SYSORM 2019

2ND SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

El Escorial
5-7 June 2019

Conference proceedings

UNIVERSITAS COMPLVTENSIS MATRITENSIS

EURO THE ASSOCIATION OF EUROPEAN OPERATIONAL RESEARCH SOCIETIES

uc3m Instituto Flores de Lemus

UNIVERSIDAD CARLOS III DE MADRID homo homini sacra res

MAPAL software

Instituto de Matemática Interdisciplinar

S E I O

Contents

Welcome letter

Dear SYSORM participant,

On behalf of SEIO, the (Spanish) Statistics and Operations Research Society, welcome!

We are about 700 hundred members all around Spain (and also some from other European and Latin American countries) sharing our interest in Statistics and Operations Research as a profession, with either an academic or an industrial perspective.

Our Society is rather singular, since both disciplines, Statistics and Operations Research, appear under the same umbrella. This singularity has turned into an exceptional opportunity in the Internet of the Things era, when any decision process has a large and complex data set behind, and any data analysis is calling for the use of rather sophisticated optimization algorithms as well.

Our Society is also singular since, besides its regular conferences –with a 1.5-year periodicity–, it organizes a conference specific for researchers in the earlier stages of their career. The 1st Spanish Young Statisticians and Operational Researchers Meeting (SYSORM) took place in Granada in September 2017. This one in El Escorial will be the second –and definitely not last– SYSORM. I am sure this new edition of SYSORM will be a great scientific success and will create a solid cooperation network, reinforcing the academic links among young researchers from many different places in Spain and abroad.

SYSORM will allow you all to learn a lot from the plenaries and participants at the same time that you make new friends. This will be a singular and unforgettable experience. Enjoy!

Emilio Carrizosa

President of SEIO

Committees

Editors: Organizing Committee

Organizing Committee

F. Javier Martín-Campo (**Chair**) - Complutense University of Madrid

Aida Calviño Martínez (Co-Chair) - Complutense University of Madrid

Elena Castilla González - Complutense University of Madrid

Javier León Caballero - Complutense University of Madrid

Inmaculada Flores García - Complutense University of Madrid

Adán Rodríguez Martínez - Complutense University of Madrid

María Sierra Paradinas - IDOM - King Juan Carlos University

Paula Terán Viadero - MAPAL Software

Gregorio Tirado Domínguez (Co-Chair) - Complutense University of Madrid

Guillermo Villarino Martínez - Complutense University of Madrid

Scientific Committee

M. Carmen Aguilera-Morillo - Carlos III University of Madrid

Eduardo García Portugués - Carlos III University of Madrid

Vanesa Guerrero Lozano - Carlos III University of Madrid

Beatriz Sinova Fernández - University of Oviedo

Organizers

UNIVERSITAS COMPLVTENSIS MATRITENSIS

S E
I O

EURO
THE ASSOCIATION OF
EUROPEAN OPERATIONAL
RESEARCH SOCIETIES

Co-organizers

Instituto de
Matemática
Interdisciplinar

uc3m
Instituto Flores
de Lemus

uc3m Universidad
Carlos III
de Madrid

Collaborators

imus

APDIO
ASSOCIAÇÃO PORTUGUESA
DE INVESTIGAÇÃO OPERACIONAL

MAPAL
software

ROADEF

Cooperation

ASPRODES
Plena
inclusión
Castilla y León

Schedule

Wednesday, June 5

8.30 - 9.30	Breakfast
9.30 - 10.00	Opening Session
10.00 - 11.20	Session 1
11.30 - 12.00	Coffee break
12.00 - 13.00	Plenary session 1:
	Giovanni Righini
13.00 - 14.00	Session 2
14.00 - 15.20	Lunch
15.20 - 16.20	Plenary session 2:
	Richard Samworth
16.20 - 17.00	Session 3
17.00 - 17.30	Coffee Break
17.30 - 18.50	Session 4
18.50 - 19.00	Break
19.00 - 20.20	Session 5
21.30	Welcome reception

Thursday, June 6

8.30 - 9.30	Breakfast
9.40 - 11.00	Session 6
11.00 - 11.30	Data Science in Action: Mapal Software
11.30 - 12.00	Coffee Break
12.00 - 14.00	Guided Visit to the Royal Monastery of San Lorenzo de El Escorial
14.00 - 15.20	Lunch
15.20 - 16.20	Plenary session 3: Elena Fernández
16.20 - 17.00	Session 7
17.00 - 17.30	Coffee Break
17.30 - 18.50	Session 8
18.50 - 19.00	Break
19.00 - 20.00	Session 9
22.00	Gala dinner

Friday, June 7

8.30 - 9.30	Breakfast
9.30 - 10.30	Plenary session 4:
	Pedro Delicado
10.30 - 11:30	Session 10
11.30 - 12.00	Coffee break
12.00 - 13.40	Session 11
13.40 - 14.00	Closing Session
14.00 - 15.00	Lunch

Plenary speakers

Speakers

Giovanni Righini

Affiliation University of Milan

Talk *An introduction to column generation (with some examples)*

Date Wednesday, June 5, 12.00

Chair Vanesa Guerrero Lozano

Bio Giovanni Righini is full professor of Operations Research at the Department of Computer Science of the University of Milan. His research interest focuses on combinatorial optimization and mathematical programming, particularly in vehicle routing problems, including both exact and heuristic algorithms. He is the founder and director of the Operations Research Laboratory "OptLab" at the University of Milan (since 1998). He belonged to the board of AIRO (Associazione Italiana di Ricerca Operativa) between 2005 and 2011, received an IBM Faculty Award in 2010 and was "Profesor Visitante Distinguido" at Complutense University of Madrid (December 2011-February 2012).

According to Scopus, he has published 57 documents, with an h-index equal to 15 and 1286 citations (accessed March 4, 2019). Among these publications, his contributions in the journals *Transportation Research C and E* (1 and 1), *Transportation Science* (2), *Computers & Operations Research* (4), *European Journal of Operational Research* (4) and *Operations Research* (3) could be highlighted.

He has been responsible of 2 public funded research projects and 10 research projects with private companies. Moreover, he has been advisor of 7 Ph.D. students and 7 post-doc fellowships.

Richard Samworth

Affiliation University of Cambridge

Talk *Classification with imperfect training labels*

Date Wednesday, June 5, 15.20

Chair Eduardo García Portugués

Bio Richard Samworth is Professor of Statistical Science and Director of the Statistical Laboratory at the University of Cambridge, and Alan Turing Institute Fellow. His main research lines are shape-constrained estimation, nonparametric statistics, high-dimensional statistical inference, resampling methods, and applications of statistics. He is present co-editor of the renowned *Annals of Statistics* (2019-) and, among others, is or has been associate editor of prestigious journals such as *Annals of Statistics* (2013-), *Biometrika* (2011-2014), *Journal of the American Statistical Association* (2017-), *Journal of the Royal Statistical Society – Series B* (2006-2014), and *Statistical Science* (2017-).

Among the many honours and awards he has received during his career, the recent COPSS President's Award for 2018 is to be highlighted. This is one of the highest awards in statistics that is jointly conceded by, among others, the American Statistical Association (ASA) and the Institute of Mathematical Statistics (IMS). Other highlighted honours and awards include the prestigious Adams Prize 2017 from the University of Cambridge for UK-based mathematicians, the elected fellowships for the ASA and IMS, and the two Early Career Fellowships (starting + extension) from the UK's Engineering and Physical Sciences Research Council (EPSRC).

According to Scopus, he has published 32 documents, with an h-index equal to 15 and 921 citations in total (accessed January 11, 2019). More importantly than these metrics is the outstanding quality of his publications, reflected in contributions to internationally-renowned statistical journals such as *Annals of Statistics* (11), *Journal of the Royal Statistical Society – Series B* (7), *Statistical Science* (3), *Biometrika* (2), *Journal of the American Statistical Association* (1) or *Bernouilli* (1).

Elena Fernández Aréizaga

Affiliation Polytechnique University of Catalonia

Talk *Revisiting Some Location/Routing Problems*

Date Thursday, June 6, 15.20

Chair Beatriz Sinova Fernández

Bio Elena Fernández Aréizaga is full professor at the Department of Statistics and Operations Research of the Polytechnique University of Catalonia. Her research interest focuses on discrete optimization, mainly problems in the areas of discrete location, vehicle routing and network design. She is associated editor of *TOP* (1995-2000 and from 2007). She also belongs to the editorial board of *Computers & Operations Research* (from 2003) and to the advisory board of *EURO Journal on Computation Optimization* (from 2007). She has also been invited editor of a special issue of *Annals of Operations Research.*

According to Scopus, she has published 78 documents, with an h-index equal to 26 and 1606 citations (accessed January 11, 2019); according to Google Scholar, her h-index and citations raise to 31 and 3205, respectively (accessed January 11, 2019). Among these publications, her contributions in the journals *European Journal of Operational Research* (18), *Computers & Operations Research* (13), *Transportation Science* (4), *Omega* (4) and *Operations Research* (3) could be highlighted.

She has been the principal investigator of 19 public funded research projects. She has supervised 11 PhD thesis.

Pedro Delicado

Affiliation Polytechnique University of Catalonia

Talk *Understanding complex models with Ghost Variables*

Date Friday, June 7, 09.30

Chair M. Carmen Aguilera-Morillo

Bio Pedro Delicado is Doctor in Economics from Carlos III University of Madrid (Extraordinary Doctorate Award). Now, he is associate professor at the Department of Statistics and Operations Research of the Polytechnique University of Catalonia. His most recent management positions include director of the Inter-University Master's Degree in Statistics and Operational Research at the Polytechnique University of Catalonia and the University of Barcelona. His research interest focuses on resampling methods, nonparametric statistics and smoothing techniques, principal curves, dimensionality reduction, distance based statistical methods, functional data analysis, spatial statistics for functional data and Big Data. He has been a member of the BBVA Foundation's Committee for the Evaluation of Grants for Scientific Research Teams in Big Data 2017 and evaluator for research agencies such as the Agencia Nacional de Evaluación y Prospectiva (ANEP) and the National Science Foundation (USA).

According to Scopus, he has published 38 documents, with an h-index equal to 12 and 498 citations (accessed January 11, 2019); according to Google Scholar, his h-index and number of citations raise to 16 and 1148, respectively (accessed January 11, 2019). Among his publications, the contributions published in the *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (1), *Journal of the Royal Statistical Society: Series A (Statistics in Society)* (1), *Biometrika* (1), *Journal of Multivariate Analysis* (2), *Computational Statistics and Data Analysis* (5), *Computational Statistics* (5) could be highlighted.

He has been the principal researcher of 6 public funded research projects and 5 research projects with private companies. Moreover, he has supervised 5 PhD thesis.

Session 1 (Wednesday 10:00)

Session talks

Estimating marginal likelihoods by Gibbs sampling

Fernando Llorente [*] Luca Martino [†] David Delgado [‡]

Fernando Llorente (PhD Student) Fernando Llorente Fernández is a Phd student in statistics at Carlos III University of Madrid (campus of Leganés). His main interests are Monte Carlo methods, and more specifically, those based on Markov chains, i.e. Markov chain Monte Carlo (MCMC), as well as their variants such as adaptive MCMC, scalable MCMC, etc.; and machine learning, especially pattern recognition. He received his B.S. degree in physics from Autonomous University of Madrid in 2016. He earned his master's degree in mathematical engineering (major in statistics) from Carlos III University of Madrid in 2018. He started his PhD thereafter under the supervision of Prof. David Delgado (Carlos III University of Madrid) and Prof. Luca Martino (King Juan Carlos University).

Bayesian inference has become very popular in different fields such as machine learning and signal processing [1, 2]. Given a probabilistic model linking the data y with the variable of interest x (i.e., the conditional density $p(y|x)$ a.k.a. the likelihood function) and a prior density $p(x)$ decided by the user in advance, the Bayesian methods relies on the study of the posterior density $p(x|y)$. However, in many applications, the analytical study of the posterior distributions is difficult or impossible. Moreover, in general, only the product $p(x)p(y|x)$ (proportional to the posterior density $p(x|y)$) can be evaluated while the normalizing constant, i.e. $p(y) = \int p(x)p(y|x)dx$, is unknown. This normalizing constant, usually denoted by Z, is called marginal likelihood or Bayesian evidence, and it is useful for model selection purposes (i.e., to compare different models by means of Bayes' factors).

To overcome these problems, one possibility is to employ conjugate models where prior and posterior distributions belong to the same family. However, conjugate models are not available or suitable in many realistic scenarios. An alternative is the use of Monte Carlo methods for approximating the posterior density by random samples [3]. More specifically, in the last decades, Markov chain Monte Carlo sampling methods, such as the Metropolis-Hastings algorithm and the Gibbs sampler, have been successfully applied to perform approximate Bayesian inference, helping to popularize this paradigm [1, 4]. For instance, MCMC methods allow estimating moments of the posterior distribution by drawing correlated samples. The only requirement is being able to evaluate some function proportional to the posterior. However, the computation of the marginal likelihood Z is not straightforward using MCMC algorithms. This task becomes harder and harder as the dimension of the inference space grows. In this work, we describe a novel scheme for estimating the marginal likelihood Z by means of a Gibbs procedure. More precisely, given a vector $\mathbf{x} = (x_1, \dots, x_D)$, we use the D full-conditional distributions, i.e. the distribution of each x_d conditional on the other $D-1$ components, and their normalizing constants Z_d $(d = 1, \dots, D)$ to estimate Z. The novel technique produces samples distributed according to the posterior based on the Recycling Gibbs approach [1]. Moreover, it yields a sequence of estimators of Z which can be properly combined providing a final unique estimator.

Keywords: Bayesian inference; Markov Chain Monte Carlo (MCMC); Marginal likelihood; Model evidence; Model selection; Gibbs sampling.

References

[1] Martino, L., Casarin, R., Leisen, F., Luengo, D. (2018). Adaptive independent sticky MCMC algorithms. *EURASIP Journal on Advances in Signal Processing*, **1**, p. 5.

[2] Martino, L., Elvira, V., Camps-Valls, G. (2018). The Recycling Gibbs Sampler for Efficient Learning. *Digital Signal Processing*, **74**, pp. 1–13.

[3] Robert, C., Casella, G. (2004). *Monte Carlo statistical methods.* Springer-Verlag, New York 2004.

[4] Martino, L., Read, J., Luengo, D. (2015). Independent Doubly Adaptive Rejection Metropolis Sampling within Gibbs Sampling. *IEEE Transactions on Signal Processing*, **63**, pp. 3123–3138.

[*] Department of Statistics, Carlos III University of Madrid, Madrid, Spain. Email: felloren@est-econ.uc3m.es

[†] Department of Signal processing, King Juan Carlos University, Madrid, Spain. Email:luca.martino@urjc.es

[‡] Department of Statistics, Carlos III University of Madrid, Madrid, Spain. Email: ddelgado@est-econ.uc3m.es

How to apply Location Theory on the Portfolio Optimization Problem

Justo Puerto [*] Andrea Scozzari [†] Moisés Rodríguez-Madrena [‡]

M. Rodríguez-Madrena (PhD student) Moisés Rodríguez-Madrena is a PhD student at the University of Seville. His main fields of interest are continuous and discrete location problems and other related problems. He received his four-years Bachelor's Degree in mathematics from the University of Seville in 2016. He earned his master's degree in mathematics from the University of Seville in 2017. After finalising his master's degree, he enrolled in the doctoral program in mathematics at the University of Seville. He is a member of the following mathematical societies or groups: Red de localización y problemas afines, SEIO and EUROYoung.

Given a set of assets and an investment capital of a stilized investor, the Portfolio Optimization Problem consists on determining the amount of the investment capital to share in each asset in order to build the most profitable portfolio. The Portfolio Optimization Problem is classically modeled as a mean-risk bi-criteria optimization problem (see the seminal paper of Markowitz in 1952 [1]):

$$\max\{[\mu(\mathbf{x}), -\varrho(\mathbf{x})] : \quad \mathbf{x} \in Q\},$$

where the mean rate of return $\mu(\mathbf{x})$ of the portfolio is maximized and a risk measure $\varrho(\mathbf{x})$ is minimized, being Q the set of feasible portfolios.

New mathematical programming models and techniques are still needed in order to efficiently solve the Portfolio Optimization Problem. A relatively recent promising line of research is to exploit clustering information of an assets network in order to develop new portfolio optimization paradigms [2, 3]. In this work we endow the assets network with a metric [4] and we show how classical location problems on networks can be used for asset clustering (for the interpretation of the p-median problem in terms of cluster analysis the reader if referred to [5]). In particular, we add a new criterion to the Portfolio Optimization Problem which measures, by means of an objective function of a classical location problem, the degree of representation of the selected assets with respect to the non-selected ones.

We propose a Mixed-Integer Linear Programming formulation for dealing with this problem. The usefulness of our approach is validated reporting some preliminary computational experiments.

Keywords: Portfolio Optimization; Location Problems on Networks; Multicretiria Optimization; Mathematical Programming.

Acknowledgements The first and the third authors have been partially supported by MINECO Spanish/FEDER grant number MTM2016-74983-C02-01. The third author would also like to thank the Erasmus+ Programme.

References

[1] Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, **7**, pp. 77–91.

[2] Beraldi, P., Bruni, M.E. (2014). A clustering approach for scenario tree reduction: an application to a stochastic programming portfolio optimization problem. *TOP*, **22**, pp. 934–949.

[3] Tola, V., Lillo, F., Gallegati, M., Mantegna, R.N. (2008). Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control*, **32**, pp. 235–258.

[4] Mantegna, R.N. (1999). Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, **11**, pp. 193–197.

[5] Hansen, P., Brimberg, J., Urosevic, D., Mladenovic, N. (2009). Solving large p-median clustering problems by primal-dual variable neighborhood search. *Data Mining and Knowledge Discovery*, **19**, pp. 351–375.

[*] IMUS, Universidad de Sevilla, Sevilla, Spain. Email: puerto@us.es

[†] Università degli Studi Niccolò Cusano, Roma, Italy. Email: andrea.scozzari@unicusano.it

[‡] IMUS, Universidad de Sevilla, Sevilla, Spain. Email: madrena@us.es

What can we do by combining a bunch of points, a couple of covariates, a few nonparametric techniques and some effort?

María Isabel Borrajo *

M.I. Borrajo (Senior Research Associate) Maribel Borrajo is a Senior Research Associate at Lancaster University, where she is part of the Data Science and the Natural Environment (DSNE) project. Previoously to that, she received her Bsc in mathematics in 2012 and her master's in statistical techniques in 2014, both from the Universidade de Santiago de Compostela. After being involved for one year in an environmental project with Endesa S.A., Maribel started a phD on "Nonparametric inference on point processes with covariates" at Universidade de Santiago de Compostela, which she finished with honours in February 2018. Her main research interests are spatial statistics, extreme value analysis and their applications to environmental sciences. As part of the multidisciplinary DSNE project, she is now working on three different environmental challenges: air quality, land-use and ice-sheet melting, developing innovative statistical methodology to tackle those grand challenges and influence in a future policy-making.

The general answer would be "many things", but focusing on my own personal experience, I present here my results after putting that recipe into practice. I came out with three innovative methodologies for spatial point processes with covariates: an intensity estimation method, a goodness-of-fit test and a two-sample comparison procedure.

Point processes are a branch of spatial statistics interested in the occurrence of events: in time (one-dimensional point processes), in space (two or higher-dimensional point processes), in space and time (spatio-temporal point processes)... I am particularly interested in the scenario where extra information is provided by covariate values, and how this can improve inference in this context. First-order intensity is one of the characteristic functions in point processes, which accounts for the expected number of points per unit time/area: $\lambda(x) = \frac{E(N(x))}{|dx|} \quad x \in W \subset \mathbb{R}^2$, where N denotes the counting measure and $|dx|$ the area of the infinitesimal region dx. Focusing on this function, extensive literature exists on defining parametric models as well as testing the effect of covariates under linear dependence assumption. However, less has been done from the nonparametric point of view, where under a model stating that

$$\lambda(x) = \rho(\mathbf{Z}(x)), \tag{1}$$

with $\mathbf{Z} : W \subset \mathbb{R}^2 \to \mathbb{R}^d$ denoting the covariates, only a couple of proposals (with not much mathematical formalisation) have been presented, see for instance [2] and [1].

In this work I try to fill this existing gap in the methodological development of nonparametric techniques for the intensity function of point processes with covariates. I derive a new consistent kernel intensity estimator with the corresponding theoretical framework, including several ad-hoc bandwidth selectors and a new bootstrap procedure. We also define a goodness-of-fit test for model (1) as well as a two-sample test under this model. All the procedures are supported by the corresponding theoretical results and simulation studies. Moreover, those new contributions are applied to the analysis of wildfire data in Canada leading to interesting conclusions in terms of covariate influence and temporal variations of the spatial distribution.

Keywords: Spatial statistics; Point processes; Kernel methods; Environmental applications.

Acknowledgements Work supported by project MTM2013-41383P and MTM2016-76969P, and PhD grant FPU2013/00473. The author would also like to thank Wenceslao González-Manteiga and María Dolores Martínez-Miranda for their contributions and support.

References

[1] Baddeley, A., Chang, Y. M., Song, Y., and Turner, R. (2012). Nonparametric estimation of the dependence of a spatial point process on spatial covariates. *Statistics and Its Interface*, **5**, pp. 221–236.

[2] Guan, Y. (2008). On consistent nonparametric intensity estimation for inhomogeneous spatial point processes. *Journal of the American Statistical Association*, **103**(483), pp. 1238–1247.

*Department of Mathematics and Statistics, Lancaster University, United Kingdom. Email: m.borrajogarcia@lancaster.ac.uk

Parallel surrogate-assisted optimization based on *q-merit* functions

José Carlos García-García [*] Ricardo García-Ródenas [†] Esteve Codina Sancho[‡]

J.C. García-García (PhD student) José Carlos García-García is a PhD student at the University of Castilla-La Mancha. His main interests are machine learning and optimization. He received his B.E. from University of Castilla-La Mancha in 2015 in Computing Engineering as well as his master's degree in Computing Engineering in 2017. In that year, he enrolled in the PhD program in Information Technology Advanced at the University of Castilla-La Mancha with an FPU fellowship. García-García was the recipient of the prize awarded to the most outstanding graduate in both degrees.

The field of parallel surrogate-assisted optimization is gaining popularity [1]. We present a class of parallel algorithms for the global optimization of expensive black functions derived from the so-called *q-merit* functions. The proposed approach samples the function to be optimized at q new points at each iteration. The *q-merit* functions determine the quality of the samples, and their optimization allows the development of multi-point infill criteria.

A special instance of this class is *q-qualSolve* which generalizes the *qualSolve* [2] algorithm to multi point sampling. We prove the convergence of *q-qualSolve* under weak conditions. However, the quality function still presents the problem that the computational burden of the calculation of integrals grows exponentially with the dimension of the problem. For this reason, we have developed an alternative schema, called *q-localQual*, within the framework of the *q-merit* functions, which avoids the computation of multiple integrals. *q-localQual* also guarantees convergence to global optima.

In parallel surrogate-assisted optimization, the number of generated points can make updating the surrogate models time-consuming. For this reason, we have developed a heuristic strategy, named *pruning strategy*, which restricts the number of points retained during the optimization process.

We have carried out numerical experiments showing that *q-localQual* significantly improves the elapsed time regarding the algorithm *q-qualSolve*, EGO and MSRS.

Keywords: Parallel Optimization; Simulation Optimization; Surrogate Model; Black Box Function; Response Surface.

Acknowledgements This work has been supported by *Ministerio de Economía, Industria y Competitividad - FEDER EU* with grant number TRA2016-76914-C3-2-P and the predoctoral FPU fellowship from the *Ministerio de Educación, Cultura y Deportes* with number 16/00792.

References

[1] Haftka, R.T., Villanueva, D., Chaudhuri, A. (2016). Parallel surrogate-assisted global optimization with expensive functions - a survey. *Structural and Multidisciplinary Optimization*, **54**, pp. 3–13.

[2] Jakobsson, S., Patriksson, M., Rudholm, J., Wojciechowski, A. (2010). A method for simulation based optimization using radial basis functions. *Optimization and Engineering*, **11**, pp. 501–532.

[*] Escuela Superior de Informática, Universidad de Castilla-La Mancha, Ciudad Real, Spain. Email: josecarlos.garcia@uclm.es
[†] Escuela Superior de Informática, Universidad de Castilla-La Mancha, Ciudad Real, Spain. Email: ricardo.garcia@uclm.es
[‡] Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, Barcelona, Spain. Email: esteve.codina@upc.edu

Session 2 (Wednesday 13:00)

Session talks

Advances for indirect questioning techniques

Beatriz Cobo [*] María del Mar Rueda [†]

B. Cobo (Interim substitute professor) Beatriz is PhD in Mathematical and Applied Statistics from the University of Granada and Technical Engineer in Management Computing from the University of Jaén. She is currently an interim substitute professor at the Department of Statistics and Operational Research of the University of Granada, and her research focuses on the use of auxiliary information in surveys for indirect questioning techniques of sensitive questions and their computational treatment.

A survey is a research method that is based on questioning a sample of individuals. The interest in sample surveys studies often focuses on sensitive or confidential aspects to the interviewees. Because of this, the typical problem that arises is social desirability, which is defined as the tendency of respondents to answer based on what is socially acceptable. For this reason, many respondents refuse to participate in the survey or provide false answers or conditioned answers, altering the accuracy and reliability of the estimations in a major way.

Randomized Response (RR) Technique (RRT) introduced by [3] is a possible solution for protecting the anonymity of the respondent and is used to reduce the risk of escape or no response to sensitive questions. Warner's study generated a rapidly-expanding body of research literature on alternative techniques for eliciting suitable RR schemes in order to estimate a population proportion. Standard RR methods are used primarily in surveys which require a binary response to a sensitive question, and seek to estimate the proportion of people presenting a given (sensitive) characteristic. Nevertheless, some studies have addressed situations in which the response to a sensitive question results in a quantitative variable.

To contribute to the development of the RRT, in recent years we have made a series of methodological advances, such as including auxiliary information in the calculation of the estimators, considering more than one sampling frame and developing software for the calculation of the estimations considering complex sampling designs.

Warner's work originated a large amount of literature and has been used in many areas, but these techniques have difficulties and limitations. Due to this, other indirect techniques emerged as an alternative to RRT, among them we found the item counting technique (ICT) [2, 1]. This technique was designed for surveys that require the study of a qualitative variable, but many practical situations can deal with sensitive variables that are quantitative in nature. To do this, the item sum technique (IST) was proposed as a generalization of ICT.

To contribute to the development of the IST in real studies, we suggest some methodological advances, such as the estimation of the IST under a generic sampling design, the use of auxiliary information to improve the efficiency of the estimates, the estimation of more than one sensitive variable at the same time and how to perform the division of the total sample into two subsamples.

Finally, we use indirect survey techniques to investigate some sensitive variables in real studies.

Keywords: Indirect questioning techniques; Randomized response technique; Item sum technique; Sensitive questions; Social desirability bias.

Acknowledgements Work supported by project MTM2015-63609-R.

References

[1] Miller, J.D. (1984). *A New Survey Technique for Studying Deviant Behavior.* Ph.D. Thesis, The George Washington University.

[2] Raghavarao, D. and Federer, W.F. (1979). Block total response as an alternative to the randomized response method in survey. *Journal of the Royal Statistical Society-B*, **41**, pp. 40–45.

[3] Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, **60**, pp. 63–69.

[*] Department of Statistics and Operational Research, University of Granada, Spain. Email: beacr@ugr.es

[†] Department of Statistics and Operational Research, University of Granada, Spain. Email: mrueda@ugr.es

A new approach in edge detection based in global evaluation with segments

Pablo A. Flores-Vidal [*] Daniel Gómez [†]

Pablo A. Flores-Vidal (PhD student) is a last-year PhD student doing his research at Complutense University of Madrid. His main research interests are related to image processing and more specifically to edge detection. He has a degree in applied statistics and a master degree in education of mathematics for secondary education. After a two-year period working for the National School of Sociologist first, and the Ministery of Education after, he enrolled in the doctoral program in Data Science at the Faculty of Statistical Studies, where he has studied statistics under Daniel Gómez. He has published already two papers in JCR journals (*Soft Computing* and *IJCIS*) and three proceedings. Nowadays, he is getting ready his thesis in order to read it soon.

Humans tend to recognise the important changes in the luminosity following global rules [1]. Due to this, Venkatesh and Rosin [2] developed a methodology that analyses this luminosity changes in a digital image taking the information of connected structures of pixels that we have called *segments* [3, 4, 5]. Thus, this methology allows a better performance in the edge extraction task compared with the more simple approach of the local evaluation where only individual information of the pixels is used to take the decision if the pixel should be an edge.

We have developed an edge detection algorithm based in this novelty global evaluation approach inspired by human vision. In order to select the relevant segments to be retained we have used both, a non-supervised and a supervised approach. For the first case we have use fuzzy clustering techniques. This fuzzy clustering of segments presented a higher performance compared to other standard edge detection algorithms. The second approach was more complex as we had to create an specific human reference made of segments in order to learn which segments were the most relevant ones -the "good" segments- to be retained. After building this new set of referenced images made of segments a few well-know classification algorithms (Random Forest, CART, Gradient Boosting and XGBoost) were employed. The supervised approach showed a higher performance of global evaluation approach compared with local evaluation. Two different sets of human referenced images were employed for the comparisons, the Berkeley data set [6] and the South Florida's.

Keywords: Edge detection; Global Evaluation; Non-supervised classification; Supervised classification; Fuzzy Clustering; Edge segments.

Acknowledgements This research has been partially supported by the Government of Spain, grant TIN2015-66471-P and Complutense University of Madrid research group GR3/14-910149.

References

[1] Goldstein, E.B. (2009). *E.B. Sensación y percepción Sexta ed.*. Thomson Editores, Spain 2009.

[2] Venkatesh, S., Rosin, P. L. (1995). Dynamic threshold determination by local and global edge evaluation. *Graphical Models and image processing.*, **57**, pp. 146–160.

[3] Flores-Vidal, P. A., Olaso, P., Gómez, D., Guada, C. (2019). A new edge detection method based on global evaluation using fuzzy clustering. *Soft Computing.* **23**, Springer Berlin Heidelberg, pp. 1809–1821.

[4] Flores-Vidal, P. A., Montero, J. Gómez, D., Villarino, G. (2018). A new edge detection method based on global evaluation using supervised classification algorithms. *International Journal of Computational Intelligence Systems.* **11**, pp. 367–378.

[5] Flores-Vidal, P. A., Martínez, N., Gómez, D. (2018). Post-processing in edge detection based on segments. *Proceedings of the 13th International FLINS Conference (FLINS 2018).* pp. 1425–1432.

[*] Department of Statistics an Operational Research, Faculty of Mathematics of Complutense University of Madrid, Spain. Email: pflores@ucm.es

[†] Statistics and Data Science, Faculty of Statistical Studies of Complutense university of Madrid, Spain. Email: dagomez@estad.ucm.es

Enhancing the Lasso by adding performance constraints

Rafael Blanquero Bravo [*] Emilio Carrizosa Priego [†] Pepa Ramírez Cobo [‡]
M. Remedios Sillero Denamiel [§]

M. Remedios Sillero Denamiel (PhD student) M. Remedios Sillero Denamiel is currently a PhD student at the University of Seville (Spain), where she got the BSc and MSc of Mathematics. Before starting her PhD, she was hired by two different research projects, during which she was co-author of two papers published in distinguished JCR journals. She has recently started the third year of her PhD, under the supervision of Prof. Rafael Blanquero (University of Seville), Prof. Emilio Carrizosa (University of Seville) and Prof. Pepa Ramírez-Cobo (University of Cádiz). Her PhD project is based on applications of mathematical programming tools to deal with statistical problems related to multivariate analysis. This line of research arises from the deep insight on multivariate analysis she acquired during one of the research projects she worked for, when she performed classification/regression on large medical datasets.

The Lasso [1] has become a benchmark data analysis procedure, and it has been studied in depth and extended by many authors. Although the Lasso formulations are stated so that overall prediction error is optimized, no full control over the accuracy prediction on certain individuals of interest is allowed.

In this work we propose a novel version of the Lasso (constrained Lasso) in which performance constraints are added to Lasso-based objective functions, in such a way that threshold values are set to bound the prediction errors in the different groups of interest. As a result, a constrained sparse regression model is obtained, addressed by solving a nonlinear optimization problem. This methodology has a direct application in heterogeneous samples where data are collected from distinct sources, as it is standard in many biomedical contexts. Theoretical properties such as the existence of a unique solution and the asymptotic behaviour in λ have been explored. In addition, consistency properties of the solution are derived in this work using the *Sample Average Approximation* theory (see [2]). Finally, empirical studies concerning the new method as well as an illustration on a real heterogeneous dataset related to gene expression data have been presented.

Keywords: Performance Constraints; Sparse Solutions; Sample Average Approximation; Heterogeneity.

Acknowledgements Research partially supported by research grants and projects MTM2015-65915-R (Ministerio de Economía y Competitividad, Spain), P11-FQM-7603, FQM-329 (Junta de Andalucía, Spain), Fundación BBVA and EC H2020 MSCA RISE NeEDS Project (Grant agreement ID: 822214).

References

[1] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), pp. 267–288.

[2] Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2009). *Lectures on stochastic programming: modeling and theory.* SIAM.

[*] Department of Statistics and O.R. - IMUS, University of Seville, Seville, Spain. Email: rblanquero@us.es

[†] Department of Statistics and O.R. - IMUS, University of Seville, Seville, Spain. Email: ecarrizosa@us.es

[‡] Department of Statistics and O.R., University of Cádiz, Cádiz - IMUS, University of Seville, Seville. Spain. Email: pepa.ramirez@uca.es

[§] Department of Statistics and O.R. - IMUS, University of Seville, Seville, Spain. Email: rsillero@us.es

Session 3 (Wednesday 16:20)

Session talks

A methodology for monitoring traffic flow and air pollution in urban areas

José Ángel Martín-Baos [*] Ricardo García-Ródenas [†] Luis Rodríguez-Benítez [‡]

José Ángel Martín-Baos (PhD Student) José Ángel Martín Baos has studied a BSc. in Computer Science in the University of Castilla-La Mancha, where he was awarded as the most outstanding graduate. Currently, he is studying a MSc in Computer Science and starting the doctoral thesis. He is also working as research assistant in the MAT research group on the University of Castilla-La Mancha. His main interests are machine learning, artificial intelligence and optimization. This work has been awarded as "Best BSc. thesis 2017-2018" by the Chair of Innovation and Cooperative and Business Development "Fundación Eurocaja Rural" and also by "Aula SMACT Avanttic".

Road transportation has become the main source of air pollution in urban areas, which has a major impact on local air quality and human health. For this reason, it is increasingly necessary to accurately estimate the contribution of road transport to air pollution in the cities, so that pollution-reduction measures can be properly designed and implemented appropriately [1]. These pollution reduction measures are becoming increasingly important due to the continued growth in vehicle use and the deterioration of driving conditions (traffic congestion). Authorities find it difficult to meet their environmental objectives and, therefore, reliable mathematical emission models are needed to accurately predict the impact of road transport on air pollution.

Nowadays, intelligent cities are essential to prevent high-level pollution situations and to act when such situations occur. Cities must anticipate pollution peaks and take mitigating measures, such as restricting traffic to a certain number of vehicles or according to their license plates, closing traffic on certain streets, reducing speed limits, etc. In addition, traffic flows must be monitored as they affect pollution levels in that city.

Typical pollution monitoring and control systems often consist of large and expensive devices that are only limited to a few points in the city, hence, they provide information for vast areas and sometimes these systems are not scalable. However, cities are distributed environments where events occur in real time and on a massive scale. Therefore, cities should rely on a low-cost Internet of Things (IoT) infrastructure connected to a cloud platform that supports this type of systems, as well as sensor-based big data applications [2]. These pollution control systems can be combined with a traffic monitoring infrastructure to provide a complete system that can be used as a Decision Support System (DSS) to help authorities make decisions about the environmental impacts caused by pollution before they occur.

In this work, a methodology has been developed for determining the traffic flow and the air pollution in several streets of a city using low-cost devices. An artificial intelligence heuristic algorithm has been employed to process video images from a camera and estimate the traffic flow in the street. This algorithm uses statistical techniques to process the motion vectors generated by the GPU when the video is encoded. Moreover, the computational time required by the algorithm is in the order of 2.5 milliseconds, which allows to count the number of vehicles in real time. This algorithm has been tested proving 90% of accuracy. Finally, this information is uploaded to a cloud service where machine learning techniques can be applied to predict the pollution levels in the city or recommend palliative actions.

Keywords: Artificial Intelligence; Transportation; Internet of Things; Traffic pollution monitoring.

Acknowledgements This work was supported by the Project TRA2016-76914-C3-2-P of the Spanish Ministry of Science and Innovation, co-funded by the European Regional Development Fund.

References

[1] Smit, R., Ntziachristos, L., and Boulter, P. (2010). Validation of road vehicle and traffic emission models - A review and meta-analysis. *Atmospheric Environment*, **44**, pp. 2943–2953.

[2] Elias Bibri, S. (2018). The IoT for smart sustainable cities of the future: An analytical framework for sensor-based big data applications for environmental sustainability. *Sustainable Cities and Society*, **38**, pp. 230–253.

[*] Mathematics Department. University of Castilla-La Mancha. Ciudad Real, Spain. Email: JoseAngel.Martin@uclm.es
[†] Mathematics Department. University of Castilla-La Mancha. Ciudad Real, Spain. Email: Ricardo.Garcia@uclm.es
[‡] Information Systems and Technologies Department. University of Castilla-La Mancha. Ciudad Real, Spain. Email: Luis.Rodriguez@uclm.es

Evaluating the efficiency of the Portuguese National Health System: A network data envelopment analysis approach

Miguel Alves Pereira [*†] Diogo Cunha Ferreira [‡] José Rui Figueira [§] Rui Cunha Marques [¶]

Miguel Alves Pereira (PhD student) Miguel Alves Pereira is a PhD student at the Instituto Superior Técnico of the University of Lisbon (IST-UL). He is also a research fellow at the same institution. His main interests concern performance management and its applications to the healthcare sector. He earned his integrated master's degree in biomedical engineering from the IST-UL in 2018. He is a member of the Portuguese Association of Operational Research, the European Working Group on Multiple Criteria Decision Aiding, the Association of European Operational Research Societies, and the International Society on Multiple Criteria Decision Making.

In contrast to conventional data envelopment analysis (DEA), where the relative efficiency of a system is measured by assuming it as a "black-box", network DEA takes into account its internal structure in order to generate more significant and enlightening results [3]. Among the various types of models, it goes without saying that putting network DEA in practice is natural- and progressively rarer as the complexity of a system's structure increases. In particular, its employment in healthcare is not an exception [2]. Thus, we applied a slacks-based model to measure the efficiency of the secondary healthcare providers bearing in mind their internal services. This unprecedented application to static mixed systems with a matrix-type structure, given the heterogeneity of the relationships between hospital services [1], allowed the evaluation of the effects that policy reforms had in the Portuguese National Health System.

Keywords: Network data envelopment analysis; Matrix-type structure; Hospital services; Efficiency evaluation.

Acknowledgements This research was generously supported by the project 'hSNS: Portuguese public hospital performance assessment using a multi-criteria decision analysis framework' (PTDC/EGE-OGE/30546/2017), funded by the Portuguese Foundation for Science and Technology (FCT, from the Portuguese abbreviation of *Fundação para a Ciência e a Tecnologia*). José Rui Figueira also acknowledges the support from the FCT grant SFRH/BSAB/139892/2018 under the POCH (the Portuguese abbreviation for the *Programa Operacional Capital Humano*).

References

[1] Ozcan, Y. A. (2008). *Health Care Benchmarking and Performance Evaluation: An Assessment using Data Envelopment Analysis (DEA)* volume 120 of *International Series in Operations Research & Management Science*. Springer, Boston.

[2] Kao, C. (2017). *Network Data Envelopment Analysis: Foundations and Extensions* volume 240 of *International Series in Operations Research & Management Science*. Springer, Boston.

[3] Kao, C. (2014). Network data envelopment analysis: A review. *European Journal of Operational Research*, **239**, pp. 1–16.

[*] CEG-IST, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais 1, 1049-001, Portugal

[†] CERIS-IST, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais 1, 1049-001, Portugal. Email: miguelalves-pereira@tecnico.ulisboa.pt

[‡] CERIS-IST, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais 1, 1049-001, Portugal. Email: diogo.cunha.ferreira@tecnico.ulisboa.pt

[§] CEG-IST, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais 1, 1049-001, Portugal. Email: figueira@tecnico.ulisboa.pt

[¶] CERIS-IST, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais 1, 1049-001, Portugal. Email: rui.marques@tecnico.ulisboa.pt

Session 4 (Wednesday 17:30)

Session talks

An omnibus goodness-of-fit test for the functional linear model with functional response

Gonzalo Álvarez-Pérez [*][†] Javier Álvarez-Liébana [‡] Eduardo García-Portugués [§]
Manuel Febrero-Bande [¶] Wenceslao González-Manteiga [‖]

G. Álvarez-Pérez (PhD student) Gonzalo Álvarez-Pérez is a PhD student at University of Oviedo. His main interests are Functional Data Analysis and Quantum Nanophotonics. He received his dual BSc in Physics and Mathematics from the University of Oviedo in 2017. The following year he earned his MScs degrees in Big Data Analytics and Nanoscience, from the Carlos III University of Madrid and the University of the Basque Country, respectively. Last September, he enrolled in the PhD program in Condensed Matter Physics, Nanoscience and Biophysics back at University of Oviedo, and has recently joined the Nanomaterials and Nanotechnology Research Center.

Functional data analysis is nowadays established as a powerful tool to exploit the complexity and richness of data measured over continuous domains. When two functional random variables are available, it may be useful to determine their relation by means of a regression model $\mathcal{Y} = m(\mathcal{X}) + \mathcal{E}$, where m is often assumed to be a Hilbert–Schmidt operator. The functional linear model with functional response (FLMFR) emerges when $m = m_\beta$ is a *linear* Hilbert–Schmidt operator between two L^2 spaces:

$$\mathcal{Y}(t) = m_\beta(\mathcal{X}(s)) + \mathcal{E}(t) = \int \mathcal{X}(s)\beta(s,t)ds + \mathcal{E}(t). \tag{1}$$

We propose a novel Goodness-of-Fit (GoF) test for the null (composite) hypothesis of the FLMFR, $H_0 : m = m_\beta$, against a general, unspecified alternative, leading to a novel omnibus GoF test (previous tests were only available for testing the no-effects hypothesis $m = 0$). The test statistic is a generalization of [1] and, by adapting the geometrical arguments found therein, is formulated in terms of a quadratic norm over the doubly-projected empirical empirical process:

$$R_n\left(u, \gamma_{\mathcal{X}}, \gamma_{\mathcal{Y}}\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \langle \widehat{\mathcal{E}}_i, \gamma_{\mathcal{Y}} \rangle \mathbb{I}_{\{\langle \mathcal{X}_i, \gamma_{\mathcal{X}} \rangle \le u\}},$$

where $\gamma_{\mathcal{X}}$ and $\gamma_{\mathcal{Y}}$ are unit-norm elements in both L^2 spaces that act as projectors, and $\{\widehat{\mathcal{E}}_i\}_{i=1}^n$ are the residuals of (1) obtained from the sample $\{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^n$. The resulting test statistic is easy to compute, interpret and calibrate on its distribution via a wild bootstrap on the functional residuals. A flexible hybrid approach involving LASSO regularization and linearly-constrained least-squares is used to perform the selection of the number of dimensions when estimating the residuals of (1).

The accuracy, computational advantages and finite sample behaviour of the test, regarding power and size, are illustrated via a complete simulation study, under different linear models and alternative hypotheses; a comparative study, which focuses on the significance test; and an application to several real datasets.

Keywords: Functional Data Analysis; Functional Linear Model; Functional Response; LASSO; Goodness-of-Fit; Random Projections; Wild Bootstrap.

Acknowledgements We would like to thank Prof. César Sánchez-Sellero, for his kindness in sharing with us some of the code required for the comparative study. The second author and the third, fourth, and fifth authors acknowledge financial support from grants MTM2015-71839-P and MTM2016-76969-P, respectively, from the Spanish Ministry of Economy and Competitiveness.

References

[1] García-Portugués, E.; González-Manteiga, W. and Febrero-Bande, M. (2014). A Goodness-of-Fit Test for the Functional Linear Model with Scalar Response. *J. Comput. Graph. Stat.*, 23, pp. 761–778.

[*] Department of Physics, University of Oviedo, Spain. Email: gonzaloalvarez@uniovi.es
[†] Nanomaterials and Nanotechnology Research Center (CSIC), Spain. Email: gonzalo.alvarez@cinn.es
[‡] Department of Statistics and Operations Research, University of Oviedo, Spain. Email: alvarezljavier@uniovi.es
[§] Department of Statistics, Carlos III University of Madrid, Spain. Email: edgarcia@est-econ.uc3m.es
[¶] Department of Statistics, M. Analysis and Optimization, University of Santiago de Compostela, Spain. Email: manuel.febrero@usc.es
[‖] Department of Statistics, M. Analysis and Optimization, University of Santiago de Compostela, Spain. Email: wenceslao.gonzalez@usc.es

A study on the financial risk assessment in the design and scheduling of industrial plants

Miguel Vieira [*] Helena Paulo[*] [†] Tânia Pinto-Varela [*] Ana Paula Barbosa-Póvoa [*]

M. Vieira (Postdoctoral Researcher) Miguel Vieira obtained his PhD in Leaders for Technical Industries in 2017 awarded by the MIT Portugal, an international program hosted by the IST- Universidade de Lisboa, EEUM - Universidade do Minho and FEUP - Universidade do Porto in partnership with the Massachusetts Institute of Technology. Currently he is a postdoctoral research fellow at the Centre for Management Studies at IST, where he has been developing several projects regarding optimization solution methods to address industrial decision support on supply chain, design, production planning and scheduling problems involving manufacturing systems.

The increasing dynamic structure of production systems requires that most strategic and operational decisions involve the assessment of optimal solutions to face the challenges of global markets. These industrial companies are seeking for highly flexible operational solutions to deliver mass customized products at optimized process metrics in their complex design and scheduling management. With Industry 4.0 digital transformation across multiple enterprise levels, technology has defined a greater relevance on model-based decision-support systems, in order to potentiate production competitiveness under operational uncertainty and market variability [1]. In the last decades, the research community has studied several exact/non-exact methods and techniques regarding design and scheduling decision-making to address real industrial problems, with particular relevance to the uncertainty impact of resource's availability, equipment efficiency and capacity reliability. The design flexibility of multipurpose plants in uncertain environments has been addressed, mostly with the use of stochastic models to compare expected performance in a deterministic approach [2]. However, most of these works do not provide the analysis of the alternative possible outcomes by assessing different risk management profiles, since stochastic models optimize the total expected performance assuming that the decision maker is risk-neutral. By allowing to evaluate alternative solution policies for different scenarios, it becomes a highly valuable tool to identify the best strategy according to the decision makers' attitude toward risk.

In this work, a study is proposed to assess decision support solutions for the design and scheduling of multipurpose industrial plants under demand uncertainty by considering a financial risk measure, the Conditional Value at Risk (CVaR). The CVaR allows to evaluate the likelihood that a specific loss or gain will exceed a certain value at risk, given for a specified confidence level. Extending the two-stage mixed integer linear programming (MILP) model proposed by Pinto-Varela et al. (2009) [2], the goal is to maximize the annualized profit of the plant operation under a set of scenarios while minimizing the associated financial risk, evaluated by the CVaR. A bi-objective model is formulated using the augmented ε-constraint method to generate an approximation to the Pareto-optimal curve, illustrating the trade-offs between plant profit (with the corresponding design and scheduling decisions) and the associated risk. The conclusions highlight the advantages of the proposed approach in the support of the decision-making process by considering the explicit risk measure while assessing different solutions for expected financial outcomes.

Keywords: Industrial plants, Uncertainty, Conditional value at risk, Stochastic programming.

Acknowledgements The authors would like to acknowledge the support from the Centre for Management Studies of IST (CEG-IST) and from Associação Portuguesa de Investigação Operacional (APDIO).

References

[1] Vieira, M., Pinto-Varela, T., Barbosa-Póvoa, AP. (2019). A model-based decision support framework for the optimisation of production planning in the biopharmaceutical industry. *Comput. Ind. Eng.*, **129**, pp. 354–367.

[2] Pinto-Varela, T., Barbosa-Póvoa APFD., Novais, AQ, (2009). Design and scheduling of periodic multipurpose batch plants under uncertainty. *Ind. Eng. Chem. Res.*, **48**, pp. 9655–9670.

[*] CEG-IST, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal. Email: migueljvieira@tecnico.ulisboa.pt, tania.pinto.varela@tecnico.ulisboa.pt, apovoa@tecnico.ulisboa.pt

[†] ISEL, IPL, Rua Conselheiro Emídio Navarro, 1959-007 Lisboa, Portugal. Email: hpaulo@deq.isel.ipl.pt

Spatio-temporal smoothing methods in disease mapping for ‘very’ large datasets

Aritz Adin [*] Tomás Goicoa [†] María Dolores Ugarte [‡]

A. Adin (Assistant Professor) Aritz Adin is an Assistant Professor at the Public University of Navarre. His main research work has been focused on the field of spatio-temporal disease mapping, developing new methods for the estimation of mortality/incidence risks. He received his B.S. in Mathematics and his master’s degree in Modelling and Mathematical Research, Statistics and Computing from the University of the Basque Country in 2010 and 2014, respectively. He got a Ph.D. in Mathematics and Statistics from the Public University of Navarre in 2017.

Several models have been proposed in the disease mapping literature to smooth mortality or incidence risks borrowing information from space and time. Possibly, the non-parametric models based on conditional autoregressive (CAR) priors for space, random walk priors for time, and different types of space-time interactions described by Knorr-Held [1] are the most used models in space-time disease mapping. Other approaches based on multidimensional P-splines have been also proposed in this field (see for example Ugarte et al. [2]). However, one key question to answer is: are these smoothing methods feasible when the number of small areas is large?

In this work, we compare both CAR and spline-based models to estimate cancer mortality risks in around 8000 municipalities in Spain during the period 1990-2015. Model fitting and inference will be carried out using the integrated nested Laplace approximation (INLA) [3] approach, and a scalable method for generalized additive models implemented in the `bam` function of the R package `mgcv` [4].

Keywords: Disease mapping; Massive data; Smoothing; Spatio-temporal models.

Acknowledgements Work supported by project MTM2017-82553-R (AEI/FEDER, UE), and project UPNA2018-11-INC (Public University of Navarre).

References

[1] Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, **19**, pp. 2555–2567.

[2] Ugarte, M.D., Adin, A., and Goicoa, T. (2017). One-dimensional, two-dimensional, and three-dimensional B-splines to specify space-time interactions in Bayesian disease mapping: model fitting and model identifiabbility. *Spatial Statistics*, **22**, pp. 451–468.

[3] Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society Series B*, **71**, pp. 319–392.

[4] Wood, S.N., Li, Z., Shaddick, G., and Augustin N.H. (2017). Generalized additive models for gigadata: modelling the UK black smoke network daily data. *Journal of the American Statistical Association.* **112**, pp. 1199–1210.

[*] Department of Statistics, Computer Science and Mathematics, Public University of Navarre, Spain. Email: aritz.adin@unavarra.es

[†] Department of Statistics, Computer Science and Mathematics, Public University of Navarre, Spain. Email: tomas.goicoa@unavarra.es

[‡] Department of Statistics, Computer Science and Mathematics, Public University of Navarre, Spain. Email: lola@unavarra.es

GRASP algorithm for the unrelated parallel machine scheduling problem with setup times and additional resources

Juan Camilo Yepes-Borrero [*] Maria Fulgencia Villa [†] Federico Perea [‡]

Juan Camilo Yepes-Borrero (PhD student) Graduate of industrial engineering and PhD student at the Universitat Politècnica de València. His main interests are Combinatorial optimization problems, scheduling, heuristics and metaheuristics. He recibed his master's degree in statistics from the Universitat Politècnica de València in 2017. After graduating from industrial engineering in 2010, he worked in different companies of the public and private sector for 3 years.

In this talk, we study the unrelated parallel machine scheduling problem with setup times and additional resources in the setups (UPMSR-S), with makespan minimization criterion. This is a more realistic extension of the traditional problem in which the setups are assumed to be done automatically. We here propose a new problem, in which the setups must be done by some limited extra resources (e.g. workers). We propose three metaheuristics to solve the problem following two approaches: one approach that ignores the information about the additional resources in the constructive phase, and another approach that takes into account this information about the resources. Computation experiments are carried out over a benchmark of small and large instances. After the computational analysis we conclude that the second approach shows an excellent performance, outperforming the first approach.

Keywords: Parallel machine; Scheduling; Sequence dependent setup times; Makespan; Additional resources.

Acknowledgements The authors would like to acknowledge the support from Spanish "Ministerio de Economia y competitividad" throughout grant number MTM2016-74983 and grant "SCHEYARD – Optimization of Scheduling Problems in Container Yards" (No. DPI2015-65895-R) financed by FEDER funds.

References

[1] Avalos-Rosales, O., Angel-Bello, F., and Alvarez, A. (2015). Efficient metaheuristic algorithm and reformulations for the unrelated parallel machine scheduling problem with sequence and machine-dependent setup times. *International Journal of Advanced Manufacturing Technology*, 76, pp. 1705-1718.

[2] Diana, R. O. M., F., d. M., de Souza, S. R., and de Almeida Vitor, J. F. (2014). An immune-inspired algorithm for an unrelated parallel machines' scheduling problem with sequence and machine dependent setup-times for makespan minimisation. *Neurocomputing*, 163, pp. 94-105.

[*] Grupo de Sistemas de Optimización Aplicada, Universitat Politècnica de València, Valencia, Spain. Email: juayebor@posgrado.upv.es
[†] Grupo de Sistemas de Optimización Aplicada, Universitat Politècnica de València, Valencia, Spain. Email: mfuvilju@eio.upv.es
[‡] Grupo de Sistemas de Optimización Aplicada, Universitat Politècnica de València, Valencia, Spain. Email: perea@eio.upv.es

Session 5 (Wednesday 19:00)

Session talks

An enhanced formulation for coloring graphs with the Douglas–Rachford algorithm

Francisco J. Aragón Artacho* Rubén Campoy† Veit Elser‡

R. Campoy (Postdoctoral Researcher) Rubén Campoy received his bacherlor's degree in Mathematics from the University of Alicante in 2013. He earned his master's degree in Mathematical Engineering in 2015 from the Carlos III University of Madrid, where he was a fellow at the Department of Statistics. After having worked for 7 months as a Pricing Analyst for a business company, he got a grant by the Spanish government for PhD students. In December 2018, Rubén obtained his PhD in Mathematics from the University of Murcia. His thesis was entitled "Contributions to the Theory and Applications of Projection Algorithms". He is currently a Postdoctoral Researcher at the Department of Mathematics of the University of Alicante. His research interest lies in Convex Analysis and Splitting Algorithms, particularly Projection Methods.

Projection algorithms are powerful tools for finding common points in a collection of sets. The probably two most well-known algorithms within this family are the method of alternating projections and the Douglas–Rachford method. The convergence of these algorithms to a point in the intersection is guaranteed, provided that the involved sets are convex. In the last years, the Douglas–Rachford algorithm has received much attention, due to the good performance of the method in nonconvex scenarios. Despite a lack of convergence results, the algorithm has been successfully employed in a wide list of combinatorial problems. We shall concern on the graph vertex-coloring problem. Given a graph and a number of colors, the goal is to find a coloring of the vertices so that all adjacent vertex pairs have different colors. In the recent work [1], the DR algorithm was shown to be a successful heuristic for solving a wide variety of graph coloring instances, when the problem was cast as a feasibility problem on binary indicator variables. In this talk, we consider a different formulation, originally proposed in [2], which is based on semidefinite programming. The much improved performance of the DR algorithm, with this new approach, is demonstrated through various numerical experiments.

Keywords: Projection Methods; Douglas–Rachford Algorithm; Graph Coloring; Feasibility Problem; Nonconvex Constraints.

Acknowledgements This work was supported by MINECO of Spain and ERDF of EU (project MTM2014-59179-C2-1-P). The author was supported by MINECO of Spain and ESF of EU (PhD grant BES-2015-073360) under the program "Ayudas para contratos predoctorales para la formación de doctores 2015".

References

[1] Aragón Artacho, F.J., Campoy, R. (2018). Solving graph coloring problems with the Douglas–Rachford algorithm. *Set-Valued and Variational Analysis*, **26**(2), pp. 277–304.

[2] Karger, D., Motwani, R., Sudan, M. (1998). Approximate graph coloring by semidefinite programming. *Journal of the ACM (JACM)*, **45**(2), pp. 246–265.

*Department of Mathematics, University of Alicante, Spain. E-mail: francisco.aragon@ua.es

†Department of Mathematics, University of Alicante, Spain. E-mail: ruben.campoy@ua.es

‡Department of Physics, Cornell University, New York, United States. E-mail: ve10@cornell.edu

Defective branching processes in a varying environment

Carmen Minuesa [*] Götz Kersting [†]

C. Minuesa (Postdoctoral Researcher) Carmen Minuesa Abril (Berlanga, 1989) received her BSc in Mathematics (2012), BSc in Statistics (2012), MSc in Science Research (2013) and PhD in Mathematics (2018) at University of Extremadura. She is currently working as a Researcher at the Department of Mathematics of the University of Extremadura. Her research mainly considers aspects related to stochastic processes, especially to branching processes. In particular, her contributions focused on the statistical inference of the main parameters of such models and the study of new branching models in varying environments.

The classical branching process, known as the Bienaymé-Galton-Watson process or simply Galton-Watson process, is a discrete-time Markov chain which models populations where the reproduction of each individual is independent of the others and the probability distribution governing the reproduction, $p = \{p_k\}_{k\in\mathbb{N}_0}$ (where p_k is the probability that one individual gives birth to k offspring), is common to all the individuals.

A defective Galton-Watson process (DGWP) having defective reproduction laws (that is, for some $\varepsilon \in (0,1)$, $\sum_{k=0}^{\infty} p_k = 1-\varepsilon$) was studied in [2]. In this work ([1]), we present a more general model resulting from letting the defective reproduction law change over the time. This process is called *defective Galton-Watson process in a varying environment (DGWPVE)* $v = \{f_1, f_2, \ldots\}$, where f_n denotes the probability generating function of the possibly defective reproduction law in the n-th generation of the process. The defect of the distribution defined by f_n, $1 - f_n(1)$, is interpreted as the probability that at a given generation n, an individual sends the process to an absorbing graveyard state Δ at generation $n+1$, where it remains forever. As a consequence, the state space of the process is $\mathbb{N}_\Delta = \mathbb{N}_0 \cup \{\Delta\}$, where two of these states are absorbing: 0 and Δ. We focus our attention on the asymptotic behaviour of DGWPVEs. Two main results will be presented: the almost sure convergence of the process to a random variable with values in $\mathbb{N}_\Delta \cup \{\infty\}$ and two characterisations of the duality extinction - absorption at Δ which holds for the DGWPs.

Keywords: Branching Process; Limit Theorems; Defective Distribution; Galton–Watson Process With Killing; Varying Environment

Acknowledgements This research was developed when Carmen Minuesa was visiting the Institute of Mathematics, Goethe University Frankfurt, in Frankfurt am Main (Germany).

This research has been supported by the Ministerio de Educación, Cultura y Deporte (grants FPU13/03213 and EST16/00404), the Ministerio de Economía y Competitividad (grant MTM2015-70522-P), the Junta de Extremadura (grants IB16099 and GR18103) and the Fondo Europeo de Desarrollo Regional.

References

[1] Kersting, G. and Minuesa, C. (2019). Defective Galton-Watson processes in varying environment. *Work in progress.*

[2] Sagitov, S. and Minuesa, C. (2017). Defective Galton-Watson processes. *Stoch. Models*, **33**, pp. 451–472.

[*] Department of Mathematics, University of Extremadura, Badajoz, Spain. Email: cminuesaa@unex.es

[†] Institute of Mathematics, Goethe University Frankfurt, Frankfurt, Germany. Email: kersting@math.uni-frankfurt.de

A link between Game theory and Statistics: sampling techniques to estimate coalitional values

Alejandro Saavedra-Nieves*

A. Saavedra-Nieves (Postoctoral researcher) Alejandro Saavedra-Nieves received his bachelor's degree in Mathematics and his M.S. in Statistics and Operations Research from the University of Santiago de Compostela (Spain). He is PhD in Statistics and Operations Research by the University of Vigo (Spain) from March 2019. His main interests are Cooperative Game Theory, Operations Research and the computational problems arisen in these fields.

Cooperative game theory deals with the analysis of those conflictive situations where a group of players decides to distribute the profits/costs resulting from their cooperation. In particular, it focuses on defining mathematical tools for proposing allocation vectors that are "acceptable" by the players. A *coalitional value* is a map that associates an allocation vector to every TU-game. The *Shapley value* and the *Banzhaf value* are probably the most important coalitional values in the literature.

The TU-games with a priori unions incorporate information about the affinities of players. A system of a priori unions is a partition of the player set describing a structure of a priori coalitions. The Shapley value was extended to games with a priori unions (see [5]); this extension is known as the *Owen value*. Besides, [6] proposes the *Banzhaf-Owen value* to extend the Banzhaf value to games with a priori unions.

The calculation of these coalitional values becomes a difficult task when the number of players is large. Sampling techniques (see [3]) are an alternative tool for their approximation. In fact, most of coalitional values are averages and then sampling theory ensures good results when estimating them. References in [4] and [2] describe a sampling procedure to estimate the Shapley value, based on simple random sampling with replacement, that is useful in those problems with large player sets. An analogous procedure to approximate the Banzhaf value is introduced in [1].

In this talk, we describe other sampling methodologies for the estimation of coalitional values for TU-games with a priori unions. We analyse our proposals from a statistical perspective and evaluate them in several well-known examples in the literature, with positive results.

Keywords: Games with a priori unions; Coalitional values; Sampling techniques.

Acknowledgements Work supported by MTM2017-87197-C3-2-P project, from Ministry of Economy, Industry and Competitiveness (MINECO), State Research Agency (AEI), and Regional Development Fund (FEDER) and UE and by ERDF (*Grupos de Referencia Competitiva*) ED431C 2016-040 project, from Xunta de Galicia.

References

[1] Bachrach, Y., Markakis, E., Resnick, E., Procaccia, A. D., Rosenschein, J. S., & Saberi, A. (2010). Approximating power indices: theoretical and empirical analysis. *Autonomous Agents and Multi-Agent Systems*, **20**, pp. 105–122.

[2] Castro, J., Gómez, D., & Tejada, J. (2009). Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research*, **36**, pp. 1726–1730.

[3] Cochran, W. G. (2007). *Sampling techniques.* John Wiley & Sons 2007.

[4] Fernández-García, F., & Puerto-Albandoz, J. (2006). *Teoría de juegos multiobjetivo.* Imagraf Impresores SA, Sevilla.

[5] Owen, G. (1977). Values of games with a priori unions. In Henn, R. and Moeschlin, O. (Eds.), *Mathematical Economics and Game Theory* (pp. 76–88). Berlin: Springer.

[6] Owen, G. (1982). Modification of the Banzhaf-Coleman index for games with a priori unions. In M.J. Holler (Ed.), *Power, Voting and Voting Power* (pp. 232–238). Heidelberg: Physica-Verlag.

*SiDOR Group. Department of Statistics and Operations Research, Universidade de Vigo. Email: asaavedra@uvigo.es

The Flight Trajectory Optimization Problem

Ralf Borndörfer [*] Pedro Maristany [†]

Pedro Maristany is a PhD student at the Zuse Institute Berlin, an interdisciplinary research institute for applied mathematics. He recieved his bachelor's and master's degree in mathematics from the Technische Universität Berlin and has been working for the last two years in the project *Flight Trajectory Optimization on Airway Networks*. This project is carried out in cooperation with Lufthansa Systems, an IT company specialized in the development of flight planning related software. Pedro's main research topics are: time dependent and multicriteria shortest paths, sensitivity analysis in combinatorial problems, and approximation algorithms.

The *Flight Trajectory Optimization Problem* (FTOP) is a Time Dependent Shortest Path Problem (TDSPP) with extra constraints that arise from weather conditions, overflight costs, and safety measures.

As *weather* has to be considered in flight planning, during optimization different prognoses are given and predefined interpolation rules have to be applied to get the actual weather conditions at each visited arc. On every arc in the graph best and worst possible weather conditions can be computed in a preprocessing step, thus getting a cost interval for every arc. However, no regularity assumption on the arc costs lying in this interval can be made. This fact naturally leads to the problem of finding shortest paths on directed graphs with interval costs which is closely related to the evaluation of the stability of a shortest path w.r.t. to changes in the input parameters. Some first insights on the impact of weather can be read in [1].

Overflight Costs constitute a further cost component in the objective function of the classical TDSPP. The input graph in the FTOP is divided into regions. A region is a set of arcs and crossing it will incur extra costs for a path. This cost component is interesting from a theoretical point of view because the overflight costs for a path crossing a region are not defined on the arcs of the region but on the euclidean distance between the region's entry and exit nodes. If a region's overflight function is linear, the Shortest Path Problem with Overflight Costs remains polynomial but if the overflight function is constant or piecewise constant, the problem is $\mathcal{NP}$-hard. The case where overflight functions are linear for every region was introduced and studied in [2].

Safety measures have to be applied to every flight route in order to guarantee the successful landing of an aircraft even in unpredictable situations. Working slots for crew members, mandatory fuel remainders, or maximum distance to the closest airport are some examples. Considering all of them simultaneously to get one single exact solution leads to prohibitive running times or difficult-to-trace infeasiblity results. Our approach is to solve a multicriteria shortest path problem with one objective set to the total costs of a path and other objectives mirroring isolated time or fuel costs. The output pareto frontiers can be used by decision makers to trace back a feasible route.

All in all the FTOP is a broad optimization problem that arises from a very intuitive practical problem and demands a variety of solution approaches to extend the classical TDSPP. In this talk we will mainly focus on the handling of overflight costs and discuss the regularity of the pareto frontiers we observe when solving instances of the FTOP.

Keywords: Time Dependent Shortest Path, Multiobjective Shortest Path, Flight Trajectory Optimization

Acknowledgements The authors would also like to thank Lufthansa Systems GmbH & Co. KG for their support.

References

[1] Blanco, M. et al. (2016) Solving Time Dependent Shortest Path Problems on Airway Networks Using Super-Optimal Wind. *16th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems (ATMOS 2016)* Vol. 54.

[2] Blanco, M. et al. Cost Projection Methods for the Shortest Path Problem with Crossing Costs. *17th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems (ATMOS 2017)*, Vol. 59.

[*] Zuse Institute Berlin Email: borndoerfer@zib.de
[†] Zuse Institute Berlin Email: maristany@zib.de

Session 6 (Thursday 09:40)

Session talks

Optimal exercise for American options under stock pinning

Bernardo D'Auria [*] Eduardo García-Portugués [†] Abel Guada-Azze [‡]

Abel Guada Azze is currently enrolled in the PhD program in Mathematical Engineering at the University Carlos III of Madrid, Spain, supported by a grant from the Department of Statistics. He is mainly interested in Optimal Stopping Theory. He received in 2014 his BSc in Mathematics from the University of Havana, Cuba. After two years at the Technological University of Havana "José Antonio Echeverría" as a Professor in Training, he moved to Spain where he obtained his master's degree in Mathematical Engineering at Carlos III University of Madrid, supported by a scholarship from the Department of Statistics.

We address the problem of optimally exercising American options based on the assumption that the underlying stock's price follows a Brownian bridge whose final value coincides with the strike price. In order to do so we solve, for put options, the following discounted optimal stopping problem via the free-boundary problem approach (see [4]):

$$V(t,x) = \sup_{0\leq\tau\leq\ T-t} \mathbb{E}_{t,x}\left[e^{-\lambda\tau}G(X_{t+\tau})\right], \tag{1}$$

endowed with the gain function $G(x) = (S-x)^+$ and a Brownian bridge whose final value equals S, that is, $(X_{t+s})_{s=0}^{T-t}$ satisfies the stochastic differential equation

$$\mathrm{d}X_{t+s} = \frac{S-X_{t+s}}{T-t-s}\,\mathrm{d}s + \sigma\,\mathrm{d}B_s, \quad 0\leq s\leq T-t, \tag{2}$$

where $(B_s)_{s\geq 0}$ is a standard Brownian motion. We then show how to easily obtain the solution for the call option case via the establishment of a put-call parity.

This work comes up as a first approach of optimally exercising an option within the so-called "stock pinning" scenario (see [2]), *i.e.,* the phenomenon describing the tendency of the price of *optionable* stocks (stocks with available options) to end up near the strike price of some of its underlying options at expiration date.

The optimal stopping boundary for problem (1) is proved to be the unique solution, up to certain conditions, of an integral equation, which is then numerically solved by an algorithm hereby exposed [1]. We face the case where the volatility σ in (2) is unspecified by providing an estimated optimal stopping boundary that, alongside with pointwise confidence intervals, provide alternative stopping rules.

Finally we demonstrate the usefulness of our method within the stock pinning scenario through a comparison with the optimal exercise time based on a geometric Brownian motion (see [3]). We base our comparison on the contingent claims and the 5-minutes intraday stock price data of Apple and IBM for the period 2011–2018.

Keywords: American option; Brownian bridge; Free-boundary problem; Optimal stopping; Option pricing; Put-call parity; Stock pinning.

Acknowledgements Work supported by projects MTM2017-85618-P, MTM2015-72907-EXP and MTM2016-76969-P from the Spanish Ministry of Economy, Industry and Competitiveness, and the European Regional Development Fund.

References

[1] D'Auria, B., García-Portugués, E., and Guada-Azze, A. (2019). Optimal exercise of American options under stock pinning. *arXiv:1903.11686.*

[2] Ni, S. X., Pearson, N. D., and Poteshman, A. M. (2005). Stock price clustering on option expiration dates. *J. Financial Econ.*, **78**, pp. 49–87.

[3] Peskir, G. (2005). On the American option problem. *Math. Financ.*, **15**, pp. 169–181.

[4] Peskir, G. and Shiryaev, A. (2006). *Optimal Stopping and Free-Boundary Problems.* Lectures in Mathematics. Birkhäuser, Basel.

[*] Department of Statistics, Carlos III University of Madrid, Spain. Email: bdauria@est-econ.uc3m.es
[†] Department of Statistics, Carlos III University of Madrid, Spain. Email: edgarcia@est-econ.uc3m.es
[‡] Department of Statistics, Carlos III University of Madrid, Spain. Email: aguada@est-econ.uc3m.es

Application of destroy and repair heuristic strategies on a humanitarian last mile distribution problem

José M. Ferrer [*] Gregorio Tirado [†] M. Teresa Ortuño [‡]

J.M. Ferrer (Lecturer) José María Ferrer is a Lecturer at the Universidad Complutense de Madrid (UCM) as well as a member of the UCM investigation group "Decision Aid Models for Logistics and Disaster Management (Humanitarian Logistics)". His main interests are decision models in humanitarian logistics and applications of mathematical programming and metaheuristics. He received his B.S. in 1999 in mathematics and his PhD in 2016, both from UCM.

This work faces a multicriteria humanitarian distribution problem under uncertain and unsecure conditions. It is aimed to be used during the response phase after the ocurrence of a disaster. The available vehicles that transport the aid from supply nodes to populations in need must travel together forming convoys to avoid assaults. The complexity of the problem and the strong connections between the elements of any feasible solution make it difficult to use heuristics based on local search, because each small change in a solution can make it infeasible. Instead, constructive heuristics such as GRASP or Ant Colony Optimization have been successfully applied [1].

In order to improve the quality of the solutions obtained by the existing solution methods, in this work the authors will explore the Large Neighborhood Search metaheuristic [2]. In this approach the current solution is partly destroyed and then repaired to obtain a new solution. Different destroy and repair strategies will be tested and the computational results obtained for some test cases will be compared.

Keywords: Humanitarian Logistics: Last Mile Distribution; Multi-criteria Decision Making; Metaheuristics.

Acknowledgements This work has been supported by the Government of Spain, grant MTM2015-65803-R. The authors would also like to thank the International Campus of Excellence of Moncloa, that allowed them to use the computation cluster EOLO.

References

[1] Ferrer, J. M., Ortuño, M. T. and Tirado, G. (2016). A GRASP metaheuristic for humanitarian aid distribution. *Journal of Heuristics*, **22**, pp. 55–87.

[2] Shaw, P. (1998). Using Constraint Programming and Local Search Methods to Solve Vehicle Routing Problems. In *CP-98 (Fourth International Conference on Principles and Practice of Constraint Programming)*, *Lecture Notes in Computer Science*, **1520**, pp. 417–431.

[*] Department of Financial and Actuarial Economics & Statistics, UCM. Email: jmferrer@ucm.es

[†] Department of Financial and Actuarial Economics & Statistics, UCM. Email: gregoriotd@ucm.es

[‡] Department of Statistics and Operational Research, UCM. Email: mteresa@ucm.es

Stochastic models applied to Music Composition

Jose Manuel Valero Esteban [*] Arminda Moreno Díaz [†] Yolanda Pastor Ruiz[‡]

José Manuel Valero Esteban (PhD student) José Manuel Valero is a part-time PhD student at Universidad Rey Juan Carlos in the field of Social Psychology. He holds a B.S. and a M.S. in Computer Science by Universidad Politécnica de Madrid (Escuela Superior de Ingenieros Informáticos), as well as a B.A. in Music with a Major in guitar interpretation by the Real Conservatorio Superior de Música de Madrid. His major interests include music, its psychology and automated composition.

Music and Probability are two disciplines that are often considered together. Historically, there is a long tradition of composers using chance and automation in their compositions [3, 6]. In some cases, the aim has been to synthesize new music in an existing style [4, 1]. In other cases, stochastically generated music is filtered using explicit rules to match a given style [2]. Yet in some other cases, stochastic methods are used to develop new musical styles and languages [5, 7]. But, have numbers, maths or probability got a sound? Could anybody create music from them? If so, what are the variables that need to be considered in such compositions? Could it be possible for a human to distinguish its origin? As an attempt to shed some light on these questions, several probabilistic models have been implemented to get a rough idea of how probability models sound. Starting with basic models, the possibilities of this artificially created music to mimic existing styles are explored. Choosing different scales to define the state space of the stochastic processes simulated leads to unique pieces of music resembling, in their structure and musical features, others originally created by real composers. The arguable aesthetic value of this artificial music is not considered here.

Keywords: Music; Composition; Probability; Stochastic process.

Acknowledgements Work supported by project MTM2017-86875-C3-3-R.

References

[1] Brooks, F. P., Hopkins, A. L., Neumann, P. G., & Wright, W. V. (1957). An experiment in musical composition. *IRE Transactions on Computers*, **EC-6(3)**, pp. 175–182.

[2] Hiller, L. A., & Isaacson, L. M. (1959). *Experimental Music: Composition with an Electronic Computer.* McGraw-Hill, New York 1959.

[3] Mosteller, F., & Youtz, C. (1968). *Mozart's Random Music. Memorandum Ex-2.* Department of Statistics, Harvard University, Cambridge, Massachusetts. Retrieved from http://www.stat.harvard.edu/Site_Content/mosteller_celebration.html

[4] Pinkerton, R. C. (1956). Information theory and melody. *Scientific American*, **194(2)**, pp. 77–87.

[5] Pritchett, J. (1989). *The development of Chance Techniques in the Music of John Cage, 1950-1956.* New York University, 1989.

[6] Temperley, D. (2007). *Music and Probability.* The MIT Press, Cambridge, Massachusetts 2007.

[7] Xenakis, I. (1971). *Formalised Music.* University Press, Indiana 1971.

[*] PhD student. Department of Medicina y Cirugía, Psicología, Medicina Preventiva y Salud Pública Inmunología y Microbiología Médica, Enfermería y Estomatología. Universidad Rey Juan Carlos.

[†] Department of Artificial Intelligence. Escuela Superior de Ingenieros Informáticos. Universidad Politécnica de Madrid.

[‡] Department of Medicina y Cirugía, Psicología, Medicina Preventiva y Salud Pública Inmunología y Microbiología Médica, Enfermería y Estomatología. Universidad Rey Juan Carlos.

Sale of Objects with Bimodal Density Functions in Real Situations. A New Auction

Inmaculada Gutiérrez [*] Javier Castro [†] Rosa Espínola [‡] Daniel Gómez [§]

I. Gutiérrez (PhD Student) Inmaculada Gutiérrez is a PhD student at Complutense University of Madrid. She has the role *hired predoctoral and research workforce in training of the UCM*, in the department of *Estadística y Ciencia de los Datos* of the faculty *Estudios Estadísticos*, of UCM. She was graduated in Mathematics in 2015, and then she did a master in *Data Mining and Business Intelligence*, both in the UCM. She has also a wide laboral experience at working in a private company.

Her main lines of research are Game Theory, focusing on auctions and on problems associated with the processing of fuzzy information; and problems about clustering, networks and graphs.

The importance and widespread use of auctions are well established. In both the public and private spheres, a lot of economic transactions are conducted through various types of auctions (see [3, 5, 8] for further details).

In this study we have defined a new auction, called the draw auction, which is based on the implementation of a draw if a minimum value of sale is not reached. We find that bidding the true valuation of the object comprises a Bayesian Nash Equilibrium strategy [1, 2] for the draw auction. Furthermore, we show that the expected profit [4, 6, 9] for the seller in the draw auction is greater than in the second-price, greater than in the second-price with minimum price of sale auction and greater than in Myerson's auction [7] for some specific situations that will be detailed along the paper. We make this affirmation for objects whose valuation can be modeled as a bimodal density function in which first mode is much greater than the second one, such as seized or inherited objects All of these results have been shown using computational tests, so that we have defined an algorithm to calculate Myerson's auction.

Keywords: Auctions and Bidding; Bimodal Distribution; Myerson Auction; Second-Price Auction; Draw Auction

Acknowledgements This research was partially supported by the Government of Spain, Grant Plan Nacional de I+D+i, MTM2015-70550-P, and TIN2015-66471-P.

References

[1] Carbonell-Nicolau, O., McLean, R. (2018). On the existence of Nash equilibrium in bayesian games. *Mathematics of Operation Research*, **43**, pp. 100-129.

[2] Gibbons, R. (1992). *Game Theory for Applied Economists.* New Jersey. Princeton University.

[3] Kagel, J., Roth, A. (1995). *The Handbook of Experimental Economic.* Princenton University.

[4] Krishna, V. (2010). *Auction Theory.* Elsevier.

[5] Lorentziadis, P. (2016). Optimal bidding in auctions from a game theory perspective. *European Journal of Operational Research*, **248**, pp. 347-371.

[6] Milgrom, P., Weber, R. (1982). A theory of auctions and competitive biddings. *Econometrica*, **50**, pp. 1089-1122.

[7] Myerson, R. (1981). Optimal auction design. *Mathematics of Operations Research*, **6**, pp. 58-73.

[8] Rose, C., Madlener, R. (2013). An auction design for local reserve energy markets. *Decission Support System*, **56**, pp. 168-179.

[9] Wolfstetter, E. (1996). Auctions: An introduction. *Journal of Economics Surveys*, **10**, pp. 367-420.

[*] Faculty of Statistical Studies, Complutense University of Madrid, Email: inmaguti@ucm.es
[†] Faculty of Statistical Studies, Complutense University of Madrid, Email: jcastroc@estad.ucm.es
[‡] Faculty of Statistical Studies, Complutense University of Madrid, Email: rosaev@estad.ucm.es
[§] Faculty of Statistical Studies, Complutense University of Madrid, Email: dagomez@estad.ucm.es

Session 7 (Thursday 16:20)

Session talks

Exact and heuristic methods to solve the premarshalling problem with crane time minimization objective

Consuelo Parreño-Torres Ramón Álvarez-Valdés Rubén Ruiz Kevin Tierney [*†‡]

C. Parreño-Torres (Ph.D. student) Consuelo Parreño-Torres is a Ph.D. student in Statistics and Operations Research at the University of Valencia. Her research subject is Operations Planning in Container Terminals, and more specifically, the study of problems related to Storage Yard Operations (Pre-marshalling, Gantry Crane Routing and Scheduling). Before starting her Ph.D. research, she obtained a Degree in Mathematics, and a Master's Degree in Business Process Planning and Management, obtaining the extraordinary award. Her Master's Thesis dealt with a related problem, concerning the Stowage Planning Problem in Container Ships.

The average berthing time of ships is one of the main parameters to measure port efficiency. In such a competitive sector, terminals attempt to reduce these times by optimizing their processes. The correct reshuffling of the container yard dramatically reduces unloading/loading times and can be performed before the arrival of a ship, when the workload at the terminal is at a minimum.

The pre-marshalling problem (CPMP) seeks to transform the initial layout of a bay into a final layout without any containers blocking, facilitating the later retrieval of containers without unproductive moves, that reduce the terminal throughput at busy periods. The classic objective of the CPMP consists in minimizing the number of moves to transform the initial layout of a bay into a final layout without any containers blocking the removal of others [1, 2]. The number of moves has been used as an indicator of the time employed by the crane to rearrange the bay. However, this study shows that the number of moves is not entirely representative and presents a more realistic objective that minimizes the real crane time (CPMPCC).

We set the times of the crane according to the speeds and technical features of rubber-tired gantry cranes (RTGCs) used in the Noatum terminal of Valencia's port. We perform an extensive analysis showing the variability in the time used by the crane that alternative solutions of an equal number of moves present, as well as the overlap of times between solutions with a different number of movements, for four instances belonging to datasets highly studied in the literature.

With respect to exact methods, we propose integer programming models to solve the CPMPCC and present upper bounds for the number of moves to solve that problem. We also propose a beam search metaheuristic approach to solve the most challenging instances. An extended computational analysis is carried out over well-known premarshalling datasets, testing both exact models and the heuristic approach.

Keywords: Logistics; Container pre-marshalling; Maritime applications; Terminal operations

Acknowledgements This study has been partially supported by the Spanish Ministry of Education, Culture, and Sport, FPU Grant A-2015-12849 and by the Spanish Ministry of Economy and Competitiveness, under projects DPI2014-53665-P and DPI2015-65895-R, partially financed with FEDER funds.

References

[1] Consuelo Parreño-Torres, Ramon Alvarez-Valdes, and Rubén Ruiz. Integer programming models for the premarshalling problem. *European Journal of Operational Research*, 274(1):142–154, 2019.

[2] S. Tanaka and K. Tierney. Solving real-world sized container pre-marshalling problems with an iterative deepening branch-and-bound algorithm. *European Journal of Operational Research*, 264(1):165 – 180, 2018.

[*] Department of Statistics and Operations Research, University of Valencia, Doctor Moliner 50, Burjassot, 46100, Valencia, Spain. Email: consuelo.parreno@uv.es

[†] Grupo de Sistemas de Optimización Aplicada, Instituto Tecnológico de Informática, Ciudad Politécnica de la Innovación, Edifico 8G, Acc. B. Universitat Politècnica de València, Camino de Vera s/n, 46021, València, Spain. Email: rruiz@eio.upv.es

[‡] Decision and Operation Technologies Group, Bielefeld University, Universitätsstraße 25D-33615, Bielefeld, Germany.

Attraction-Repulsion clustering with applications to fairness

Hristo Inouzhe [*] Eustasio del Barrio [†] Jean-Michel Loubes [‡]

Hristo Inouzhe (PhD Candidate) I'm a PhD candidate in Mathematics in the University of Valladolid, while I'm also a member of IMUVA. My main interests are in clustering, optimal transport and fairness. I graduated in 2013 in Physics and in 2014 in Mathematics in the Universidad Autónoma de Madrid. I earned my masters degree in Mathematics and Applications in 2015 in Universidad Autónoma de Madrid. My thesis advisors in Valladolid are Eustasio del Barrio and Carlos Matrán.

Clustering techniques are increasingly more influential in people's life since they are used in credit scoring, article recommendation, risk assessment, spam filtering, sentencing recommendations in courts of law, etc... Therefore, there are justified concerns about the fairness of the outcomes of such automatic procedures. Nonetheless, if the the data at hand reflects a real world bias, learning algorithms can pick on this behaviour and emulate it.

Recently, concerns about fairness have received an increasing attention, resulting into two main strategies to address it. Transform the data in order to avoid correlation between the set of sensitive attributes and the rest of the data [2] or modify the objective functions of the algorithms in a way that eliminates or reduces unfairness [3].

In our setting, we have an i.i.d. sample (X_1, S_1), $\dots, (X_n, S_n) \sim (X, S)$, where $S \in \mathbb{R}^p$ represents the sensitive attributes and $X \in \mathbb{R}^d$ represents the rest of variables of interest. We assume that X and S are not independent. A fair clustering of the data $X_1, \dots, X_n$, is a partitioning such that the proportion of every sensitive (protected) class $S_{\cdot,1}, \dots, S_{\cdot,p}$ in each cluster (group) is the same as the corresponding proportion in the whole sample. Achieving fairness with constrains on group proportions is computationaly demanding [1]. In this work, we aim to achieve groups with proportions of the sensitive attributes more similar to that of the whole dataset, introducing dissimilarities that allow a repulsion between elements of the same protected class and/or an attraction between elements of different classes. We do this in a way that is computationally efficient and that can be adapted to standard clustering tools.

One of the dissimilarities we have introduced is

$$\delta\left((X_1, S_1), (X_2, S_2)\right) = \left(1 + ue^{-v\|S_1 - S_2\|^2}\right)\|X_1 - X_2\|^2$$

with $u, v \geq 0$, which is a multiplicative perturbation of the squared Euclidean distance. u controls the maximum perturbation achievable, while v modulates how fast we diverge from this maximum perturbation when S_1 is different to S_2. Let us fix $S_1, S_2 \in \{-1, 1\}$, $u = 0.1$ and $v = 100$. When $S_1 \neq S_2$ we have approximately $\|X_1 - X_2\|^2$, while when $S_1 = S_2$ we have $1.1\|X_1 - X_2\|^2$, introducing a repulsion between elements of the same class.

Our methods are based on a dissimilarity matrix $\Delta_{i,j} = \delta((X_i, S_i), (X_j, S_j))$. We can use Δ for a hierarchical clustering procedure. Even more, we can use multidimensional scaling to transform $(X_1, S_1), \dots, (X_n, S_n)$ into $X'_1, \dots, X'_n \in \mathbb{R}^{d'}$, where $D_{i,j} = \|X'_i - X'_j\|$ is similar to $\Delta_{i,j}$. Hence, we can apply any clustering procedure suited for the Euclidean space on the transformed data. When using clustering based on Δ, which increases distances between elements with the same class, we expect clusters to be more heterogeneous in the protected class.

Keywords: Fair Clustering; Multidimensional Scaling; Hierarchical Clustering; Kernel-Trick.

References

[1] del Barrio, E. Gamboa, F. Gordaliza, P. and Loubes, J-M. (2018). Obtaining fairness using optimal transport theory. Retrieved from https://arxiv.org/pdf/1806.03195.pdf.

[2] Chierichetti, F. Kumar, R. Lattanzi, S. and Vassilvitskii, S. (2017). Fair clustering through fairlets. *Advances in Neural Information Processing Systems*, **30**, pp. 5029–5037.

[3] Zafar, M. B. Valera, I. Rodriguez, M. G. and Gummadi, K. P. (2017). Fairness constraints: mechanisms for fair classification. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, **54**, pp. 962–970.

[*] Departamento de Estadística e Investigación Operativa and IMUVA, Universidad de Valladolid, Spain. Email: hristo.inouzhe@uva.es

[†] Departamento de Estadística e Investigación Operativa and IMUVA, Universidad de Valladolid, Spain. Email: tasio@eio.uva.es

[‡] Université de Toulouse, Institut de Mathématiques de Toulouse, Toulouse, France. Email: loubes@math.univ-toulouse.fr

Session 8 (Thursday 17:30)

Session talks

Embedding uncertainty to integrated depth measures for partially observed functional data

Antonio E. Fernández [*] Laura M. Sangalli [†] Anna M. Paganoni [‡] Raúl J. Jiménez [§]

Antonio E. Fernández (PhD student) Antonio Elías Fernández is a PhD student at Universidad Carlos III de Madrid, Department of Statistics. His research interest is in functional data and, particularly, he tries to make the most out of the potential applicability of depth measures. The main part of his thesis is devoted to develop a depth-based method for forecasting functional time series and for function reconstruction problems. In addition, he has worked on exploring functional data sets trough networks and visualizing very high-dimensional functional outliers.

The classical literature on Functional Data Analysis (FDA) focus on the analysis of curves that are commonly observed in a continuous and compact domain. However, the presence of data sets where the functions are not completely recorded is becoming more recurrent in real applications and, unfortunately, this issue invalidates many of the methodologies for FDA.

Methodologies for dealing with this issue have been proposed for principal components analysis, clustering, classification and functional reconstruction. However, and up to our knowledge, there is not a suitable depth ([1]) for this setting, i.e., a measure that allows to order a partially observed or sparse functional data set from the center to outwards. Depth measures are an important tool for robust and nonparametric statistics; They are not only useful for ranking data sets, but for visualization, outlier detection, classification, forecasting and missing values imputation, among others.

In this work, we propose a building-block depth definition that allows to incorporate the uncertainty associated with the censoring process. The first block is an integrated functional depth and the second one a weighting function that penalizes the domain where the sample is poorly observed. The validity of our proposal is studied theoretically and tested by simulation under different censoring settings in terms of expected loss, number of missing intervals and statistical relationship with the process of interest. We show that the depth values obtained by our proposal on the partially observed sample are comparable with the values obtained from the fully observed data set. Once we have a suitable definition, we are able to unlock well-known techniques that require the use of a depth measure.

As a case study we consider the AneuRisk dataset that has been analysed for evaluating the role of vascular geometry and hemodynamics in the pathogenesis of cerebral aneurysms [2]. Our goal is to discriminate patients with aneurysms in different districts by using the geometric features of the internal carotid artery expressed by its radius profile and ceterline curvature. These two variables are partially observed and many analysis had to be restricted to the domain where all the individuals were recorded. We show that, using our definition on the partially observed sample and DD-classifiers [3], we achieve smaller classification errors than other studies published before.

Keywords: Functional Data Analysis; Depth Measures; Censored Data; Classification; AneuRisk.

Acknowledgements Supported by the Spanish Ministerio de Educación, Cultura y Deporte under grant FPU15/00625 and partially supported by the Spanish Ministerio de Economía y Competitividad under grant ECO2015-66593-P.

References

[1] Nagy, S. and Gijbels, I. (2017). On a general definition of depth for functional data. *Statistical Science*, **32**, pp. 630–639.

[2] Sangalli, L. M., Secchi, P., Vantini, S. and Veneziani, A. (2009). A Case Study in Exploratory Functional Data Analysis: Geometrical Features of the Internal Carotid Artery. *JASA*, **104**, pp. 37–48.

[3] Cuesta-Albertos, J. A., Febrero-Bande, M. and Oviedo de la Fuente, M. (2017). The DD^G-classifier in the functional setting. *TEST*, **26**, pp. 119–142.

[*] Department of Statistics, Universidad Carlos III de Madrid, Spain. Email: aelias@est-econ.uc3m.es.

[†] Mox - Department of Mathematics, Politecnico di Milano, Italy. Email: laura.sangalli@polimi.it

[‡] Mox - Department of Mathematics, Politecnico di Milano, Italy. Email: anna.paganoni@polimi.it

[§] Department of Statistics, Universidad Carlos III de Madrid, Spain. Email: rjjimene@est-econ.uc3m.es

Inventory Movement Optimization Problem

Abdessamad Ouzidan [*†] Marc Sevaux [†] Eduardo G. Pardo [‡] Bérenger David [*]
Alexandru-Liviu Olteanu [†]

A. Ouzidan (PhD student) Abdessamad Ouzidan is a PhD student working currently for the company Fives Syleps and the Lab-STICC, Université Bretagne Sud, both located in the city of Lorient in France. He earned his master's degree in applied mathematics from Université Bretagne Sud in 2017. Abdessamad presented his work in several international and national conferences, such as: the International Symposium on Mathematical Programming (ISMP 2018) and the national conference ROADEF (2018 & 2019).

Inventory Movement Optimization Problem (IMOP) is a new problem derived from the industry which looks for an efficient way of processing orders within a warehouse, by minimizing the number of movements of single-item containers from the storage zone to the processing zone (see Figure 1).

In the order picking process, a group of orders are positioned in the processing area which has a limited number of slots. Then, one or multiple auto-guided vehicles move single-item containers from the storage area to the processing zone. When a container arrives to the processing area, an operator feeds all orders that are in the processing area from this single-item container. The container then returns to the storage area, as it is depicted in Figure 1.

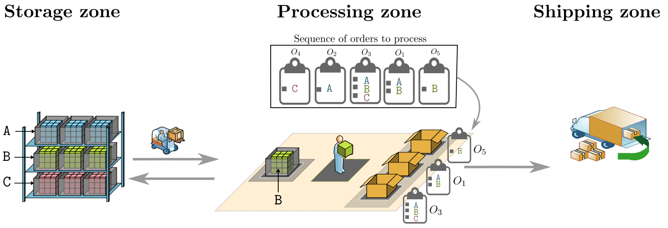

Figure 1: Warehouse areas distribution

Usually, the order picking process is performed in a mode called *Batching*. In this mode, we look for a batch of orders that have similar needs where the number of orders does not exceed the size of the processing zone. These orders will be processed simultaneously and will be sent to the shipping area when all the orders of the batch are fully processed.

Unlike the *Batching* mode, where we wait for all the orders of a batch to be fully processed, in the IMOP, when an order is fully processed, it is sent immediately to the shipping area, thus releasing a slot in the processing zone.

The IMOP looks for (a) a sequence in which orders should be processed and (b) a sequence in which items should be retrieved from the storage area to satisfy the orders, with the objective of minimizing the size of the latter. Notice that the number of orders that can be processed at a time is defined by the maximum number of slots available in the processing zone, where each slot can handle one order.

The benefits of minimizing the number of movements are:

- reducing the number of needed vehicles in the warehouse that are in charge of moving single item containers from the storage area to the processing zone;
- reducing the size of the processing zone;
- reducing energy consumption.

All these factors have an economical impact for the company.

Since we are dealing with a new problem that, to the best of our knowledge, has not yet been tackled in the literature [1], we are working on proving its complexity [2] and proposing Mixed-Integer Linear Programming approaches to solve it to optimality as well as efficient heuristics to tackle larger, real instances of the company.

Keywords: Combinatorial Optimization; Mathematical Programming; Warehouse; Scheduling; Complexity.

Acknowledgements Work supported by the company Fives Syleps, Association Nationale de la Recherche et de la Technologie (France) with PhD grant ref. 2017/1525 and by Ministerio de Economía y Competitividad (Spain) with grant ref. TIN2015-65460-C2-2-P.

References

[1] Yalaoui, A. and Chehade, H. and Yalaoui, F. and Amodeo, L. (2012). *Optimization of logistics.* John Wiley & Sons.

[2] Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W. H. Freeman.

[*] Fives Syleps, Lorient, France. Email : {abdessamad.ouzidan,berenger.david}@fivesgroup.com
[†] Lab-STICC, Université Bretagne Sud, Lorient, France. Email: {abdessamad.ouzidan,alexandru.olteanu,marc.sevaux}@univ-ubs.fr
[‡] Universidad Politécnica de Madrid, Dept. Sistemas Informáticos, Madrid, Spain. Email: eduardo.pardo@upm.es

Linear frontiers induced by the diagnostic problem

Sonia Pérez-Fernández [*] Pablo Martínez-Camblor [†] Norberto Corral [‡]

S. Pérez-Fernández (PhD student) Sonia Pérez Fernández is a PhD student at the University of Oviedo. Her main interests are Multivariate Analysis and the Receiver Operating Characteristic curve. She received her Bachelor's Degree in Mathematics from the University of Oviedo in 2014 and one year later, she earned her Master's Degree in Mathematical Modelling and Research, Statistics and Computing (Interuniversity: six Spanish Universities involved). She enrolled at the Doctorate Program in Mathematics and Statistics in the University of Oviedo in 2015 and since then she has done several research stays in the Technical University of Vienna (Austria), with Peter Filzmoser, as well as in The Dartmouth Institute (NH, United States), with Pablo Martínez-Camblor, both co-supervisors.

The Receiver Operating Characteristic (ROC) curve is a commonly used graphical tool to analyse the capacity of a continuous variable (marker) to distinguish between two populations. Particularly, it displays the trade-off between the probabilities of the two types of error induced by any binary classification when the cut-off point taken to define the categorisation varies along the domain of the marker.

Consideration of a single cut-off point to classify is based on the assumption that larger values of the marker are associated with a higher probability of belonging to one class, while lower values are linked to the other one. However, there are situations where both extremes (higher and lower) values are related to a higher probability of belonging to one class. The natural extension of the ROC curve to those scenarios is to consider two thresholds instead of one, resulting in the so-called generalized ROC (gROC) curve [1].

Both ROC and gROC curves are defined for univariate markers, but it is clear that taking more than one marker simultaneously may enable substantial improvements in the classification accuracy. It is known that the optimal transformation, in terms of achieving the highest sensitivity for any fixed specificity (defined by a single cut-off point), is the likelihood ratio, i.e, the rate of the density functions of the marker in each population [2]. However, a classification rule based on the transformation of the multivariate marker by its likelihood ratio may lead to classification regions which are difficult to interpret for the practitioner.

In this work we consider the gROC curve approach in the multivariate setting to analyze the impact of the restriction of the problem under study to classification regions defined by two linear frontiers or hyperplanes on the resulting gROC curve, compared to the aforementioned optimal ROC curve, for different parametric scenarios.

On the other hand, in order to deal with non-parametric estimators of the ROC curve, the empirical ROC curve is the most used one, by using the empirical distribution function in one population and the empirical quantile function in the other. The asymptotic properties of this estimator were first developed by [3], who showed that it converges to the sum of two independent Brownian bridges. In this work, some properties of the empirical gROC are derived by using the theory of empirical processes.

Keywords: Empirical processes; Generalized ROC curve; Linear combinations; Multivariate marker.

Acknowledgements Work supported by project MTM2015-63971-P from the Spanish Ministerio of Economía y Competitividad and by FC-GRUPIN-IDI/2018/000132 and PhD grant Severo Ochoa BP16118 from the Principado de Asturias.

References

[1] Martínez-Camblor, P. et al. (2017). Receiver operating characteristic curve generalization for non-monotone relationships. *Stat. Methods Med. Res.*, **26**(1), pp. 113–123.

[2] McIntosh, M. W. and Pepe, M. S. (2002). Combining several screening tests: optimality of the risk score. *Biometrics*, **58**(3), pp. 657–664.

[3] Hsieh, F. and Turnbull, B. W. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Ann. Statist.*, **24**(1), pp. 25–40.

[*] Department of Statistics and Operational Research and Mathematics Didactics, University of Oviedo, Spain. Email: perezsonia@uniovi.es

[†] The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine at Dartmouth, New Hampshire, United States. Email: Pablo.Martinez.Camblor@Dartmouth.edu

[‡] Department of Statistics and Operational Research and Mathematics Didactics, University of Oviedo, Spain. Email: norbert@uniovi.es

Uncovering key nodes in networks: A Mixed-Integer Programming approach

Emilio Carrizosa [*] Alfredo Marín [†] Mercedes Pelegrín [‡]

M. Pelegrín (PhD student) Mercedes Pelegrín García is a predoctoral researcher at the University of Murcia. Her main interests are Integer Programming, Locational Analysis and Computer Science. She received her B.S. from University of Murcia in 2015 in mathematics and engineering informatics; she earned her master's degree in advanced mathematics from the University of Murcia a year after. In October 2016, she enrolled in the doctoral program in mathematics at the same university, where she currently studies combinatorial optimization under Alfredo Marín. She is a member of the National Society of Statistics and Operational Research and the EURO Working Group on Locational Analysis. Pelegrín was a participant of the 6th Heidelberg Laureate Forum and was the recipient of several honors, including a secondary award in the national undergraduate competition XIV Certamen Arquímedes and the University of Murcia award to academic excellence in double-degree programs.

Identifying key members in social networks is a crucial step to understand and modify the underlying systems, including biological organisms and human society [4]. Centrality measures such as degree, closeness, betweenness or eigenvector centrality, have been introduced to spot most relevant individuals, but they fail to discern directly the relevance of a group of nodes. Some of these measures have been adapted to address group centrality [1, 2]; however, eigenvector centrality remains unexplored.

Here, we adapt eigenvector centrality to identify the group of p most relevant nodes in a network. Eigenvector computation is embedded in a clustering procedure to design a method that guarantees key nodes detection while preventing their spheres of influence from overlapping. Modeling this idea with mathematical optimization variables involves highly non-linear equations, which are linearized to produce a Mixed-Integer Linear Programming formulation for the problem. Our model uncovers the group of p most relevant nodes in the network together with their spheres of influence, which are obtained as a byproduct of the optimization and coincide with the clusters. The method divides the network in p clusters, in such a way that sum of the highest eigenvector centralities in each cluster is maximized. The introduction of a mixing parameter in the model guarantees cluster cohesion and allows to reinterpret clusters as network communities.

Our computational experience shows that our formulation is effective and of practical value, being able to find optimal solutions on networks of several hundreds of nodes and thousands of links. In addition, the model correctly identifies the community structure of the networks of Lancichinetti, Fortunato and Radicchi [3].

Keywords: Networks; Eigenvector Centrality; Clustering; Mixed-Integer Programming.

Acknowledgements Work supported by project MTM2015-65915-R and PhD grant FPU15/05883.

References

[1] Borgatti S.P. (2006). Identifying sets of key players in a social network. *Computational and Mathematical Organization Theory*, **12**, pp. 21–34.

[2] Everett, M. G., Borgatti, S. P. (1999). The centrality of groups and classes. *The Journal of Mathematical Sociology*, **23**(3), pp. 181–201.

[3] Lancichinetti, A., Fortunato, S., Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. *Physical Review E*, **78**(4), 046110.

[4] Newman, M., Barabási, A. L., Watts, D. J. (2006). *The structure and dynamics of networks* (Vol. 19). Princeton University Press.

[*] Department of Statistics and Operational Research and Instituto de Matemáticas (IMUS), Universidad de Sevilla, Spain. Email: ecarrizosa@us.es

[†] Department of Statistics and Operational Research, Universidad de Murcia, Spain. Email: amarin@um.es

[‡] Department of Statistics and Operational Research, Universidad de Murcia, Spain. Email: mariamercedes.pelegrin@um.es

Session 9 (Thursday 19:00)

Session talks

Nonparametric methods to estimate the probability of default

Rebeca Peláez Suárez [*] Ricardo Cao Abad [†] Juan Manuel Vilar Fernández [‡]

R. Peláez Suárez (PhD student) Rebeca Peláez Suárez is a PhD student in the doctoral program in Statistics and Operational Research at the University of A Coruña. Her main interests are nonparametric statistics, censored data and financial risk. She received a degree in Mathematics from the University of Oviedo in 2017. She earned her master's degree in Statistical Techniques from the University of A Coruña in 2019 and then she enrolled in the doctoral program where she studies under the supervision of Ricardo Cao and Juan Vilar.

Once the credit scoring assigned by a financial institution to a client who enjoys a personal credit is known, it is interesting to find the probability that the borrower declares himself unable to face the debt contracted with the bank after some time (for example, one year) of its formalization. The main aim of this work is to propose models to estimate this probability, known as the probability of default and denoted by PD.

The probability of default conditional to the credit scoring can be written as a transformation of the conditional survival function of the variable "time to default". This property is used to develop new PD estimators. Throughout the study of a set of credits, the default is not observed for all of them and the variable "time to default" is censored. As a consequence, censored data and survival analysis will be used.

Three estimators for the conditional survival function have been considered: Beran's generalized product-limit estimator [1], Cai's estimator [2] and Van Keilegom-Akritas' estimator [3]. They have been transformed to obtain the corresponding PD estimators. Using the asymptotic bias, variance and normality results of the Cai [2] and Van Keilegom-Akritas [3] survival estimators, the asymptotic bias, variance and normality of the PD estimators are obtained. These are complex expressions that make it difficult to obtain an approximation of the MSE or MISE since they depend on several unknown parameters.

In order to analyze the behavior of the three resulting PD estimators by simulation, several models have been considered. In the first studies carried out, Beran's estimator provides the best results because it gives the best approximations (smaller mean squared error) and less computation time. The Cai's and Van Keilegom-Akritas' estimators for the distribution function turn out to be better than Beran's with heavy censorship. A simulation study is in progress to check whether this property is inherited by the corresponding PD estimators. These three estimators involve a smoothing parameter in the covariate (credit scoring); the choice of this bandwidth parameter is being examined in the simulation study.

Another fact observed in the simulation is that the PD estimators conditional to the credit scoring present excessive variability even using a large bandwidth to smooth the covariate. Therefore, a modification of the proposed estimators based on smoothing in the time variable is proposed. The first analyses seem to indicate that these estimators have a better behavior.

Keywords: censored data; conditional survival function; credit risk; nonparametric regression; probability of default

Acknowledgements This research has been supported by MINECO grant MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the ERDF.

References

[1] Beran, R. (1981). Nonparametric regression with randomly censored survival data. *Technical report*, University of California.

[2] Cai, Z. (2003). Weighted local linear approach to censored nonparametric regression. *Recent Advances and Trends in Nonparametric Statistics*, Michael G. Akritas and Dimitris N. Politis, pp. 217–231.

[3] Van Keilegom, I. and Akritas, M.G. (1999). Transfer of tail information in censored regression models. *The Annals of Statistics*, **27**:5, pp. 1745–1784.

[*] Research Group MODES, Department of Mathematics, CITIC University of A Coruña, A Coruña, Spain. Email: rebeca.pelaez@udc.es

[†] Research Group MODES, Department of Mathematics, CITIC University of A Coruña and ITMATI, A Coruña, Spain. Email: ricardo.cao@udc.es

[‡] Research Group MODES, Department of Mathematics, CITIC University of A Coruña and ITMATI, A Coruña, Spain. Email: juan.vilar@udc.es

An adversarial risk analysis approach for differential games: A botnet defense model

Jorge González-Ortega [*] Antonio Gómez-Corral [†] David Ríos Insua [‡]

J. González-Ortega (Postdoctoral researcher) Jorge González Ortega is a postdoctoral researcher at Instituto de Ciencias Matemáticas. His main interests are decision analysis and Bayesian statistics. He received his bachelor in mathematics from Universidad Complutense de Madrid (UCM) in 2013 and earned his master's degree in mathematical engineering also from UCM in 2014. After one year at Management Solutions as an assistant consultant, he enrolled in the IMEIO PhD program at UCM, where he studied adversarial risk analysis under David Ríos Insua.

Differential Games (DGs) are mathematical models concerning strategic interactions between two or more agents that control the evolution of a system over time [3]. Its applications are relevant in finance, cyber-security, and predator-prey models. Their methods may be seen as a combination of game theory and optimal control theory.

Essentially, DGs consider that the evolution of a system's state in time is dictated by a system of differential equations that depend on the state itself and the decisions made by two or more decision-makers, who receive payments that depend on the system's evolution, the final state and the implemented decisions. As a fundamental solution concept, using and extending optimal control concepts (such as Pontryagin's maximum principle), Nash equilibria are computed [1]. This requires some common knowledge assumptions so that each agent knows about the status of the others, which is hardly sustainable in many of the above applications.

Adversarial Risk Analysis (ARA) provides a way forward, as common knowledge is no longer required. Nash equilibria notions are abandoned and a single decision-maker (defender) is supported against the others (attackers), minimising her subjective expected costs while treating the attackers' decisions as random variables. Under assumptions about their rationality, ARA tries to assess the attackers' probabilities and utilities to predict their optimal actions, with the uncertainty in the assessments leading to probability distributions over them. Thus, the original DG is transformed to a set of stochastic control problems in which the stochastic parameters relate to the different strategic decisions that the agents make.

Our approach is illustrated through a botnet defense example developed in [2] under a game theoretical perspective. Botnets are computer networks infected with malicious programs that allow cyber-criminals (botnet herders) to control the infected machines remotely without the user's knowledge. In the example, a botnet herder pursues economic profits by intensifying his intrusion in a network of computers while a defender tries to mitigate it. The percentage of infected computers in the network evolves according to a modified SIS epidemic model. We provide an ARA solution and compare it to the game theoretical one.

Keywords: Non-cooperative games; Decision analysis; Nonzero-sum games; Cybersecurity.

Acknowledgements JGO's research is financed by MINECO under FPI SO grant agreement BES-2015-072892. Besides, DRI is supported by the MICIN program MTM2014-56949-C3-1-R and the AXA-ICMAT Chair on Adversarial Risk Analysis. This work has also been partially supported by MINECO through program SEV-2015-0554 and project MTM2015-72907-EXP; and the EU's H2020 Programme under grant agreements no. 285223 and 740920.

References

[1] Basar, T., & Li, S. (1989). Distributed computation of Nash equilibria in linear-quadratic stochastic differential games. *SIAM Journal on Control and Optimization*, **27(3)**, 563–578.

[2] Bensoussan, A., Kantarcioglu, M., & Hoe, S. (2010). A game-theoretical approach for finding optimal strategies in a botnet defense model. In Alpcan T., Buttyán L., Baras J.S. (Eds.). *Decision and game theory for security. GameSec 2010* (1st ed., pp. 135–148). Berlin, Germany: Springer.

[3] Nisio, M. (2014). *Stochastic control theory: Dynamic programming principle* (2nd ed.). Tokyo, Japan: Springer.

[*] Instituto de Ciencias Matemáticas, CSIC-UAM-UC3M-UCM, Madrid, Spain. Email: jorge.gonzalez@icmat.es
[†] Departamento de Estadística e Investigación Operativa, UCM, Madrid, Spain. Email: antonio_gomez@mat.ucm.es
[‡] Instituto de Ciencias Matemáticas, CSIC-UAM-UC3M-UCM, Madrid, Spain. Email: david.rios@icmat.es

Nonparametric estimation of the conditional survival function when cure is partially known

Wende Clarence Safari [*] Ignacio López-de-Ullibarri [†] María Amalia Jácome [‡]

W.C. Safari (PhD student) Wende Clarence Safari is currently a PhD student at the University of A Coruña, Spain. Her main interest is in developing practical statistical methods which can contribute to solve real life problems. Particularly, she works on nonparametric mixture cure models with cure partially known. Her original training was in Tanzania and Belgium, having a B.S. in Applied Statistics (Mzumbe, Tanzania), and an MSc in Biostatistics (UHasselt, Belgium).

Standard analyses of time-to-event data assume that every individual in the study will eventually experience the event of interest if followed for long enough. However, some individuals may be cured or long-term survivors, and they will not experience the event of interest, no matter how long they are followed. Let Y be a time to the event and C a censoring time independent of Y given a covariate X. It is of interest to estimate $S(t|x) = P(Y > t|X = x)$, the conditional survival function.

A fully nonparametric estimator of $S(t|x)$ without cure which is the generalized product-limit (PL) estimator was proposed in [1]. Approaches to estimating $S(t|x)$ with cures have been developed mainly based on (semi)parametric models. In the mixture cure models (MCM), [3] considered the nonparametric PL estimator in [1]. It is customary to assume no additional information on the cure status, thus, to model it as a latent variable. However, there might be situations when extra information about cure status is available, e.g., an individual is assumed to be cured or a long-term survivor if the observed survival time is greater than the cure threshold; or based on diagnostic tests. When cure is partially known a semiparametric estimator of $S(t|x)$ was proposed in [2]. Here we propose a nonparametric estimator of $S(t|x)$.

With cure partially known the observations are $\{(X_i, T_i, \delta_i, \xi_i, \xi_i\nu_i) : i = 1, \ldots, n\}$ where $T = \min(Y, C)$ is the observed time, $\delta = \mathbf{1}(Y \le C)$ is the uncensoring indicator, ξ is a binary variable indicating whether cure status is known ($\xi = 1$) or not ($\xi = 0$), and ν is the cure indicator. Thus, $\xi\nu = 1$ indicates that the individual is known to be cured. The conditional survival function $S(t|x)$ can be estimated nonparametrically by

$$\widehat{S}_h(t|x) = \prod_{i=1}^{n} \left(1 - \frac{\delta_{[i]} B_{h[i]}(x) \mathbf{1}\left(T_{(i)} \le t\right)}{B_{h[i]}(x) + \sum_{j=i+1}^{n} B_{h[j]}(x) \mathbf{1}\left(\xi_{[j]}\nu_{[j]} = 0\right) + B_h^c(x)} \right)$$

where $B_h^c(x) = \sum_{j=1}^{n} B_{h[j]}(x) \mathbf{1}\left(\xi_{[j]}\nu_{[j]} = 1\right)$ is the sum of the weights of all the individuals known to be cured, $B_{h[i]}(x) = \frac{K_h\left(x - X_{[i]}\right)}{\sum_{j=1}^{n} K_h\left(x - X_{[j]}\right)}$ are the Nadaraya-Watson weights, with $K_h(.) = K(./h)/h$ a rescaled kernel with bandwidth h. Finally, $\delta_{[i]}, X_{[i]}, \xi_{[i]}$ and $\nu_{[i]}$ are the concomitants of the ordered observed times $T_{(1)} < T_{(2)} < \ldots < T_{(n)}$.

Some theoretical properties of $\widehat{S}_h(t|x)$ have been derived, and a simulation study has been conducted to assess its performance. Data were simulated from a logistic-exponential mixture cure model. The mean integrated squared error was computed as a function of the bandwidth, and compared to that of the conditional product-limit estimator in [1] and to the semiparametric estimator in [2].

Keywords: Bandwidth; Cure models; Local weights.

Acknowledgements This work has been supported by MINECO grant MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the ERDF.

References

[1] Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical Report, University of California, Berkeley.

[2] Bernhardt, P. (2016). A flexible cure rate model with dependent censoring and a known cure threshold. *Statistics in Medicine*, **25**, pp. 4607–4623.

[3] López-Cheda, A. Jácome, M.A. and Cao, R. (2017). Nonparametric latency estimation for mixture cure models. *TEST*, **2**, pp. 353–376.

[*] Universidade da Coruña, MODES group, Department of Mathematics, A Coruña, Spain. Email: wende.safari@udc.es
[†] Universidade da Coruña, MODES group, Department of Mathematics, Ferrol, Spain. Email: ilu@udc.es
[‡] Universidade da Coruña, CITIC, MODES group, Department of Mathematics, A Coruña, Spain. Email: maria.amalia.jacome@udc.es

Session 10 (Friday 10:30)

Session talks

DChaos: An R Package for Detecting Chaotic Signals inside Time Series

Julio E. Sandubete [*]   Lorenzo Escot [†]

Julio E. Sandubete (PhD Student) completed his undergraduate degree in Economics and dual master's degree in International Financial Markets and Social Science Research Methods before moving on to study a PhD in Data Science in the Faculty of Statistical Studies at Complutense University of Madrid. His main interests are Chaos theory, Nonlinear time series analysis, Topological data analysis and Econometrics.

We present the R package DChaos [1] which contains several algorithms for the purpose of detecting chaotic signals inside univariate stationary time series. We have focused on methods derived from chaos theory which estimate the complexity of a dataset through exploring the structure of the attractor.

An attractor is a set of points towards which the trajectories of a dynamical system $f : \mathbb{R}^n \to \mathbb{R}^n$ converge. Λ is an attractor defined by $\Lambda = \cap f^t\left(\bar{A}\right)$ for $t = 1, 2, 3, \ldots, N$ where $\bar{A}$ is the attractor's basin of attraction. That is, it is the region of the phase space such that any of its points will eventually be iterated into that region.

We have considered the Lyapunov exponents λ as an ergodic measure of the system. The existence of at least one positive Lyapunov exponent implies the presence of a chaotic behaviour, that is the sensitivity on the initial conditions of the trajectories within the attractor. The Lyapunov exponents are measures of the average rate of divergence of each iterated orbit $f^t\left(x_t\right)$ into the attractor Λ corresponding to the map $x_t = f\left(x_{t-1}\right)$ for $\{x_t\}_{t=1}^{N}$. The k-th Lyapunov exponents are defined as

$$\lambda_k = \lim_{t\to\infty} \frac{1}{t} \log\left(\mu_k\left(\left|Df^t\right|\right)\right)$$

where $k = 1, 2, 3, \ldots, n$ and μ_k is the k-th largest eigenvalue of Df^t where $Df^t = Df\left(x_t\right) \cdot Df\left(x_{t-1}\right) \cdot \ldots \cdot Df\left(x_1\right)$ and $Df\left(\right)$ is the Jacobian. Then, our null hypothesis is $H_0 : \lambda_k \geqslant 0$ for at least one k. Not reject H_0 means that the dynamical system f from time series has a chaotic attractor. Regarding the estimation of the Lyapunov exponents there are two kind of methods. On the one hand the direct approach are based on the calculation of the growth rate of the divergence between two neighbouring trajectories with an infinitesimal difference in their initial conditions. On the other side, indirect methods try to estime the Lyapunov exponents indirectly through the estimation of the Jacobian Df [2].

We have implemented the Jacobian method by a fit through neural networks in order to estimate both the largest and the spectrum of the Lyapunov exponents. We have taken into account the full sample and three different methods of subsampling by blocks (non-overlapping, equally spaced and bootstrap) to estimate the Lyapunov exponents [3]. Some remarkable advantages of this method over the direct method are their robustness to the presence of noise and their satisfactory performance in moderate sample sizes.

In addition, our method allows to make inference about them and test if the estimated Lyapunov exponents values are or not statistically significant. This library can be used with time series whose observations are sampled at fixed or variable time intervals. The current released package version is available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=DChaos`.

Keywords: Chaos detection; Lyapunov exponent; Jacobian method; Neural networks; R.

Acknowledgements This work was supported by the Government of Spain (grant RTI2018-094901-B-I00); the Complutense University of Madrid - Faculty of Statistical Studies (grant CT45/15) and the Data Analysis in Social and Gender Studies and Equality Policies Research Group (`https://www.ucm.es/aedipi`).

References

[1] Sandubete, J.E., Escot, L. (2019). DChaos: Chaotic Time Series Analysis. *Comprehensive R Archive Network (CRAN).*

[2] Eckmann, J.P., Ruelle, D. (1985). Ergodic Theory of Chaos and Strange Attractors. *Reviews of Modern Physics*, **57**, pp. 617–656.

[3] Shintani, M., Linton, O. (2004). Nonparametric Neural Network Estimation of Lyapunov Exponents and a Direct Test for Chaos. *Journal of Econometrics*, **120**, pp. 1–33.

[*] Faculty of Statistical Studies, Complutense University of Madrid, Spain. Email: jsandube@ucm.es

[†] Faculty of Statistical Studies, Complutense University of Madrid, Spain. Email: escot@ucm.es

Multiple-criteria decision-making for recruitment according to company's reality

Pablo A. Pinto De la Cadena*

P. A. Pinto De la Cadena (PhD student) is an industrial engineer specialized in process management. He has a master's degree in business process planning and management, obtained at the University of Valencia. He is currently a PhD candidate in the statistics and optimization program of the University of Valencia. He has worked professionally as a senior consultant in the improvement and redesign of processes in financial entities in Ecuador. He has worked in the development of technological innovation projects in financial entities participating as a developer and leader of the project. He also has experience in the production area in textile, footwear and printing companies. There he has worked as a plant leader in the management of production. In addition to being responsible for the planning of production and demand forecast.

We deal with a real company's personnel selection problem using and comparing four multiple criteria based decision methods. These methods are: TOPSIS, OWA, OWA combined with expert assessment and an approach relaying on the similarity to an ideal profile prefixed by the company.

We will define the problem in the following way: A company has n candidates for $R_0 < n$ jobs. The selection of the most suitable candidates for each job will be based on their assessment in m competences. Once the candidates have been assessed in each competence they will be sorted according to an overall score and the first R_0s will be chosen as the most suitable. This process will be followed for each of the four proposed methods.

Each method is the most adequate given the real context of the company in each moment. TOPSIS is the most adequate method for those situations in which the company has not got its own prefixed definition of the ideal profile of the candidate for a specific job. In this scenario, the most common behaviour is to find the best qualified candidate in terms of all the considered competence, the ideal candidate. A candidate like this rarely exists due to the usual degree of conflict among the competences. Therefore, this candidate is usually a fictitious profile. When this happens a rational decision is to try to find the candidate more similar to the ideal profile. TOPSIS goes further, and it also takes into account the anti-ideal profile where the fictitious candidate has the worst qualifications in all the competences and thus, similarity with it becomes undesired.

In the second method, the company has not got a prefixed own defined ideal candidate. However, in this occasion the selection will be made with another criterion. The candidate that excels in a specific competition will not be sought. This approach looks for those candidates that excel in the majority even in their lowest grades. Weights in this method are not associated to the competences, but to a rearrangement of these.

In the third case, we will introduce an expert assessment to the previous scenario. This assessment will be conducted using a small group of candidates. This expert global assessment and skills qualifications allows replication to a greater number of candidates using a simple model of minimum squares which means for the company savings in time and money.

In the fourth case, it will be assumed that the company has an ideal defined profile for the position to be hired. Based on this profile, the selection of the best candidates will be made. These will be the candidates whose qualifications are closest to the ideal profile established by the company.

Keywords: Multiple-criteria decision-making; Competences; OWA operators; Personnel Selecction.

Acknowledgements I would like to thank University of Valencia with the program of scholarships Luisa Cardona.

References

[1] Canós, L., Liern, V. (2008). Soft computing-based aggregation methods for human resource management. *European Journal of Operational Research*, **189**, pp. 669–681.

[2] Yager, R. R. (2004). Generalized OWA aggregation operators. *Fuzzy Optimization and Decision Making*, **3**, pp. 93–107.

[*] Department of Business Mathematics, Universitat de València, Spain. Email: pinpa@alumni.uv.es.

On a projection-based class of uniformity tests on the hypersphere

Paula Navarro-Esteban [*] Eduardo García-Portugués [†] Juan Antonio Cuesta-Albertos [‡]

Paula Navarro Esteban (Associate Lecturer) Paula Navarro Esteban is an Associate Lecturer in the Department of Economics in the University of Cantabria since 2018. Her main research lines are random projections and outlier detection. She received her BSc in Mathematics from University of Zaragoza in 2011. She earned her MSc in Mathematical Modelling, Statistics, and Computation from the University of Zaragoza. She is currently doing her PhD in the University of Cantabria. From 2018, she is the coordinator of the Functional Data Analysis working group of the Spanish Society of Statistics and Operations Research.

Testing uniformity of a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of a random vector $\mathbf{X}$ supported on the unit hypersphere $\Omega_q := \{\mathbf{x} \in \mathbb{R}^{q+1} : \mathbf{x}'\mathbf{x} = 1\}$ of $\mathbb{R}^{q+1}$, with $q \geq 1$ is one of the first steps when analysing multivariate data for which only the directions (and not the magnitudes) are of interest – the so-called *directional data*. This kind of data arise in many applied disciplines, such as astronomy, biology, etc.

In this work a projection-based class of uniformity tests on the hypersphere Ω_q is proposed. The inspiration comes from the projection-based test of [2], which is based on the fact that the distribution of $\mathbf{X}$ is determined by that of a one-dimensional *random* projection, $\boldsymbol{\gamma}'\mathbf{X}$. For each $\boldsymbol{\gamma}$ (uniformly distributed on Ω_q and independent of the sample), [2] considered a Kolmogorov–Smirnov test statistic on the projected sample $\boldsymbol{\gamma}'\mathbf{X}_1, \ldots, \boldsymbol{\gamma}'\mathbf{X}_n$. This test clearly depends on $\boldsymbol{\gamma}$, which [2] mitigates by taking k random directions $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_k$ and combining the p-values associated to each of the k tests.

Differently from [2], we consider for each $\boldsymbol{\gamma}$ the well-known weighted quadratic norm by [1]:

$$(1) \quad Q^w_{n,q,\boldsymbol{\gamma}} := n \int_{-1}^{1} \left(F_{n,\boldsymbol{\gamma}}(x) - F_q(x)\right)^2 w(F_q(x))\, \mathrm{d}F_q(x),$$

where w is a weight function, $F_{n,\boldsymbol{\gamma}}$ and F_q are the empirical cumulative distribution function and the cumulative distribution function of the projected sample, respectively. In addition, instead of drawing several random directions and aggregating afterwards the outcomes of the associated tests, our statistic itself gathers information from all the directions on Ω_q: it is defined as the *expectation* of (1) with respect to $\boldsymbol{\gamma}$. The new class of uniformity tests is thus the one indexed by the weights w.

Using this formulation, simple expressions for several test statistics are obtained for the circle and sphere, and relatively tractable forms for higher dimensions. Despite their different origins, the proposed class and the well-studied Sobolev class of uniformity tests (see [3]) are shown to be related. Our new parametrization proves itself advantageous by allowing to derive new tests for hyperspherical data that neatly extend the circular tests by Watson, Ajne, and Rothman, and by introducing the first instance of an Anderson–Darling-like test in such context. The asymptotic distributions and the local optimality against certain alternatives of the new tests are obtained. A simulation study corroborates the theoretical findings. Finally, a real data example illustrates the usage of the new tests.

Keywords: Circular data; Directional data; Hypersphere; Sobolev tests; Uniformity.

Acknowledgements Work supported by grants MTM2016–76969–P and MTM2017–86061–C2–2–P from the Spanish Ministry of Economy, Industry and Competitiveness.

References

[1] Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. *J. Amer. Statist. Assoc.*, **49**, pp. 765–769.

[2] Cuesta-Albertos, J. A., Cuevas, A. and Fraiman, R. (2009). On projection-based tests for directional and compositional data. *Stat. Comput.*, **19**, pp. 367–380.

[3] Prentice, M. J. (1978). On invariant tests of uniformity for directions and orientations. *Ann. Statist.*, **6**, pp. 169–176.

[*] Department of Mathematics, Statistics and Computation, University of Cantabria, Spain. Email: paula.navarro@unican.es

[†] Department of Statistics, University Carlos III of Madrid, Spain. Email: edgarcia@est-econ.uc3m.es

[‡] Department of Mathematics, Statistics and Computation, University of Cantabria, Spain. Email: juan.cuesta@unican.es

Session 11 (Friday 12:00)

Session talks

Quantile regression: an adaptive penalized estimation

Álvaro Méndez Civieta [*] M. Carmen Aguilera-Morillo [†] Rosa E. Lillo [‡]

A. Méndez Civieta (PhD student) Álvaro Méndez is a PhD student at the University Carlos III of Madrid. His PhD thesis is centered in variable selection in high dimensional scenarios. He received his B.S. in mathematics from University of Oviedo in 2015. He earned his master's degree in Big Data Analytics from the University Carlos III of Madrid in 2016. After a year working as a data analyst at the consulting company Accenture, he enrolled in the doctoral program in mathematical engineering at the University Carlos III of Madrid, where he is currently working under the supervision of Rosa E. Lillo and M. Carmen Aguilera-Morillo.

Along years, regression has become a key method in statistics. Ordinary least squares (OLS) regression estimates the conditional mean response of a variable as a function of the covariates. However, OLS estimators rely on certain hypothesis over the first two moments that are not always verified in practical applications. Quantile regression models [1] allow a relaxation of the classical first two moment conditions over the model error, and it provides robust estimators capable of dealing with heteroscedasticity and outliers. They can also estimate different quantile levels of a response variable, giving a precise insight of the relation between response and covariates at upper and lower tails.

In recent years, high-dimensional data have become increasingly common. This problem can be found in many different areas like pattern recognition, climate data, or genetic data among others. In these scenarios, variable selection gains special importance offering sparse modeling alternatives that help identifying significant covariates and enhancing prediction accuracy. Sparse group LASSO (SGL) [2] is a penalization technique used in regression problems where the covariates have a natural grouped structure, and provides solutions that are both between and within group sparse. To the best of our knowledge, the SGL technique has not been studied in the framework of QR models, so we first address this gap extending the SGL penalization to quantile regression.

[3] was the first to propose the usage of specific weights for each variable on LASSO penalization as a way to increase the model flexibility. This idea, generally known as the adaptive idea, was then extended to many other penalizations. The weights of the adaptive idea are typically defined in the literature based on the results of non-penalized models. This definition is a key step for the demonstration of the oracle properties of the estimators [4], but it is restrictive in the sense that it limits the usage of the adaptive penalizations just to the case in which solving a non-penalized model is a feasible first step.

Our main contribution lies here. We define an adaptive sparse group lasso (ASGL) for quantile regression estimator, and we center our efforts on enabling the usage of our ASGL estimator in high-dimensional scenarios (with $p \gg N$), and with this objective in mind, we propose four alternatives for the weight calculation. We want to remark that our weight calculation alternatives can be used not only in the case of the ASGL estimator, but also in the rest of the adaptive-based estimators available in the literature.

Keywords: Quantile Regression; Group Variable Selection; Adaptive Sparse Group LASSO; High Dimension; Weight Calculation

References

[1] Koenker, R. (2005). *Quantile regression.* Cambridge University Press.

[2] Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22:231–245.

[3] Zou, H. (2006). The adaptive lasso and its oracle properties. *American statistical association.*

[4] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

[*] Department of Statistics, University Carlos III of Madrid. Email: alvaro.mendez@uc3m.es

[†] Department of Statistics, University Carlos III of Madrid. Email: maguiler@est-econ.uc3m.es

[‡] UC3M-Santander Big Data Institute, Department of Statistics, University Carlos III of Madrid. Email: lillo@est-econ.uc3m.es

Cost-sensitive classification and class probability estimation for Support Vector Machines

Sandra Benítez-Peña [*] Rafael Blanquero [†] Emilio Carrizosa [‡] Pepa Ramírez-Cobo [§]

S. Benítez-Peña (PhD Student) Sandra Benítez-Peña (Mairena del Alcor, 1993) is a 3rd year PhD student at the University of Sevilla (Spain), where she received both her BSc (2015) and MSc (2016) in Mathematics. Her research, under the supervision of Prof. Rafael Blanquero Bravo (University of Sevilla), Prof. Emilio Carrizosa (University of Sevilla) and Prof. Pepa Ramírez Cobo (University of Cádiz), is focused on the area of Mathematical Optimization for Supervised Classification.

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.

On the other hand, SVM does not provide probabilities as other classifiers do in a natural way. Many attempts have been carried out in order to obtain those values, as in [1] and [2]. Here, a bootstrap-based method yielding class probabilities and confidence intervals is proposed for a novel version of the SVM, namely, the cost-sensitive SVM in which misclassification costs are considered by incorporating performance constraints in the problem formulation. This is important in many contexts as credit scoring and fraud detection where misclassification costs may be different in different classes. In particular, our target is to seek the hyperplane with maximal margin yielding misclassification rates below given threshold values.

Keywords: Feature Selection; Mixed Integer Quadratic Programming; Probabilistic Classification; Sensitivity/Specificity trade-off; Support Vector Machines.

Acknowledgements This research is financed by projects EC H2020 MSCA RISE NeEDS Project (Grant agreement ID: 822214), FQM329 and P11-FQM-7603 (Junta de Andalucía, Andalucía) and MTM2015-65915-R (Ministerio de Economía y Competitividad, Spain). The last three are cofunded with EU ERD Funds. The authors are thankful for such support.

References

[1] Sollich, P. (2002). Bayesian methods for support vector machines: Evidence and predictive class probabilities. *Machine learning*, **46**, pp. 21–52.

[2] Platt, J. and others (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, **10**, pp. 61–74.

[*] Instituto de Matemáticas de la Universidad de Sevilla (IMUS), Departamento de Estadística e Investigación Operativa, Universidad de Sevilla, Sevilla, España. Email: sbenitez1@us.es

[†] Instituto de Matemáticas de la Universidad de Sevilla (IMUS), Departamento de Estadística e Investigación Operativa, Universidad de Sevilla, Sevilla, España. Email: rblanquero@us.es

[‡] Instituto de Matemáticas de la Universidad de Sevilla (IMUS), Departamento de Estadística e Investigación Operativa, Universidad de Sevilla, Sevilla, España. Email: ecarrizosa@us.es

[§] Departamento de Estadística e Investigación Operativa, Universidad de Cádiz, Cádiz, España e Instituto de Matemáticas de la Universidad de Sevilla (IMUS), Universidad de Sevilla, Sevilla, España. Email: pepa.ramirez@uca.es

A variable selection approach for high-dimensional Cox regression

Juan C. Laria [*] M. Carmen Aguilera-Morillo [†] Rosa E. Lillo [‡§]

Juan C. Laria (PhD Student) Juan C. Laria is a PhD student in Mathematical Engineering at UC3M. Graduated in Mathematics from the University of Havana in 2015. Master in Mathematical Engineering, in the field of Statistics, in 2017 by UC3M. He has worked on topics of Reliability, Stochastic Comparisons, and Computerized Adaptive Tests. His research currently focuses on variable selection algorithms in high dimensional regression, applied to genetic models, unsupervised learning for transcriptome clustering, and survival models.

In the last decade, numerous studies have shown that including sparsity restrictions in the solution contributes to better results and interpretability in high-dimensional regression problems. The l_1 penalty (lasso)[1] has become very popular in recent years, being used not only in linear regression problems, but also in others such as logistic regression, Cox regression, or even to train weights in neural networks.

For problems in which covariates are grouped and sparse structures are desired, both at group and within group levels, the sparse-group lasso (SGL) regularization method [3] has proved to be very efficient.

The solution obtained with this method depends on several hyper-parameters that control the penalization on the coefficients. Most of the applications treat this problem as a minor issue, and the parameters are either fixed or trained merely in a minimal grid of possible values, using cross-validation. For the lasso penalty, a grid search is affordable, since there is only one hyper-parameter, but the sparse-group lasso depends on, at least, as many as groups plus one.

Recently, we introduced the iterative sparse-group lasso [2], a coordinate descent algorithm to train the hyper-parameters of the sparse-group lasso for linear and logistic regression. It has many computational advantages over traditional grid-search methodologies, obtaining solutions that are equivalent or even better than searching in a grid of 100 values in each dimension.

In this talk, we introduce the extension of this algorithm to Cox regression. The advantages of this methodology are illustrated with an application to the prediction of survival of triple-negative breast cancer patients after surgery.

Keywords: Gradient Descent; High-dimension; Optimization; Regularization; Random-search.

References

[1] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58(1)**, pp. 267–288.

[2] Laria, J.C., Aguilera-Morillo M.C., Lillo R.E. (2019). An iterative sparse-group lasso. *Journal of Computational and Graphical Statistics*, doi:10.1080/10618600.2019.1573687

[3] Simon N., Friedman J., Hastie T., Tibshirani R. (2013) A sparse-group lasso. *Journal of Computational and Graphical Statistics*, **22(2)**, pp. 231–245.

[*] Department of Statistics, University Carlos III of Madrid, Leganés, Spain. Email: juancarlos.laria@uc3m.es

[†] Department of Statistics, University Carlos III of Madrid, Leganés, Spain. Email: maguiler@est-econ.uc3m.es

[‡] UC3M-BS Institute of Financial Big Data, Madrid, Spain

[§] Department of Statistics, University Carlos III of Madrid, Getafe, Spain. Email: lillo@est-econ.uc3m.es

Optimal randomized decision trees

Rafael Blanquero [*] Emilio Carrizosa [†] Cristina Molero-Río [‡] Dolores Romero Morales [§]

Cristina Molero-Río (PhD student) Cristina Molero-Río is a PhD student at the University of Seville (Spain), where she got the BSc and MSc of Mathematics. During this period, she was hired as a researcher by the following projects: *Metaheuristics for model selection in the context of portfolio optimization*, *Cost-sensitive classification. A mathematical optimization approach. COSECLA (FUNDBBVA/016/005)* and *Mathematical Optimization for Data Visualization and Decision Making (MTM2015-65915-R)*. She currently works as a substitute professor in the Department of Statistics and Operations Research at the University of Seville. She has recently started her second year of PhD, under the supervision of Prof. Rafael Blanquero (University of Seville), Prof. Emilio Carrizosa (University of Seville) and Prof. Dolores Romero Morales (Copenhagen Business School). Her PhD project is related to apply Mathematical Optimization on the area of Supervised Classification. In particular, she is focused on decision trees.

Classic decision trees [1] are defined by a set of orthogonal cuts, i.e., the branching rules are of the form variable X not lower than threshold c. The variables and thresholds are obtained by a greedy procedure. The use of a greedy strategy yields low computational cost, but may lead to myopic decisions. Although oblique cuts, with at least two variables, have also been proposed, they involve cumbersome algorithms to identify each cut of the tree. The latest advances in Optimization techniques have motivated further research on procedures to build optimal decision trees, with either orthogonal or oblique cuts. Mixed-Integer Optimization models have been recently proposed to tackle this problem, see for instance [2]. Although the results of such optimal decision trees are encouraging, the use of integer decision variables leads to hard optimization problems. In this talk, we propose to build optimal decision trees by solving nonlinear continuous optimization problems, thus avoiding the difficulties associated with integer decision variables. This is achieved by including a cumulative density function that will indicate the path to be followed inside the tree. Numerical results show the usefulness of this approach: using one single tree, we obtain better accuracies than classic decision trees, being much more flexible that those since sparsity or preference of performance in a subsample can be easily controlled.

Keywords: Classification and Regression Trees; Nonlinear Programming; Cost-sensitive Classification; Sparsity

Acknowledgements Work supported by research projects EC H2020 MSCA RISE NeEDS (Grant agreement ID: 822214), COSECLA - Fundación BBVA, MTM2015-65915R, Spain, P11-FQM-7603 and FQM-329, Junta de Andalucía, the last three with EU ERF funds. This support is gratefully acknowledged.

References

[1] Breiman, L. and Friedman, J. and Stone, C. J. and Olshen, R. A. (1984). *Classification and regression trees.* CRC press, New York 1984.

[2] Bertsimas, D. and Dunn, J. (2017). Optimal classification trees. *Machine Learning*, **106(7)**, pp. 1039–1082.

[*] Department of Statistics and Operations Research - IMUS, University of Seville, Spain. Email: rblanquero@us.es
[†] Department of Statistics and Operations Research - IMUS, University of Seville, Spain. Email: ecarrizosa@us.es
[‡] Department of Statistics and Operations Research - IMUS, University of Seville, Spain. Email: mmolero@us.es
[§] Department of Economics, Copenhagen Business School, Denmark. Email: drm.eco@cbs.dk

Directional differentiability for supremum-type functionals: Statistical applications

Javier Cárcamo [*] Antonio Cuevas [†] Luis-Alberto Rodríguez [‡]

L.-A. Rodríguez (PhD student) Luis Alberto Rodríguez Ramírez is a PhD student at Autonomous University of Madrid. His main interests are empirical process theory and functional data analysis. He received his B.S. from Autonomous University of Madrid in 2017 in mathematics. He earned his master's degree in mathematics and applications from Autonomous University of Madrid one year later. In 2018, he enrolled in the doctoral program in mathematics at Autonomous University of Madrid where he has been studying statistics under the supervision of Javier Cárcamo and Antonio Cuevas.

The supremum norm has been used in statistics in a wide variety of situations. One of the best known is the goodness-of-fit Kolmogorov-Smirnov test, where the supremum norm is utilized to measure the discrepancy between the distribution functions.

Let $\mathfrak{X}$ be a non-empty set (usually $\mathbb{R}$, $\overline{\mathbb{R}}^d$ with $\overline{\mathbb{R}} = [-\infty, \infty]$, or $\mathcal{F}$ a class of functions) and $\ell^\infty(\mathfrak{X})$ the space of bounded functions $f : \mathfrak{X} \longrightarrow \mathbb{R}$ equipped with the supremum norm $\|f\|_\infty = \sup_{x\in\mathfrak{X}} (|f(x)|)$. We denote for $f \in \ell^\infty(\mathfrak{X})$

$$\delta(f) = \sup_{x\in\mathfrak{X}} (|f(x)|) \quad \sigma(f) = \sup_{x\in\mathfrak{X}} (f(x)).$$

Suppose that we wish to estimate $\varphi(q)$ for $q \in \ell^\infty(\mathfrak{X})$ and $\varphi \in \{\delta, \sigma\}$. If there exists $\{\mathcal{Q}_n\}_{n\in\mathbb{N}} \subset \ell^\infty(\mathfrak{X})$ and $\{r_n\}_{n\in\mathbb{N}}$ sequence of real numbers such that $r_n \longrightarrow \infty$ and

$$r_n\left(\mathcal{Q}_n - q\right) \rightsquigarrow \mathcal{Q}$$

as $n \longrightarrow \infty$, where $\rightsquigarrow$ denotes weak convergence and $\mathcal{Q}$ is a tight element of $\ell^\infty(\mathfrak{X})$, it seems natural to use $\varphi\left(\mathcal{Q}_n\right)$ to estimate $\varphi(q)$. In other words, if q can be estimated in $\ell^\infty(\mathfrak{X})$ by $\mathcal{Q}_n$, it is reasonable to use the plug-in estimator to approximate $\varphi(q)$. The aim of this work is dealing with the asymptotics of

$$D_n(\varphi) = r_n\left(\varphi\left(\mathcal{Q}_n\right) - \varphi(q)\right),$$

with $\varphi \in \{\delta, \sigma\}$.

To the best of our kowledge, the first remarkable result in this direction was obtained by [2]. The proofs provided in [2] are essentially based on a careful analysis of the behaviour of the empirical process. However, we explore an alternative and more general approach: the Functional Delta Method. In this work we are going to analyze the Hadamard directional differentiability, which in some sense is the most general notion of differentiability in order to apply the Functional Delta Method (see [3]). As particular examples of this framework we show an extension of the results in [2] and the solution of an open question about Berk-Jones statistic (see [1, Question 2, p.329]), among others.

Keywords: Berk-Jones statistic; empirical process; Functional Delta Method; Kolmogorov-Smirnov distance.

Acknowledgements Research supported by the Spanish MCyT grant MTM2016-78751-P.

References

[1] Jager, L., & Wellner, J. A. (2004). On the "Poisson boundaries" of the family of weighted Kolmogorov statistics. In *A festschrift for Herman Rubin* (pp. 319-331). Institute of Mathematical Statistics.

[2] Raghavachari, M. (1973). Limiting distributions of Kolmogorov-Smirnov type statistics under the alternative. *The Annals of Statistics*, *1*(1), 67-73.

[3] Shapiro, A. (1991). Asymptotic analysis of stochastic programs. *Annals of Operations Research*, *30*(1), 169-186.

[*] Department of Mathematics, Autonomous University of Madrid, Madrid (SPAIN). Email: javier.carcamo@uam.es

[†] Department of Mathematics, Autonomous University of Madrid, Madrid (SPAIN). Email: antonio.cuevas@uam.es

[‡] Department of Mathematics, Autonomous University of Madrid, Madrid (SPAIN). Email: luisalberto.rodriguez@uam.es