SYMPOSIUM ON BIG DATA IN FINANCE, RETAIL AND COMMERCE

Statistical and Computational Challenges

BOOK OF ABSTRACTS

and

EXTENDED ABSTRACTS

2-3 November, 2017

Hotel Jupiter

Lisbon, Portugal

SYMPOSIUM ON BIG DATA IN FINANCE, RETAIL AND COMMERCE:

Statistical and Computational Challenges

Book of Abstracts
and
Extended Abstracts

Lisbon, 2-3 November 2017

Editors
Manuel Scotto
Lisete Sousa
Patricia de Zea Bermúdez
Daniel Peña

Editors: Manuel Scotto, Lisete Sousa, Patricia de Zea Bermúdez and Daniel Peña

Title: Symposium on Big Data in Finance, Retails and Commerce: Statistical and Computational Challenges

ISBN: 978-989-733-059-9

Depósito Legal: 433165/17

Local Organizing Committee

K. F. Turkman - Universidade de Lisboa
Lisete Sousa - Universidade de Lisboa
Patrícia de Zea Bermudez - Universidade de Lisboa
Manuel Scotto - Universidade de Lisboa
M. A. Amaral Turkman - Universidade de Lisboa

Scientific Committee

Daniel Peña - Universidad Carlos III de Madrid
Emílio Carrizosa - President of the SEIO, Universidad de Sevilla
Maria Eduarda Silva - President of the SPE, Universidade do Porto
Nuno Crato - Universidade de Lisboa and JRC of the EC, Italy
Isabel Fraga Alves - Universidade de Lisboa
K. F. Turkman - Universidade de Lisboa

Sponsors

Centro de Estatística e Aplicações - Universidade de Lisboa
Faculdade de Ciências - Universidade de Lisboa
Universidad Carlos III de Madrid
Cátedra Luis de Camoens UC3M
Centro Internacional de Matemática - CIM
Sociedade Portuguesa de Estatística
Sociedad de Estadística e Investigación Operativa
Fundação para a Ciência e a Tecnologia - FCT

Foreword

This book contains abstracts and extended abstracts (invited and contributed) presented at the *Symposium on Big Data in Finance, Retail and Commerce: Statistical and Computational Challenges* that will take place in Lisbon (Portugal) from 2-3 November 2017. The Symposium is a joint organization of the Institute of Financial Big Data, Universidad Carlos III de Madrid, Centro de Estatística e Aplicações, Universidade de Lisboa (CEAUL), Sociedad Española de Estadística e Investigación Operativa (SEIO) and Sociedade Portuguesa de Estatística (SPE).

The objective of the symposium is to present new approaches to deal with big data applications in finance, retail and commerce bringing together professionals from companies in these areas and researchers who generate sound methods and tools to handle such data. It is hoped that such an interaction may help in understanding the new needs and trends in the emerging field of Data Sciences.

We would like to express a very special thanks to the invited speakers, Peter Diggle from Lancaster University (UK), João Marques Silva from Universidade de Lisboa (Portugal), Cristina San José from Banco Santander (Spain), Marc Halin from Université Libre de Bruxelles (Belgium), Tiago Pereira Durão from Deloitte (Portugal), Vicente Calzado from El Corte Inglés (Spain) and Nuno Crato from Universidade de Lisboa (Portugal) and Joint Research Centre of the European Commission (Italy). We would also like to extend our thanks to all the participants with special emphasis to those who will enhance the quality of this symposium with their contributed talks and posters.

Finally the realization of this symposium was only possible thanks to the generous contribution of our sponsors, the Institute of Financial Big Data, Universidad Carlos III de Madrid, Centro de Estatística e Aplicações, Universidade de Lisboa (CEAUL), Sociedad Española de

Estadística e Investigación Operativa (SEIO), Sociedade Portuguesa de Estatística (SPE), Cátedra Luis de Camoens UC3M, Centro Internacional de Matemática (CIM), Fundação para a Ciência e a Tecnologia, Portugal (FCT) and Faculdade de Ciências da Universidade de Lisboa.

The Editors
Lisbon, November of 2017

INVITED SPEAKERS

Peter Diggle - Lancaster University, UK

João Marques Silva - Universidade de Lisboa, Portugal

Cristina San José - Banco Santander, España

Marc Hallin - Université Libre de Bruxelles, Belgium

Tiago Pereira Durão - Deloitte, Portugal

Vicente Calzado - El Corte Inglés, España

Nuno Crato - Universidade de Lisboa, Portugal, and Joint Research Centre of the European Commission, Italy

Contents

Contributed Talks

Posters

PROGRAMME

Programme

Thursday 2

8:30 - 9:15	Registration
9:15 - 9:30	Welcome Session
9:30 - 10:30	**Invited lecture 1** Chair: Nuno Crato
	Peter Diggle
	Statistics: a data science for the twenty-first century
10:30 - 11:00	Coffee break
11:00 - 12:00	**Invited lecture 2** Chair: Eduarda Silva
	João Marques Silva
	Recent trends in machine learning and data mining
12:00 - 13:00	**Invited lecture 3** Chair: Daniel Peña
	Cristina San José
	Opportunities for machine learning in financial institutions
13:00 - 14:20	Lunch
14:30 - 16:10	**Contributed Talks:** Big Data Methodology
	Chair: Lisete Sousa
	João Branco
	High dimensionality: big trouble for big data scientists
	Laura Borrajo
	Nonparametric mean estimation for big-but-biased data

	Luísa Canto e Castro
	Education's big data: management and statistical challenges
	Maria Costa
	Maximum entropy in inhomogeneous large-scale data
	Paula Brito
	Analysing (micro) big data at the macro level: the case of risk categories of loan data
16:10 - 17:10	**Invited lecture 4** Chair: Isabel Fraga Alves
	Marc Hallin
	Time series in high dimension
	General dynamic factor models
17:15 - 19:00	**Cocktail and Poster Session**
	Ersin Yilmaz
	Fernanda Figueiredo
	Joaquim Costa
	Matilde Oliveira
	Raquel Fonseca
	Rômulo Soares
	Sílvia Rebouças

Friday 3

9:00 - 10:00	**Invited lecture 5** Chair: Feridun Turkman
	Tiago Pereira Durão
	How financial institutions are leveraging big data to change they interact with customers
10:00 - 11:00	**Invited lecture 6** Chair: Emilio Carrizosa
	Vicente Calzado
	The value of data for the retail and commerce business
11:00 - 11:30	Coffee break
11:30 - 13:10	**Contributed Talks:** Financial Applications and Time Series Chair: Patrícia de Zea Bermudez
	Alvaro Gómez Losada
	Forecasting on imbalanced big time series using ensembles of classifiers and minimum information: application to ozone data
	Carlos Trucíos
	Forecasting conditional covariance matrix via principal volatility components in the presence of additive outliers
	Lara Quijano-Sanchez
	The big chase: a decision support system for client acquisition applied to financial networks
	Raquel Gaspar
	Empirical simulation analytics in financial engineering
	Renata Alcoforado
	Text mining and ruin theory: a case study on risk models with dependence

13:10 - 14:30	Lunch
14:30 - 16:10	**Contributed Talks:** Miscellaneous Applications
	Chair: Manuel Scotto
	João Lopes
	Mismatch between jobs and skills in the EU
	João Romacho
	Mutual fund competition in European union
	M. Remedios Sillero-Denamiel
	Sparse and constrained naïve Bayes for cost-sensitive classification
	Sandra Benítez Peña
	Sparse support vector machines with performance constraints
	Albert Satorra
	Exploratory (big) data analysis
16:10 - 17:10	**Invited lecture 7** Chair: Daniel Peña
	Nuno Crato
	The role of big data and administrative data in statistical social research and counterfactual impact assessment
17:15 - 18:00	Coffee break and prize for the best oral communication

INVITED LECTURES

Statistics: a data science for the twenty-first century

Peter Diggle
CHICAS, Lancaster University Medical School, UK,
p.diggle@lancaster.ac.uk

Abstract
The rise of data science could be seen as a potential threat to the long-term status of the statistics discipline. I will argue that, although there is a threat, there is also a much greater opportunity to re-emphasize the universal relevance of statistical method to the interpretation of data. I will give a short historical outline of the increasingly important links between statistics and information technology. I then describe several applications from my own field of research, biostatistics, through which I hope to demonstrate that statistics makes an essential, but incomplete, contribution to the emerging field of "electronic health" research. Finally, I offer personal thoughts on how statistics might best be organized in a research-led university and on what we should teach our students.

Recent trends in machine learning and data mining

João Marques Silva
Faculdade de Ciências, Universidade de Lisboa, Portugal,
jpms@ciencias.ulisboa.pt

Abstract
Artificial intelligence, mostly embodied by the advances in machine learning, is revolutionizing our lives, and will continue to do so in years to come.

Many universities, mainly through their computing and statistics departments, are adapting to this new reality, offering new graduate and undergraduate courses, but also reshaping their research focus.

This talk briefly overviews the ongoing advances in machine learning, and outlines a number of novel research directions. These not only reveal far reaching challenges, but are also of key relevance for companies.

Concretely, the talk overviews ongoing research in explainable artificial intelligence, highlights the issues with adversarial attacks on machine learning models, and also summarizes the efforts for deploying machine learning models in resource-constrained settings.

Opportunities for machine learning in financial institutions

Cristina San José
Banco Santander, Spain,
cristina.sanjosebrosa@gruposantander.com

Abstract
A practitioner's view on Big Data as an enabler of Financial Institutions' strategic priorities – what are the opportunities, how to accelerate and the roadblocks on the way.

Time series in high dimension General dynamic factor models

Mark Hallin
ECARES, Universite Libre de Bruxelles, Belgium,
mhallin@ulb.ac.be

Abstract
The analysis of high-dimensional data in the past few years has become one of the most active subjects of modern statistical methodology, and a central problem in the *big data* revolution. The reason is that information increasingly often takes the form of T observations with values in n-dimensional real spaces, where n is quite large, often of the same magnitude as T. This is true, particularly, in Economics and Econometrics, and in Finance: a macroeconometric analysis of the European Union, for instance, requires a joint study of at least 20 economic series in 28 member states, yielding a time series in dimension 560; the Standard & Poor's 500 is a panel of 500 stocks selected by economists as a representative subset of U.S. equities—hence a time series in dimension 500. Even the simplest parametric approaches in this context are helpless: a VAR(1) model in dimension 500 requires the estimation of 250,000 autoregressive parameters, along with a 500×500 innovation covariance matrix. The situation is even worse in finance, where volatilities are the main point of interest. All domains of applications, however, are facing similar challenges: genetics, chemometrics, environmental studies, image analysis, etc. A successful class of methods has emerged, mainly from the econometric literature: the so-called Dynamic Factor Models. In this talk, we provide a non-technical presentation of the methodological foundations of those models, contrasting such concepts as commonality and idiosyncrasy, factors and common shocks, dynamic and static principal components. Econometric and financial illustrations will be provided.

How financial institutions are leveraging big data to change they interact with customers

Tiago Pereira Durão
Deloitte, Portugal, *tdurao@deloitte.pt*

Abstract
Modern Organizations are facing challenges and opportunities imposed by market changes and strong competition. Rapid growth, disparate solutions and the increased demand for information, all complicate the ability to build enterprise-wide momentum. The challenge is further complicated by the very nature of business: the amount of data generated and the need to support rapid decision-making.

The investment in analytical solutions is essential to create the ability to process near real-time data and measure operators business and technical performance insightful and make critical decisions that directly impact global strategy.

Financial services institutions are going through a rough period, with declining profit margins, increasing regulation, still managing the impacts of the recent financial crises and making their best to answer all new customer expectations brought by a new digital type of customer who values customer experience.

At Deloitte we have the opportunity and privilege of working with the top companies in Financial Services and help them address these topics. This presentation will share our experience and our innovation initiatives for this industry.

The value of data for the retail and commerce business

Vicente Calzado
El Corte Inglés, Spain, *vicente_calzado@iecisa.com*

Abstract
The Retail industry is one of the most affected by the constant changes and challenges it faces, in a scenario defined by:

- The need to respond quickly to changes in consumer habits in new digital customers, highly informed, hyper connected and very demanding
- The convergence of traditional channels with new on-line commerce channels
- The emergence of a new digital competition derived from new e-commerce players
- The high price pressure with its impact on margins and the need to reduce costs.

In order to face the challenges of this scenario, it is necessary to carry out a digital transformation of the business in which one of the main strategic objectives must be to take advantage of the value that can be provided by the increasing amount of internal and external data that are produced in the retail business. Retailers have been using data analytics to generate business intelligence for years, however the high complexity of the data that is currently generated in business needs new solutions and tools.

As retailers include multiple channels of sales and communication with their customers, digital advertising and social media, the amount of information both internal and external increases exponentially. Extracting the value that such data can bring to the business can

be the key to competing in this complex and dynamic scenario. Big Data Analytics applications in this sector have to focus on four complementary lines of action:

1. Customer engagement based on the customer's knowledge, preferences and behavior analysis to be able to provide a satisfactory purchase experience through all possible channels and engage in effective digital marketing and promotions.

2. Configuration of the offer, through the selection of products and services that are adapted at all times to the preferences of customers, with price optimization techniques that guarantee competitiveness and a suitable design of both physical and online stores

3. Efficiency in operations, particularly in the supply chain and inventory management, making predictions based on stock data in stores and improving logistics.

4. Strategic planning, management and sales forecasts that allow a comprehensive optimization of operations and management for a more appropriate design of the "go-to-market" strategy.

In short, Retail is a data-driven business in which extracting the value of the huge amount of data generated in commercial activity and its application to all business processes is a key strategy towards competitiveness and growth.

Keywords: Retail; e-commerce; Analytics; Customer experience; Customer engagement; Inventory management; Digital marketing; Supply chain; Sales forecasting

The role of big data and administrative data in statistical social research and counterfactual impact assessment

Nuno Crato
University of Lisbon, Portugal, and European Commission Joint Research Center, Italy,
Nuno.CRATO@ec.europa.eu

Abstract
National administrations, European institutions and public organisations collect, supervise, and keep track of extremely varied and extensive types of data. Modern technologies and better organized civil lives have facilitated the collection and custody of these data on a scale previously unknown. This opens novel perspectives to our daily lives, but also allows for a much more detailed and sound knowledge of our economies and our societies. In what follows, I will deal with an important issue within this general framework: the case for the collection, treatment, availability and use of micro-data, in particular administrative data (admin-data). I will urge to take action on the use of these data for a better, easier and more cost-effective evaluation of policies. In fact, there is already an incredible wealth of available data that opens the door to a better knowledge of our economies and our societies. If these data are well kept, organized, complemented, and linked, if data are regularly updated and used for the knowledge of the economic and social situation, if policy measures are recurrently evaluated and adjusted on the basis of this information, and if appropriate scientific methods are used, then our societies can make a better use of their resources and our policy measures can be more efficient. At a moment of increasing attention to the efficacy of public spending, of increasing scrutiny over the effects of policies, it is more important than ever to be able to understand our society and to know how policy measures are impacting over our lives. We have the data, we have the means, and

we have the necessary scientific methods. We have to do it.

CONTRIBUTED TALKS

High dimensionality: big trouble for big data scientists

João Branco
CEMAT and IST, Lisbon, *jbranco@math.ist.utl.pt*
Ana Pires
CEMAT and IST, Lisbon, *apires@math.ist.utl.pt*

Abstract
The recent massive production of high-dimensional data has brought great difficulties and concomitant challenges to statistics since its usual methods were not designed to cope with such kind of data. High dimensionality triggers the curse of dimensionality and unexpected behavior of some statistical tools may surprise even those aware of the intricacies of multidimensional spaces with a large number of dimensions.

We look at the Mahalanobis distance, a tool that is crucial to the functioning of the traditional multivariate statistical methods, and see that when the number of variables p approaches the number of observations n it becomes degenerated and loses its properties, so essential for the analysis of multivariate data. We conclude that if scientists are going to ignore fundamental results arrived at in this research and blindly use software to analyze data, the results of their analyses may not be trustful, and the findings of their experiments may never be validated.

Keywords: Curse of dimensionality; High dimensional data; Mahalanobis distance

Nonparametric mean estimation for big-but-biased data

Laura Borrajo
Research Group MODES, Department of Mathematics, CITIC, Campus de Elviña, Universidade da Coruña, 15071 A Coruña, Spain,
laura.borrajo@udc.es

Ricardo Cao
Research Group MODES, Department of Mathematics, CITIC and ITMATI, Campus de Elviña, Universidade da Coruña, 15071 A Coruña, Spain,
ricardo.cao@udc.es

Abstract
Crawford [2] has recently warned about the risks of the sentence *with enough data, the numbers speak for themselves.* Some of the problems coming from ignoring sampling bias in big data statistical analysis have been recently reported by Cao [1]. The problem of nonparametric statistical inference in big data under the presence of sampling bias is considered in this work. The mean estimation problem is studied in this setup, in a nonparametric framework, when the biasing weight function is known (unrealistic) as well as for unknown weight functions (realistic). Two different scenarios are considered to remedy the problem of ignoring the weight function: (i) having a small sized simple random sample of the real population and (ii) having observed a sample from a doubly biased distribution. In both cases the problem is related to nonparametric density estimation. Asymptotic expressions for the mean squared error of the estimators proposed are considered. This leads to some asymptotic formula for the optimal smoothing parameter. Some simulations illustrate the performance of the nonparametric methods proposed in this work.

Keywords: Bias correction; Big data; Kernel method; Mean esti-

mation; Nonparametric inference

References

[1] Cao, R. (2015) Inferencia estadística con datos de gran volumen. *La Gaceta de la RSME* **18**, 393–417

[2] Crawford, K. (2013) The hidden biases in big data. Harvard Business Review, April 1st. Available at https://hbr.org/2013/04/the-hidden-biases-in-big-data

Education's big data: management and statistical challenges

Luísa Canto e Castro
CEAUL and University of Lisbon, and Directorate-General for Education and Science Statistics, *luisa.castro@dgeec.mec.pt*

Abstract
The presentation will report on the challenges the Directorate-General for Education and Science Statistics (DGEEC) is facing to manage the multiple databases of education and training and the efforts it has made to make them, as a whole, an integrated and coherent information system that allows useful analyzes and readings to the political decision, transparency in the publicity to the public and agile procedures in the availability to the researchers. Examples will be given in any of DGEEC's areas of intervention - initial education (preschool to upper secondary), higher education and science - and the main potentialities of the integration of the different data sources strategy will be explained, namely in the benchmarking of schools, courses, training areas and research areas, in the study of employability and continuation of studies, in the discontinuation of surveys with the consequent reduction of costs and burden on the respondents. In addition, the main impacts of non-matching of records during the cross-database processes will be mentioned, especially in relation to the calculation of statistical indicators where total matching is especially important, such as abandonment rates, courses completion rates in the envisaged time, continuing education rates and employability rates.

Keywords: Data base management; Raw data; Data Integration; Data Exploration; Cross-databases; Non-matching problems

Maximum entropy in inhomogeneous large-scale data

Maria da Conceição Costa
University of Aveiro and CIDMA, *lopescosta@ua.pt*
Pedro Macedo
University of Aveiro and CIDMA, *pmacedo@ua.pt*

Abstract
The term *Big Data* usually refers to datasets that are large in different ways: there many observations, many variables, or both, or data is recorded in different time regimes or taken from multiple sources. In this context, retaining optimal, or, at least, reasonably good statistical properties with a computationally efficient analysis becomes a challenge. Another difficult issue relates to dealing with inhomogeneous data that does not fit in the classical framework: data is neither i.i.d., exhibiting outliers or not belonging to same distribution, nor stationary, as time-varying effects my be present. Standard statistical (linear) models fail to capture inhomogeneity structure in data, compromising estimation and interpretation of model parameters, and, of course, prediction. On the other hand, statistical approaches for dealing with inhomogeneous data, such as varying-coefficient models, mixed effects models, mixture models or clusterwise regression models, are typically very computationally cumbersome. Sub-sampling and aggregation procedures may lead to a reduction in computational burden. Several aggregation procedures have been already proposed in literature, but it was not until recently that an aggregation procedure was proposed to deal with inhomogeneous large data-sets [1].

A new approach is proposed in this work, with the introduction of the concepts of Info-Metrics to the analysis of inhomogeneous large-scale data. As the science and practice of information processing with finite, noisy or incomplete information, Info-Metrics provides

the suitable mathematical foundation for inference in the above scenario [3], [4]. The framework of information-theoretic estimation methods is presented, along with some information measures. A new aggregation procedure in proposed, based on the normalized entropy [2]. A simulation study is presented and preliminary results clearly indicate that normalized entropy methods provide very satisfactory solutions, when compared to standard aggregation techniques.

Keywords: Big data; Info-Metrics; Maximum entropy

References

[1] Bühlmann, P. and Meinshausen, N. (2016) Magging: maximin aggregation for inhomogeneous large-scale data. *Proceedings of the IEEE* **104**, 126–135.

[2] Golan, A., Judge, G. and Miller, D. (1996)*Maximum Entropy Econometrics: Robust Estimation with Limited Data*. John Wiley & Sons, Chichester.

[3] Golan, A. (2006) Information and Entropy Econometrics - A Review and Synthesis. *Foundations and Trends in Econometrics* **2**, 1–145.

[4] Golan, A. (2012) On the Foundations and Philosophy of Info-Metrics. *Lecture Notes in Computer in Computer Science* 7318, 238–245.

Analysing (micro) big data at the macro level: the case of risk categories of loan data

Paula Brito
Fac. Economia & LIAAD-INESC TEC, Univ. Porto, Portugal,
mpbrito@fep.up.pt
A. Pedro Duarte Silva
Católica Porto Business School, & CEGE, Univ. Católica Portuguesa, Porto, Portugal,
psilva@porto.ucp.pt

Abstract
In Data Mining, Multivariate Data Analysis and classical Statistics, data is usually represented in a data array where each row represents a "case", or "individual", for which one single value is recorded for each numerical or categorical variable (in columns). This representation model is however restricted when the data to be analysed comprises variability, that is the case when the entities under analysis are not single elements, but groups formed on the basis of some given common properties. Then, for each descriptive variable, the observed variability inherent to each group should be taken into account, to avoid an important loss of pertinent information. This is relevant in Data Mining applications where huge sets of data are collected, but data should be analysed at a higher level - e.g. take the case of large department stores, which record data on each purchase made (amount spent, items purchased, etc.), but where the focus does not lie on individual purchases but rather on consumer behaviour, and therefore information about the purchases of each client, or specific groups of clients, must be somehow aggregated. Symbolic Data Analysis [1] provides a framework for the representation and analysis of such data, which comprehends inherent variability. New variable types have been introduced whose realizations

are not single real values or categories, but sets, intervals, or, more generally, distributions over a given domain. Methods for the (multivariate) analysis of such symbolic data have been developed which allow taking into account the variability expressed in the data representation.

We focus here on interval-valued data, i.e., where for each entity under analysis an interval is recorded. In [2] a parametric modelling for interval data, assuming multivariate Normal or Skew-Normal distributions for the MidPoints and Log-Ranges of the interval variables is proposed. The intrinsic nature of the interval variables leads to special structures of the variance-covariance matrix. This modelling allows for multivariate parametric analysis ; in particular M(ANOVA) [2], discriminant analysis [4], model-based clustering [3] and outlier detection [5] have been addressed under this framework.

In this talk we analyse a large dataset of loan data, with more than half a million records, where our main units of interest are risk categories. The microdata have thus been aggregated in the form of intervals, for the variables considered. Multivariate analysis of the resulting interval-valued data reveal insights that we not easily captured at the original microdata level.

Keywords: Interval Data; Multivariate data analysis; Symbolic data analysis

References

[1] Brito, P. (2014) Symbolic data analysis: another look at the interaction of Data Mining and Statistics. *WIREs Data Mining and Knowledge Discovery* **4**, 281–295.

[2] Brito, P. and Duarte Silva, A.P. (2012) Modelling interval data with Normal and Skew-Normal distributions. *Journal of Applied Statistics* **39**, 3–20.

[3] Brito, P., Duarte Silva, A.P. and Dias, J.G. (2015) Probabilistic clustering of interval data. *Intelligent Data Analysis* **19**, 293–313.

[4] Duarte Silva, A.P. and Brito, P. (2015) Discriminant analysis of interval data: an assessment of parametric and distance-based approaches. *Journal of Classification* **32**, 516–541.

[5] Duarte Silva, A.P., Filzmoser, P. and Brito, P. Outlier detection in interval data. *Under revision.*

Forecasting on imbalanced big time series using ensembles of classifiers and minimum information: application to ozone data

Alvaro Gómez-Losada
European Commission, Joint Research Centre (JRC), Economics of Climate Change, Energy and Transport, Seville, Spain, *alvaro.gomez-losada@ec.europa.eu*

Abstract
Ozone (O_3) is a transboundary air pollutant of major concern on both the European and global scale. At ground level, ozone can be formed by photochemical reactions between anthropogenic nitrogen oxides and volatile organic compounds in the presence of sunlight. O_3 concentrations reveal a strong spatial and temporal variability, with a significant differentiation between rural, suburban and urban sites. The target value for O_3 established by the European Union Air Quality Directive [1] for the protection of human health and vegetation is frequently exceeded in large regions of Europe, including urban agglomerations.

Air quality forecasters often use all available meteorological variables to forecast O_3. Variables retrieved from air quality agencies or meteorological services typically include temperature, solar radiation, surface wind speed and direction, cloudiness and vertical temperature gradient until certain height in atmosphere, to cite a few. However, on many occasions such information is not readily available. Equipment at meteorological or air quality monitoring stations can be limited, or its maintenance programs not duly observed. As a consequence, the data provided are scarce, of poor quality, or both. Furthermore, when variables' availability for forecasting O_3 is not ensured, many of the forecasting models found in literature are not reproducible.

The field of air quality is just one of many that utilizes forecasting models within a data classification framework. Traditional classification algorithms perform poorly on imbalanced data sets [2]. These algorithms fail to properly represent the distributive characteristics of the data and as a result, provide unfavourable accuracies across the different classes. Also, usual metrics used to evaluate the performance of classifiers in imbalanced learning are not appropriate since the misclassification error of the minority class is far costlier than that of the majority class.

More advanced forecasting models use ensembles of classifier approaches. Generally speaking, an ensemble of classifiers combines the prediction of individual classifier to often produce more accurate results than that of any single classifier. This is due to the ability ensembles have to produce generalization, which is usually higher than that of base classifiers. The computational cost of ensembles is higher than individual learners, but also the model complexity.

The aim of this study is evaluate the O_3 forecasting ability of nine base classifiers and several derived ensembles, using a reduced number of predictors and long time series with severe class imbalance. These predictors were extracted using the time of collecting O_3 hourly concentrations at monitoring stations as the only information; no other additional information related to the O_3 genesis was used.

O_3 hourly concentrations from five monitoring stations from three urban environments in Andalusia (Spain) were obtained for this study, from 2005 to 2015, allowing a range of local atmospheric conditions to be considered. This study adopts the O_3 forecast problem using a binary classification task, where the positive (minority) class included all the instances which values exceeded the limit value set for O_3 according to the Directive 2008 [1], and the negative class, otherwise. In the studied time series from the monitoring sites, the

proportion of instances of the minority class to the total of instances ranged from 0.8% to 3.2%.

The model selection was performed after tuning the parameters of the base classifiers using a 10-fold cross-validation scheme. The strategy for dealing with the unbalanced classification tasks was to randomly under-sample the majority class in the training sets before learning the base classifiers. Afterwards, the decision boundary for the positive and negative classes was estimated using the test set and the final forecasting results obtained using the validation set. To evaluate the classification accuracy of ensembles and base classifiers, three specific metrics for measuring performance in class imbalanced learning were used, namely, area under the curve (AUC), precision-recall area (PRA) [3] and Matthew's correlation coefficient (MCC). Additionally, to study the learning curve of base classifiers and ensembles, base classifiers were trained on subsets of different size obtained from initial training data sets, preserving in them the class imbalance from the original training data. The classification accuracy was later evaluated on the original validation sets using the three metrics indicated above.

The classification accuracy for base classifiers ranged from 0.78 to 0.95 (AUC), 0.78 to 0.94 (PRA) and 0.14 to 0.42 (MCC). The ensemble algorithms achieved similar results, from 0.68 to 0.95 (AUC), 0.63 to 0.96 (PRA) and 0.08 to 0.51 (MCC). Methodology designed shows that no base classifier uniformly outperforms over all the time series studied at different locations, and that gaining accuracy of ensembles is not significant. However, the learning curves of base classifiers and ensembles indicated a faster convergence of the base classifiers. In addition, they indicated that similar accuracy is obtained when training base classifiers, and therefore ensembles, using training data set sizes smaller than those used initially.

Keywords: Classification; Time series; Imbalanced learning; En-

semble methods; forecasting

Disclaimer
The views expressed are purely those of the author and may not in any circumstances be regarded as stating an official position of the European Commission.

References

[1] Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe [Internet]. [cited 2017 Jul 13].

[2] He, H. and Ma, Y. (2013) *Imbalanced Learning. Foundations, Algorithms and Applications*. IEEE Press-Wiley, Canada.

[3] Saito, T. and Rehmsmeier, M. (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced data sets. *Plos One* **10**, 1-21.

Forecasting conditional covariance matrix via principal volatility components in the presence of additive outliers

Carlos Trucíos
São Paulo School of Economics, FGV, *ctrucios@gmail.com*
Luiz K. Hotta
Dep. of Statistics, University of Campinas, *hotta@ime.unicamp.br*
Pedro Valls
São Paulo School of Economics, FGV, *pedro.valls@fgv.br*

Abstract
In this work, we analyse a recently procedure called principal volatility components. This procedure overcome several difficulties in modelling and forecasting the conditional covariance matrix in large dimensions. We show that outliers have a devastating effect on the construction of the principal volatility components and on the forecast of the conditional covariance matrix. We propose a robust procedure and analyse its finite sample properties by means of Monte Carlo experiments and present an empirical application. The robust procedure outperforms the classical method in contaminated series and has a similar performance in uncontaminated ones.

Keywords: Conditional covariance matrix; Constant volatility; Curse of dimensionality; Jumps; Principal components

Acknowledgments

The authors acknowledge financial support from São Paulo Research Foundation (FAPESP) grants 2016/18599-4, 2013/00506-1 and 2013/22930-0 respectively. The first two authors also acknowledges support from Laboratory EPIFISMA while the third author is also grateful with the National Council for Scientific and Technological Development (CNPq) grant 309158/2016-8.

The big chase: a decision support system for client acquisition applied to financial networks

Lara Quijano-Sanchez
UC3M-BS Institute of Financial Big Data, Universidad Carlos III de Madrid, Spain, *laraquij@inst.uc3m.es*

Federico Liberatore
UC3M-BS Institute of Financial Big Data, Universidad Carlos III de Madrid, Spain, *fliberat@inst.uc3m.es*

Abstract
Bank agencies daily store a huge volume of data regarding clients and their operations. This information, in turn, can be used for marketing purposes to acquire new clients or sell products to existing clients. A Decision Support System (DSS) can help a manager to decide the sequence of clients to contact to reach a designed target. In this paper we present the BIG CHASE, a DSS that translates bank data into a reliability graph. This graph models relationships based on a probability of traversal function that includes social measures. The proposed DSS, developed in close collaboration with *Banco Santander, S.A.*, fits the parameters of the probability function to explicit solution evaluations given by experts by means of a specifically designed Projected Gradient Descent algorithm. The fitted probability function determines the reliabilities associated to the edges of the graph. An optimization procedure tailored to be efficient on very large sparse graphs with millions of nodes and edges identifies the most reliable sequence of clients that a manager should contact to reach a specific target. The BIG CHASE has been tested with a case study on real data that includes *Banco Santander, S.A.* 2015 Spain bank records. Experimental results show that the proposed DSS is capable of modeling the experts' evaluations into probability function with a small error.

Keywords: Client acquisition; Financial networks; Social modelling; Projected gradient descent; Maximum reliability path

Empirical simulation analytics in financial engineering

Raquel M. Gaspar
ISEG and CEMAPRE, Universidade de Lisboa,
Rmgaspar@iseg.ulisboa.pt

Abstract
The financial industry has always been data-intensive. Nowadays, the amount of financial information available to market participants is huge, leading to enormous challenges on how to efficiently use it. Computer capacity is also at a stage when simulations can be easily performed and any resulting data stored without problems. For overviews on Big Data in finance see [5] or [4], and references therein.

Still, some financial services seem not to have adapted fast enough to the new data related facilities available. This study focuses on the financial engineer service – the development of investment products and/or strategies. Most of these products and/or strategies are still developed relying on theoretical models that, no matter their degree of sophistication are, by definition, a simplification of reality. They are sold to investors relying on model-based analytics. Most these products are only taken to test in a "real life" context after they are sold to investors.

In today's Big Data context this is hard to understand. We advocate that financial engineers, should back-test their proposals' design using empirical (past) information, or empirically simulated data. Risk analytics statistics resulting from empirical data are both realistic and intuitive, from the common investor point of view. On the importance of empirical data properties see also [2].

We use a well-known portfolio insurance strategy – the constant proportion portfolio insurance (CPPI) – to illustrate our point of

view. Previous studies on CPPIs by the same author, although not based upon empirical data are [3] and [1].

Keywords: Empirical simulation; Big data; Risk analytics

Acknowledgments

This research is partially financed by project CEMAPRE–UID/ MULTI/ 00491/ 2013 financed by the Portuguese Science Foundation (FCT/ MCE) through national funds.

References

[1] Carvalho, J., Gaspar, R.M., and Sousa, J.B. (2016) On path–dependency of constant proportion portfolio insurance strategies. *Available at SSRN, 2016*.

[2] Cont, R. (2001) Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance* **1**, 223–236.

[3] Costa, J. and Gaspar, R.M. (2014) Portfolio insurance–a comparison of naive versus popular strategies. *Insurance Markets and Companies: Analyses and Actuarial Computations* **5**, 53–82.

[4] Fang, B. and Zhang, P. (2016) Big data in finance. *In: Big Data Concepts, Theories, and Applications*. Springer, pp. 391–412.

[5] Seth, T. and Chaudhary, V. (2015) Big data in finance. *In: Big Data Concepts, Theories, and Applications* ch. 17, pp. 329–356.

Text mining and ruin theory: a case study on risk models with dependence

Renata G. Alcoforado
ISEG & CEMAPRE, Universidade de Lisboa & Universidade Federal de Pernambuco, *alcoforado.renata@gmail.com*
Alfredo D. Egídio dos Reis
ISEG & CEMAPRE, Universidade de Lisboa, *alfredo@iseg.ulisboa.pt*

Abstract
This work aims to analyse unstructured data using a text mining approach. In our study, the subject is composed by 27 published papers of the risk and ruin theory topic, area of actuarial science, that were coded in 32 different categories. For the purpose, all data was analysed by using the software *NVivo 11 plus*. Software *NVivo* is a specialized tool in analysing unstructured data.

Keywords: Big data; Unstructured data; Text mining; Risk theory; Ruin probability; Dependence

1 Introduction

Big Data is an area of great development in statistics. We can define Big Data as "a phenomenon defined by the rapid acceleration in the expanding volume of high velocity, complex, and diverse types of data. Big Data is often defined along three dimensions – volume, velocity, and variety" ([3]). According to [2] data mining is the process of mining through large amount of data to extract meaningful information, knowledge. It's also treated by many people as a synonym for knowledge discovery from data, or KDD. Text mining in an analogous manner as data mining, aims to extract information from data, but in this case the data comprehend to texts and does

it through identification and exploration of interesting patterns ([1]).

The manuscript is organized as follows: In Section 2 we speak about the collected data to be analysed. Section 3 is about the coding of the data, the coding matrix, the relationship between the nodes, that is, we plotted the nodes hierarchically. Then, we present the cluster analysis for the nodes and the sources (papers), the comparison diagrams and to finalize, a structural matrix. To finalize we exhibit some conclusions.

2 The Data

In our particular work we uploaded 27 papers in the platform, codified and then analised all data. These papers are references for a particular research project in development in risk theory.

These papers are: Afonso et al. (2017), Ammeter (1948), Asmussen and Albrecher (2010), Bergel and Egídio dos Reis (2016), Constantinescu et al. (2011), Constantinescu et al. (2012); Constantinescu et al. (2016), Czado et al. (2011), Frees and Wang (2006), Frees et al. (2011), Frees et al. (2016), Garrido et al. (2016), Gschlößl and Czado (2007), Jasiulewicz (2001), Jorgensen and Paes De Souza (1994), Krämer et al (2013), Kreer et al. (2015), Li et al. (2015), Maume-Deschamps et al. (2017), Ni et al. (2014a), Ni et al. (2014b), Quijano Xacur and Garrido (2015), Renshaw (1994), Rolski et al. (1999), Schulz (2013), Shi et al. (2015), Song et al. (2009). The software builds a *word cloud* composed by the the most pertinents words in our entire data base to use in our study. After removing all the verbs, articles and non-meaningful wording, the words are then gathered according to their *stem*, making possible to obtain the *cloud* in Figure 1.

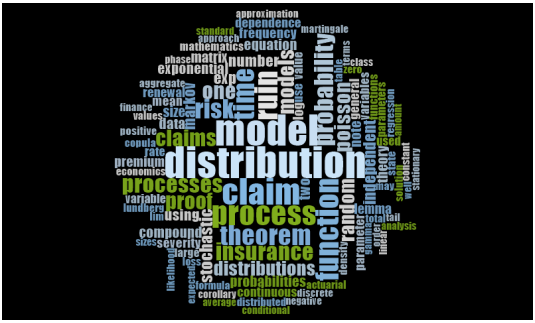

Figure 1: Word Cloud Source: The authors

3 The Coding and Analysis

Data mining assumes that the data is already stored in an structured way, whereas text mining assumes that the data is unstructured and still needs coding. ([1]).

In sequence, all the unstructured data was codified, that is, we put in each one of the categories the respective parts from text to be able to analyse it in a mathematical way. In other words, after coding we get a structure to be able to analyse with clusters and matrices. Then, plotting the data which was not possible previously. The categories were selected after extensive reading and observing the *word cloud*.

In our particular exercise the *codes* are: Actuarial; Aggregate Claims Model; Claim Frequency; Claim Severity; Compound Poisson; Conditional; Copulas; Covariates; Dependence; Exponential; Formula; Function; Gamma; Independence; Insurance; Joint Distribution; Loss; Markov; Martingale; Mixed-Poisson; Parameters; Prediction; Premium; Randomness; Regression; Renewal; Risk Theory; Ruin Probability; Simulation; Spatial; Stationary and Stochastic Process.

After the code and data organizing, we constructed the *coding matrix* presented on the left graph of Figure 2, which shows both in numbers and in graph what is the relationship between the coded categories. A *word tree* in each category is plotted to see the connection from that word (or expression) in the sentence where it belongs.

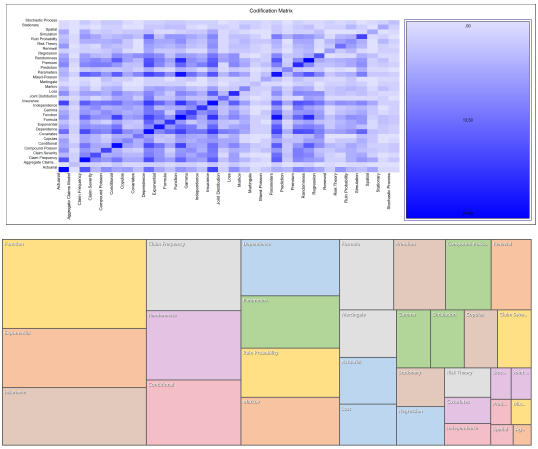

Figure 2: Coding Matrix - Heat and Nodes Hierarchically respectively

We plotted the *nodes* hierarchically presented on the right graph of the Figure 2 to observe which categories are most frequent, the most important among the data available. The *cluster analysis* was performed in cluster by word similarity using Pearson's correlation coefficient as the similarity metric. We made it for both the categories and the sources to see how they relate. The cluster analysis for the nodes is presented in Figure 3, a diagram on the left and a circle graph on the right.

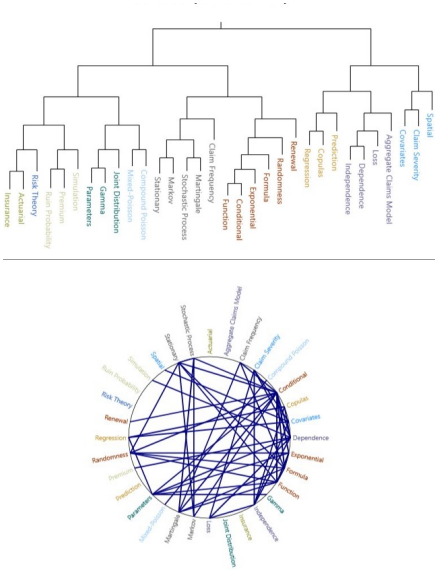

Figure 3: Cluster Analysis of the Nodes

4 Conclusions

Our source intended to talk about the calculation of premiums and ruin probabilities for insurance application, also how to associate the claim frequency with their severity. Some authors use copulas, others use covariates in a regression model, and others try to find a distribution that can capture that dependence. The result showed to be interesting to compare respective categories and plot comparison diagrams, for instance, comparing *Dependence* with *Independence*; *Simulation* with *Formula*; *Copulas* with *Covariates*; *Regression* with *Copulas*; *Claim Severity* with *Claim Frequency* among others. To finalize, we built the *structural matrix* where each row shows each paper and each columns the category mentioned above, in order to

identify subtle connections which can allow a thorough and rigorous study.

References

[1] Feldman, R. and Sanger, J. (2006) *The Text Mining Handbook*. Cambridge University Press, New York, USA.

[2] Han, J., Kamber, M. and Pei, J. (2012) *Data mining. Concepts and Techniques*. Elsevier, Waltham, USA, third edition.

[3] TechAmerica Foundation's Federal Big Data Commission (2012). Demystifying Big Data: A Practical Guide To Transforming The Business of Government. Technical report, TechAmerica Foundation's. Retrieved July 10, 2017, from https://www.attain.com/sites/default/files/take-aways-pdf/Solutions_Demystifying Big Data - A Practical Guide To Transforming The Business Of Government.pdf

Mismatch between jobs and skills in the EU

João Lopes
Instituto Nacional de Estatística, *joao.lopes@ine.pt*
Marco Moura
Instituto Nacional de Estatística, *marco.moura@ine.pt*
Sónia Quaresma
Instituto Nacional de Estatística, *sonia.quaresma@ine.pt*

Abstract
In order to study the mismatch between the available skills of labour force and the skills required by labour market, we developed the concept of Labour Market Attractiveness. This concept consisted of the combination of a set of variables from 6 Eurostats datasets on different subjects (i.e. Demographics; Earnings structure; Education and training; Life conditions; Employment and unemployment; and National accounts). The impact of this combined dataset on Skills Mismatch, as well as on Labour Market Mobility and Emigration, was assessed using various data mining techniques, particularly, clustering analysis, model selection analysis using multivariate regression and weighted network correlation analyses. We showed that Labour Market Attractiveness is able to form consistent clusters at country-level, which can be well defined using only the variables "Youth Unemployment" and "GDP". Furthermore, from this combined dataset we defined 6 Eigenvariables, namely, "Unemployment", "Poverty", "Ageing Population", "Secondary Education (Adults)", "Employment" and "Earnings structure". Considering these Eigenvariables, we found that: Skills Mismatch is negatively associated to "Employment" and "Secondary Education (Adults)", while being positively associated to "Poverty" and "Unemployment"; Labour Market Mobility is associated to "Earnings structure" and "Employment"; and Emigration is negatively associated to "Secondary Education (Adults)" and "Ageing Population". From model selection,

we showed that: Skills Mismatch is best explained by "Proportion of employed youth" and "Proportion of employed youth working at NACE M-N"; Labour Market Mobility is best explained by "Proportion of employed adults working at NACE L and K"; and Emigration is best explained by "Proportion of employed adults working at NACE K", "Proportion of employed youth with Higher Education", "Proportion of employed youth working at NACE B-E and O-Q" and "Total Population size".

Keywords: Labour Market Attractiveness; Labour Market Mobility; Skills Demand; Skills Supply

1 Introduction

The European Big Data Hackathon took place in March 2017 - in parallel with the conference New Techniques and Technologies for Statistics (NTTS). This event was organised by the European Commission (Eurostat) and gathered 22 teams from 21 European countries. The aim was to compete for the best data product combining official statistics and Big Data to support policy makers in a pressing policy question, namely, *How to tackle the mismatch between jobs and skills at regional level in Europe?* Indeed, the mismatch between the available skills of the labour force and the skills required by the labour market entail significant economic and social costs for individuals and firms [1]. Furthermore, a strong education and an efficient development of skills are essential for thriving in the emerging new economy and fast-changing labour market [1]. Nonetheless, a survey from 2014 showed that skills mismatch (i.e. over-qualification, under-qualification) remains at 45% in the European Union [2]. This led to the publication of the European Union Guidelines for the employment policies of the Member States in 2015, which called for enhancing labour supply, skills and competences [3]. In order to better support policy makers in solving this skills mis-

match problem, the data product was required to be supported by relevant data, statistical analysis and visualization. Teams were also invited to use provided datasets (including European Employment Services (EURES) data on jobseekers and on job vacancies [4]), and additional publicly available data sources with international applicability (e.g. Eurostat online database [5]).

The development of our data product was focused on three main ideas: 1) combine official statistics data from Eurostat at NUTS2 level (i.e. solid data sets known for being well-structured, clean and accurate, but also characterized by a morose collect and release process) with real-time unstructured Big Data; 2) explore the notion of Labour Market Attractiveness as an important factor in the mismatch between skills demand and supply, in labour market mobility, and in emigration; 3) create a flexible, interactive and user-friendly product that allows for customization of the answers in order to be used either by policy makers or by both citizens searching for help on jobseeking and enterprises looking for particular labour market characteristics.

The definition of Labour Market Attractiveness has to be considered carefully, thus, our approach should be seen has a first-step towards a more mature definition. We considered 17 variables from 6 Eurostat datasets, namely "reg-demo" for demographics data, "earn" for earnings structure data, "edtr" for data on education and training, "ilc" for life conditions information, "employ" for employment/unemployment data, and "na10" for national accounts data. These variables were broken by several categorical levels (e.g. "age groups", "level of education", "qualifications", "occupations") originating more than 70 variables. Several data mining techniques were then considered to analyse this compiled Labour Market Attractiveness dataset. Using the datasets we calculated distances between regions and visualize those using social networks algorithms. We further clustered the regions using a Partition Around Medoids method on those distances creating a categorical variable with grouping information [6]. This

created variable, along with collected variables on Skills mismatch, Labour Market Mobility and Emigration, were used separately as dependent variables on model selection using multivariate linear and non-linear regression analyses with the Labour Market Attractiveness dataset as independent variables [7]. We further constructed Eigenvariables from the considered set and performed Weighted Correlation Network Analysis on the dependent variables [8].

In our work we assumed two major simplifications in the construction of the skills mismatch indicator, however, these simplifications do not affect our product in terms of proof-of-concept and can be dropped in later developments. The first one was to use previously cleaned and treated data on job vacancies and education attainment from the Eurostat's "labour" and "edtr" data sets, respectively. Instead, a better approach would be to use the freshly collected EURES data provided, but the use of this data would have two caveats: a) the cleaning and structuring of the data requires a considerable expertise on the subject; b) the normalizing of the data, using for example marginal calibration techniques, requires detailed demographic data at the required regional level in order to successfully capture the populations considered. The second simplification was to use an ad hoc mapping between qualifications (classified using ISCED-F 13) and the cross between occupations (defined using ISCO-08) and economic activity (defined using NACE Rev. 2). Nonetheless, a formal mapping will be released in mid-2017 by European Skills, Qualifications and Occupations (ESCO) from the EC.

Statistical analyses were carried out at country-level and at NUTS1- and NUTS2-level. They were performed in R using libraries *cluster*, *glmulti*, *Hmisc*, *MASS*, *nnet*, *sna* and *WGCNA*.

2 Conclusions

At country-level, the compiled Labour Market Attractiveness dataset is able to form consistent clusters (clusters separation between 2.31 and 4.68). Moreover, these clusters can be well defined even using only a data subset of "Youth unemployment" and "GDP" (RMcFadden = 0.84). Furthermore, the Labour Market Attractiveness dataset can be reduced to 6 Eigenvariables, namely, "Unemployment", "Secondary Education (Adults)", "Poverty", "Ageing Population", "Employment" and "Earnings structure". Considering these Eigenvariables, we found that Skills Mismatch is very negatively associated to "Employment" ($\rho = -0.69$, $p = 0.058$) and moderately negatively associated to "Secondary Education (Adults)" ($\rho = 0.38$, $p = 0.352$), while being moderately associated to "Poverty" ($\rho = 0.38$, $p = 0.352$) and "Unemployment" ($\rho = 0.36$, $p = 0.385$). Labour Market Mobility is strongly associated to "Earnings structure" ($\rho = 0.59$, $p = 0.002$) and moderately associated to "Employment" ($\rho = 0.35$, $p = 0.082$). Lastly, Emigration is very negatively associated to Secondary Education (Adults) ($\rho = -0.50$, $p = 0.007$) and moderately negatively associated to Ageing Population ($\rho = -0.36$, $p = 0.063$). Finally, the model selection using multivariate linear regression shows that for Skills Mismatch the most important explanatory variables are "Proportion of employed youth" (Importance = 0.84) and "Proportion of employed youth at NACE M-N" (Importance = 0.78). For Labour Market Mobility they are "Proportion of employed adults at NACE L" (Importance = 1.00), "Proportion of employed adults at NACE K" (Importance = 0.98) and "Proportion of employed youth at NACE B-E" (Importance = 0.67). Regarding Emigration the variables are "Proportion of employed adults at NACE K" (Importance = 0.93), "Proportion of employed youth with Higher Education" (Importance = 0.89), "Proportion of employed youth at NACE O-Q" (Importance = 0.86), "Total population size" (Importance = 0.74) and "Proportion of employed youth at NACE B-E" (Importance = 0.55).

References

[1] https://ec.europa.eu/commission/publications/skills-education-and-lifelong-learning-european-pillar-social-rights-en

[2] CEDEFOP (2015) Skills, qualifications and jobs in the EU: the making of a perfect match? *Publications Office of the European Union*, Luxembourg.

[3] Council Decision (EU) 2015/1848 of 5 October 2015

[4] https://ec.europa.eu/eures/public/homepage

[5] http://ec.europa.eu/eurostat/data/database

[6] Reynolds, A.P., Richards, G., de la Iglesia, B. and Rayward-Smith V.J. (1992) Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms* **5**, 475–504.

[7] Calcagno, V. and de Mazancourt, C. (2010) glmulti: an R package for easy automated model selection with (generalized) linear models. *Journal of Statistical Software* **34**, 1–29.

[8] Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis BMC. *Bioinformatics* **9**, 559.

Mutual fund competition in European union

João Romacho
Interdisciplinary Coordination for Research and Innovation (C3i), Polytechnic Institute of Portalegre, Portugal,
jromacho@estgp.pt
Cristina Dias
CMA-Center for Mathematics and Applications of the Nova University of Lisbon, Polytechnic Institute of Portalegre, Portugal,
cpsilvadias@gmail.com

Abstract
This study analyses the competition/strategic behaviour among the mutual funds of several countries from the European Union. To achieve this aim, it is used the Brown, Harlow and Starks (1996) methodology applied in different settings. Thus, the competition/strategic behaviour is analysed in subperiods with the same duration and in subperiods that correspond to different market cycles, the characteristics of the funds are also contemplated (the dimension of their portfolio and its period of activity), as well as the possible effect of the survivorship bias.

The outcomes obtained show the existence of strategic behaviour among the mutual funds from the European Union, this is stronger among the funds from Belgium, Spain and United Kingdom. Furthermore, with the exception of the United Kingdom, this behaviour is stronger among the funds with a smaller period of activity and in the most recent period. So, it seems that, on the one hand, the greatest strategic interaction among funds with smaller period of activity can occur from its greatest audacity, unlike the most experienced funds that tend to reveal higher caution. On the other hand, the growth of the markets from the European Union concerning the number of funds seems to contribute to the increase of the strategic

behaviour.

Keywords: Competition; Strategic behaviour; Mutual funds

1 Introduction

The industry of asset management is an essential sector for economic growth, which has assumed, in the last decades, an increasing importance in the US and Europe. So, the competition in the mutual fund industry study is of paramount importance for several reasons. For the quality, variety and cost of financial products, as well as by understanding the volatility of the portfolios in response to competition in the sector, so "the competition in the mutual fund industry may therefore have far ranging and long lasting implications for the wealth of investors" (Ramos, 2009, pp. 176).

2 Mutual fund competition studies and methodology

Brown, Harlow e Starks (1996) (BHS) were the first to study the mutual fund competition and to confirm their hypothesis that de mutual fund with worse performance in the middle of the year tend to increase more the portfolio risk in the last part of the year (competition behaviour). Others studies that used the BHSÂ´ methodology, and alternative ones, also identify competition in the mutual fund industry, such as the case of Busse (2001) and Schwarz (2008). Contrary to previous studies, Qiu (2003) and Elton, Gruber, Blake, Krasny and Ozelge (2010) identify strategic behaviour, that is, they checked that the mutual funds with higher performance are the ones which have more incentives to increase their risk.

In the present work it is applied the BHS' methodology. With the

goal to test the above hypothesis, it's necessary to calculate the loser funds return standard deviation (with return below the median) and the winner funds (with return above the median) in the first and second part of the year. In order to test if the loser funds increase their risk greater than the winner funds from a given moment in the year itÂ´s necessary to define two variables. The first, the Accumulated Return (AR) determined until month M of the year, which allows to split the mutual funds in winners and losers. The second, the Risk Adjusted Ratio (RAR) allow to compare the volatility of each fund before and after month M, that is, the relationship between the standard deviation of the second and the first part of the year.

3 Empirical analysis

The sample selected for this study consists of 1485 global stock mutual funds, including surviving and extinct funds, from seven countries of European Union (EU) (Germany, Belgium, Spain, France, Italy, UK and Sweden), for the period from January/1994 to December/2009. Figure 1 shows, for each country and for total sample, the results of the application of the BHS' methodology in order to identify competition/strategic behaviour.

In the Figure 1 ten settings are used: they include surviving and extinct mutual funds throughout the total sample period (1) and only surviving mutual funds (2); the total period is divided into two of equal duration, subperiod 1 (3) and subperiod 2 (4); the mutual funds are separated which present the lower (5) and higher (6) number of months of quotation throughout all sample period; the mutual funds are separated which present the lower (7) and higher (8) portfolio size; and, they identify bull (9) and bear (10) market periods. The number of mutual funds in each setting is presented too.

Countries	Evaluation period	Settings / Behaviour									
		Survivorship bias		Subperiods		Age		Dimension		Market phases	
		Surv.Ext. (1)	Surviv. (2)	SP1 (3)	SP2 (4)	New (5)	Old (6)	Small (7)	Big (8)	*Bull* (9)	*Bear* (10)
Germany	(4,8)	-	S	-	-	S	-	-	S	-	S
	(5,7)	-	-	-	-	S	-	-	-	-	S
	(6,6)	-	-	-	S	SS	-	-	-	-	-
	(7,5)	S	-	-	-	SS	-	-	-	S	-
	(8,4)	S	-	-	-	-	-	S	-	SS	-
	Nº of funds:	*[112]*	*[175]*	*[110]*	*[112]*	*[106]*	*[106]*	*[85]*	*[85]*	*[111]*	*[112]*
Belgium	(4,8)	SS	S	-	SS	-	-	-	SS	-	SS
	(5,7)	-	-	-	-	-	-	-	-	C	SS
	(6,6)	SS	S	-	SS	-	SS	-	-	-	SS
	(7,5)	SS	SS	-	SS	SS	-	SS	-	-	SS
	(8,4)	S	SS	-	S	-	S	-	S	-	SS
	Nº of funds:	*[201]*	*[284]*	*[53]*	*[201]*	*[201]*	*[201]*	*[87]*	*[88]*	*[198]*	*[196]*
Spain	(4,8)	-	-	-	-	-	-	-	-	-	-
	(5,7)	SS	SS	-	SS	S	-	S	-	SS	-
	(6,6)	SS	SS	-	SS	SS	S	SS	-	SS	SS
	(7,5)	SS	S	-	SS	SS	S	SS	-	S	-
	(8,4)	S	S	-	S	-	-	-	-	S	-
	Nº of funds:	*[83]*	*[53]*	*[57]*	*[83]*	*[47]*	*[41]*	*[17]*	*[26]*	*[52]*	*[83]*
France	(4,8)	-	-	-	-	-	-	-	-	-	-
	(5,7)	-	-	-	-	-	-	-	-	-	SS
	(6,6)	-	-	-	-	-	-	-	-	-	SS
	(7,5)	-	-	-	S	SS	-	-	-	-	-
	(8,4)	-	-	-	-	-	-	-	-	-	-
	Nº of funds:	*[142]*	*[152]*	*[158]*	*[142]*	*[129]*	*[129]*	*[125]*	*[125]*	*[147]*	*[105]*
Italy	(4,8)	-	-	-	-	S	-	-	-	-	-
	(5,7)	-	-	-	-	S	-	-	-	-	-
	(6,6)	-	-	-	-	SS	-	-	-	-	-
	(7,5)	-	-	-	-	SS	-	-	-	-	-
	(8,4)	S	-	-	S	-	-	-	-	-	S
	Nº of funds:	*[93]*	*[83]*	*[98]*	*[93]*	*[98]*	*[97]*	*[89]*	*[89]*	*[95]*	*[87]*
UK	(4,8)	-	-	-	-	-	-	-	-	CC	SS
	(5,7)	-	-	SS	-	-	-	-	-	-	SS
	(6,6)	-	-	SS	-	-	-	-	-	-	-
	(7,5)	SS	SS	SS	SS	S	SS	SS	SS	SS	-
	(8,4)	SS	SS	SS	S	-	SS	-	S	SS	-
	Nº of funds:	*[307]*	*[333]*	*[231]*	*[307]*	*[253]*	*[254]*	*[118]*	*[129]*	*[306]*	*[289]*
Sweden	(4,8)	-	-	-	-	-	-	-	-	C	S
	(5,7)	-	-	-	SS	-	-	-	-	-	SS
	(6,6)	-	-	CC	-	-	-	-	-	-	-
	(7,5)	-	-	C	S	-	-	-	-	-	-
	(8,4)	-	-	-	-	-	-	-	-	-	-
	Nº of funds:	*[58]*	*[59]*	*[84]*	*[58]*	*[54]*	*[54]*	*[29]*	*[28]*	*[58]*	*[57]*
TOTAL SAMPLE	*(4,8)*	–	–	*S*	–	*S*	–	–	–	*CC*	*SS*
	(5,7)	–	–	*S*	–	–	–	–	–	*CC*	*SS*
	(6,6)	–	–	–	–	*S*	–	–	–	*CC*	*SS*
	(7,5)	*SS*	*SS*	*SS*	*SS*	*SS*	*SS*	*SS*	*SS*	*SS*	–
	(8,4)	*SS*	*S*	–	*SS*	*S*	*S*	–	–	*SS*	–
	Nº of funds:	*[1485]*	*[1701]*	*[683]*	*[1485]*	*[743]*	*[742]*	*[500]*	*[500]*	*[1477]*	*[1439]*

Figure 1: Competition/strategic behaviour in different settings. C (CC): Period with statistically significant competition behaviour at 5% (1%) significance level S (SS): Period with statistically significant strategic behaviour at 5% (1%) significance level. S (SS): Period with statistically significant strategic behaviour at 5% (1%) significance level

4 Conclusions

Given the different settings analysed, it seems there will be strategic behaviour among EU mutual funds, with higher strength in Belgium, Spain and UK. This behaviour will be, except in the UK, more clear in mutual funds with smaller period of activity and in the most recent period. In these three countries, the most favourable settings for strategic behaviour will be: to Belgium, in the subperiod 2 and in bear market; to Spain, in the subperiod 2, in bull market, to younger mutual funds and to smaller portfolios; and, to UK, in the subperiod 1 and to older mutual funds.

This results show that, on the one hand, the greatest interaction among younger mutual funds may come from their greater fearlessness, opposition to older mutual funds that seem to reveal greater careful foresight. On the other hand, the EU market growth in terms of mutual funds (2002 to 2009) seems to contribute to increase the strategic behaviour.

References

[1] Brown, K., Harlow, W. and Starks, L. (1996) On tournament and temptations: an analysis of managerial incentives in the mutual fund industry. *Journal of Finance* **51**, 85-110.

[2] Busse, J. (2001) Another look at mutual fund tournaments. *Journal of Financial and Quantitative Analysis* **36**, 53-73.

[3] Elton, E., Gruber, M., Blake, C., Krasny, Y. and Ozelge, S. (2010) The effect of holdings data frequency on conclusions about mutual fund behavior. *Journal of Banking & Finance* **34**, 912-922.

[4] Ramos, S. (2009) The size and structure of the world mutual fund industry. *European Financial Management* **15**, 145-180.

[5] Qiu, J. (2003) Termination risk, multiple managers and mutual fund tournaments. *European Finance Review* **7**, 161-190.

[6] Schwarz, C. (2008) Mutual fund tournaments: the sorting bias and new evidence. *Working Paper*, University of California at Irvine.

Sparse and constrained naïve Bayes for cost-sensitive classification

M. Remedios Sillero-Denamiel
Department of Statistics and Operations Research, University of Seville, Spain IMUS, Instituto de Matemáticas de la Universidad de Sevilla, Spain, *rsillero@us.es*

Rafael Blanquero
IMUS, Instituto de Matemáticas de la Universidad de Sevilla. Departamento de Estadística e Investigación Operativa, Universidad de Sevilla. Spain, *rblanquero@us.es*

Emilio Carrizosa
IMUS, Instituto de Matemáticas de la Universidad de Sevilla. Departamento de Estadística e Investigación Operativa, Universidad de Sevilla. Spain, *ecarrizosa@us.es*

Pepa Ramírez-Cobo
IMUS, Instituto de Matemáticas de la Universidad de Sevilla. Departamento de Estadística e Investigación Operativa, Universidad de Cádiz. Spain, *pepa.ramirez@uca.es*

Abstract
Naïve Bayes is a tractable and remarkably efficient approach to classification learning. However, as it is common in real clasification contexts, datasets are often characterized by a large number of features and, in addition, there could exist an imbalance between the correct classification rates of different classes. On the one hand, it may complicate the interpretation of the results as well as slow down the method's execution. On the other hand, classes are often not equally important and making a misclassification in one of them leads to undesirable consequences that can be avoided by controlling the correct classification rate in that particular class. In this work we propose a sparse and constrained version of the Naïve Bayes in which a variable reduction approach, that takes into account the dependen-

cies among features, is embedded into the classification algorithm. Moreover, a number of constraints over the performance measures of interest are embedded into the optimization problem which estimates the involved parameters. Unlike typical approaches in the literature modifying standard classification methods, our strategy allows the user to control simultaneously the different performance measures that are considered. Our findings show that, under a reasonable computational cost, the number of variables is significantly reduced obtaining competitive estimates of the performance measures. Furthermore, the achievement in the different individual performance measures under consideration is controled.

Keywords: Conditional independence; Dependence measures; Variable selection; Heuristics; Probabilistic classification, Constrained optimization; Efficiency measures

Sparse support vector machines with performance constraints

Sandra Benítez Peña
IMUS, Instituto de Matemáticas de la Universidad de Sevilla. Departamento de Estadística e Investigación Operativa, Universidad de Sevilla. Spain, *sbenitez1@us.es*

Rafael Blanquero
IMUS, Instituto de Matemáticas de la Universidad de Sevilla. Departamento de Estadística e Investigación Operativa, Universidad de Sevilla. Spain, *rblanquero@us.es*

Emilio Carrizosa
IMUS, Instituto de Matemáticas de la Universidad de Sevilla. Departamento de Estadística e Investigación Operativa, Universidad de Sevilla. Spain, *ecarrizosa@us.es*

Pepa Ramírez-Cobo
IMUS, Instituto de Matemáticas de la Universidad de Sevilla. Departamento de Estadística e Investigación Operativa, Universidad de Cádiz. Spain, *pepa.ramirez@uca.es*

Abstract
Support Vector Machine (SVM) is a powerful tool to solve binary classification problems. Many realworld classification problems, such as those found in credit-scoring or fraud prediction, involve misclassification costs which may be different in the different classes. Providing precise values for such misclassification costs may be hard for the user, whereas it may be much easier to identify acceptable misclassification rates values. Hence, we propose here a novel SVM model in which misclassification costs are considered by incorporating performance constraints in the problem formulation. In particular, our target is to seek the hyperplane with maximal margin yielding misclassification rates below given threshold values.

This novel model is extended by performing Feature Selection (FS), which is a crucial task in Data Science, making thus the classification procedures more interpretable and more effective.

The reported numerical experience demonstrates that our model gives the user control on the misclassification rates in addition to the usefulness of the proposed FS procedure. Indeed, our results on benchmark data sets show that a substantial decrease of the number of features is obtained, whilst the desired trade-off between false positive and false negative rates is achieved.

Keywords: Constrained classification; Feature selection; misclassification costs; mixed integer quadratic programming; Sparsity; Sensitivity/Specificity trade-off; Support vector machines

Exploratory (big) data analysis

Albert Satorra
UPF, Spain, *albert.satorra@upd.com*
Catia Nicodemo
Oxford University, UK, *catia.nicodemo@gmail.com*

Abstract
The collection of large amounts of databases, big data, (clinician, social security, facebook...), is becoming more common, as the research studies that use them. Usually this type of data are quite rich and with many potential benefits. However, the description and the analysis of big datasets could be hard to performance without the right techniques.

A primary goal of this paper is to present clearly and efficiently via statistical graphics, plots and information graphics to describe and explore big datasets.

Effective data visualization helps to analyse and understand about data and evidence. It makes complex data more accessible, understandable and usable like to make comparisons or understanding causality. Charts are used to show patterns or relationships in the data for one or more variables facilitating the task to figure out the description and the possible correlation in the data.

In this paper we use the statistical software *R* that provides new tools to display in real-time changes and more illustrative graphics of the big databases, thus going beyond pie, bar and other charts. These illustrations veer away from the use of hundreds of rows, columns and attributes toward a more artistic visual representation of the data. In this paper we focus our attention in ***ggplot2*** is an R package for data visualization. It provides a powerful model of graphics that makes it easy to produce complex multi-layered graphics.

To design our study, we use two big databases: the Spanish Social Security data (Muestra Continua de Vidas Laborales) hereafter MCVL) in 2010[1] and the Hospital Episode Statistics data (HES) for UK [2].

The MCVL data comes from the register of the Social Security System (SS) in Spain for active people in the labor market, representing more than one million people each year. The data set gives information of all of the historical relationships of any individual with the Social Security System (in terms of work and unemployment benefits). The data will include more than 15 million of observations per year.

The HES data provide information concerning all inpatients and outpatients admitted to NHS hospitals from 1989-90 onwards. It includes private patients treated in NHS hospitals, patients resident outside of England, and care delivered by Treatment Centres (including those in the independent sector) funded by the NHS. Each patient record contains detailed information, including: clinical information, patient characteristics, such as age and gender, and administrative and location information, such as method of admission, and the geography of treatment and residence. Since our focus is on GP influence upon admissions, our analysis concerns only the 'first admission' to the hospital, which the GP is most likely to influence, rather than admissions for continuing treatments. The database contains more than 80 million of observations per year.

Big data are data on a massive scale in terms of volume, intensity, and complexity that exceed the capacity of standard software tools. They present opportunities as well as challenges to statisticians. There are several statistical methods to analyse the big data

[1] $http://www.seg-social.es/prdi00/groups/public/documents/binario/190489.pdf$
[2] $http://content.digital.nhs.uk/hes.$

like learning machine, Lasso, etc,. However very few evidence is on how to describe databases with huge number of observations.

The analysis in this study consist first in preparing the databases and after using *ggplot2* to explore the data and present several factors. In particular, we will study from the social security data the correlation between wage and age conditional on the fact to have a permanent or temporary contact, to be male or female, and the level of education. The scope is to analyse the difference across young and old people in a three-dimensional way. Traditional graphs, like plotting the two variable again each other, are useless. This analysis allow us to explore more deeply the data and present them in an easy way to understand from a general audience like policy makers, stakeholder, etc.

The health data are explored to look at the correlation between the referrals (a specialist visit in the hospital) and the treatment (a surgery) in the hospital at practice level. Among general practice in the UK exist a huge variation in terms of people referred to a hospital and people treated in the hospital. Before to analyse these data with statical complex models we want to present trough "ggplot2" the variation across practices and understand the possible co-factors that could driven certain results, like for example if the practice is in a poor or reach area.

Our study will bring a reference for people interested in exploring the data before to think which is the best model to predict the outcomes.

Keywords: Social Security Data; Health Data Big Data

POSTERS

Estimating partially linear model with ridge type smoothing spline for high dimensional data

Ersin Yilmaz
Mugla Sitki Kocman University, *yilmazersin13@hotmail.com*
Dursun Aydin
Mugla Sitki Kocman University, *duaydin@hotmail.com*

Abstract
Let consider the partially linear model

$$Y_i = X_{ij}\beta_i + f(z_i) + \epsilon_i, 1 \leq i \leq n, 1 \leq j \leq p, \quad (1)$$

where Y_i's are the observations of response variable, X_{ij}'s are the values of explanatory variable, β_i's are the regression coefficients for parametric component of the model, $f(\cdot)$ is the smooth function to be estimated, z_i is a nonparametric variable and ϵ_i's are the identically distributed and independent random error terms with mean zero and constant variance.

In this paper, we discuss the estimation problem in which the number of variable p is much bigger than observation number n (i.e.,$n >> p$). It is well known that high variance and overfitting are a major concern in this setting. As a result, simple, highly regularized approaches often become the methods of choice. So, we fit a partially linear model (1) with quadratic regularization (i.e., ridge regression [1], [2]) on the coefficients. Also, there are several proposed alternative methods such as partial least squares (PLS), least absolute shrinkage and selection operator (LASSO; [3]) and various improved versions of these three methods can be considered.

In here, we estimated the model (1) with using smoothing spline estimator ([4]; [5]; [6]) based on ridge regression. Note also that

the basic idea of paper is to reduce the dimension by singular value decomposition (SVD) method and to estimate components of the model (1).

To fit the model (1) to data, we can use ridge regression that shrinks the regression coefficients by imposing a penalty on their size. This procedure can be related to the idea of hints due to [7], where the parameter vector $\boldsymbol{\beta}$ can be obtained by minimizing the quadratic regularization criterion (QRC)

$$\begin{aligned} QRC(\boldsymbol{\beta}; f; \lambda) &= \sum_{i=1}^{n} (\tilde{Y}_i - \tilde{X}_i\boldsymbol{\beta})^2 + k\sum_{j=1}^{n}\beta_j^2 \\ &= (\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta})'(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}) + k||\mathbf{0} - \boldsymbol{\beta}||^2, \end{aligned} \tag{2}$$

where $k \geq 0$ is the shrinkage parameter that controls the magnitude of the penalty term $\tilde{\boldsymbol{X}} = (\mathbf{I} - \mathbf{S}_\lambda)\boldsymbol{X}$ and $\tilde{\boldsymbol{Y}} = (\mathbf{I} - \mathbf{S}_\lambda)\boldsymbol{Y}$, where $\mathbf{S}_\lambda$ is a symmetric and positive definite smoothing matrix (see, [8]). Direct algebraic computations show that the ridge type smoothing spline solution of minimization problem (2)

$$\hat{\boldsymbol{\beta}} = (\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}} + k\mathbf{I})^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{Y}} \text{ and } \hat{\mathbf{f}}_R(k) = \mathbf{S}_\lambda(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_R(k)) \tag{3}$$

However, (3) cannot be used directly due to high dimensional problem. To overcome this problem we consider ridge method based on SVD. Given the $n \times p$ data matrix , let be

$$\boldsymbol{X} = \mathbf{SVD}' = \mathbf{RD}', \tag{4}$$

the SVD of $\boldsymbol{X}$. Where $\mathbf{D}$ is the $p \times n$ orthonormal matrix, $\mathbf{S}$ is the $n \times n$ orthogonal matrix and $\mathbf{V}$ is a diagonal matrix. Also, assume that $\mathbf{R} = \mathbf{SV}$ is a $n \times n$ matrix to make required calculations. Replacing $\tilde{\boldsymbol{X}}$ in (3) by $\mathbf{RD}'$ and after some further manipulations, this can be shown to equal

$$\hat{\boldsymbol{\beta}}_{ridge} = \mathbf{D}(\mathbf{R}'\mathbf{R} + k\mathbf{I})^{-1}\mathbf{R}'\tilde{\boldsymbol{Y}} \text{ and } \hat{\mathbf{f}}_{ridge} = \mathbf{S}_\lambda(\boldsymbol{Y} - \boldsymbol{RD}'\hat{\boldsymbol{\beta}}_{ridge}) \tag{5}$$

1 Simulation Study

In here, we made a short simulation study to illustrate how the estimation method works on high dimensional data. As we said before, purpose of our study is estimating the with partially linear model for high dimensional data ($p > n$). To realize our aim we generate the partially linear model as follows

$$Y_i = X_{ij}\beta_i + f(z_i) + \epsilon_i, 1 \leq i \leq 100, 1 \leq j \leq 100, \tag{6}$$

where $X_{ij}\beta_i$ is the parametric component of the model and it contains the high-dimensional X which is formed by huge number of variable. The ϵ_i's are the random error terms generated from N($\mu = 0, \sigma = 0.1$). z_i is the nonparametric variable defined by

$$\left\{ z_i = \left(\frac{i - 0.5}{n} \right), \, i = 1, \ldots, 50 \right\}$$

and $f(\cdot)$ is the smooth function to be estimated given below

$$f(z_i) = \sin(2z_i) + 2e^{-16z_i^2}.$$

Matrix and vector form of model (6) could be written as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \mathbf{f} + \boldsymbol{\epsilon}. \tag{7}$$

This section, we estimated the function $\mathbf{f}$ and the coefficient vector $\boldsymbol{\beta}$ with smoothing spline method. We made this short simulation study for only one sample size (50) and three different number of parameters (300, 500 and 1000) with 1000 repetition. In addition to that we used mean square error (MSE) as a performance measurement for estimated function $\hat{\mathbf{f}}$ and fitted values ($\hat{Y}$) given as respectively

$$MSE\left(\hat{f}(z_i)\right) = \frac{1}{n}\sum_{i=1}^{n}\left(f(z_i) - \hat{f}(z_i)\right)^2$$

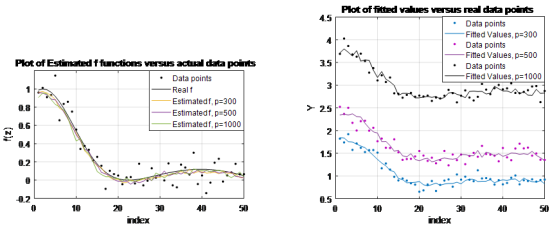

Figure 1: Plots for estimated functions and fitted values.

and

$$MSE\left(\hat{Y}_l\right) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_l\right)^2.$$

Outcomes of study are summarized with following tables and figures. Table 1 shows that suggested method works well on high-dimensional

n	p	$MSE(\hat{f})$	$MSE(\hat{Y}_l)$
	300	0.00074	0.0089
50	500	0.00095	0.0079
	1000	0.00160	0.0092

Table 1: MSE values for estimated f and fitted values

data. It can be clearly seen that, when number of parameters is low, smaller MSE values are obtained. Moreover, we can say that, in large number of parameters, MSE values are still convincing.

For satisfaction, Figure 1 is given below which includes two panels. Left panel illustrates the estimation of nonparametric component of the model and right panel is the results of fitted values. In Figure 1, we can see that when number of parameter is getting higher, quality of estimation is falling in both panel as we expected. These are the basic results for this experiments.

We should say that, we will extend this paper with some real data experiments such as time-series, censored data and so on.

Keywords: High dimensional data; Smoothing spline; Ridge regression; Partially linear models

References

[1] Hoerl, A.E. and Kennard, R.W. (1970a) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55-67.

[2] Hoerl, A.E. and Kennard, R.W. (1970b) Ridge regression: Applications to nonorthogonal problems. *Technometrics* **12**, 69-82.

[3] Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267-288.

[4] Engle, R.F., Granger, C.W., Rice, J. and Weiss, A. (1986) Semiparametric estimates of the relation between weather and electricity sales. *Journal of The American Statistical Association* **81**, 310-320.

[5] Eubank, R.L. (1988) *Spline Smoothing and Nonparametric Regression*. Marcel-Dekker, New York.

[6] Wahba, G. (1990) *Spline Model For Observational Data*. Siam, Philadelphia, PA: SIAM.

[7] Speckman P. (1988) Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B* **50**, 413-436.

[8] Green, P.J. and Silverman, B.W. (1994) *Nonparametric Regression and Generalized Linear Model*. Chapman & Hall.

An overview of big data applications

Fernanda Figueiredo
CEAUL, Universidade de Lisboa, and Faculdade de Economia, Universidade do Porto,
otilia@fep.up.pt

Adelaide Figueiredo
Faculdade de Economia and LIAAD-INESC TEC-Porto, Universidade do Porto, and CEAUL, Universidade de Lisboa,
adelaide@fep.up.pt

M. Ivette Gomes
DEIO and CEAUL, Universidade de Lisboa,
ivette.gomes@fc.ul.pt

Abstract
In the 21st century, the possibility of using sophisticated and powerful computers and data warehouses together with advanced technology has increased substantially, and allows that a very large amount of more detailed information is accessible to be processed and analyzed. Although several traditional statistical methods can be used to analyze big and complex data sets, different powerful tools are required and need to be developed. In this work we present some areas of big data applications, with some illustrative examples found in the literature, namely in the area of statistical quality control. We also mention some data mining, machine learning and statistical tools that are being used in the context of big data, and we refer to some possible statistical methodologies that need to be developed.

Keywords: Big data; Data analysis; Monitoring and surveillance; Sampling; Statistical quality control

Acknowledgments

Research partially supported by National Funds through **FCT** - Fundação para a Ciência e a Tecnologia, project UID/MAT/00006/2013.

A spline-based approach for the clustering of high dimensional data

Joaquim Costa
Department of Mathematics, Faculty of Sciences, University of Porto, Centre of Mathematics of the University of Porto,
jpcosta@fc.up.pt
A. Rita Gaio
Department of Mathematics, Faculty of Sciences, University of Porto, Centre of Mathematics of the University of Porto,
argaio@fc.up.pt

Abstract
Model based clustering for high dimensional data, and in particular sparse data, remais a substantial challenge. The number of variables is sometimes much larger than the number of observations ($p >> N$) and thus the number of parameters far exceeds the number of observations. We propose a Gaussian model-based clustering approach that greatly reduces this number of parameters. The variables are initially ordered according to their means; then, in each component of the mixture, the mean vector is modelled by a linear spline that is dependent on the order of the variables. Within each component, the variables are taken to be independent and the variances to be piecewise constant. The estimation of the parameters is performed by the Expectation-Maximization algorithm. We illustrate the use of these technique with a well-known dataset on cancer, consisting of 62 individuals. The knots in the above mentioned splines are taken to be for instance in a geometric progression, in case the mean values present a geometric behavior, that refines the variables with the highest means (also presenting the highest slopes). This is because the initial variables (the ones with lowest means) are relatively uniform with respect to their means. For q knots and K components, the total number of parameters is $K(1 + (q + 2) + (q + 1)) - 1$.

Estimation of Markov transition probabilities via clustering

Matilde Castro de Oliveira
Faculty of Sciences and Technology of the New University of Lisbon, *msp.oliveira@campus.fct.unl.pt*

Manuel L. Esquível
Department of Mathematics, Faculty of Sciences and Technology of the New University of Lisbon & CMA/FCT/UNL, *mle@fct.unl.pt*

Susana Nascimento
Department of Informatics, Faculty of Sciences and Technology of the New University of Lisbon, *snt@fct.unl.pt*

Hugo R. Lopes
National Scool of Public Health of the New University of Lisbon, *hugo.ramalheira.lopes@gmail.com*

Gracinda R. Guerreiro
Department of Mathematics, Faculty of Sciences and Technology of the New University of Lisbon & CMA/FCT/UNL, *grg@fct.unl.pt*

Abstract
We report on the clustering analysis of a database of continuing care in 2015 in Portugal, rich of 120 000 records with 70 variables each. Our main goal was to recognize a small number of dependence states in the general population and to estimate transition probabilities between every pair of states.

Keywords: Clustering; K-medoids; Markov transition probabilities; long term care

Introduction

This work is part of a project in the area of *Long Term Care* (LTC) insurance. With the transition probabilities determined via clustering analysis we were able to calibrate the intensities of a continuous

time Markov process, that was taken as a model of a multiple dependance state population evolution scheme. By Monte Carlo simulation of this Markov continuous time process – done after numerically integrating the correspondent Kolmogorov forward equations – it was possible to obtain values for premiums of different kinds of LTC insurance contracts (see [1]).

The *PCAHS* datebase

The *Portuguese Central Administration of Health System* (*PCAHS*) of 2015 is summarily described next. The system used by the RNCCI professionals to assess the LTC individuals is called Integrated Biopsychosocial Assessment Instrument. This instrument is used to register information regarding individuals' sociodemographic characteristics, care process and dependence level. This instrument incorporates different international scales for the assessment of the cognitive and physical status. Regarding the dependence levels, each individual is classified into one of four levels.

Clustering Analysis of *PCAHS* datebase

Based on previous work we had the following goals to the performed clustering analysis (see [4] and [2]).

1. To use data on the items *Cognition – Physical and Mental Control*, *Daily Activities* and *Locomotion Capabilities* in order to identify distinct dependency states of the population.

2. Verify the stability of these states not only through time – one year long period – but also over the whole set of records of the database.

3. To determine the probability transition matrices between states by considering different evaluations of the same subject at different dates.

In order to achieve these goals the general methodology was the following.

1. The clustering analysis using *PAM* algorithm was first applied to subsets of dimension 20 000 of the whole set of records of the database in order to identify the possible medoids as main representatives of possible dependence states; the preassigned number of medoids was run for 3 to 8 and the stability of these medoids on at least three large disjoint subsets of the database was confirm ed. It was also noticed that in an appreciable number of situations, the same medoids would remain when the number of clusters increased. As a consequence, a detailed analysis was made on the most significant medoids found and the final number of **four** clusters was chosen to account for the following dependence states: **healthy**, **light dependence**, **moderate dependence** and **heavy dependence**.

Cognition										Daily Life Activities								Locomotion		
Year	Month	Day	Esta	Week day	Country	District	City	House	Floor	Washing	Dressind	Toilet	Laying	Sitting	Urine	Fezes	Feeding	House	Street	Stairs
3	3	3	3	3	3	3	3	3	3	2	2	3	3	3	3	3	3	2	2	2
3	3	3	3	3	3	3	3	3	3	1	1	1	1	1	2	2	2	1	1	0
0	0	0	0	0	3	3	3	0	0	1	1	1	1	1	1	1	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

2. The database records were then separated by user and evaluation dates associated to each user; the number of transitions from one dependence state to the other was determined by using the cluster classification associated to the representative medoids chosen. The result obtained was a set of transition matrices associated with five different class ages with a similar number of instances in each of them. The age sets were cho-

Age sets	**Observations**
[60, 71]	20 752
[72, 77]	20 613
[78,81]	18 424
[82,86]	24 231
[87,107]	21 374
Total	**105 394**

sen so that all the numbers of observations in these sets are

comparable. This is desirable for having similar approximation errors in the coefficients of the Markov probability transition matrices.

3. It was observed an abnormal low number count of deaths that biased the transition probabilities to the state of death; further analysis showed that a significant number of the deaths occurring in the population – as expected by present day mortality laws – were not recorded in the database due to the fact that only when a person is using one of the institutions of the health system is, his or her death registered in the database. As so, the transition probability matrices were corrected by the 2015 Portuguese Mortality able of the National Institute of Statistics. Finally, by an averaging procedure – over the five age classes – a global probability transition matrix was obtained.

The final five probability matrices – corresponding, respectively, to age classes [60,71], [72,77], [78,81], [82,86] and [87,107] and the averaged one – according to the number of records in each class as in formula just below, are as follows.

$$
\begin{aligned}
{}_tP_x = {}_t\,P_{x\in[60\,;\,71]}\frac{20752}{105394} \;+\; {}_tP_{x\in[72\,;\,77]}\frac{20613}{105394} + {}_t\,P_{x\in[78\,;\,81]}\frac{18424}{105394} + \\
+ {}_t\,P_{x\in[82\,;\,86]}\frac{24231}{105394}\,{}_tP_{x\in[87\,;\,107]}\frac{21374}{105394}
\end{aligned}
$$

$$
\begin{pmatrix}
83.47\% & 11.05\% & 1.44\% & 0.66\% & 3.38\% \\
18.49\% & 68.66\% & 4.61\% & 2.97\% & 5.28\% \\
6.72\% & 20.48\% & 45.85\% & 17.50\% & 9.45\% \\
1.21\% & 6.44\% & 10.04\% & 67.86\% & 14.45\% \\
0.00\% & 0.00\% & 0.00\% & 0.00\% & 100.00\%
\end{pmatrix}
\begin{pmatrix}
81.63\% & 11.88\% & 1.71\% & 0.57\% & 4.20\% \\
16.57\% & 67.68\% & 6.26\% & 3.30\% & 6.20\% \\
3.53\% & 18.41\% & 50.12\% & 15.16\% & 12.78\% \\
0.91\% & 5.43\% & 10.01\% & 63.65\% & 20.00\% \\
0.00\% & 0.00\% & 0.00\% & 0.00\% & 100.00\%
\end{pmatrix}
$$

$$
\begin{pmatrix}
76.60\% & 13.80\% & 2.70\% & 0.84\% & 6.06\% \\
13.82\% & 66.47\% & 7.87\% & 3.95\% & 7.89\% \\
3.29\% & 15.63\% & 45.72\% & 17.48\% & 17.88\% \\
0.52\% & 3.74\% & 8.90\% & 58.60\% & 28.24\% \\
0.00\% & 0.00\% & 0.00\% & 0.00\% & 100.00\%
\end{pmatrix}
\begin{pmatrix}
68.41\% & 15.69\% & 4.89\% & 1.02\% & 9.99\% \\
11.43\% & 64.26\% & 8.87\% & 3.97\% & 11.47\% \\
2.97\% & 13.13\% & 40.26\% & 14.06\% & 29.57\% \\
0.34\% & 2.64\% & 6.70\% & 45.39\% & 44.93\% \\
0.00\% & 0.00\% & 0.00\% & 0.00\% & 100.00\%
\end{pmatrix}
$$

$$\begin{pmatrix} 54.15\% & 12.63\% & 6.48\% & 1.48\% & 25.26\% \\ 7.53\% & 52.54\% & 9.33\% & 4.17\% & 26.43\% \\ 1.22\% & 6.57\% & 23.08\% & 8.45\% & 60.69\% \\ 0.09\% & 0.85\% & 2.23\% & 18.06\% & 78.77\% \\ 0.00\% & 0.00\% & 0.00\% & 0.00\% & 100.00\% \end{pmatrix} \begin{pmatrix} 72.50\% & 13.08\% & 3.53\% & 0.92\% & 9.97\% \\ 13.45\% & 63.80\% & 7.44\% & 3.68\% & 11.63\% \\ 3.52\% & 14.72\% & 40.76\% & 14.41\% & 26.59\% \\ 0.60\% & 3.76\% & 7.48\% & 50.15\% & 38.00\% \\ 0.00\% & 0.00\% & 0.00\% & 0.00\% & 100.00\% \end{pmatrix}$$

Remark 1 The transition probabilities matrix obtained by this clustering analysis further corrected by the mortality tables is comparable, yet different, to a previous determined probability transition matrix using a set of 600 records of a different population (see for full details [3]).

Conclusion

The clustering algorithms *PAM* and *CLARA* were applied to a medium size database to obtain probability transition matrices among dependence states corresponding to stable clusters around the cluster's medoids.

Acknowledgments We express our gratitude to the *Portuguese Central Administration of Health System* for the availability of the database 2015. This work was partially supported by the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the project UID/MAT/00297/2013 (Centro de Matemática e Aplicações).

References

[1] Manuel L. Esquível, Matilde Castro de Oliveira, Gracinda Guerreiro, Cristina Nobre (2017) Calibration and Simulation of a Continuous Time Markov Chain Model for Long Term Care. *Proceedings of Second International Conference on Computational Finance 2017*, 136–141.

[2] Kaufman, L., Rousseeuw, P.J. (1987) Clustering by means of Medoids. In *Statistical Data Analysis Based on the L^1–Norm and Related Methods*, edited by Y. Dodge, North-Holland, pp. 405–416.

[3] Matilde Castro de Oliveira (2017) Calibração e Simulação num Modelo de Cadeias de Markov para *Long Term Care*, Master Thesis, Mestrado em Matemática e Aplicações, Faculty of Science and Technology, New University of Lisbon.

[4] Hae-Sang Park, Chi-Hyuck Jun, (2009) A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications* **36**, 3336-3341

Highway traffic analytics - detecting mobility patterns in a Portuguese operator big data system

Raquel João Fonseca
CMAFCIO, Universidade de Lisboa, *rjfonseca@ciencias.ulisboa.pt*
J.M. Pinto Paixão
CMAFCIO, Universidade de Lisboa, *jmpaixao@ciencias.ulisboa.pt*
J. Telhada
Faculdade de Ciências, Universidade de Lisboa,
joao.telhada@ciencias.ulisboa.pt

Abstract
With traffic counters located at strategic points along a busy suburban highway, data concerning the number of vehicles, speed, and weight, just to name a few, is collected continuously. Given such volume of data available, it is in the interest of the highway operator to develop an automatic and self-learing mechanism in order to construct a number of business management key performance indicators and assist in the decision-making process.

With that in mind, this project aims to identify and model patterns regarding the utilization of the highway by its drivers, particularly as far as their entry and exit points are concerned. Preliminary results estimating traffic flow seem to point towards the need to produce and apply methodologies that will enable the detection of similarities and distinctions among different time periods in a day, and in between different days.

Keywords: Pattern recognition; Big data; Highway traffic management

Effect of age, state of survival and proximity to death on the care costs of the beneficiaries of a health care operator

Rômulo Alves Soares
Universidade Federal do Ceará, *romuloalves61@gmail.com*
Silvia Pedro Rebouças
Universidade Federal do Ceará, *smdpedro@gmail.com*
Clever de Souza Gondim
Unimed Fortaleza, *cleversg@gmail.com*

Abstract
The objective of this article is to evaluate the effect of age, survival status and proximity to death on care costs in a health plan operator (OPS). For this, information about 300,000 beneficiaries of a large OPS in the state of CearÃ¡, between the years of 2014 and 2015, was used. The information collected was related to the monthly costs of these users in the analyzed period, their age, whether they remained alive or not in the period, their sex and the number of months until the death for the beneficiaries who died. The statistical analyzes conducted involved T and Wilcoxon tests for the comparison of the annual costs of surviving and non-surviving beneficiaries and quantum regression to verify the influence of the independent variables on the cost of care. In addition, the classification and regression tree and random forest techniques were used to verify the degree of importance of these variables in determining the cost. The results show that both age and survival status are able to influence cost when considering all beneficiaries, which is also observed for age and proximity to death when only beneficiaries who have died are taken into account. However, it is important to note that the importance given to age in relation to other variables (survival status and proximity to death) is quite low, while sex in some of the analyzes is not even significant. Based on the results obtained, it

is important that the state of survival and proximity to death be taken into account for the projection of healthcare costs in OPS, in order to guarantee its sustainability in a sensitive and increasingly competitive market.

Keywords: Care costs; Near death; Survival status; Random forest; Classification and regression tree

Analysis of the determinants of profitability and loyalty of the beneficiaries of a dental plan using classification and regression trees

Silvia Pedro Rebouças
Universidade Federal do Ceará, *smdpedro@gmail.com*
Aline Rodrigues Martins
Adtalem Educacional do Brasil, *aliner-martins@yahoo.com.br*
Rômulo Alves Soares
Universidade Federal do Ceará, *romuloalves61@gmail.com*

Abstract
The present study aims to verify the main determinants of the profitability and loyalty of the beneficiaries of a given Brazilian dental plan. The sample used is comprised of 42,784 beneficiaries who joined the plan between January 1, 2006 and December 31, 2012, which were active on April 30, 2013 and belonged to the Fortaleza branch. The interval between December 2012 and April 2013 represents the grace period contemplated in the dental plans marketed by this operator. The beneficiaries who were in the grace period were not included. To achieve the objectives of the research, we used Classification and Regression Trees (CART). This method is based on the successive binary division of data based on the sampling results of independent variables, seeking the creation of subsets that are more homogeneous with respect to the dependent variable. Data analysis also included descriptive statistics, t-tests for independent samples, ANOVA, correlation analysis and Chi-Square tests to characterize the sample, to compare groups and to analyze relations between variables. The CART multivariate technique allowed the tracing of customer profiles with similar profitability or loyalty and quantifying the importance of each determinant in the regression, in the case of profitability, or in the classification, in the case

of loyalty. Profitability was assessed by the 12-month accumulated contribution margin of each beneficiary. The means of payment, the segment in which the beneficiary operates and contract time are the main determinants of profitability. To predict customer loyalty, the key determinants are contract time, contribution margin and means of payment. It is concluded that this operator should focus on loyalty actions in the months of the first year of the contract.

Keywords: Dental plans; Profitability; Loyalty; Classification; Regression trees

Authors Index

Joint organization of the Institute of Financial Big Data, Universidad Carlos III de Madrid, Centro de Estatística e Aplicações, Universidade de Lisboa (CEAUL), Sociedad Española de Estadística e Investigación Operativa (SEIO) and Sociedade Portuguesa de Estatística (SPE)

Financial support: FCT Portugal UID/MAT/00006/2013