

**INNOVATION, ENTREPRENEURSHIP AND MANAGEMENT SERIES**

**BIG DATA, ARTIFICIAL INTELLIGENCE AND DATA ANALYSIS SET**



**Volume 11**

# **Data Analysis and Related Applications 3**

*Theory and Practice  
New Approaches*

**Edited by  
Yiannis Dimotikali  
Christos H. Skiadas**

**ISTE**

**WILEY**



## Data Analysis and Related Applications 3





**Big Data, Artificial Intelligence  
and Data Analysis Set**

coordinated by  
Jacques Janssen

Volume 11

---

**Data Analysis and Related  
Applications 3**

---

*Theory and Practice – New Approaches*

*Edited by*

Yiannis Dimotikalis  
Christos H. Skiadas

**ISTE**

**WILEY**

First published 2024 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc.

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms and licenses issued by the CLA. Enquiries concerning reproduction outside these terms should be sent to the publishers at the undermentioned address:

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

[www.iste.co.uk](http://www.iste.co.uk)

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

[www.wiley.com](http://www.wiley.com)

© ISTE Ltd 2024

The rights of Yiannis Dimotikalis and Christos H. Skiadas to be identified as the authors of this work have been asserted by them in accordance with the Copyright, Designs and Patents Act 1988.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s), contributor(s) or editor(s) and do not necessarily reflect the views of ISTE Group.

Library of Congress Control Number: 2023949006

---

British Library Cataloguing-in-Publication Data  
A CIP record for this book is available from the British Library  
ISBN 978-1-78630-962-4

---

---

# Contents

---

<b>Part 1. Data Analysis Classification</b>	1
<b>Chapter 1. Walking into the Digital Era: Comparison of the European Union Countries in the Last Decade</b>	3
Fernanda Otilia FIGUEIREDO and Adelaide FIGUEIREDO	
1.1. Introduction	3
1.1.1. The dataset.	4
1.1.2. Preliminary analysis of the data	5
1.2. Double principal component analysis: a brief description	7
1.3. Representation of the trajectories.	8
1.4. Some final conclusions	11
1.5. Acknowledgements	12
1.6. References	12
<b>Chapter 2. Multivariate Kernel Discrimination Applied to Bank Loan Classification</b>	13
Mark Anthony CARUANA and Gabriele LENTINI	
2.1. Introduction	13
2.2. Multivariate kernel density estimation.	14
2.2.1. Kernel density estimator	15
2.2.2. The optimal bandwidth.	16
2.3. Kernel discriminant analysis	17
2.4. Applying kernel discriminant analysis to local data	18
2.4.1. Description of data	18
2.4.2. Variable selection.	20
2.5. Conclusion	25
2.6. References	25

<b>Chapter 3. Introducing an Ontology of Adolescents’ Digital Leisure</b> . . . . .	29
George FILANDRIANOS, Aggeliki KAZANI, Dimitrios PARSANOGLOU, Maria SYMEONAKI and Giorgos STAMOU	
3.1. Introduction . . . . .	29
3.2. Related work . . . . .	32
3.3. Ontology . . . . .	35
3.4. Description of the ontology construction . . . . .	36
3.5. Data annotation . . . . .	39
3.6. Conclusions . . . . .	41
3.7. Acknowledgments . . . . .	41
3.8. References . . . . .	41
 <b>Chapter 4. Blackjack and the Kelly Bet: A Simulation Assessment of Selected Playing Strategies</b> . . . . .	43
Jim FREEMAN and Haoyu MIAO	
4.1. Introduction . . . . .	43
4.2. Blackjack . . . . .	44
4.2.1. Rules of the game . . . . .	44
4.3. Kelly bet . . . . .	46
4.4. Simulation study . . . . .	47
4.4.1. Modeling assumptions . . . . .	47
4.5. Empirical results . . . . .	48
4.6. Conclusions . . . . .	49
4.7. References . . . . .	50
 <b>Part 2. Estimators</b> . . . . .	51
 <b>Chapter 5. An Evaluation of the Efficiency of a Shape Parameter Estimator for the Log-logistic Distribution</b> . . . . .	53
Frederico CAEIRO and Ayana MATEUS	
5.1. Introduction . . . . .	53
5.2. Estimation of the shape parameter . . . . .	56
5.2.1. Estimators under study . . . . .	56
5.2.2. Asymptotic properties of the estimators . . . . .	57
5.3. Simulation study . . . . .	58
5.4. References . . . . .	60

<b>Chapter 6. Restricted Minimum Density Power Divergence Estimator for Step-stress ALT with Nondestructive One-shot Devices</b> . . . . .	63
Narayanaswamy BALAKRISHNAN, María JAENADA and Leandro PARDO	
6.1. Introduction . . . . .	63
6.2. Minimum density power divergence estimator and restricted minimum density power divergence estimator . . . . .	65
6.3. Influence function of the restricted minimum density power divergence estimator. . . . .	68
6.4. Applications of the restricted MDPDE. . . . .	69
6.4.1. Rao-type test statistics . . . . .	69
6.4.2. DPD-based statistics . . . . .	72
6.5. Simulation study . . . . .	73
6.6. Conclusions . . . . .	75
6.7. References . . . . .	75
<b>Part 3. Finance</b> . . . . .	77
<b>Chapter 7. Properties of American Options Under a Semi-Markov Modulated Black-Scholes Model</b> . . . . .	79
Kouki TAKADA, Marko DIMITROV, Lu JIN and Ying NI	
7.1. Introduction . . . . .	79
7.2. American option pricing . . . . .	81
7.3. Exercising strategies. . . . .	83
7.3.1. Setting parameters . . . . .	83
7.3.2. Relationship between the option price and “economic situation, asset price and maturity” . . . . .	84
7.3.3. Relationship between option price and transition probability matrix. . . . .	86
7.3.4. Consideration of the optimal exercise region . . . . .	89
7.3.5. Condition of transition probability matrix for option price transition to be monotonous . . . . .	91
7.4. Conclusion . . . . .	93
7.5. References . . . . .	93
<b>Chapter 8. Numerical Studies of Implied Volatility Expansions Under the Gatheral Model</b> . . . . .	97
Mohammed ALBUHAYRI, Marko DIMITROV, Ying NI and Anatoliy MALYARENKO	
8.1. Introduction . . . . .	97
8.2. Previous results . . . . .	99
8.3. Accuracy of the asymptotic expansions . . . . .	102
8.4. Numerical analysis. . . . .	105
8.5. Conclusion and future work. . . . .	106

8.6. Acknowledgments . . . . .	106
8.7. References . . . . .	106

**Chapter 9. Constructing Trinominal Models Based on Cubature Method on Wiener Space: Applications to Pricing Financial Derivatives.** . . . . .

Hossein NOHROUZIAN, Anatoliy MALYARENKO and Ying NI 109

9.1. Introduction and outline of this chapter . . . . .	109
9.2. Cubature formula in Black–Scholes and Black’s models. . . . .	110
9.2.1. Black–Scholes model via cubature formula . . . . .	111
9.2.2. Black’s model via cubature formula. . . . .	112
9.3. Constructing a trinomial tree via cubature formula . . . . .	113
9.3.1. A trinomial tree approximation for the Black–Scholes model . . . . .	113
9.4. Convergence to geometric Brownian motion . . . . .	115
9.4.1. Convergence of finite-dimensional distribution . . . . .	117
9.4.2. Relative compactness . . . . .	122
9.5. Martingale probability measure . . . . .	122
9.6. Extension of the results and examples . . . . .	124
9.6.1. A trinomial tree approximation for Black’s model . . . . .	126
9.6.2. Pricing American options using the trinomial tree . . . . .	127
9.7. Discussion . . . . .	129
9.8. References . . . . .	129

**Chapter 10. A Bayesian Approach to Measuring Risk on Portfolios with Many Assets** . . . . .

Samuel BONELLO, David SUDA and Monique BORG INGUANEZ 131

10.1. Introduction . . . . .	131
10.2. Dynamic principal component analysis . . . . .	133
10.3. The Bayesian GARCH(1,1) model . . . . .	135
10.4. Modeling a portfolio with many assets . . . . .	137
10.5. Forecasting and risk estimation . . . . .	140
10.6. Measuring predictive ability. . . . .	143
10.7. Conclusion . . . . .	144
10.8. Acknowledgments . . . . .	145
10.9. References . . . . .	145

**Chapter 11. Financial Management of Four Hellenic Public Health Units. Analysis and Evaluation through Numerical Indicators** . . . . .

Maria SACHINIDOU and George MATALLIOTAKIS 147

11.1. Introduction . . . . .	148
11.2. Theoretical framework. . . . .	148
11.3. Methodology . . . . .	149

11.4. Research results . . . . .	149
11.4.1. Liquidity indicators . . . . .	149
11.4.2. Activity ratios . . . . .	153
11.4.3. Profitability ratios . . . . .	157
11.4.4. Capital adequacy and viability ratios . . . . .	159
11.5. Conclusions . . . . .	162
11.6. References . . . . .	163
<b>Part 4. Health Services . . . . .</b>	<b>165</b>
<b>Chapter 12. Lean Management as an Improvement Factor in Health Services. The Case of Venizeleio General Hospital of Crete, Greece . . . . .</b>	<b>167</b>
Eleni GENITSARIDI and George MATALLIOTAKIS	
12.1. Introduction . . . . .	168
12.2. Theoretical framework . . . . .	168
12.3. Purpose of the research . . . . .	172
12.4. Methodology . . . . .	172
12.5. Research results . . . . .	172
12.6. Conclusions . . . . .	179
12.7. References . . . . .	180
<b>Chapter 13. Satisfaction of Employees in Primary and Secondary Healthcare Structures During the Pandemic Period in the Prefecture of Magnesia . . . . .</b>	<b>183</b>
Sofia TRIKALLIOTI and George MATALLIOTAKIS	
13.1. Introduction . . . . .	184
13.2. The Covid-19 pandemic and its effects . . . . .	184
13.3. Job satisfaction of healthcare professionals . . . . .	185
13.4. Healthcare professionals and the burnout syndrome . . . . .	186
13.5. Research purpose . . . . .	188
13.6. Research methodology . . . . .	188
13.7. Research conclusions . . . . .	189
13.8. References . . . . .	193
<b>Chapter 14. A Parametric Analysis of OpenFlow and P4 Protocols Based on Software Defined Networks . . . . .</b>	<b>197</b>
Lincoln S. PETER, Hlabishi I. KOBO and Viranjay M. SRIVASTAVA	
14.1. Introduction . . . . .	198
14.2. Motivation . . . . .	200
14.3. Challenges to the existing models . . . . .	202
14.3.1. A comparison with OpenFlow . . . . .	202

14.3.2. System: P4Consist. . . . .	203
14.4. Conclusions and future recommendations . . . . .	205
14.5. References . . . . .	205
<b>Chapter 15. A Dynamic Neural Network Model for Accurate Recognition of Masked Faces . . . . .</b>	<b>209</b>
Oladapo T. IBITOYE and Viranjay M. SRIVASTAVA	
15.1. Introduction . . . . .	210
15.2. Methodology . . . . .	213
15.2.1. Data collection and preprocessing . . . . .	214
15.2.2. Model training and validation . . . . .	215
15.2.3. System implementation . . . . .	216
15.2.4. System evaluation . . . . .	217
15.3. Results and discussion . . . . .	217
15.3.1. Training accuracy . . . . .	217
15.3.2. Validation accuracy . . . . .	218
15.3.3. Training loss . . . . .	219
15.3.4. Validation loss . . . . .	219
15.3.5. Overall accuracy. . . . .	220
15.4. Conclusions and future recommendations . . . . .	220
15.5. References . . . . .	221
<b>Chapter 16. An Action-based Monitoring Tool for Processes Subject to Multiple Quality Shifts . . . . .</b>	<b>225</b>
Konstantina TSIOTA and Konstantinos A. TASIAS	
16.1. Introduction . . . . .	225
16.2. Problem definition and assumptions . . . . .	227
16.3. Conventional versus proposed SPC approach . . . . .	230
16.3.1. Conventional SPC approach . . . . .	230
16.3.2. Proposed SPC approach. . . . .	231
16.4. Markov chain models . . . . .	232
16.4.1. Transition probability matrix . . . . .	232
16.4.2. Transition probabilities . . . . .	234
16.4.3. Expected cost per time unit. . . . .	236
16.5. Optimization problem . . . . .	238
16.6. Numerical analysis . . . . .	239
16.6.1. Numerical examples. . . . .	239
16.6.2. Performance comparisons . . . . .	239
16.7. Conclusions. . . . .	242
16.8. References . . . . .	242



---

<b>Chapter 17. Phi Divergence and Consistent Estimation for Stochastic Block Model</b> . . . . .	243
Cyprien FERRARIS	
17.1. Introduction . . . . .	243
17.2. Statement of the problem . . . . .	244
17.3. Results. . . . .	247
17.4. Simulations . . . . .	248
17.5. Study of real dataset: email-Enron . . . . .	250
17.6. Summary and discussion. . . . .	251
17.7. Acknowledgment . . . . .	252
17.8. Appendix . . . . .	252
17.9. References . . . . .	257
<b>List of Authors</b> . . . . .	259
<b>Index.</b> . . . . .	263



PART 1

# Data Analysis Classification



---

# Walking into the Digital Era: Comparison of the European Union Countries in the Last Decade

---

This study allows us to understand how European Union countries are more or less prepared for confronting the digital era. In particular, we aim to understand whether the least developed countries are close or fairly distant from the others. For this study, data for some variables associated with digital skills, digital economy and digital society have been collected from the Eurostat database during the period 2010–2020. To analyze the data, we applied a double principal component analysis.

These results allow us to identify the differences and similarities between countries and indicators from 2010–2020, and, more precisely, to study the countries' evolution trends and the evolution of the relations between the different indicators.

## 1.1. Introduction

This study allows us to understand how the countries in the European Union are more or less prepared for confronting the digital era. People and societies have to deal with these new challenges not only to gain knowledge but also to promote growth, development and opportunities for everyone. In particular, it is interesting to see whether the least developed countries are close or relatively distant from the others in this route, in order to develop new strategies and to overcome detected fragilities.

---

Chapter written by Fernanda Otilia FIGUEIREDO and Adelaide FIGUEIREDO.

For a color version of all the figures in this chapter, see [www.iste.co.uk/dimotikalis/data3.zip](http://www.iste.co.uk/dimotikalis/data3.zip).

G1: Activity and employment status of individuals V1 – % employed persons in the universe of active persons with ICT education V2 – % males in the universe of employed persons with ICT education V3 – % persons with tertiary education among employed persons with ICT education V4 – % persons aged 15–34 among the employed persons with ICT education
G2: Internet at households with population aged 16–74 years V5 – % households with Internet access V6 – % households with broadband Internet connection
G3: Internet use by individuals V7 – % individuals with daily Internet access in the last three months before the survey V8 – % individuals with Internet use in the last three months before the survey V9 – % individuals with Internet use in the last 12 months before the survey V10 – % individuals with Internet use more than a year ago before the survey V11 – % individuals who have never used Internet before the survey
G4: Online purchase by individuals V12 – % individuals with last online purchase in the last three months before the survey V13 – % individuals with last online purchase in the last 12 months before the survey
G5: Reasons of the individuals for using Internet V14 – % individuals using Internet for finding information about goods and services V15 – % individuals using Internet for looking for a job or sending a job application V16 – % individuals using Internet for selling goods and services
G6: Enterprises use for selling (having e-commerce) V17 – % enterprises with e-commerce sales of at least 1% turnover
G7: Enterprises that provided training to develop/upgrade ICT skills, excluding enterprises of the financial sector V18 – % small enterprises provided training to ICT/IT specialists V19 – % medium enterprises provided training to ICT/IT specialists V20 – % SMEs provided training to ICT/IT specialists V21 – % large enterprises provided training to ICT/IT specialists V22 – % small enterprises provided training to other employed persons V23 – % medium enterprises provided training to other employed persons V24 – % SMEs provided training to other employed persons V25 – % large enterprises provided training to other employed persons V26 – % small enterprises provided training to their personnel V27 – % medium enterprises provided training to their personnel V28 – % SMEs provided training to their personnel V29 – % large enterprises provided training to their personnel

**Table 1.1.** *Variables considered in the study grouped by topic*

### 1.1.1. The dataset

For this study, a large number of indicators associated with digital skills, digital economy and digital society (29 variables) have been considered, as well as all European countries for which we have data available (30 individuals). The data

for these indicators were collected from the Eurostat database during the period 2010–2020.

The variables are grouped by topic (groups G1–G7), and are listed in Table 1.1. When we refer to ICT (Information and Communication Technology) Education, in the general sense, it means providing users with a diverse set of technological tools, definitions and resources to create, store, communicate, manage and optimize information. The database aggregates the ICT education levels into two categories: upper-secondary and post-secondary non-tertiary (levels 3 and 4), and tertiary (levels 5–8). Concerning employed persons with ICT education, they are divided by two classes of age: 15–34 and 35–74. The enterprises are classified accordingly to the number of employees and self-employed persons as follows: small (10–49), medium (50–249), SMEs (10–249) and large (250 or more).

Of course, it would be interesting to include other indicators such as the percentage of the ICT sector in GPD, the R&D personnel in the ICT sector as a percentage of total R&D personnel or the business expenditure on R&D in the ICT sector as a percentage of total R&D expenditure, for instance. However, we do not have information for all countries and years during the period.

The countries (individuals) considered in this study are presented in Table 1.2. We also further analyzed six data tables, corresponding to the years 2010, 2012, 2014, 2016, 2018 and 2020, each one with the same countries and variables. As the United Kingdom was not part of the EU in 2020, we considered the values obtained in 2019, instead of 2020, so we could still include it in the analysis.

Countries	Countries	Countries
EU27 – European Union 27	FR – France	AT – Austria
BE – Belgium	HR – Croatia	PL – Poland
BG – Bulgaria	IT – Italy	PT – Portugal
CZ – Czechia	CY – Cyprus	RO – Romania
DK – Denmark	LV – Latvia	SI – Slovenia
DE – Germany	LT – Lithuania	SK – Slovakia
EE – Estonia	LU – Luxembourg	FI – Finland
IE – Ireland	HU – Hungary	SE – Sweden
GR – Greece	MT – Malta	UK – United Kingdom
ES – Spain	NL – Netherlands	NO – Norway

**Table 1.2.** *Countries considered in the study*

### 1.1.2. Preliminary analysis of the data

Before proceeding with a multivariate data analysis, we carried out an exploratory analysis in order to gain insight into the data. In particular, we represented the boxplots

by year and for all of the variables (not presented here) and among other conclusions, it is worth mentioning here the huge number of outliers observed in the data, and reported in Table 1.3. The countries with an \* are severe outliers. Some countries appear as outliers with respect to some variables during several years of the period 2010–2020, such as GR, RO and BG: GR is a severe lower outlier in variable V1 (the percentage of employed persons in the universe of active persons with ICT education is small comparatively to the corresponding percentage in the other countries); RO and BG are lower outliers in variables V21, V25 and V29 (the percentage of large enterprises providing training in ICT skills to their personnel, specialists or other employed persons is small when compared to what happens in other countries). We also mention, for instance, DK and FI, who appear during some years of the period as upper outliers in variable V15 (with a large percentage of individuals using the Internet to look for a job or sending a job application compared to other countries), and countries such as BE and FI, who in 2020 were upper outliers in variables V19, V23 and V27 (i.e. they have a large percentage of medium enterprises providing ICT training when compared with other countries). Other details can be found in Table 1.3.

Year	Variables and associated outliers
2010	Lower V1(GR), V3(IT), V21(RO), V29(RO)
2012	Lower V1(ES,GR*), V21(RO), V29(RO) Upper V10(CZ)
2014	Lower V1(PT,GR*), V7(RO), V21(RO), V29(RO) Upper V10(BG)
2016	Lower V1(GR*), V2(CY), V7(RO), V21(RO), V25(BG,RO), V29(BG,RO) Upper V15(DK), V17(IE)
2018	Lower V1(GR*,IT,PT), V3(PT), V21(RO), V25(BG,RO), V29(RO) Upper V26(NO)
2020	Lower V1(GR*,ES), V2(DK), V14(IT,RO), V19(RO), V21(RO), V23(BG,RO), V25(BG,RO), V27(BG,RO), V29(RO) Upper V10(BG,IT), V15(DK,FI), V19(BE), V23(FI), V27(FI)

**Table 1.3.** *Variables that present several outliers along the period*

After an introduction, aiming to present this study and detailed descriptions of the dataset, as well as a preliminary data analysis, this chapter is organized as follows. Section 1.2 gives a small description of the double principal component analysis method, the statistical methodology applied in this work, and also includes part of the results obtained from the application of this method to the data. The representation and interpretation of the trajectories are provided in section 1.3, and section 1.4 concludes the chapter with some final conclusions.



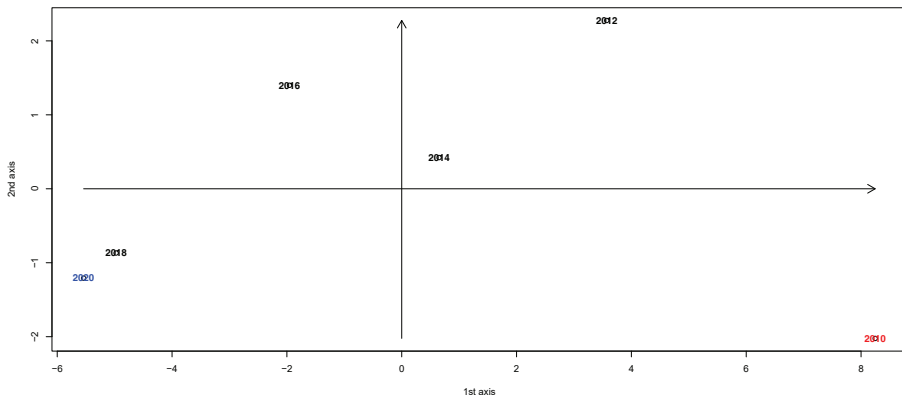
## 1.2. Double principal component analysis: a brief description

To analyze the data, we applied a double principal component analysis (DPCA), a method of multivariate data analysis introduced in Bouroche (1975) to analyze three-way data with quantitative variables. This method can be modified to allow the analysis of categorical data. See, for instance, Lera et al. (2006).

DPCA is an extension of the principal component analysis (PCA) method, and allows us to jointly analyze several data tables with information collected on the same variables and individuals, in several instants of time, for instance. In this study, we have six data tables, corresponding to the years 2010, 2012, 2014, 2016, 2018 and 2020, all with the same countries and variables.

The application of the DPCA method comprises four steps: (1) interstructure, i.e. the analysis of the global evolution of the data along the period of time; (2) analysis of the clouds of individuals; (3) intrastructure, i.e. the choice of the best common space to represent the individuals and the variables; and (4) representation and interpretation of the trajectories of the individuals and variables.

From Figure 1.1, we can observe that the years appear in chronological order along the first axis. As can be expected, there are big changes between the periods of 2010 and 2020.



**Figure 1.1.** *Representation of the interstructure*

The analysis of individuals' clouds and, in particular, the percentage of explained inertia when they are projected in a subspace of a smaller dimension, lead us to select the plan of the first two principal axes obtained from the PCA of the year 2014 to represent the individuals and variables for all years. This system of principal axes minimizes the average loss of information, i.e. it globally maximizes the percentage of explained inertia. The projection of the countries and variables in this plan globally

explains 76.6% of the total inertia. If we consider three axes, the percentage of the explained inertia increases to 81%, but the balance between the gain and the increase of complexity to interpret the trajectories lead us to choose only two axes. The results corresponding to the last step of the method are presented in the next section.

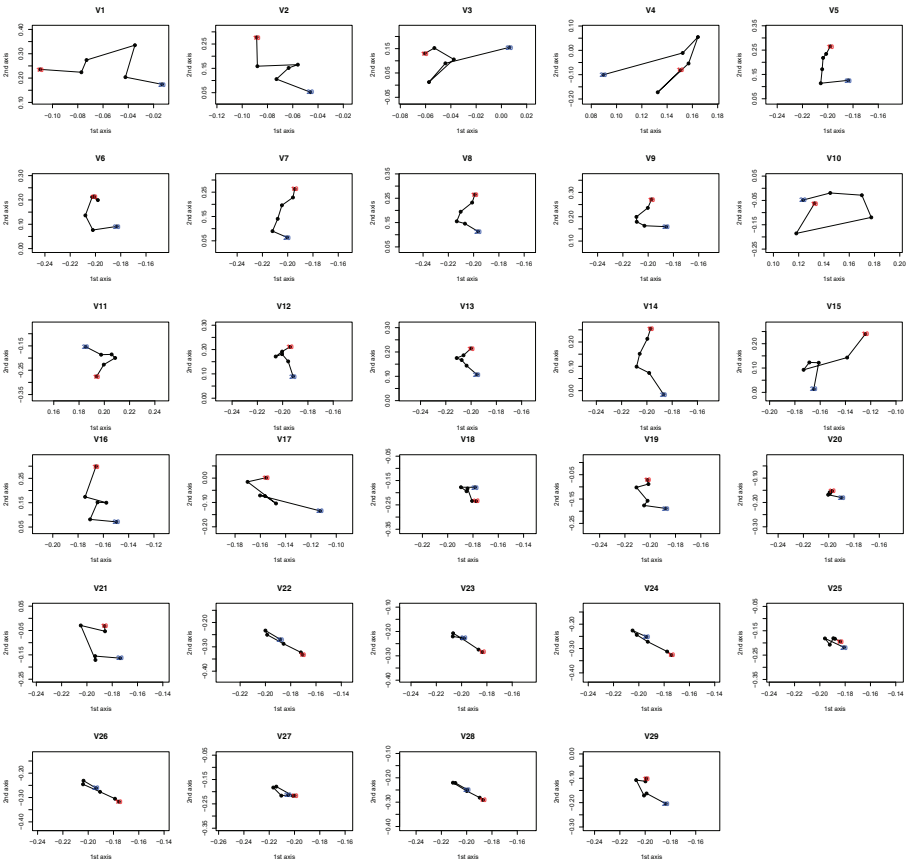
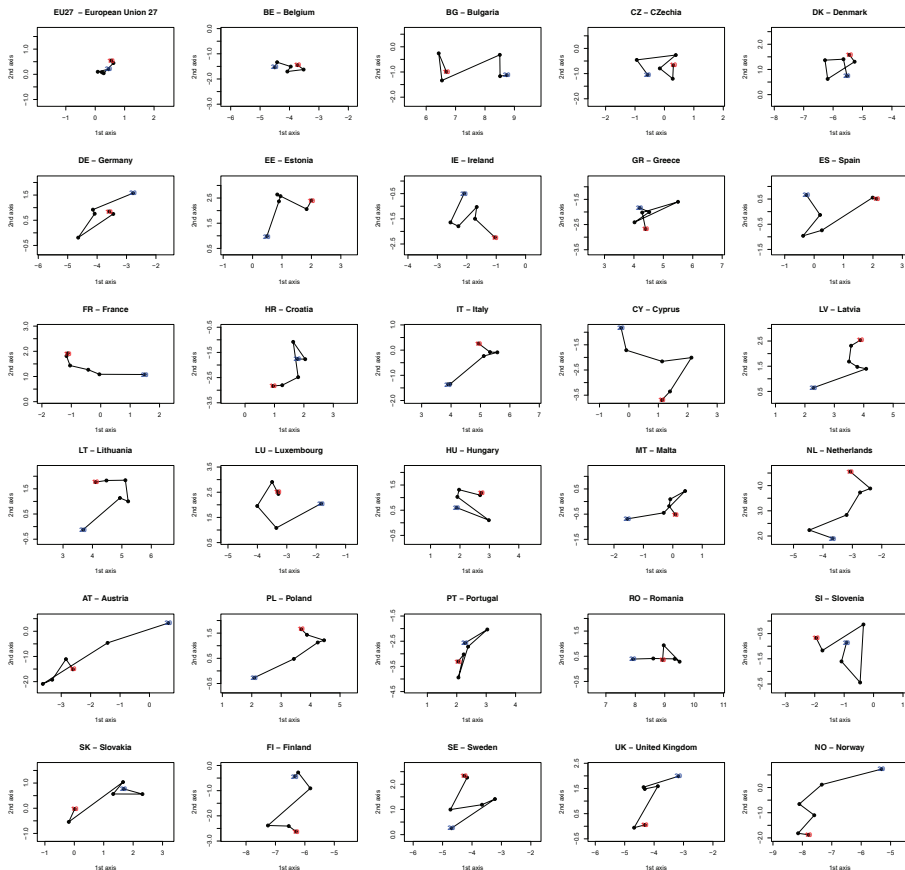


Figure 1.2. Variables' trajectories

1.3. Representation of the trajectories

The representation and interpretation of the trajectories is an important step in DPCA. It allows us to identify the differences and similarities between individuals (countries), and analyze the evolution of the correlations between the different variables during the 2010–2020 period. These trajectories are presented in Figures 1.2 and 1.3. All of the graphics have the same scale, i.e. the range in the first axis is the same, as well as the range in the second axis, to have comparable graphics. We have highlighted the year 2010 in red and the year 2020 in blue.

Looking at Figure 1.2, among the variables that present closed trajectories (i.e. they are variables with correlation values with the other variables stable along the period), we highlight the variables V18, V20, V25 and V27, related to the percentage of small or medium enterprises providing training in ICT skills to specialists or to their personnel employees or large enterprises who provide training to other employed persons. All of the other variables have extended trajectories, i.e. they are variables with relevant changes in their correlations with the others along the period. Among them, we highlight the variables V1, V2, V3 and V4 from group 1 (Activity and employment status of individuals), the variable V10 – % of individuals with Internet more than a year ago, the variables V15 and V16 from group 5 (Reasons for using Internet) and the variable V17 – % enterprises with e-commerce having received orders online, at least 1%.

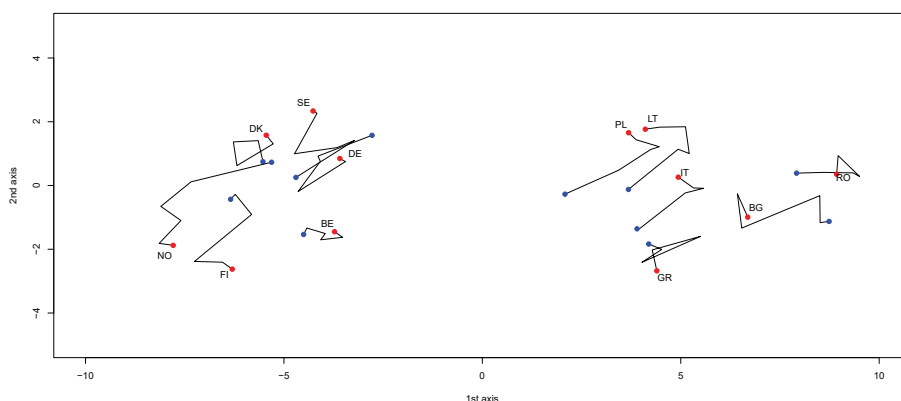


**Figure 1.3.** *Countries' trajectories*

As shown in Figure 1.3, in general, most of the countries present large or extended trajectories, at least along one of the axes, and in particular, we highlight the trajectories of BG, ES, FR, CY, NL, AT, PL and NO. With very closed trajectories, we highlight EU27 and BE. The countries PT and GR also exhibit some instability in their trajectories.

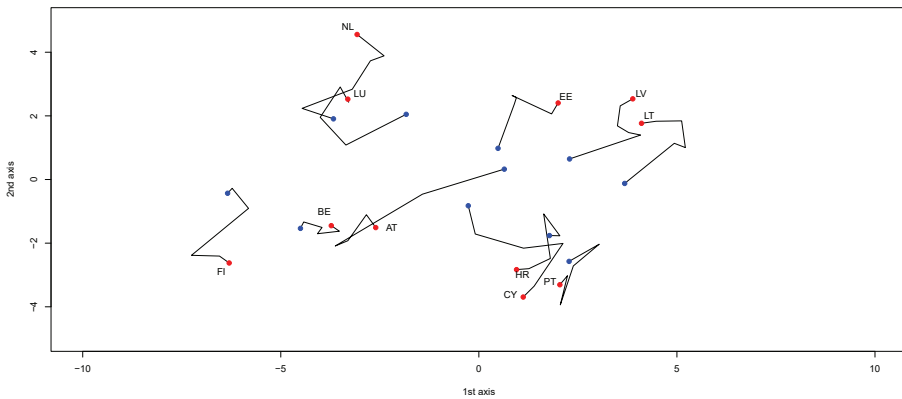
For the interpretation of the trajectories, we only considered the countries that contributed the most to the axes, with their trajectories represented together in Figures 1.4–1.5, and the variables correlated more so with them.

Concerning the first axis, which explains 65.3% of the total inertia, almost all of the variables have high correlations with the axis in all years, except V1, V2 and V3. Therefore, we only highlight the variables with correlations above 0.8 in absolute value. The first axis opposes the group of countries NO, FI, DK, SE, DE and BE that have large values in the variables V5–V7, V9, V12–V14 and V19–V29 and small values in variable V11, to the group of countries RO, BG, IT, GR, LT and PL that present large values in the variable V11 and small values in the others.



**Figure 1.4.** *Interpretation of trajectories in the first axis (65.3% inertia)*

The second axis only explains 11.3% of the total inertia, and few variables have significant correlations with this axis. Thus, we highlight the variables with correlations above 0.5 in absolute value. The second axis opposes the group of countries PT, HR, FI, CY, AT and BE, all of which have high values in the variables V24, V26 and V28 (years 2010, 2012 and 2014), V23–V24 (years 2010 and 2012) and small values in variable V1 (years 2010, 2012, 2014 and 2016), compared to the group of countries NL, LU, EE, LT and LV, all of which present high values in variable V1 and small values in the others.



**Figure 1.5.** *Interpretation of trajectories in the second axis (11.3% inertia)*

#### 1.4. Some final conclusions

The results of this study allow us to identify some existing differences or similarities between groups of European countries during the period 2010–2020 under analysis, concerning ICT skills and employment, Internet usage and digital development.

As expected, this study confirms our perception that the countries of the north and central Europe are very prepared for confronting the challenges of the digital age, since they have already greatly invested in providing training in ICT skills to their populations and how everyone is comfortable using the Internet for a variety of purposes: communication, purchase, selling, finding, and so on. Among these countries, the most similar are the following subgroups of countries: (DE and SE), (NO and DK), FI and BE.

On the other hand, the countries of the East and South of Europe are less prepared, having made little investment in ICT education and, in most of them, a large part of the population has never used the Internet. Among these countries, the closest are the subgroups (PL and LT), (IT and GR), BG and RO.

We also observe that in countries such as PT, AT, CY, HR, BE and FI, where the percentage of employed persons with ICT education is yet small, a large percentage of small and medium enterprises provide training to their personnel or to other employed persons to develop or upgrade ICT skills. The opposite happens, for instance, in the countries LT, LV, EE, LU and NL.

## 1.5. Acknowledgements

This work was supported by the National Funds through the Portuguese funding agency, FCT – Fundação para a Ciência e a Tecnologia, within the framework of the projects UIDB/00006/2020 (CEAUL) and LA/P/0063/2020 (INESC TEC).

## 1.6. References

- Bouroche, J.M. (1975). Analyse des donnés ternaires : la double analyse en composantes principales. Third-cycle PhD Thesis, Université de Paris VI, Paris.
- Lera, L., Pérez, R., Bouquet, A. (2006). El doble análisis em componentes principales para datos categóricos y su aplicación en un estudio de migración. *Revista Colombiana de Estadística*, 29(1), 17–34.

---

## Multivariate Kernel Discrimination Applied to Bank Loan Classification

---

The purpose of this chapter is to apply a kernel discriminant analysis to classify bank loans and determine which loans are at risk of default. This study begins by introducing the concept of kernel density estimation, which is a widely used non-parametric technique to obtain an estimate for the probability density function. This procedure is based on two main parameters: the kernel function and the bandwidth, the latter being the crucial parameter. The multivariate kernel density estimator is later applied to discriminant analysis to obtain kernel discrimination. This is a method that classifies observations into a predetermined number of distinct and disjointed classes. Finally, we apply multivariate kernel discriminant analysis to a sample of bank loans to determine which loans can be classified as defaulted. This model can help predict the likelihood that future loans may default.

### 2.1. Introduction

As the name implies, credit risk analysis is used to assess the likelihood of a borrower's repayment failure and the loss caused to the financier when default occurs. This concept greatly affects the long-term success of any bank or financial institution. In Malta, the non-performing loan ratio stood at 3% in December 2019 (Central Bank of Malta). This ratio reflects the country's credit quality of loans. Banks need to determine the probability of non-performing loans of companies to decrease the prospect of weakened liquidity. The aim of this study is to create a model that predicts whether a company will be able to pay its outstanding loan, based on several variables. To obtain such a model, we apply the multivariate kernel discriminant analysis. This technique uses kernel density estimation, as well as non-parametric discriminant analysis.

Rosenblatt (1956), Whittle (1958), Parzen (1962) and Watson and Leadbetter (1963) studied and presented a number of results regarding univariate kernel density estimation. Moreover, Cacoullos (1966) extended Parzen's results to the multivariate case. Loftsgaarden and Quesenberry (1965) and Epanechnikov (1969) also applied multivariate non-parametric kernel density estimation.

The quality of multivariate kernel density estimation depends mainly on the choice of the bandwidth matrices. The univariate plug-in bandwidth estimator provided by Sheather and Jones (1991) was extended to the multivariate case by Wand and Jones (1993). Moreover, a number of authors, including Chacón and Duong (2010, 2012) and Horová et al. (2013), also contributed to the field of optimal bandwidth selection by applying a number of novel techniques.

Multivariate kernel density estimation has been applied to discriminant analysis. The properties of this technique feature in Murphy and Maron (1986) and Silverman (1986). Over the past several years, kernel discriminant analysis has been applied to several fields. For example, Liberati et al. (2012) and Widiharhi et al. (2018) both used kernel discriminant analysis to address the problems of credit risk analysis in banking and making efficient credit granting decisions. Some further applications of kernel discriminant analysis include Hand (1983) and Chen et al. (2005) in medicine, Martin (2011) in agriculture and Farida and Aidi (2016) in education.

This section concludes by giving a brief overview of the structure of this chapter. In section 2.2, we discuss multivariate kernel density estimation. In section 2.3, we outline the properties of kernel discriminant analysis. In section 2.4, we apply these techniques to a local dataset that contains not only the data of a number of local companies but also the status of their bank loan (whether the company is in default or otherwise). Further details concerning the said dataset are provided in section 2.4. As mentioned earlier, the main aim will be to create a model which uses kernel discriminant analysis, which can identify firms that are close to bankruptcy. In this section, we split the data into the training set and test set. We train the model using the former dataset and test the model performance using the latter dataset. Section 2.5 contains some concluding remarks and suggestions for future research.

## 2.2. Multivariate kernel density estimation

Let  $p$  denote the total number of variables in the dataset and let  $n$  denote the total number of companies in the dataset.

Let  $X_i$  be a random variable representing the  $i^{\text{th}}$  variable. Moreover, let  $\mathbf{X}$  be a random vector, where  $\mathbf{X} = (X_1, X_2, \dots, X_p)^T$  and  $\mathbf{x}_n = (x_{n1}, x_{n2}, \dots, x_{np})^T$  will be



the vector that contains all of the observations of the  $n^{\text{th}}$  company. Let  $\mathbf{x}$  be the  $n \times p$  matrix which contains all the observations of the dataset, such that:

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_{11} & \cdots & \mathbf{x}_{1p} \\ \vdots & \ddots & \vdots \\ \mathbf{x}_{n1} & \cdots & \mathbf{x}_{np} \end{pmatrix}.$$

Multivariate kernel density estimation is composed of two main components: the kernel function, which will now be denoted by  $K$ , and the bandwidth, which will now be denoted by  $\mathbf{H}$ . In a multivariate setting, the bandwidth is a  $p \times p$  matrix, where  $p$  represents the dimension of the data, i.e. the number of variables being studied. The approach chosen for multivariate kernel density estimation highly depends on the type of bandwidth matrix considered. When  $\mathbf{H}$  is taken to be diagonal, this is usually referred to as a constrained bandwidth matrix. Otherwise, it is referred to as an unconstrained bandwidth matrix. The diagonal bandwidth matrix is thus a special case of the unconstrained bandwidth matrix.

### 2.2.1. Kernel density estimator

For the multivariate case with a bandwidth matrix  $\mathbf{H}$ , the kernel density estimator is defined as follows:

$$\hat{f}(\mathbf{x}; \mathbf{H}, \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n) = \frac{1}{n} \sum_{i=1}^n K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i),$$

where  $n$  represents the sample size and  $K_{\mathbf{H}}(\mathbf{x})$  is said to be the scaled kernel, which is defined as:

$$K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-\frac{1}{2}} K\left(\mathbf{H}^{-\frac{1}{2}} \mathbf{x}\right),$$

where  $|\mathbf{H}|^{-\frac{1}{2}}$  is the inverse of the square root of the determinant of the matrix  $\mathbf{H}$  and  $\mathbf{H}^{-\frac{1}{2}}$  is the inverse of the matrix square root. Therefore, the kernel density estimator can be written as:

$$\hat{f}(\mathbf{x}; \mathbf{H}, \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n) = \frac{1}{n|\mathbf{H}|^{\frac{1}{2}}} \sum_{i=1}^n K\left(\mathbf{H}^{-\frac{1}{2}}(\mathbf{x} - \mathbf{X}_i)\right),$$

where  $K$  is the multivariate kernel function.  $K$  is assumed to be the symmetric function. Furthermore,  $\int_{\mathbb{R}^p} K(\mathbf{z}) d\mathbf{z} = 1$  and  $K(\mathbf{z}) \geq 0 \forall \mathbf{z} \in \mathbb{R}^p$ , implying that  $K$  satisfies the properties of a multivariate probability density function. There are various types of functions that satisfy the above-mentioned properties, which include the standard multivariate normal kernel density function and the multivariate