. Of the three parameters, F indicates the location of the accident hump, E represents its spread and D indicates its severity.

The third component corresponds to a Gompertz exponential that represents the aging or the deterioration of the body. G is the base level of senescent morality and H is the rate of increase in that mortality.

One of the advantages of such a model is that it can describe, relatively accurately, the changes in mortality by age and also, as seen before, all of its parameters have a demographic interpretation, thus enhancing its applicability in demographic analysis and evaluation. However, among the disadvantages of the model is the relatively high number of the parameters used, not to mention some deviations that may occur when applying the model on empirical data.

The first of these deviations is connected with the spread of the mortality hub because of a systematic error of the fit of the eight parameters Heligman–Pollard formula on adult mortality, which has been addressed by Kostaki (1992). According to her interpretation, a better fit of the model is achieved if:

$$\frac{q_x}{p_x} = \begin{cases} A^{(x+B)^C} + De^{-E_1(lnx-lnF)^2} + GH^x, & \text{for } x \le F \\ A^{(x+B)^C} + De^{-E_3(lnx-lnF)^2} + GH^x, & \text{for } x > F \end{cases}. \quad [15.2]$$

The only difference between the formulas [15.1] and [15.2] deals with the parameter E, which has been replaced with the relevant E_1 and E_2, representing the spread of the accident hump to the left and right of its top (its location denoted by the parameter F), respectively.

Siler (1979, 1983) extended once more the Gompertz model by proposing a new one, including a negative Gompertz function and the Makeham–Gompertz model. This model has three components, one for the prematurity period, which includes a novel exponentially decreased hazard, a constant hazard, which is dominant during the period of maturity, and the conventional Gompertz hazard, which is dominant during senescence.

Delaportas *et al*. (2001) used a form of parametric model of Heligman–Pollard and afterward they adopted a Bayesian analysis by applying a Monte Carlo simulation for the further processing of the results. Sharrow *et al*. (2013) also used Bayesian methods along with the Heligman–Pollard model in order to investigate the age-specific changes in the mortality of a prospectively monitored rural population in South Africa. Zafeiris and Kostaki (2017) used the Heligman–Pollard model along with three subsequent cubic splines in order to study the mortality trends in Greece (1961–2014).

Also, Rocha *et al*. (2013) propose an additional model on the basis of a mixing of a half normal and a skew bimodal normal distribution. Bongaarts (2005) uses a logistic model for evaluating the future mortality trends of mortality. In other methods, Ouellette and Bourbeau (2011) use P-splines to refine monitoring that has occurred in age at death distribution, which in fact is a non-parametric fitting approach. In any case, both p-splines and b-splines have been widely used for smoothing mostly the probability of death distributions (see Wood 2016).

15.4. Mortality and health

Life table analysis has long been (Graunt 1676; Halley 1692) used to estimate probabilities of survival and life expectancy in different ages during the course of human life. However, if applied in its classical form – as a technique aimed to describe the patterns of mortality and survival – it fails to give an overall picture of the total health status of a population and its changes with age and time, especially, if health is positively defined not simply as the absence of a disease or infirmity but in a broader way as "a state of complete physical, mental and social well-being" (WHO 1948). Several solutions have been given to this problem.

In order to mathematically describe the state of health of a population, Chiang (1965) introduced the term "Index of health H_x" based on the probability distribution of the number and the duration of illness and the time lost due to death (time of death) calculated from data from the Canadian Sickness Survey (1950–1951). Sanders (1964) used life table techniques to construct tables of "*effective life years*" as a measure of the current health of the population based on mortality and morbidity rates. In that case, morbidity was measured by the functional adequacy of an individual to fulfill the role, which a healthy member of his age and sex is expected to fulfill in his society.

Sullivan (1966) criticized these approaches on the grounds of the methodology used and its effectiveness in measuring the health status of a population and later (1971) he used life table techniques to calculate two related indices: the expectation of life free of disability and the expectation of disability based on published data of the current abridged life tables and surveys conducted by the National Center for Health Statistics.

Torrance (1976) developed a health status index model for the determination of the amount of health improvement created by a health care program, calculating a point-in-time health index (H_t), a period-of-time health index ($H_{t1,t2}$) and a health life expectancy index (E_x). The first measures the instantaneous health of all individuals in the population at a single point in time and averages these to give a group's index. The second is similar but it refers to a period of time, that is, a year. The third is calculated by a method that uses the actual mortality and morbidity experience of the population to determine a table of age- and sex-specific health expectancy figures.

Recently, another method was developed by the WHO. This method combines mortality data from several sources with the findings of the Global Burden of Disease Study (see Murray *et al*. 2012, 2015), including self-reporting data on health and disability.

According to WHO[1], the Global Burden of Disease Study measures the "burden of disease using the disability-adjusted-life-year (DALY). This time-based measure combines years of life lost due to premature mortality and years of life lost due to time lived in states of less than full health". However, several limitations of the method, besides its extremely high complexity (see Das and Smarasekera 2013) and the fact that it cannot be used by any other than those who carried it out, include the lack of reliable data on mortality and morbidity for several countries. Also, there is a problem with the comparability of the self-reported data from health interviews and the measurement of health-state preferences for such self-reporting[1]. Thus, the uncertainty of the findings of this method must be considered significant. This is obvious in Salomon *et al*. (2012) and Murray *et al*. (2015).

In the method used by the WHO, the years lost due to disability (YLD) are estimated across a comprehensive set of disease and injury causes (see Vos *et al*. 2012; WHO 2013). Then, the per capita fraction of YLD for all causes is calculated for every age group, sex and country, after adjusting for independent comorbidity. Based on that fraction, the lost years of healthy life are calculated for each age group and the healthy life expectancy at age x is the sum of healthy life years from the age x up to the open-ended interval of the life table divided by the survivors in each age x (WHO 2013, 2014).

Meanwhile, Jansen and Skiadas (1995) used the general theory of dynamic models in order to model human life. By definition, a first-hitting-time model or a first exit time model in stochastic analysis has two basic components: a parent stochastic process and a boundary, barrier or threshold, which indicates a stopping condition for the process under consideration (see Lee and Whitmore 2006).

In the case of mortality or in the survival context, the parent stochastic process is human health that follows a stochastic and thus unpredictable path for each individual, and the barrier is denoted by a person's death that occurs when health status falls below that barrier. In this context, even if a person's health is almost totally unpredictable, as the mortality curve provided from the death and population data can be modeled with simple or more sophisticated models, the same can be done with health.

Then, the problem is to search for the health state curve of a population by extracting the main health state information from the death and population data sets (see Skiadas and Skiadas 2015a). Although the development and detailed presentation of the relevant theory and the necessary mathematical calculations are beyond the scope of this paper and can be found in detail in Skiadas and Skiadas (2010, 2012, 2014, 2015a, 2015b, 2017, 2018) and Jansen and Skiadas (1995), it can

1 See http://www.who.int/topics/global_burden_of_disease/en/.

be proven that the death probability density function $g(x)$ of a life table can be modeled as a function of age (x) and of the health state $H(x)$. The *death probability density function g(x)* is given by:

$$g(x) = k(x)^{-\frac{3}{2}} e^{-\frac{H_x^2}{2x}}$$

and the health state function H(x) is given by:

$$H_x = a_1 + ax^4 - b\sqrt{x} + lx^2 - cx^3$$

where k, a_1, a, b, l and c are parameters that should be estimated for each of the populations studied (see Jansen and Skiadas 1995).

The health state function describes the average levels of health by age in a population (see Figure 15.2). Health increases quite rapidly until puberty when a disturbance is observed because of a transient excess of mortality especially in males. Later on, health is still improving until the age of maximum health, when the degradation of the vitality of the organisms gradually leads health to zero values at the older ages.

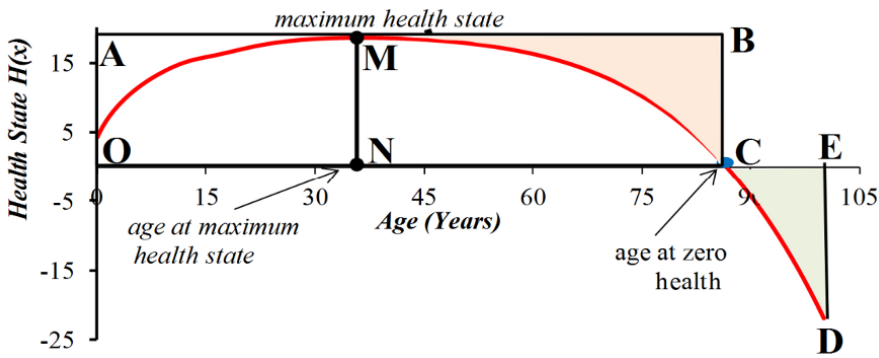


Figure 15.2. *The health state function and some health indicators. For a color version of this figure, see www.iste.co.uk/skiadas/data1.zip*

Several health indicators can be estimated on the basis of health state function (Figure 15.2; see Skiadas and Skiadas 2010, 2012, 2014, 2015a, 2015b). The *total health state* (THS) represents the area covered by the line OMC and thus the overall health attained by the population. The *maximum health state* (Hmax) denotes the maximum value of health attained by the population (point M) in the *age at maximum health state* (point N). The *age at zero health* will be another indicator (point C).

Also, the *healthy life expectancy under moderate and severe disabilities* can be estimated (*HLEB3*). In fact, the *health state function* (Figure 15.2) has an increasing stage of health during the human life cycle, which denoted by the rectangle AMNO (Skiadas and Skiadas 2012, pp. 69–92). The point M corresponds to the maximum vitality of the organisms. The white area within the rectangle MBCN represents the deterioration phase of human health until its zero point.

If no-deterioration mechanism was present, or the repairing mechanism of the human body was perfect during that phase, then the health state would continue following the straight line AMB, which is parallel to the x axis. This is not the case of course and that leads to the gradual disruption of human health. The problem is how to estimate the "lost healthy life years" *LHLY1* during the deterioration phase of the human life cycle. If THD_{ideal} is the ideal total dynamics of the population, a solution can be given as:

$$LHLY1 = \lambda \frac{OABC}{THD_{ideal}} \frac{THD_{ideal}}{MBCM} = \lambda \frac{OABC}{MBCM},$$

where λ is a parameter expressing years and should be estimated for every case and MBCM the gray area of the rectangle MNCB. It was found that for purposes of multiple comparison of countries, λ could be set to be 1 year.

However, the above formula has to be expanded further if the health state of the people living beyond the age at zero health is taken into consideration. In fact, these people contributed to the health state of the population and for the sake of visualization they are represented in the area ECD. Then the equation above can be expanded in order for a new estimation to be named *LHLY3* to be calculated as:

$$LHLY3 = \lambda \frac{OABC + ECD}{MBCM}.$$

Based on the last equation, the *healthy life expectancy HLEB3* can be simply calculated as: *LEB – LHLY3*.

Recently, Skiadas and Skiadas (2016) have proposed another, yet more parsimonious, method. The procedure is briefly described below.

If μ_x is the force of mortality in age x, then it is given as:

$$\mu_x = \left(\frac{x}{T}\right)^b,$$

where T is the age at which $\mu_x = 1$ and b is a parameter expressing the curvature of μ_x.

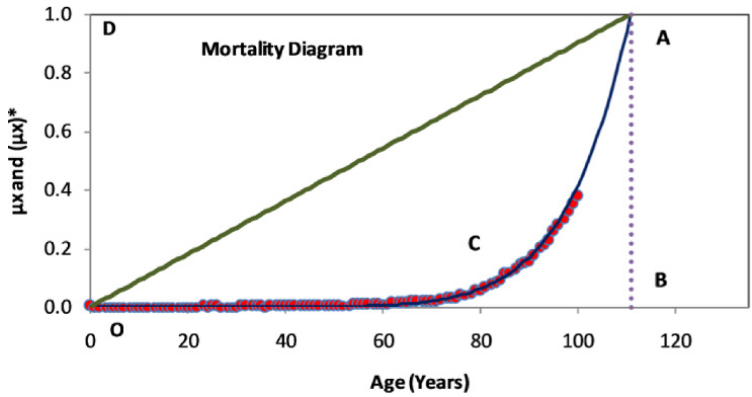


Figure 15.3. *The mortality diagram used in the μ_x-based method. For a color version of this figure, see www.iste.co.uk/skiadas/data1.zip*

The main task is to calculate the healthy life years as a fraction of surfaces in a mortality diagram (see Figure 15.3). This idea, which originates from the first exit time theory and the health state function approach (See Skiadas 2012), is to estimate the area E_x under the curve OCABO:

$$E_x = \int_0^T \left(\frac{x}{T}\right)^b d_x = \frac{T}{(b+1)}\left(\frac{x}{T}\right)^b,$$

where d_x represents the life table's death distribution. The resulting value for E_x in the interval [0, T] is given by

$$E_{mortality} = \frac{T}{(b+1)}.$$

It is also clear that the total area E_{total} for the healthy and mortality part of the life is the area included in the rectangle of length T and height 1, thus $E_{Total} = T$. Then, the healthy area is given by

$$E_{healthy} = T - E_{mortality} = T - \frac{T}{(b+1)} = \frac{bT}{(b+1)}.$$

Obviously:

$$\frac{E_{health}}{E_{mortality}} = b$$

and

$$\frac{E_{total}}{E_{mortality}} = b + 1.$$

These two indicators can describe the health status of the population, the second one being compatible with the severe and moderate causes indicator of the health state approach and thus it can be used as an estimator of the loss of healthy life years (LHLY) in the form of:

$$LHLY = \lambda\,(b + 1)$$

where λ is a correction multiplier, which for multiple comparisons can be set to be 1 year. In this way, similar results to the WHO approach are found.

Four ways for the estimation of b have been developed. In the direct estimation, without applying any model, the calculations can be made either on the mortality (m_x) curve or the probability of death curve (q_x). Then we have the m_x curve:

$$b + 1 = \frac{E_{total}}{E_{mortality}} = \frac{xm_x}{\sum_0^x m_x}$$

and

$$b = \frac{E_{health}}{E_{mortality}} = \frac{xm_x - \sum_0^x m_x}{\sum_0^x m_x} = \frac{xm_x}{\sum_0^x m_x} - 1.$$

Concerning the q_x curve, we have:

$$b + 1 = \frac{E_{total}}{E_{mortality}} = \frac{xq_x}{\sum_0^x q_x}$$

$$b = \frac{E_{health}}{E_{mortality}} = \frac{xq_x}{\sum_0^x q_x} - 1.$$

Afterward, a Gompertz model is applied on the probability density function as

$$f_x = e^{-k+bx-e^{-l+bx}},$$

where x is the age and the other letters on the right of the equation above (except e) are parameters. The parameter expressing the LHLY is l. This is also demonstrated by observing the cumulative distribution function of the form:

$$F_x = e^{-e^{-l+bx}}$$

the relevant survival function is:

$$S_x = 1 - e^{-e^{-l+bx}}$$

the probability density function is:

$$f_x = be^{-l+bx} - e^{-e^{-l+bx}}$$

and the hazard function is:

$$h(x) = \frac{f_x}{F_x} = e^{-k+bx}$$

Finally, the Weibull model is used. This model has a probability density function (b and T are parameters) given as

$$f_x = \frac{b}{T}\left(\frac{x}{T}\right)^{b-1} e^{\left(\frac{x}{T}\right)^b}$$

the hazard function is:

$$h_x = \frac{b}{T}\left(\frac{x}{T}\right)^{b-1}$$

and the cumulative hazard is given by:

$$H_x = \left(\frac{x}{T}\right)^b$$

which is precisely the form for the single mode presented earlier. The parameter b expresses the healthy life years lost.

Obviously, the four ways developed for the calculation of healthy years lost give different results. Thus, it is chosen to present the results of the analysis as the average of the four methods.

15.5. An advanced health state function form

The improvement of the health state methodology produced an advanced form for the health state presented in Figure 15.4 (related references appear in Skiadas and Skiadas 2010, 2017, 2018). The first approach is the estimation of the health state based on a simple model of the form $H(x) = 1 - (bx)^c$, which is presented by the dashed curve and then adding a corrected health state form (continuous line) to account for the gradual improvement of the health state from birth until a maximum

level at years 12–16 and then a small decline especially for male followed by a local maximum at years 30–35 and then the gradual decline. This form is analogous to that presented in Figure 15.3 with the improvement in the area of the maximum health state. The maximum mean health state is assumed to be at unity. However, as the first part of the life span is mainly a development stage for the organism and thus the important part of the deterioration process is mainly governed by the term Det = $(bx)^c$, the health state and the deterioration state can be expressed by the simple model forms providing a convenient method to estimate several important issues for estimating the health state at the retirement age or to use the deterioration function as a tool to reproduce the death distribution. The health state function is simply estimated when the death distribution function $g(x)$ is provided. The mathematical form is given by

$$g(x) = \frac{|H_x - xH'_x|}{\sigma\sqrt{2\pi x^3}}\, e^{-\frac{H_x^2}{2\sigma^2 x}}$$

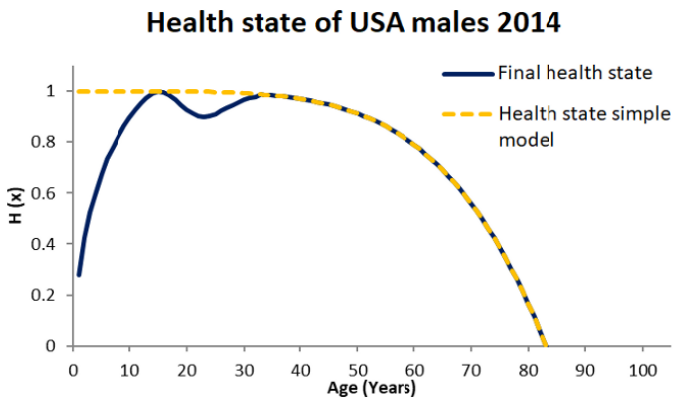
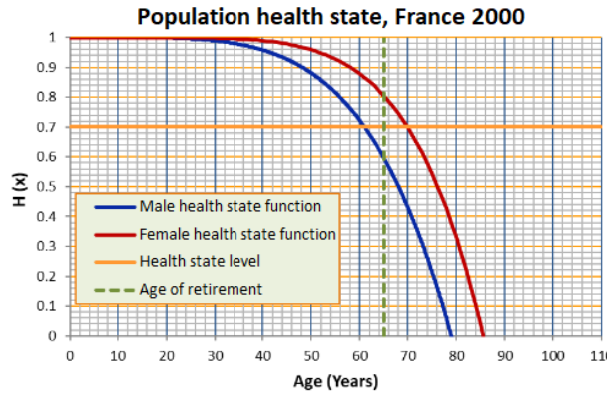


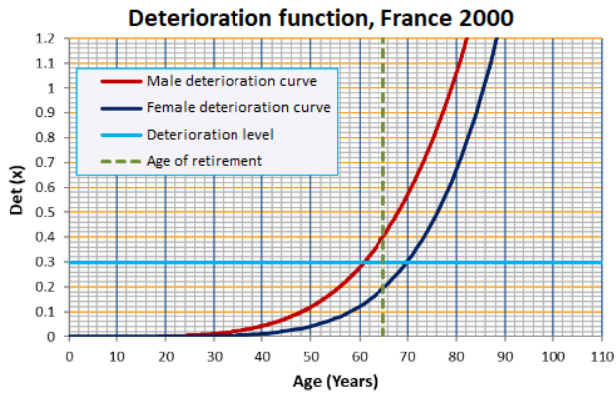
Figure 15.4. *Final health state function and the supporting curve of the simple model. For a color version of this figure, see www.iste.co.uk/skiadas/data1.zip*

The simple health state form is presented in Figure 15.5(A) for male and female in France in the year 2000. The graph is provided in the form that is convenient to find the health state at the retirement age. We set as the retirement health level the 70% of the health state while the retirement age is set at 65 years of age. There are significant differences between female and male regarding the health state at the retirement age of 65 years. The health state level at the retirement age is higher for female than male. The same results are provided in Figure 15.5(B) where the deterioration function is used to find the deterioration level for male and female at the retirement age of 65 years. In this case, the deterioration for male is larger than female by means that female have higher health level than male for the same retirement age.

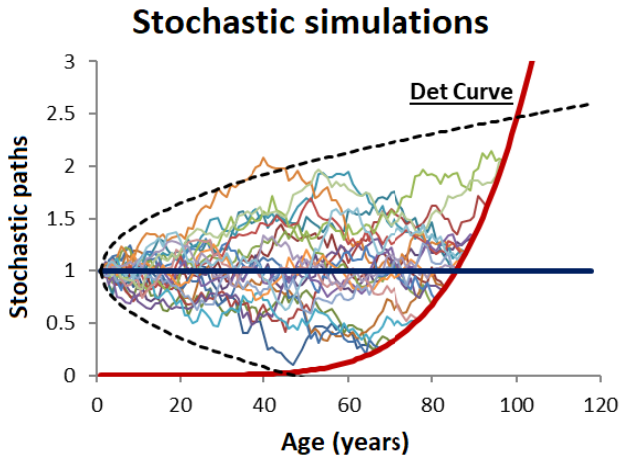
Figure 15.5(C) presents how to use the deterioration curve as the stop barrier for stochastic simulations expressing the health state for individuals. The mean health state level is set at 1 and the health state of an individual is presented by the stochastic paths represented by light lines expressing the standard Wiener process. The end of the life span occurs when the stochastic path finds the deterioration curve. Then we count the age at this incident and collect all the cases to reconstruct the death probability density function $g(x)$. As it is the usual practice, the confidence intervals are presented by the two dashed curves. Few stochastic realizations are illustrated in the graph. In the example presented, 1,200,000 simulations are done providing the graph shown in Figure 15.5(D). The estimated value of R^2 is 0.997, a very satisfactory estimation level.



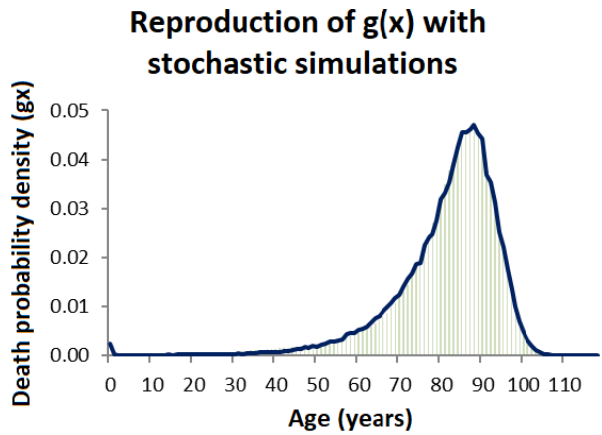
A)



B)



C)



D)

Figure 15.5. *(A) Population health state, (B) deterioration function, (C) stochastic simulations and (D) reproduction of death probability density with stochastic simulations. For a color version of this figure, see www.iste.co.uk/skiadas/data1.zip*

15.6. Epilogue

For centuries, both mortality and health have been the main topic of focus in the scientific community. The developments on both fields must be considered important and in fact the ongoing research will serve positively in both fields. Recently, another very important aspect of mortality has been studied. This deals with the study of very old people and because of the gradual increase in longevity it

is of great importance (see Allard *et al*. 1998; Coles 2004). Thus, as happens with all the scientific disciplines, new knowledge poses new questions all the time.

15.7. References

Adams, R.H. (1971). *The Parish Clerks of London*. London, Phillimore.

Allard, M., Lebre, V., Robine, J., Calment, J. (1998). *Jeanne Calment: From Van Gogh's Time to Ours: 122 Extraordinary Years*. New York: W.H. Freeman & Co.

Babbage, C. (1823). On the tables of single and annual assurance premiums. *J Instit Actuar*. 6, 185.

Beard, R.E. (1971). Some aspects of theories of mortality, cause of death analysis, forecasting and stochastic processes. *In Biological Aspects of Demography*, W. Brass (ed.) London: Taylor & Francis.

Bongaarts, J. (2005). Long-range trends in adult mortality: Models and projection methods. *Demography* 42(1), 23–49.

Chiang, C. L. (1965). *An Index of Health: Mathematical Models*. U.S. Department of HEW, Series 2, No. 5, Public Health Service, Publication No. lCXK.

Coles, L. S. (2004). Demographics of human supercentenarians and the implications for longevity medicine. *Ann NY Acad Sci.* 1019, 490–495.

Das, P. and Samarasekera, U. (2013). The story of GBD 2010: a "super-human" effort. *Lancet* 380, 2067-2070.

de Moivre, A. (1725). *Annuities upon Lives*. For the 1731 version, see https://books.google.gr/books?id=D4ziaz96PYcC&hl=el&source=gbs_similarbooks.

Dellaportas, P., Smith, A. F. M., Stavropoulos, P. (2001). Bayesian analysis of mortality data. *J R Statist Soc A*. 164(2), 275–291.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philos Trans R Soc A*. 115, 513–583.

Gompertz, B. (1860). On one uniform law of mortality from birth to extreme old age and on the law of sickness, presented to International statistical congress 1860 and reproduced in 1871. *J Inst Actuar.* 16, 329–344.

Gompertz, B. (1862). A supplement to two papers published in the Transactions of the Royal Society, "On the science connected with human morality" the one published in 1820 and the other in 1825. *Philos Trans R Soc A* 152, 511–559.

Graunt, J. (1665). *Natural and Political Observations Mentioned in a Following Index and Made upon the Bills of Mortality*, 3rd edition. London: Springer.

Halley, E. (1693). An estimate of the degrees of the mortality of mankind, drawn from curious tables of the births and funerals at the city of Breslaw; with an attempt to ascertain the price of annuities upon lives. *Philos Trans R Soc.* 17, 596–610.

Heligman, L. and Pollard, J. H. (1980). The age pattern of mortality. *J Inst Actuar.* 107, 47–80.

Jansen, J. and Skiadas, C. H. (1995). Dynamic modelling of life-table data, *Appl Stoch Models Data Anal*. 11(1), 35–49.

Jones, H. W. (1945). John Graunt and his bills of mortality. *Bull Med LibrAssoc*. 33(1), 3–4.

Kostaki, A. (1992) A nine parameter version of the Heligman–Pollard formula. *Math Popul Stud.* 3(4): 277–288.

Lee, M.-L T., Whitmore, G.A. (2006). Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. *Statist Sci*. 21(4), 501–513.

Makeham, W.M. (1867). On the law of mortality. *J Inst Actuar*. 8, 301–310.

Makeham, W.M. (1890). On the further development of Gompertz's law. *J Inst Actuar*. 28, 152–159, 185–192, 316–332.

Murray, C.J.L., Barber, R.M., *et al*. (2015). Global, regional, and national disability-adjusted life years (DALYs) for 306 diseases and injuries and healthy life expectancy (HALE) for 188 countries, 1990–2013: Quantifying the epidemiological transition. *Lancet* 386, 2145–2191.

Murray, C.J.L., Ezzati, M., Flaxman, A. D., Lim, S., Lozano, R., Michaud, C., Naghavi, M. Salomon, J.A., Shibuya, K., Vos, T, Wikler, D., Lopez, A.D. (2012). GBD 2010: design, definitions, and metrics. *Lancet* 380, 2063–2066.

Namboodiri, K. (1990). *Demographic Analysis: A Stochastic Approach*. Bingley: Emerald Group Publishing Limited.

Ouellette, N., Bourbeau, R. (2011). Changes in the age-at-death distribution in four low mortality countries: A non-parametric approach. *Demogr Res.* 25: 19, 595-628.

Preston, H., Heuveline P., Guillot, M. (2001). *Demography. Measuring and Modeling Population Processes*. Oxford: Blackwell Publishers.

Rocha, G.H., Loschi, R.H. and Arellano-Valle, R.B. (2013). Inference in flexible families of distributions with normal kernel. *Statistics* 47(6), 1184–1206.

Salomon J.A., Wang H., Freeman, M.K., Vos, T., Flaxman, A.D., Lopez, A.D., Murray, C.J. (2012). Healthy life expectancy for 187 countries, 1990-2010: a systematic analysis for the Global Burden Disease Study. *Lancet* 380, 2144–2162.

Sanders, B.S. (1964). Measuring Community Health Levels. *Am J Public Health* 54, 1063-1070.

Sharrow, D.J., Clark, S.J., Collinson, M.A., Kahn, K., Tollman S.M. (2013). The age pattern of increases in mortality affected by HIV: Bayesian fit of the Heligman-Pollard model to data from the agincourt hdss field site in rural northeast south Africa. *Demogr Res*. 29, 1039–1096.

Siler, W. (1979). A competing-risk model for animal mortality. *Ecology*, 60(4), 750–757.

Siler, W. (1983). Parameters of mortality in human populations with widely varying life spans. *Statist Med*. 2, 373–380.

Skiadas, C., Skiadas, C. H. (2010). Development, simulation and application of first exit time densities to life table data. *Commun Statist.* 39(3), 444–451.

Skiadas, C.H., Skiadas, C. (2012). Estimating the healthy life expectancy from the health state function of a population in connection to the life expectancy at birth. In *The Health State Function of a Population*, 1st edition. Skiadas, C. H., Skiadas, C. (eds.) Athens: ISAST.

Skiadas, C.H., Skiadas, C. (2014). The first exit time theory applied to life table data: the health state function of a population and other characteristics. *Commun Statist.* 34, 1585-1600.

Skiadas, C.H., Skiadas, C. (2015a). Exploring the state of a stochastic system via stochastic simulations: an interesting inversion problem and the health state function. *Methodol Comput Appl Probab*. 17, 973-982.

Skiadas, C.H., Skiadas, C. (2015b). The health state curve and the health state life table: life expectancy and healthy life expectancy estimates. *Commun Statist.* 45, https://doi.org 10.1080/03610926.2014.944662.

Skiadas, C.H., Skiadas, C. (2017). The Health-Mortality Approach in Estimating the Healthy Life Years Lost Compared to the Global Burden of Disease Studies and Applications in World, USA and Japan. In: *Exploring the Health State of a Population by Dynamic Modeling Methods*. Skiadas, C.H. and Skiadas, C. (eds) London: Springer, 67–124.

Skiadas, C.H. and Skiadas, C. (2018). The health status of a population estimated: The history of health state curves. In *Demography and Health Issues – Population Aging, Mortality and Data Analysis*, Skiadas, C.H. and Skiadas, C. (eds) Springer, Cham, Switzerland.

Sullivan, D.F. (1966). *Conceptual Problems in Developing an Index of Health, U.S. Department of HEW*. Public Health Service Publication No. 1000, Series 2, No. 17.

Sullivan, D.F. (1971) (National Center for Health Statistics): A single index of mortality and morbidity. *HSMHA Health Rep.* 86, 347–354.

Torrance, G.W. (1976). Health Status Index Models: A Unified Mathematical View. *Manage Sci.* 22(9): 990-1001.

Weibull, W. (1951). A statistical distribution function of wide applicability, *J Appl Mech*. 18, 293–297.

WHO (2014). WHO methods for life expectancy and healthy life expectancy. Global Health Estimates Technical Paper WHO/HIS/HSI/GHE/2014.5.

WHO (1948). Preamble to the Constitution of the World Health Organization as adopted by the International Health Conference, New York, 19-22 June, 1946; signed on 22 July 1946 by the representatives of 61 States (Official Records of the World Health Organization, no. 2, p. 100).

WHO, Department of Health Statistics and Information system (2013). WHO methods and data sources for the global burden of disease estimates 2000-2011. Global Health Estimates Technical Paper WHO/HIS/HSI/GHE/2013.4.

Wood, S.S. (2017). P-splines with derivative based penalties and tensor product smoothing of unevenly distributed data. *Statist Comput*. 27, 985–989.

Zafeiris, K.N. and Kostaki, A. (2017). Recent mortality trends in Greece. *Commun Statist.* https://doi.org/10.1080/0361-0926.2017.1353625

16

An Application of Data Mining Methods to the Analysis of Bank Customer Profitability and Buying Behavior

In this chapter, we use a database from a Portuguese bank, with data related to the behavior of customers, to analyze churn, profitability and next-product-to-buy (NPTB). The database includes data from more than 94,000 customers, and includes all transactions and balances of bank products from those customers for the year 2015. We describe the main difficulties found concerning the database, as well as the initial filtering and data processing necessary for the analysis. We discuss the definition of churn criteria and the results obtained by the application of several techniques for churn prediction and for the short-term forecast of future profitability. Finally, we present a model for predicting the next product that will be bought by a client. The models show some ability to predict churn, but the fact that the data concerns just a year clearly hampers their performance. In the case of the forecast of future profitability, the results are also hampered by the short timeframe of the data. The models for the next product to buy show a very encouraging performance, being able to achieve a good detection ability for some of the main products of the bank.

16.1. Introduction

The huge amounts of data that banks currently possess about their customers allow them to make better decisions concerning the efforts to obtain new customers and the types of marketing campaigns they undertake. Better decisions are beneficial to the bank, since they may lead to increased profits, but they may also be beneficial to customers, who can now be targeted just by campaigns concerning products that may interest them.

Chapter written by Pedro GODINHO, Joana DIAS and Pedro TORRES.

Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining,
First Edition. Edited by Christos H. Skiadas and James R. Bozeman.

One important piece of information that can be estimated from data in bank databases is the customer lifetime value (CLV). CLV can be understood as the total value that a customer produces during his/her lifetime (EsmaeiliGookeh and Tarokh 2013). There are many models for quantifying this value (see, for instance, Singh and Jain 2010 for a review of the most prominent models). Some existing models are based on the recency, frequency, monetary (RFM) framework (Fader *et al*. 2005) and Pareto/NBD (Schmittlein *et al*. 1987; Schmittlein and Peterson 1994) or related models (Fader *et al*. 2005, 2010). As pointed out by Blattberg *et al.* (2009), due to the uncertainty in future customer behavior, as well as in the behavior of the firm's competitors and of the firm itself, CLV is indeed a random variable and methodologies should try to compute an expected CLV.

CLV prediction in the retail banking sector is especially difficult for a number of reasons, including product diversity (which can jeopardize the use of RFM-based approaches; Ekinci *et al*. 2014), the existence of both contractual and non-contractual clients (meaning that some clients are free to leave as soon as they want, while others have long-term contracts) and even the difficulty in identifying lost customers. Despite these difficulties, several authors have addressed the estimation of CLV in retail banks. Glady *et al.* (2009) use a modified Pareto/NBD approach to estimate CLV in the retail banking sector. The authors show that the dependence between the number of transactions and their profitability may be used to increase the accuracy in CLV prediction. Haenlein *et al.* (2007) present a model with four different groups of profitability drivers based on a classification and regression tree. Clients are clustered into different groups, and a transition matrix is used to consider movements between clusters. A CLV model based on RFM and Markov chains is proposed in Mzoughia and Limam (2015). Calculating the churn probability for a given client or cluster of clients may support the estimation of CLV. Ali and Arıtürk (2014) present a dynamic churn prediction framework that uses binary classifiers. Customer churn prediction is also tackled by He *et al.* (2014) by applying support vector machines.

Another important issue in retail banking is identifying the products that a customer is most likely to buy in order to enhance the effectiveness of cross-selling strategies or marketing campaigns. This may be addressed by NPTB models, which attempt to predict "which product (or products) a customer would be most likely to buy next, given what we know so far about the customer" (Knott *et al*. 2002).

Examples of works analyzing NTBD models and cross-selling strategies in banking can be found, for example, in Knott *et al*. (2002) and Li *et al*. (2005, 2011). Knott *et al.* (2002) compare several NTBD models in the context of a retail bank. The authors compare the use of different predictor variables, different calibration strategies and different methods, including discriminant analysis, multinomial logit,

logistic regression and neural networks. The authors conclude that the use of both demographic data, information concerning the products currently owned and customer activity data increases the model accuracy, and that random sampling performs better than non-random sampling. Concerning the method, the authors do not find large differences, although neural networks seem to perform slightly better than the remaining methods, and discriminant analysis seems to perform slightly worse. Li *et al.* (2005) use a structural multivariate probit model to analyze purchase patterns for bank products. Li *et al.* (2011) use a multivariate probit model and stochastic dynamic programming in order to optimize cross-selling campaigns, aiming to offer the right product to the right customer at the right time, through the right communication channel.

In this chapter, we address the estimation of future profitability and churn probability as initial steps in CLV estimation, and we also aim at predicting the next product to be bought by a client. We rely both on econometric models and data mining techniques, choosing the one with the best predictive ability in the test set, that is, the one that performs better in a set that is independent from the one used to estimate the model.

This chapter is organized as follows. After this introduction, the database is presented and discussed in section 16.2. Section 16.3 addresses the estimation of customer profitability, and section 16.4 considers the prediction of customer churn. Section 16.5 focuses on NPTB models, and the conclusions and future research are discussed in section 16.6.

16.2. Data set

The database used in this work includes data from more than 94,000 customers of a Portuguese retail bank, incorporating all transactions and balances of bank products and bank-related activity of those customers in the year 2015. The database contains only anonymized data, guaranteeing the privacy of the data and preventing the identification of clients.

Sociodemographic data include the age, the first digits of the postcode (allowing the identification of the region in which the client resides), the marital status, the job, the way the client opened the bank account (whether in a bank branch, online or in other way) and the day the client opened the account.

All bank products are associated with checking accounts, and the database also contains the transactions and balances of all products associated with the client's account, as well as the number and value of the products of each type owned by the customer. Data are aggregated at the monthly level, meaning that balances correspond to the end of the month and transactions correspond to the accumulated

monthly activity. The bank products include different types of mutual funds, insurance products and credit products, as well as credit and debit cards, term deposits and stock market investments. Additionally, the number of online logins made by the customer to the bank site and the number of transactions made online are also available in the database. Other important pieces of data are the net profits the bank gained with each customer in each month for different categories of products. The number of records concerning transactions, balances and numbers of logins is larger than 8.5 million.

The database had to be cleaned, since it contained some obviously invalid values (e.g. invalid customer ages, including a few negative ages). Customers with invalid data were removed from the database.

Other preprocessing included aggregations in some categorical variables. The initial database included 486 different jobs and, using an official taxonomy of jobs for Portugal, we mapped them into a set of just 17 jobs. A similar procedure was performed for the marital status: initially, there were 11 different values for this variable (including different values for married customers for different types of premarital agreements). These original values were mapped into a set of five different values.

After this initial preprocessing of the data, relations between different variables were analyzed, and some expected relations were indeed found. An example is the relation between the wealth deposited in the bank and the profitability of the client for the bank. Figures 16.1 and 16.2 show this relation, for the months of January and December, as well as a trend line. It is clear that profitability tends to increase with wealth, as was to be expected.

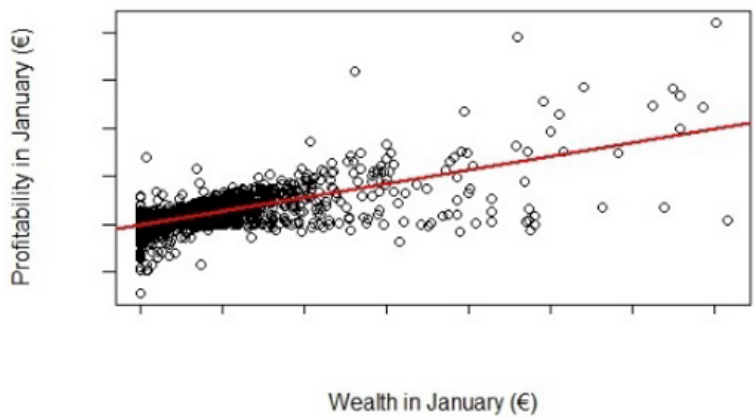


Figure 16.1. *Relation between customer wealth and profitability for the bank in January*

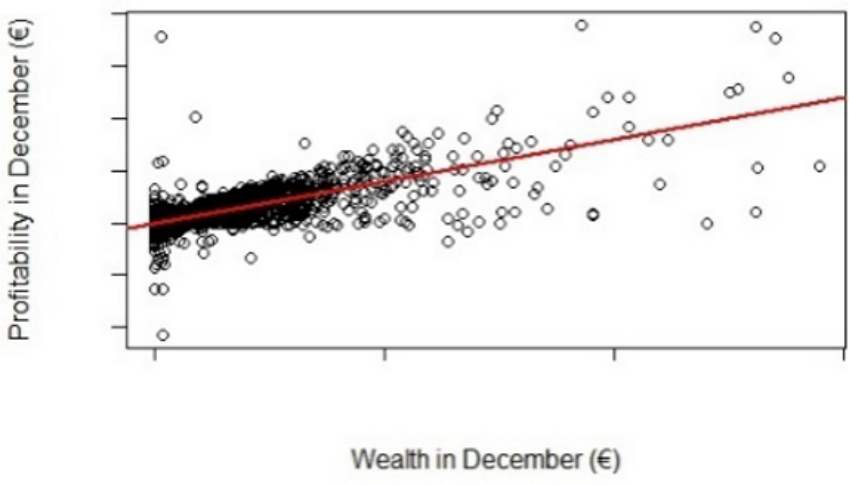


Figure 16.2. *Relation between customer wealth and profitability for the bank in December*

Three shortcomings of the database were made evident in a preliminary analysis. The first is related to outliers in customer profitability, which will be analyzed in section 16.3.

The second shortcoming is that the records that are interpreted as different customers may correspond to the same person who chooses to open different accounts: for example, someone who chooses to create an account for day-to-day transactions and another for retirement savings (retirement mutual funds, stock market investments and the like). Although this may create some bias, we do not expect this to happen in many cases, so the impact of such possibility will probably be limited.

Another, more serious, shortcoming is the existence of just 1 year of data, aggregated in monthly values. This makes it difficult to assess the medium- and long-term behavior of the customers, for example to determine whether or not a customer is in churn. It also makes it impossible to test medium- and long-term forecasts. This shortcoming is expected to cause some problems in the estimation of customer profitability and customer churn.

In order to assess the accuracy of prediction models, data were divided into two sets; 60% of the observations were used as a training set to estimate the models. The other 40% of the observations constitutes a test set, which was used to assess the prediction accuracy in data that was not used in the estimations.

16.3. Short-term forecasting of customer profitability

Profitability from a client in a given month is expected to be strongly correlated with the profitability in the previous month. This is clearly shown to be the case in Figure 16.3, which shows the relation between the profitability in January and February. As expected, the points in this graph are very close to the straight line $y = x$, showing that that profitability given by the client in a given month is a good forecast of the profitability given by the client in the next month. Therefore, we aimed at forecasting the changes in profitability instead of the profitability in order to avoid getting apparently good forecasting results just because profitability shows high persistence.

Figure 16.4 shows the relation between the profitability in June and July. Once again, the relation is close to the straight line $y = x$, for the large majority of observations, but there are several important outliers, corresponding to customers whose profitability shows a visible increase. In fact, in May and July, the profitability associated with some customers has an important increase, only to show a similar decrease in the following month (June and August, respectively). This introduces outliers in the data, harming the ability to predict future profitability. According to an analysis of this situation made with bank members, this seems to be due to the way the profitability of some (very few) products is accounted. New, more realistic ways of considering the profitability of these products will be analyzed with the bank but, meanwhile, we chose to use the existing values in order to avoid the risk of introducing biases in the data.

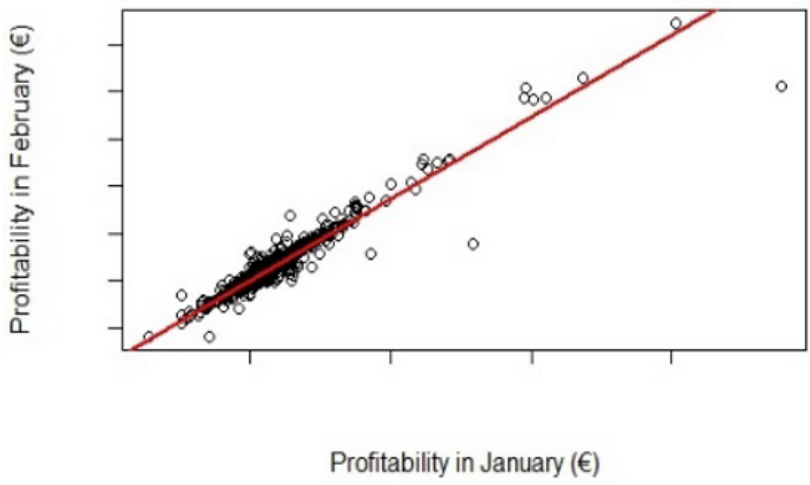


Figure 16.3. *Relation between the profitability of the clients in January and February*

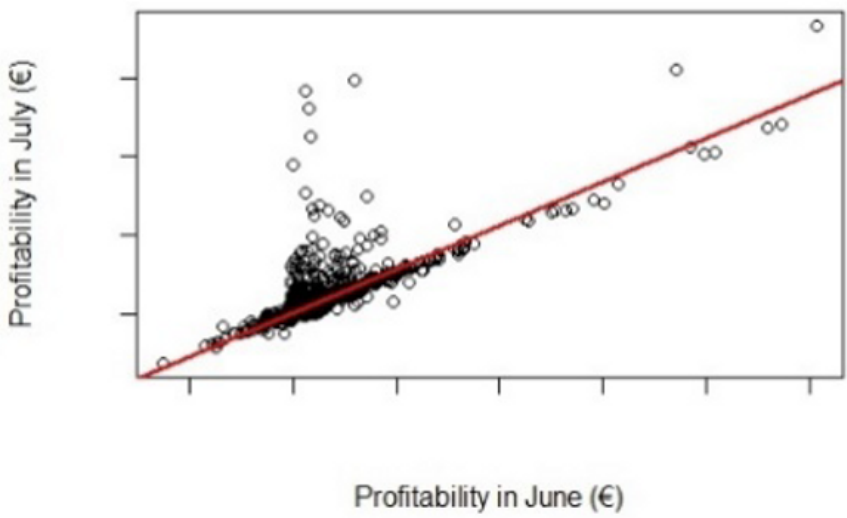


Figure 16.4. *Relation between the profitability of the clients in June and July*

Figure 16.5 shows a histogram with the monthly values of the profitability. We can see that there are a very large number of slightly negative values of the monthly profitability.

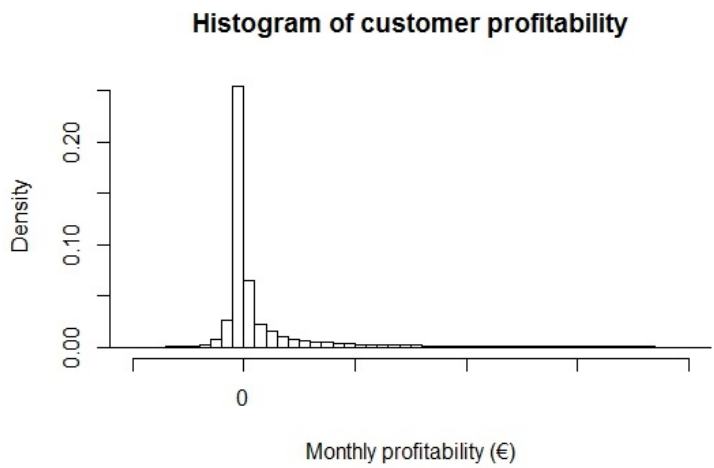


Figure 16.5. *Histogram with monthly values of customer profitability*

We started by trying to predict the change in customer profitability 1 month ahead. We used both sociodemographic data and data from the customer activity and balances in the three previous months to estimate the change in the customer profitability in the next month. For example, data concerning activity, transactions and balances from January, February and March, are used to estimate the change in profitability between March and April. The goodness-of-fit measure that we chose is the root mean square error (RMSE) and we compare the obtained forecast with the naïve forecast that assumes that the change in profitability is equal to the average change in the training period (termed ModAvg).

The first types of models to be estimated were linear models. This allowed us to get a first idea of the relevance of the different variables for explaining the changes in profitability. The non-relevant variables were iteratively removed and, in the end, the model presented an adjusted R^2 of 0.4593. The performances of the model thus obtained and of the benchmark model (ModAvg), both in the training and in the test sets, are summarized in Table 16.1.

Model	RMSE in the training set	RMSE in the test set
Linear model	12.90	15.21
ModAvg	17.67	19.72

Table 16.1. *Performance of the linear model and benchmark model*

As can be seen in Table 16.1, the linear model is better than ModAvg, both in the training set and in the test set, and model performance deteriorates in the test set.

In order to give an idea of the impacts of the different variables, we show the sign of the coefficients and their statistical significance in Table 16.2, for some of the most significant variables. Since we are considering some data from the three previous months to estimate the change in customer profitability, the coefficient signs are presented for each of these months (1, 2 and 3 months before the change in profitability we are trying to forecast).

In some cases, the coefficient signs change from 1 month to the next, while remaining very significant. This is a clear indication that not only the value of the variable is relevant to forecast the change in profitability, but the change in the value may be relevant as well. For example, wealth in the latest month has a positive sign, whereas wealth in the month before has a negative sign: this may mean that both the most recent value of the wealth and the latest change in wealth have a positive influence in the expected change in profitability.

	Sign and significance		
	One month before	**Two months before**	**Three months before**
Online logins	+***	+*	–
Number of online transactions	+***	–***	–***
Credit card transactions	+***	–***	+***
Number of different mutual funds in the account	+***	+	–***
Value of stock market holdings	–***	+***	+
Number of stock market transactions	–***	+**	–***
Total wealth	+***	–***	–***
Total value of loans	+***	+	–***
Value allocated to term deposits in the month	+***	+***	–
Value removed from term deposits in the month	–***	–***	–***
Amount of wages deposited in the bank	+***	–***	+**
Profitability from checking account	–***	+***	+
Profitability from term deposits	–***	+***	+
Profitability from home equity loans	–***	+***	+***
Profitability from other (non-home equity) loans	–***	+***	+***
Profitability from mutual funds	–***	+***	–***
Profitability from stock market holdings	–***	+***	+
Age	+***		

+,–: sign of the coefficient; *significant at the 10% level; **significant at the 5% level; ***significant at the 1% level.

Table 16.2. *Sign and significance of some of the most significant variables of the linear model*

In the cases of stock market holdings and transactions, the coefficient signs seem to be the contrary of what was expected. One possible explanation may be that customers with larger stock market holdings use the account mostly to make trades and deposit such assets (that is as an investment account), and they do not tend to buy new products that are profitable to the bank.

Another interesting result is the negative and statistically significant sign in the last month profitability, for the different categories of products. However, this has a simple interpretation: all other things remaining constant, the larger the profitability already is, the less it is expected to increase.

Finally, note that only one sociodemographic variable is significant: age. Older clients generate more profits than younger clients. The significance of age was also found on the other models that we considered.

We also applied linear models to forecast the change of profitability at 2-, 3- and 4- month horizons. The results, shown in Table 16.3, clearly show that the forecasting ability of the models decreases when the forecasting horizon becomes longer.

Forecasting horizon	RMSE in the training set	RMSE in the test set
1 month	12.90	15.21
2 months	14.63	17.29
3 months	16.39	18.62
4 months	17.56	20.33

Table 16.3. *Performance of linear models on the test set for different forecasting horizons*

After this linear model, several data mining methods were applied: regression random forests, gradient boosting, naïve Bayes and linear discriminant analysis. Although these methods perform better than the linear model in the training set (sometimes very significantly), we could never improve the predictive performance in the test set, when compared with the linear model. So, there seems to be an overfitting problem with the application of these data mining techniques to predict profitability. However, due to the long computation times associated with the application of these techniques, we tried a limited number of configurations for each one. In random forests, for example, it is possible that a different set of variables, or different numbers of trees or of candidates to each split, might lead to better results.

16.4. Churn prediction

One important initial difficulty in churn prediction was the definition of churn. The bank had no clear definition and, for this work, we chose to define churn through rules that are mostly based on common sense: there is churn if there are no relevant products and small amounts of credits and of wealth deposited in the bank. The exact rules consisted of defining that a customer was in churn if, simultaneously, he/she had no insurance contracts, no term deposits, no mutual funds and no credit or debit cards, the wealth in the bank was below 1,000 € and the loans amounted to less than 100 €.

Our goal in predicting churn was mainly to predict which customers are not currently in churn but have a high probability of churning in the future. When we mention churning in the future, we are considering a reasonable amount of time; given the fact that we have data for only 1 year, we chose to predict churn using a 6-month horizon. In order to have enough data to try using different lags, we aimed at trying to predict which customers were not in churn in June 2015, but were in churn in December 2015. The number of customers in this situation was quite low, less than 0.7% of the customers in the database.

Churn prediction was handled as a classification problem. We used both linear models (probit and logit) and several data mining techniques (Adaboost, linear discriminant analysis, classification random forests). The best results were achieved with classification random forests, which obtained a much better performance than all the other models. We will only present the results obtained by classification random forests and logit models (the linear models with the best performance).

We started to use a large number of variables in the models, both sociodemographic and related to balances, transactions and other activity. For balances, transactions and other activity, we started by using the values from January to June 2015. We defined a variable that measures the ratio between the value of wealth in June and the average wealth in the semester, and we defined a similar variable for the amount of loans. We also defined new binary variables, for several products, to define whether or not the customer had any of that product (regardless of the amount) in each month, and also for determining whether the customer had made any online logins and transactions in each month.

Several configurations were tried (mostly in the linear models) in order to assess whether the binary variables or the initial values performed better, and then the less significant variables were iteratively removed. In the end, the number of relevant variables was much smaller than for profitability prediction. In almost all the cases, we found out that only the most recent value was relevant, the main exceptions being the two new variables that measured the relation between June values and average semester values for wealth and credits.

In general, the most relevant variables were:

– age;

– wealth;

– ratio between the wealth in June and the average wealth on the semester;

– value of loans;

– ratio between the value of loans in June and the average value of the semester;

– profitability in the latest months;

– total balance of term deposits;

– existence of transactions and logins in the latest months.

The models we used define the probability that a customer is going to churn. In order to assess the prediction ability, we calculated the average probability that the models assign to churners and to non-churners. The results are shown in Table 16.4, and they show that random forests assign much higher probabilities to customers that effectively end up churning, although they also assign slightly higher probabilities to non-churners.

Model	Logit model	Random forest
Average probability assigned by the model to customers that effectively churn	3.78%	9.89%
Average probability assigned by the model to customers that end up not churning	0.66%	0.71%

Table 16.4. *Performance of the best linear model (logit) and the best nonlinear technique (random forests)*

We can see that random forests show some ability in differentiating future churners from non-churners. However, we must acknowledge that, due to limitations in the data that were mentioned in section 16.2, we cannot be sure if the customers we are identifying as churning are, in reality, churning.

16.5. Next-product-to-buy

We also tried to predict, for some products, whether or not a given product from the bank will be the given customer's next buy. Since data are monthly, we are in fact identifying whether customers buy a product in the next month in which they

acquire one or more products from the bank. This is also a classification problem: a product is classified as whether or not it will be bought in the month a next purchase is made. We used both linear models (probit and logit) and data mining techniques (random forests, Adaboost, linear discriminant analysis).

Three products, held by an important percentage of customers, were considered: term deposits, debit cards and credit cards. Purchase of such products was identified as an increase in the number of units of the product in the customer account. This allows us to avoid incorrectly classifying as purchases the cases in which a customer just changes a product he currently holds by another of the same kind (e.g. ending a term deposit and applying the capital in a new one).

The logic used for defining the training and test sets was somewhat different in this analysis. In the training set, the prediction was made for the next purchase in the months from May to August using data from the previous three months (February to April). In the test set, we intended to predict the next purchase in the 4-month period from September to December using data from the previous 3 months (June to August). Only customers making any kind of purchase in the considered 4-month period were taken into account (i.e. we try to predict what is the next product to be bought; we are not making a joint prediction of the next product and of the probability of a buying occurring).

Apart from sociodemographic variables and transactions, balances and bank-related activity in the three previous months, new binary variables were added regarding the occurrence of purchases of the different types of products for each of the three previous months.

Product	Percentage of customers for which the product is the next to be bought	Average probability estimated by the model when the product is the next to be bought	Average probability estimated by the model when the product is not the next to be bought	Percentage of customers correctly identified by the model as buying the product next	Percentage of customers, among the 5% with largest probability in the model for which the product is the next to be bought
Term deposits	51.4%	59.7%	41.9%	71.3%	90.0%
Debit cards	11.4%	20.5%	9.7%	38.2%	60.2%
Credit cards	13.7%	19.9%	11.7%	31.6%	38.0%

Table 16.5. *Performance of logit models in predicting the next product bought by a customer*

The best results in the test set were obtained with logit models. In Table 16.5, we present the performance of these models. In order to assess the predictive ability of the models, we considered the average probability given by the model when the product is the next to be bought and when it is not, the percentage of customers correctly identified by the model as buying the product next, and also the percentage of customers, among those with the top 5% probabilities estimated by the model, who effectively buy that product next. This last measure is particularly interesting for defining targeted marketing campaigns, since it allows the identification of the customers that will most probably buy the product. We also present the percentage of customers for which the product is the next to be bought – this is, in fact, the probability of a customer next buying that product, when you choose him/her at random.

We can see that the models perform quite well. In particular, the customers to whom the models assign higher probabilities really do have a high probability of next purchasing the product.

16.6. Conclusions and future research

In this chapter, we present the results of an analysis of churn, profitability and NPTB, obtained using a database concerning the behavior of customers from a Portuguese bank. If it is possible to accurately predict churn probabilities and the evolution of profitability, then it is possible to estimate CLV, which is of great importance for defining marketing strategies.

As we explained, the database has some shortcomings, including not identifying the same client with different accounts, the existence of profitability outliers and the fact of there being just 1 year of data, aggregated in monthly values.

A linear model showed a good performance in the estimation of future short-term profitability at the 1-month horizon, but the performance of the estimated models seems to deteriorate when the prediction horizon increases, even if it is only to a few months. For churn, we had no solid reference to determine when a customer churns, so we defined a rule for identifying churning customers. A random forest seems to have an interesting ability to forecast which customers will churn in the next 6 months. However, given the short time period covered by the database, we cannot be completely sure that the customers identified as having churned did, indeed, churn. Therefore, given the limitations in the results concerning profitability and churn prediction, we feel that it is not yet possible to make a credible calculation of CLV. Still, the results concerning churn are interesting and may help identify the customers whose relation with the bank is becoming very weak. The bank may thus

target these customers with marketing campaigns in order to try to avoid losing them.

The results of the models of the NPTB are very interesting and show that a logit model has a good ability to predict the next product that a customer will buy. In particular, a large percentage of the customers to whom the model predicted the top 5% largest probabilities of purchasing each of the considered products did indeed buy that product next. This opens the way to targeted marketing campaigns for selling the products that the customers are more likely to purchase.

At the outset, we expected data mining techniques to outperform the predictive ability of linear models. Although data mining techniques usually perform much better in the training set, only in the case of churn were they able to beat a logit model in the test set. Possible explanations for this may be that linear models are particularly suited to this data set, that the shortcomings of the database are especially harming the performance of data mining techniques and that different parameterizations of the techniques should be tested in order to fine-tune them to the characteristics of the data. Concerning this latter explanation, the number of tested parameterizations was indeed limited due to very long computational running times, but we will, in the future, try new parametrizations and new approaches in order to achieve better predictions.

As future work, we are already in contact with the bank to get a database covering a longer time period. This is expected to allow us to define a more credible identification of churning customers and better predictions of future profitability and NPTB. We will also address the estimation of CLVs, both using predictions of churn probability and future profitability and also using other approaches made available by a longer database. Finally, we will try to obtain better predictions of the NPTB and propose models for defining long-term market strategies based on these predictions.

16.7. References

Ali, Ö.G., Arıtürk, U., (2014). Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. *Expert Syst Appl*. 41(17), 7889–7903.

Blattberg, R.C., Malthouse, E.C., Neslin, S.A. (2009). Customer lifetime value: Empirical generalizations and some conceptual questions. *J Interact Market*. 23(2), 157–168.

Ekinci, Y., Ülengin, F., Uray, N., Ülengin, B. (2014). Analysis of customer lifetime value and marketing expenditure decisions through a Markovian-based model. *European J Oper Res*. 237(1), 278–288.

EsmaeiliGookeh, M., Tarokh, M.J. (2013). Customer lifetime value models: A literature survey. *Int J Indust Eng*. 24(4), 317–336.

Fader, P.S., Hardie, B.G., Lee, K.L. (2005). "Counting your customers" the easy way: An alternative to the Pareto/NBD model. *Market Sci*. 24(2), 275–284.

Fader, P.S., Hardie, B.G., Lee, K.L. RFM, CLV: Using iso-value curves for customer base analysis. *J. Market Res*. 42(4), 415–430.

Fader, P.S., Hardie, B.G., Shang, J. (2010). Customer-base analysis in a discrete-time noncontractual setting. *Market Sci*. 29(6), 1086–1108.

Glady, N., Baesens, B., Croux, C. (2009). A modified Pareto/NBD approach for predicting customer lifetime value. *Expert Syst Appl*. 36(2), 2062–2071.

Haenlein, M., Kaplan, A.M., Beeser, A.J. (2007). A model to determine customer lifetime value in a retail banking context. *Eur Manage J*. 25(3), 221–234.

He, B., Shi, Y., Wan, Q., Zhao. X. (2014). Prediction of customer attrition of commercial banks based on SVM model. *Procedia Comput Sci*. 31, 423–430.

Knott, A., Hayes, A., Neslin, S.A. (2002). Next-product-to-buy models for cross-selling applications. *Journal of Interact Market*. 16(3), 59–75.

Li, S., Sun, B., Montgomery, A.L. (2011). Cross-selling the right product to the right customer at the right time. *J Market Res*. 48(4), 683–700.

Li, S., Sun, B., Wilcox, R.T. (2005). Cross-selling sequentially ordered products: An application to consumer banking services. *J Market Res*. 42(2), 233–239.

Mzoughia, M.B., Limam, M. (2015). An improved customer lifetime value model based on Markov chain. *Appl Stoch Models Bus Industry* 31(4), 528–535.

Schmittlein, D.C., Morrison, D.G., Colombo, R. (1987). Counting Your Customers: Who-Are They and What Will They Do Next? *Manage Sci*. 33(1), 1–24.

Schmittlein, D.C., Peterson, R.A. (1994). Customer base analysis: An industrial purchase process application. *Market Sci*. 13(1), 41–67.

Singh, S.S., Jain, D.C. (2010). Measuring customer lifetime value. *Rev Market Res*. 6, 37–62.

List of Authors

Benard ABOLA
Mälardalen University
Västerås
Sweden

Antonio BALZANELLA
Università della Campania
"Luigi Vanvitelli"
Naples
Italy

Pitos BIGANDA
Mälardalen University
Västerås
Sweden

James R. BOZEMAN
American University of Malta
Bormla
Malta

Dominique DESBOIS
INRA-AgroParisTech,
Paris-Saclay University
France

Joana DIAS
CeBER, INESC-Coimbra
University of Coimbra
Portugal

Agnese M. DI BRISCO
University of Milano-Bicocca
Italy

Christopher ENGSTRÖM
The School of Education, Culture and Communication
Mälardalen University
Västerås
Sweden

Nikolaos FARMAKIS
Aristotle University of Thessaloniki
Greece

Jim FREEMAN
University of Manchester
United Kingdom

Evgeny GERSHIKOV
Braude Academic College
Carmiel
Israel

Francesco GIORDANO
University of Salerno
Fisciano
Italy

Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining,
First Edition. Edited by Christos H. Skiadas and James R. Bozeman.

Pedro GODINHO
CeBER
University of Coimbra
Portugal

William GÓMEZ-DEMETRIO
ICAR
Universidad Autónoma
del Estado de México
Toluca
Mexico

Christian HENNIG
Dipartimento di Scienze Statistiche
"Paolo Fortunati"
University of Bologna
Italy

Eleni KETZAKI
Department of Mathematics
Aristotle University of Thessaloniki
Greece

Samuel KOSOLAPOV
Braude Academic College
Carmiel
Israel

Ana LORGA DA SILVA
ECEO–FCSEA
Universidade Lusófona de
Humanidades e Tecnologias
Lisbon
Portugal

Francisco MARTÍNEZ-CASTAÑEDA
ICAR
Universidad Autónoma
del Estado de México
Toluca
Mexico

Sonia MIGLIORATI
University of Milano-Bicocca
Italy

Marcella NIGLIO
University of Salerno
Fisciano
Italy

Fumiaki OKIHARA
Chuo University
Tokyo
Japan

Andrea ONGARO
University of Milano-Bicocca
Italy

Ioanna PAPATSOUMA
Aristotle University of Thessaloniki
Greece

Monica RUIZ-TORRES
IIZD
Universidad Autónoma
de San Luis Potosí
Mexico

Ernesto SÁNCHEZ-VERA
ICAR
Universidad Autónoma
del Estado de México
Toluca
Mexico

Leonel SANTOS-BARRIOS
ICAR
Universidad Autónoma
del Estado de México
Toluca
Mexico

Gilbert SAPORTA
Conservatoire National des Arts et
Métiers
Paris
France

Armin O. SCHMITT
Georgia-Augusta-University of Göttingen
Germany

Sergei SILVESTROV
The School of Education, Culture and Communication
Mälardalen University
Västerås
Sweden

Christos H. SKIADAS
ManLab
Technical University of Crete
Chania
Greece

Monika A. TOMKOWICZ
Free University of Bozen–Bolzano
Italy

Pedro TORRES
CeBER
University of Coimbra
Portugal

Rosanna VERDE
Università della Campania
"Luigi Vanvitelli"
Naples
Italy

Jan Ámos VÍŠEK
Charles University
Prague
Czech Republic

Cosimo Damiano VITALE
University of Salerno
Fisciano
Italy

Henri WALLARD
Ipsos Science Center
Paris
France

Norio WATANABE
Chuo University
Tokyo
Japan

Konstantinos N. ZAFEIRIS
Democritus University of Thrace
Rhodopi
Greece

Xin ZHAO
University of Manchester
United Kingdom

Index

Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining,
First Edition. Edited by Christos H. Skiadas and James R. Bozeman.

L, M

N, P

Q, R

S

T, V