SYSØR 2017

Proceedings of

1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

NOVEMBER 13-15, 2017

IEMATH-GRANADA

(UNIVERSITY OF GRANADA)

With support from

EURO THE ASSOCIATION OF EUROPEAN OPERATIONAL RESEARCH SOCIETIES

SEIO

IEMATH-GR Instituto de Matemáticas de la Universidad de Granada

UNIVERSIDAD DE GRANADA

1ST SPANISH YOUNG STATISTICIANS
AND OPERATIONAL RESEARCHERS MEETING

Proceedings of the
1st Spanish Young Statisticians and Operational Researchers Meeting

November 13th-15th, 2017 – IEMATH-GR (Granada, Spain)

Editors: Javier Álvarez Liébana and Víctor Blanco

congresos.ugr.es/sysorm17
sysorm17@ugr.es

List of Talks

Plenary speakers

(Monday 9:00) **Juan A. Cuesta Albertos** (University of Cantabria):

Big Data world: big consensus in estimation using parallelized inference

(Monday 15:00) **Pierre Bonami** (IBM):

Two recent works in Mixed Integer Quadratic Optimization

(Tuesday 10:55) **Ingrid Van Keilegom** (University of Leuven):

Estimation of the boundary of a variable observed with symmetric error

(Wednesday 11:45) **Rafael Martí** (University of Valencia):

Adaptive Memory Programming

SYSOR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Multi-Stage Investment Models for Renewable Capacity Expansion in Power Systems Under Uncertainty

R. Domínguez Martín [1] , M. Carrión Ruiz Peinado [2]

R. Domínguez Martín (Assistant Professor).- Ruth Domínguez Martín received the Ingeniero Industrial degree and the PhD degree from University of Castilla-La Mancha, Ciudad Real (Spain) in 2010 and 2015, respectively. She is currently workin as an Assistant Professor at University of Castilla-La Mancha, Toledo (Spain). Her research interests are planning, operations, and economics of power systems.

Renewable generating capacity needs to be significantly increased in power systems if the effects of global warming aim to be mitigated. Moreover, due to the high uncertainty involved in long-term planning exercises, investment decisions are usually made in several stages as uncertainty unfolds over time. Stochastic programming was proved to be a good mathematical tool to represent the uncertainty in the decision-making process [1]. However, multi-stage stochastic-programming models usually lead to computational intractability. In this work, we propose a multi-stage investment model in renewable generating capacity in which the uncertainty related to the demand growth and the investment costs of generating units is considered, as well as the variability of the renewable production throughout the year. The proposed model is solved under four different approaches: first, a multi-stage stochastic-programming problem; second, a linear decision rule (LDR) approach [2]; third, a two-stage stochastic-programming problem solved under a rolling window procedure; and fourth, the results obtained from these models are compared with the deterministic equivalent one. We assume that the uncertain parameters, namely the investment costs and the demand growth, take values within confidence intervals whose size is determined through an uncertainty level. To make all models comparable, the scenarios used to define the values of the uncertain parameters in the stochastic programming model are determined within those confident intervals.

Numerical results are provided through a case study based on the IEEE 24-node Reliability Test System (RTS) [3] and using realistic data obtained from the system operator of Texas [4]. In this case study we consider the possibility of building renewable capacity in the power system in order to counteract the thermal capacity decommissioned during the planning horizon. The capacity expansion exercise is studied through different planning situations, i.e. it is analyzed the influence of the number of decision stages over the results. Additionally, the results obtained from each optimization approach are compared. Finally, an out-of-sample analysis is carried out in order to compare the performance of the MS and the LDR models.

The main results obtained from this work are twofold: first, the LDR approach allows to solve the multi-stage capacity expansion problem in a reasonable time, while the outcomes obtained from this approach are comparable to those of the stochastic programming approach; second, considering a large number of decision stages allows a better representation of the decision process and the uncertain parameters, which results in comparatively lower total costs.

Key words: Capacity expansion; Linear decision rules; Multi-stage planning; Renewable units; Stochastic programming.

Acknowledgements R. Domínguez and M. Carrión are partly supported by Ministry of Economy and Competitiveness of Spain through CICYT project DPI2015-71280-R.

References

[1] Birge J. R., Louveaux F. (1997). *Introduction to stochastic programming*. Springer-Verlag, New York.

[2] Kuhn D., Wiesemann W., Georghiou A. (2011). Primal and dual linear decision rules in stochastic and robust optimization. *Math Program*, **130**, pp. 199–209.

[3] Grigg, C. (1999). Reliability Test System Task Force. The IEEE reliability test system. *IEEE Trans Power Syst*, **14**, pp. 1010–1020.

[4] The Electric Reliability Council of Texas (ERCOT) (2016). `http://www.ercot.com`.

[1]Department of Electrical Engineering, University Castilla – La Mancha, Toledo, Spain. Email: ruth.dominguez@uclm.es
[2]Department of Electrical Engineering, University Castilla – La Mancha, Toledo, Spain. Email: miguel.carrion@uclm.es

SYSØR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

A unified stochastic modelling framework for the spread of nosocomial infections

M. López-García [1]

M. López-García (Assistant Professor).- Martín López García finished his graduate studies in Mathematics at the University of Alicante, in 2009. From 2009 to 2013, he held a FPI research fellowship from the Spanish Government, at the Complutense University of Madrid. He was awarded, on July 2013, a PhD degree in Mathematics, with the first-class qualification of Pass Cum Laude, under the supervision of Antonio Gómez-Corral. Until September 2016, he was a Postdoctoral Research Fellow in the Mathematical Biology and Medicine Group at the University of Leeds, as part of the project *Vascular Receptor-Ligand Programming: Stochastic Modeling of Cellular Fate*. Nowadays, he is a Lecturer in the School of Mathematics at the University of Leeds, and an MRC fellow, being the PI for the MRC funded project *Mathematical modelling of the emergence and spread of antibiotic resistant bacteria in healthcare settings: a stochastic approach*. In this project, his aim is to develop new stochastic models regarding the spread of bacteria in hospital settings.

The risk of acquiring nosocomial (*i.e.*, *hospital-acquired*) infections is a recognized problem in health care facilities worldwide. The emergence of antibiotic resistance among these pathogens has posed a second major problem, stressing the need for understanding their transmission routes in health care facilities [1]. Infection control strategies usually implemented in hospital settings include, among others, hand disinfection procedures, environmental cleaning and isolation of colonized individuals [2]. In recent years, mathematical modelling has proven to be a robust tool for understanding the role played by different factors on the emergence and spread of these pathogens in health care facilities, while measuring the impact of these strategies [1].

In this work, we propose a versatile unified stochastic modelling framework, in terms of a multi-compartment SIS epidemic model with detection, that allows one to analyse the spread dynamics of a nosocomial outbreak while accounting for spontaneous colonization of patients, patient-to-staff and staff-to-patient contamination/colonization, environmental contamination, patient cohorting, room configuration of the hospital ward, staff hand-washing compliance levels, the presence of different types of HCWs or specific staff-patient contact network structures. Moreover, we show how this unified modelling framework comprehend, as particular cases, many of the existing models in the field.

In particular, for individuals in a given hospital ward split into M sub-groups (*e.g.,* colonized/non-colonized patients, contaminated/non-contaminated health care workers (HCWs),...), we consider a multi-compartment SIS epidemic model in terms of the Markovian process $\mathcal{X} = \{(I_1(t), \ldots, I_M(t)) : \ t \geq 0\}$, where $I_j(t)$ represents the number of *infective* individuals at time $t \geq 0$ at group j. Our modelling approach allows for the exact and analytical study of the reproduction number of each agent at the hospital ward during the nosocomial outbreak. The reproduction number $R^{(i)}$ of an individual at sub-group i (*e.g.,* the reproduction number of a colonized patient in the ward, or a contaminated HCW), is defined as the number of *infections* caused by this individual during his/her *infectious period* . This is studied in terms of its probability distribution

$$\mathbb{P}(R^{(i)} = n), \quad n \geq 0$$

which we compute by means of first-step arguments, and by solving significantly large systems of linear equations in an efficient matrix-oriented way.

This unified approach also allows one to study the different infection transmission routes playing a significant role during a nosocomial outbreak. To this aim, we focus on the reproduction number $R^{(i)}(j)$ of an individual at sub-group i among individuals at sub-group j, $1 \leq j, i \leq M$, which can be analysed in a similar way. For example, if value i refers to the group of colonized patients, while j amounts to non-contaminated HCWs, quantity $R^{(i)}(j)$ is an infectiousness measure of colonized patients among HCWs. Finally, our numerical results highlight the importance of maintaining high hand-hygiene compliance levels by HCWs, support control strategies involving to increase environmental cleaning during nosocomial outbreaks, and show the potential of some HCWs to act as *super-spreaders* during these outbreaks.

Key words: Nosocomial infections; continuous-time Markov chains; summary statistics; infection control.

Acknowledgements This research is funded by the Medical Research Council (United Kingdom) through a *Skills Development Fellowship* (MR/N014855/1)

References

[1] van Kleef, E., Robotham, J- V., Jit, M., Deeny, S. R., Edmunds, W. J. (2013). Modelling the transmission of healthcare associated infections: a systematic review. *BMC Infect. Dis.*, **13**, pp. 294.

[2] Bonten, M. J. M. (2002). Infection in the intensive care unit: prevention strategies. *Curr. Opin. Infect. Dis.*, **15**, pp. 401–405.

[1] Dpt. Applied Mathematics, School of Mathematics, University of Leeds, Leeds LS2 9JT, UK. Email: m.lopezgarcia@leeds.ac.uk

SYSOR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

A new lifting theorem for vertex packing problems

M. Pelegrín [1] , A. Marín[1]

A *vertex packing* in an undirected graph $G = (V, E)$ is a subset of nodes $P \subseteq V$ such that every pair $u, v \in P$ satisfies $(u, v) \notin E$. The set of all vertex packings of G, usually denoted by $\mathcal{P}_G$, can be identified with the feasible region of the following integer program

$$\text{(VP)} \quad \max \Big\{ \sum_{v \in V} x_v \ : \ x_u + x_v \leq 1 \quad \forall (u, v) \in E,$$

$$x_v \in \{0, 1\} \quad \forall v \in V \Big\},$$

where $x_v = 1$ iff v is in the vertex packing. We call *integer polytope* the convex hull of the feasible points of (VP), i.e.:

$$\mathcal{B}_G = \text{conv}\{x \in \{0, 1\}^n \ : \ x_u + x_v \leq 1 \quad \forall (u, v) \in E\}.$$

In this work, we give a general procedure to obtain *facets* of $\mathcal{B}_G$, under certain conditions. An inequality $\pi x \leq \pi_0$ is a facet of a general polytope $\mathcal{P}$ of dimension n if:

1. (validity) every $x \in \mathcal{P}$ satisfies $\pi x \leq \pi_0$ and
2. (maximality) there exist n affinely independent vectors $x^i \in \mathcal{P}$ satisfying $\pi x^i = \pi_0$, $i = 1, \ldots, n$.

That is, a facet is a face of a polytope that has maximal dimension. We present a theorem to transform a facet of $\mathcal{B}_G$ into a facet of $\mathcal{B}_{G^*}$ when G is a subgraph of G^* and both G and G^* have a particular structure. Such transformation is usually known as *lifting*, since the initial facet is "lifted" to a facet of higher dimension. Finally, we introduce some new *facet-defining graphs* obtained with our theorem, that is, graphs that have associated facets with all coefficients strictly positive. Facet-defining graphs provide us with valid inequalities of $\mathcal{B}_G$ when they are present as subgraphs of G (which could also be lifted to obtain facets of $\mathcal{B}_G$). Some known facet-defining graphs are *cliques* and odd *holes* [1], *webs* [2], *wheels* [3] and *hanks*, *fans*, *grilles* and *ranges* [4, 5].

Key words: vertex packing; facet; facet-defining graph; lifting.

Acknowledgements The authors acknowledge that research reported here was supported by Spanish *Ministerio de Economía y Competitividad*, project MTM2015-65915-R, *Ministerio de Educación, Cultura y Deporte*, PhD grant FPU15/05883, *Fundación Séneca de la Consejería de Educación de la Comunidad Autónoma de la Región de Murcia*, project 19320/PI/14 and *Fundación BBVA*, project "Cost-sensitive classification. A mathematical optimization approach" (COSECLA).

References

[1] Padberg, M. W. (1973). On the facial structure of set packing polyhedra. *Math. Program.*, **5**, pp. 199–215.

[2] Trotter, L. E. (1975). A class of facet-producing graphs for vertex packing polyhedra. *Discret. Math.*, **12**, pp. 373–388.

[3] Cheng, E., Cunningham, W. H. (1997). Wheel inequalities for stable set polytopes. *Math. Program.*, **77**, pp. 389–421.

[4] Cánovas, L., Landete, M., Marín, A. (2000). New facets for the set packing polytope. *Oper. Res. Lett.*, **27**, pp. 153–161.

[5] Landete, M. (2001). Obtención de facetas de poliedros asociados a problemas de empaquetamiento. *PhD Thesis*, University of Murcia.

[1]Department of Statistics and O.R., University of Murcia, Murcia, Spain. Email: mariamercedes.pelegrin@um.es, amarin@um.es

SYSOR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Branch&Price approaches for the Discrete Ordered Median Problem

S. Deleplanque[1], M. Labbé[2] , D. Ponce [3] [4], J. Puerto[5] [6]

D. Ponce (Postdoctoral Researcher).- Diego Ponce studied Industrial Chemical Engineering and Management Engineering at Politécnica University of Cartagena. After a short period in which he took some Operational Research courses at University of Murcia, he went on to obtain a MSc in Advanced Mathematics at University of Sevilla. Then, he wrote his PhD thesis entitled *The DOMP revisited: new formulations, properties and algorithms*, under the joint supervision of Justo Puerto (University of Sevilla) and Martine Labbé (Université Libre of Bruxelles). The joint supervision led to my obtaining the degrees of Doctor in Mathematics from the University of Sevilla and Doctor in Sciences from the Université Libre of Bruxelles. Currently, he is working as a postdoctoral researcher at University of Sevilla, collaborating with University of Brussels. He is interested in extending his results to other areas of research such as finance, data analysis or industrial problems. His research interests has recently taken him into the field of Data Analysis (regression, clustering, prediction, cluster-wise regression, segmented regression, etc). Additionally, he enjoys doing research in a broad range of combinatorial optimization problems. In particular, he is interested in the fields of Location, Logistics and Routing.

The Discrete Ordered Median Problem (DOMP) is a modeling tool that provides flexible representations of a large variety of problems, which include most of the classical discrete location problems.

For the sake of comprehension, let us give a brief explanation of DOMP (see [1] for more information). Let C be the cost matrix, c_{ij} is the cost to serve the demand point i with facility j. We have n demand points, which at the same time are the candidate sites to be facilities. Let I be the set of demand points and let J be the set of servers, such that $J \subseteq I$, the DOMP is characterized by a set of constraints: the problem must have exactly p facilities, where $1 \leq p \leq n$ and every client must be served.

In order to define the objective function in DOMP, we need to calculate the vector c. This vector has n components, which are the sorted cost for each location, i.e. $c_i(J) = \min_{j \in J} c_{ij}$. By means of a permutation we obtain the vector $c_{\leq}$, such that it satisfies $c^1_{\leq}(J) \leq \cdots \leq c^n_{\leq}(J)$.

Then, we have to solve next problem

$$\min_J \sum_{k=1}^{n} \lambda^k c^k_{\leq}(J),$$

where λ is a n-vector with $\lambda^k \geq 0 \, \forall k$.

In this talk we present new formulations for the DOMP and a novel embedding based on set packing constraints ([2]) on a column generation approach built over a set formulation which leads to a Branch & Price method. These new formulations, allow us to gain some insights on the polyhedral structure of the DOMP that advance the knowledge of this problem over what was known previously.

Regarding the column generation approach, DOMP is formulated as a set partitioning problem using an exponential number of variables. Each variable corresponds to a set of demand points allocated to the same facility with the information of the sorting position of their corresponding costs. We develop a column generation approach to solve the continuous relaxation of this model. Then, we apply a branch-cut-and-price algorithm to solve to optimality small to moderate size of DOMP in competitive computational time.

Key words: Location problems; Discrete Ordered Median Problem; Combinatorial Optimization; Column Generation; Branch-and-cut-and-price.

Acknowledgements This work has been partially supported by the project FQM-5849 (Junta de Andalucía \FEDER) and the Interuniversity Attraction Poles Programme P7/36 COMEX initiated by the Belgian Science Policy Office.

References

[1] Nickel, S., Puerto, J. (2005) *Location Theory: A Unified Approach*. Springer-Verlag, Berlin Heidelberg 2005.

[2] Labbé, M., Ponce, D., Puerto, J. (2017). A Comparative Study of Formulations and Solution Methods for the Discrete Ordered p-Median Problem. *Computers & Operations Research*, **78**, pp. 230–242.

[1]Départment d'Informatique, Université Libre de Bruxelles, Brussels, Belgium. Email: sdelepla@ulb.ac.be
[2]Départment d'Informatique, Université Libre de Bruxelles, Brussels, Belgium. Email: mlabbe@ulb.ac.be
[3]Departamento of Estadística e Investigación Operativa, Universidad de Sevilla, Seville, Spain. Email: dponce@us.es
[4]IMUS, Universidad de Sevilla, Seville, Spain. Email: dponce@us.es
[5]Departamento of Estadística e Investigación Operativa, Universidad de Sevilla, Seville, Spain. Email: puerto@us.es
[6]IMUS, Universidad de Sevilla, Seville, Spain. Email:puerto@us.es

SYSOR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Optimization of gas transmission networks: a two-step sequential linear programming algorithm for NLP and MINLP problems

Ángel M. González-Rueda [1] , Julio González-Díaz [2], María P. Fernández de Córdoba [3]

A. M. González-Rueda (PhD Student).- Ángel Manuel González Rueda was born in 1988 in Boiro (Spain). He received his BSc in Mathematics from University of Santiago de Compostela in 2011, and his MSc in Engineering Mathematics from Complutense University of Madrid in 2012. He is a PhD candidate in Statistics and Operation Research from the University of Santiago de Compostela and he is funded by a FPU Grant (University Professors Formation). His research is focused on cooperative game theory and operations research. He collaborates on projects for the transfer of mathematics to the industry. Nowadays, he participates in a project with a company belonging to the Spanish gas transport network, and his doctoral dissertation is focused on problems related to this theme. He belongs to SaGaTh (Santiago Game Theory research group).

In this talk we present a mathematical programming model for the optimal management of gas transmission networks. We propose a two-step sequential linear programming algorithm for solving it that leads to a general heuristic approach for mixed-integer nonlinear optimization problems. Sequential linear programming (SLP) is a widely used approach to deal with complex nonlinear optimization problems. Informally, it consists of an iterative procedure which, at each step, works with a linearization of the problem around the current solution. This solution is then updated by moving, withing a given trust region, in the direction of the optimum of the linear programming (LP) problem at hand. The first stage of the algorithm we propose builds upon a natural modification of the above idea: no trust region is considered during the iterative process. At each iteration we move to the optimum of the LP problem at hand and use it for the new linearization. We study to what extent this modification preserves the theoretical properties of the classic SLP. Importantly, this modification can be immediately applied to MINLP problems. In the second step, all the integer variables are fixed according to the solution obtained in the first stage and the penalty successive linear programming algorithm ([2], [1, Section 10.3]) is applied. This heuristic has already been applied in a real-life optimization problem on gas transmission networks and has also been tested on multicommodity flow problems obtaining satisfactory results.

Key words: Gas transmission networks; Sequential Linear Programming; Mixed-Integer Nonlinear Programming.

Acknowledgements The authors acknowledge support from the Spanish Ministry for Economics and Competitivity, the Spanish Agencia Estatal de Investigación and FEDER through projects MTM2011-27731-C03 and MTM2014-60191-JIN. Support from Xunta de Galicia through projects INCITE09-207-064-PR and EM 2012/111 is also acknowledged. We also thank support from the Technological Institute for Industrial Mathematics (ITMATI). Ángel M. González-Rueda acknowledges support from the Spanish Ministry of Education through Grant FPU13/01130.

References

[1] Bazaraa, M. S., Sherali, H. D., Shetty, C. M. (2006). *Nonlinear programming: theory and algorithms*. John Wiley and Sons, New Jersey.

[2] Zhang, J. Z., Kim, N. H., Lasdon, N. H. (1985). An improved successive linear programming algorithm. *Management Sci.*, **31**, pp. 1312–1331.

[1]Department of Statistics, Mathematical Analysis and Optimization, University of Santiago de Compostela, Spain. Email: angelmanuel.gonzalez@usc.es
[2]Department of Statistics, Mathematical Analysis and Optimization, University of Santiago de Compostela, Spain. Email: julio.gonzalez@usc.es
[3]Xunta de Galicia, Spain. Email: mpfernandezdecordoba@edu.xunta.es

SYSØR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

RKHS, Mahalanobis and Variable Selection in Functional Data Classification

J.R. Berrendero[1], A. Cuevas[2], J.L. Torrecilla[3]

J. L. Torrecilla (Postdoctoral Researcher).- José Luis Torrecilla-Noguerales is a Postdoctoral Researcher at the Institute UC3M-BS of Financial Big Data. His areas of research are functional and functional and high dimensional data analysis, classification, dimension reduction and variable selection methods, measures of dependence and machine learning. His present research work includes functional variable selection, optimal classification rules and analysis of functional datasets in biomedical settings. José Luis obtained his BSc in Mathematics and Computer Science and his MSc in *Mathematics and Applications and Computer Science: Artificial Intelligence* at Autónoma University of Madrid (UAM). From 2010 to 2015, he developed his PhD Thesis entitled *On the Theory and Practice of Variable Selection for Functional Data*, under the supervision of José Ramón Berrendero and Antonio Cuevas, at the Department of Mathematics (UAM), where he was granted with post-graduate scholarships, including FPI and three months as visiting researcher at UC-Davis.
Torrecilla's research experience also includes different positions at the Institute of Knowledge Engineering (2007-2010) and a postdoctoral research position at the Machine Learning Group at the Department of Computer Science (EPS-UAM, 2016), as well as the membership to the Functional Data Analysis working group (SEIO).

Working with functional data entails several difficulties including the lack of a *natural* density, the infinite dimension and the collinearity among close points. The RKHS (Reproducing Kernel Hilbert Space) theory provides some tools in order to deal with the first problem, and an appropriate dimension reduction methodology can solve the last two.

Variable selection techniques have become a popular alternative for dimension reduction with an easy interpretation. However, we are still far from getting a standard in the functional classification framework. Here we propose a new functional-motivated variable selection methodology (RK-VS) when our final goal is classifying. This method appears as a direct consequence of looking at the functional classification problem from an RKHS point of view. In this context, under a general Gaussian model and a sparsity assumption, the optimal rules turn out to depend on a finite number of variables. These variables can be selected by maximizing the Mahalanobis distance between the finite-dimensional projections of the class means. Our RK-VS method is an iterative approximation to this. This is an easy-to-interpret and fast methodology which allows for easily adding extra information about the model. The empirical performance of RK-VS is extremely good when the considered problems fit the assumed model but it turns out to be also quite robust against partial departures from the hypotheses, typically leading to very good results in general problems.

Key words: Functional Data Analysis; Mahalanobis distance; Reproducing Kernel Hilbert Space; Supervised Classification; Variable Selection.

[1]Department of Mathematics, Universidad Autónoma de Madrid, Madrid, Spain. Email: joser.berrendero@uam.es
[2]Department of Mathematics, Universidad Autónoma de Madrid, Madrid, Spain. Email: antonio.cuevas@uam.es
[3]Institute UC3M-BS of Financial Big Data, Universidad Carlos III de Madrid, Getafe, Spain. Email: joseluis.torrecilla@uc3m.es

On sparse and constrained Naïve Bayes

R. Blanquero Bravo [1] , E. Carrizosa Priego [2] , P. Ramírez Cobo [3] , M. R. Sillero Denamiel [4]

M. R. Sillero Denamiel (PhD Student).- María Remedios Sillero Denamiel is currently working as a PhD student at University of Sevilla (Spain), where she got BSc and MSc of Mathematics. Before starting her PhD, M. Remedios 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 second year of her PhD, under the supervision of Prof. Emilio Carrizosa (University of Sevilla), Prof. Rafael Blanquero (University of Sevilla) 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 on large medical datasets."

Naïve Bayes is a tractable and remarkably efficient approach to classification learning. However, as it is common in real clasification contexts, datasets are often characterized by a large number of features and, in addition, there could exist an imbalance between the correct classification rates of different classes. On the one hand, it may complicate the interpretation of the results as well as slow down the method's execution. On the other hand, classes are often not equally important and making a misclassification in one of them leads to undesirable consequences that can be avoided by controlling the correct classification rate in that particular class. In this work we propose a sparse and constrained version of the Naïve Bayes in which a variable reduction approach, that takes into account the dependencies among features, is embedded into the classification algorithm. Moreover, a number of constraints over the performance measures of interest are embedded into the optimization problem which estimates the involved parameters. Unlike typical approaches in the literature modifying standard classification methods, our strategy allows the user to control simultaneously the different performance measures that are considered. Our findings show that, under a reasonable computational cost, the number of variables is significantly reduced obtaining competitive estimates of the performance measures. Furthermore, the achievement in the different individual performance measures under consideration is controled.

Key words: Conditional independence; Dependence measures; Variable Selection; Heuristics; Probabilistic Classification; Constrained optimization; Efficiency measures.

Acknowledgements Research partially supported by research grants and projects MTM2015-65915-R and ECO2015-66593-P (Ministerio de Economía y Competitividad, Spain), P11-FQM-7603, FQM-329 (Junta de Andalucía, Spain) and Fundación BBVA.

[1]Department of Statistics and O.R., University of Sevilla, Sevilla, Spain. Email: rblanquero@us.es
[2]Department of Statistics and O.R., University of Sevilla, Sevilla, Spain. Email: ecarrizosa@us.es
[3]Department of Statistics and O.R., University of Cadiz, Cadiz, Spain. Email: pepa.ramirez@uca.es
[4]Department of Statistics and O.R., University of Sevilla, Sevilla, Spain. Email: rsillero@us.es

SYSOR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

On the choice of the tuning parameter for the Tukey's biweight loss function in the context of fuzzy M-estimators of location

B. Sinova [1] , S. Van Aelst [2]

B. Sinova (Assistant Professor).- Beatriz Sinova graduated in Mathematics, from the University of Oviedo, in 2009, achieving both the end of degree award and special award for degree. She enjoyed an FPU grant from the Spanish Ministry of Education to do her PhD, which was supervised by María Ángeles Gil, Gil González Rodríguez and Stefan Van Aelst. In 2014 she obtained the joint PhD from the universities of Oviedo and Ghent (Belgium) and she received the doctorate special award in Sciences from the University of Oviedo. Since 2015, she is an Assistant Professor at the University of Oviedo. Her work has combined techniques, concepts and developments of Robust Statistics with the usual methodology for the statistical analysis of imprecise-valued data. She has focused on the location and, in particular, she has studied the M-estimation problem in the imprecise-valued settings. This work has received the prizes *Premio Ramiro Melendreras a jóvenes investigadores* (2016) from the Spanish Society of Statistics and Operation Research and *Premio Vicent Caselles* (2017) from the Spanish Royal Mathematical Society and FBBVA.

Fuzzy number-valued data are useful to mathematically model many real-life experiments characterized by an underlying imprecision, especially those related to human valuations (such as ratings, judgements, perceptions, etc.). Due to their interest, different statistical methodologies have already been adapted to deal with this kind of data. However, most of these procedures are based on the Aumann-type mean as location measure, which is highly sensitive to outliers as generalization of the mean of a real-valued random variable. Therefore, several robust location measures for random fuzzy numbers have been proposed in the literature (see e.g. [3, 1, 2]) and, among them, fuzzy M-estimators of location seem to keep their success from real-valued settings.

It has been shown in [3] that, under mild conditions on the loss function involved in their definition, fuzzy M-estimators of location exist and can be expressed as weighted means of the observations (this result will be called representer theorem). This property is crucial to guarantee that they indeed remain inside the space of fuzzy numbers, $\mathcal{F}_c(\mathbb{R})$, which entails the particularity of not being linear. Good news is that well-known families of loss functions, such as Huber's, Hampel's and Tukey's, fulfill the mild conditions required to establish the representer theorem (see [4, 3]). The robustness of the three alternatives has been proven by means of the computation of their finite sample breakdown point, which is a measure of the impact of global contamination on the corresponding estimates.

The empirical comparison of the finite-sample behavior of the fuzzy M-estimators based on Huber's, Hampel's and Tukey's loss functions is not complete yet. In [3] it is shown that, with usual choices for the tuning parameters, the Hampel loss function makes the corresponding fuzzy M-estimator more accurate than the Huber loss function in many situations. Naturally, other choices for the tuning parameters could be explored, but, as Hampel's loss function also allows more flexibility due to its three tuning parameters, we will focus on the comparison of fuzzy M-estimators based on Hampel's and Tukey's loss functions. Could the latter improve the performance of the Hampel fuzzy M-estimator? The aim of this paper is to propose a choice of the tuning parameter involved in Tukey's biweight loss function in order to get a Tukey M-estimator providing a better estimate than the Hampel M-estimator.

Key words: M-estimator of location; Random fuzzy number; Robustness; Tukey biweight loss function; Tuning parameter.

Acknowledgements This research has been partially supported by the Spanish Ministry of Economy and Competitiveness through the Grant MTM2013-44212-P, the Principality of Asturias/FEDER Grant GRUPIN14-101, Grant C16/15/068 of International Funds KU Leuven and IAP research network grant nr. P7/06 of the Belgian government. Their support is gratefully acknowledged.

References

[1] Colubi, A., González-Rodríguez, G. (2015). Fuzziness in data analysis: Towards accuracy and robustness. *Fuzzy Sets Syst.*, **281**, pp. 260–271.

[2] Sinova, B., Gil, M. A., Colubi, A., Van Aelst, S. (2012). The median of a random fuzzy number. The 1-norm distance approach. *Fuzzy Sets Syst.*, **200**, pp. 99–115.

[3] Sinova, B., Gil, M. A., Van Aelst, S. (2016). M-estimates of location for the robust central tendency of fuzzy data. *IEEE Trans. Fuzzy Syst.*, **24**, pp. 945-956.

[4] Sinova, B., Van Aelst, S. (2017). Tukey's biweight loss function for fuzzy set-valued M-estimators of location. In: Ferraro, M.B., Giordani, P., Vantaggi, B., Gagolewski, M., Gil, M. A., Grzegorzewski, P., Hryniewicz, O. Soft Methods for Data Science. *Advances in Intelligent Systems and Computing*, **456**, pp. 447–454.

[1]Department of Statistics and O.R. and D.M., University of Oviedo, Oviedo, Spain. Email: sinovabeatriz@uniovi.es
[2]Department of Mathematics, KU Leuven, Leuven, Belgium. Email: stefan.vanaelst@kuleuven.be

SYSØIR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Optimal experimental design for calibration in radiation biodosimetry

M. Higueras Hernáez [1] , Jesús López Fidalgo [2] , Pere Puig Casado [3]

M. Higueras Hernáez (Postdoctoral Researcher).- Manuel Higueras Hernáez got his PhD in Mathematics in 2015, entitled *Advanced Statistical Methods for Cytogenetic Radiation Biodosimetry*, at Autónoma University of Barcelona. Currently, he is postdoctoral fellow within the Applied Statistics group at Basque Center for Applied Mathematics. He has experience in statistical modelling for radiation research. More concretely, he has been working in statistical models for cytogenetic biodosimetry and radiation epidemiology. In cytogenetic biodosimetry, he has developed models for cytogenetic dose estimation, useful for predicting the derived health consequences in overexposed individuals. In the radiation epidemiology field, his task was to analyse the increased risk of leukaemia and brain tumours in paediatric patients exposed to ionising radiation by CT scans. His research interests are count data models (including overdispersion, zero-inflation and finite mixtures), Bayesian analysis, inverse regression, and excess relative risk models. He is also interested in R-project programming and Shiny RStudio apps.

Ionising radiation (IR) may be absorbed by humans, implying negative health consequences. For the clinical actions after a radiation event, absorbed dose estimates are necessary. IRs produce damage at a cellular level in form of chromosomal aberrations which are used as biomarkers of the absorbed dose. To study the effect of IR, calibration dose-response curves are built.

These curves are based on the irradiation of *in vitro* n blood samples which simulate homogeneous whole body exposures. Each sample is irradiated to a dose d_i $(i = 1 \dots n)$. After the irradiation, m_i blood cells are analysed for sample i and $y_{i,j}$ chromosomal aberrations are scored for each cell $(j = 1, \dots, m_i)$. For low LET irradiation, it is assumed that the number of chromosomal aberrations follow a Poisson distribution whose intensity is a quadratic function of the absorbed dose, *i.e.* $Y \sim \text{Pois}(\alpha_0 + \alpha_1 D + \alpha_2 D^2)$, where Y represents the number of aberrations, D is the absorbed dose and $\{\alpha_0, \alpha_1, \alpha_2\}$ is the calibration parameter set. These parameters are calculated by maximum likelihood estimation. It is important to remark that in this Poisson model the link function is the identity, instead of the usual logarithmic link. This assumption is supported by the biological process of production of IR induced chromosomal aberrations, Hall and Giaccia (2012) [1], and statistically in Oliveira *et al.* 2016 [4].

The experimental design of these curves is based on the International Atomic Energy Agency (IAEA) suggestions. The manual of the IAEA (2011) [2] states that at least 10 doses should be used in the range $(0, 5]$ Gy, plus a control sample, and at least 4 of them in $(0, 1]$ Gy. The sample size, *i.e.* the number of scored cells at each dose, should aim to detect 100 chromosomal aberrations at each dose, but for the lower doses it is suggested a number in the range $[3000, 5000]$. It is also suggested to reduce the variance of the linear term, *i.e.* α_1.

The main purpose of this work is to provide experimental designs for minimizing the variance of the dose estimation. Due this is a calibration problem, it cannot be explored by pure i-optimization, but two variants of the i-optimal criterion are defined and applied. It is also explored the c-optimization for the variance minimization of the linear dose effect parameter, α_1. As the optimal designs for the different criterions are extremal, sequences of doses and weights are optimized, analogously to López Fidalgo and Wong(2002) [3], to produce suboptimal designs for the different criterios without replicates which look similar to the IAEA directions [2]. The proposed designs are approximate and the weight for each dose represents the proportion of blood cells scored.

Key words: Retrospective biodosimetry; i-optimization; c-optimization; inverse regression.

Acknowledgements Thanks to BIOSTATNET's Young Researchers Stay Scholarship and to Ministerio de Economía y Competitividad grants SEV-2013-0323, MTM2013-47879-C2-1-P, MTM2015-69493-R.

References

[1] Hall, E. J., Giaccia, A. J. (2012) *Radiobiology for the radiologist*, 7^{th} edition. Lippincott Williams & Wilkins, Philadelphia.

[2] IAEA (2011). *Cytogenetic Dosimetry: Applications in Preparedness for and Response to Radiation Emergencies*. International Atomic Energy Agency, Vienna.

[3] López Fidalgo, J., Wong, W. K. (2002). Design for the Michaelis-Menten model. *Journal of Theoretical Biology*, **215**, pp. 1–11.

[4] Oliveira, M., Einbeck, J., Higueras, M., Ainsbury, E. A., Rothkamm, K., Puig, P. (2016). Zero-inflated regression models for radiation-induced chromosome aberration data: A comparative study. *Biometrical Journal*, **58**, 259–279.

[1]Basque Center for Applied Mathematics, Bilbao, Spain. Email: mhigueras@bcam.org
[2]Institute for Culture and Society, University of Navarre, Pamplona, Spain. Email: fidalgo@unav.es
[3]Department of Mathematics, Autonomous University of Barcelona, Barcelona, Spain. Email: ppuig@mat.uab.cat

SYSØR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Online Pickup and Delivery Problem under special constraints

S. Bsaybes [1] , A. Quilliot[2], A. Wagler[2]

S. Bsaybes (PhD Student).- Sahar Bsaybes obtained her PhD from University of Clermont Auvergne, Clermont Ferrand (France), in October 2017. She is currently working as a Lecturer in ENSIMAG, one of the engeneering schools of Grenoble INP and as a researcher in the G-SCOP laboratory in Grenoble in France. Her main research interests are Operational Research and the development of mathematical optimisation models, and intelligent decision support systems for automatically producing high-quality solutions to a wide range of real-world optimisation, especially the vehicle routing problems.

The VIPAFLEET project aims at developing a framework to manage a fleet of Individual Public Autonomous Vehicles (VIPA). We consider a fleet of cars distributed at specified stations in an industrial area to supply internal transportation, where the cars can be used in different modes of circulation (tram mode, elevator mode, taxi mode) [1]. We handled the pickup and delivery problem [2] related to the taxi mode where VIPAs run on a connected network to serve transport requests (from any start to any destination station in the network within given time windows) by means of flows in time-expanded networks [3]. This enables us to compute optimal offline solutions, to propose a replan strategy for the online situation, and to evaluate its performance in comparison with the optimal offline solution.

Key words: fleet management; offline and online pickup and delivery problem.

References

[1] Bsaybes, S., Quilliot, A., Wagler, A. (2017). Fleet management for autonomous vehicles *arXiv preprint arXiv:1609.01634*. 2016.

[2] Bsaybes, S., Quilliot, A., Wagler, A. (2017). Fleet management for autonomous vehicles: Online PDP under special constraints *arXiv preprint arXiv:1703.10565*,. 2017.

[3] Ford Jr, L. R., Fulkerson, D. R. (1958). Constructing maximal dynamic flows from static flows. *Operations research.*, **6**, pp. 419–433.

[1]Grenoble INP, GSCOP Laboratory Grenoble, France. Email: Sahar.Bsaybes@grenoble-inp.fr
[2]Université Clermont Auvergne (LIMOS UMR CNRS 6158), France. Email: alain.quilliot@uca.fr, annegret.wagler@uca.fr

SYSØR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

On the use of functional additive models for electricity demand and price prediction

P. Raña [1], J.M. Vilar [2], G. Aneiros [3]

P. Raña (Postdoctoral Researcher).- Paula Raña Míguez received the BSc degree in Mathematics, from the University of Santiago de Compostela in 2011, the MSc degree in Statistics, from the same university in 2012, and the PhD degree in Statistics, from the University of A Coruña in 2016. She has been working on several research topics, including functional data, detection of outliers, forecasting in electricity demand and price, prediction and confidence intervals with functional data and functional time series forecasting.

This study presents an application of functional additive models in the context of electricity demand and price prediction. Data from the Spanish electricity market is used to obtain pointwise predictions. Also prediction intervals and prediction densities, based on a bootstrap procedure, are computed.

Prediction in electricity markets has been extensively studied in the literature. Most of the papers studying methods of electricity demand and price prediction take information from scalar variables, but in recent years, the use of functional data has been extended in this area. Considering the daily curves of electricity demand or price as functional data, the prediction problem in electric markets can be studied taking use of functional regression methods.

The case of forecasting curves (as well as scalar values) of demand and price from functional data is studied in [1]. In that paper, nonparametric and semi-functional partial linear models are analysed within the Spanish electricity market. Non-parametric autoregressive models with functional data provide good predictions, due to its flexibility, but the results can improve by adding exogenous variables to the model. For that purpose, the use of a semi-functional partial linear model is considered, with a nonparametric autoregressive functional component and introducing other scalar covariates in a linear way.

The aim of this study is next-day forecasting of hourly values of electricity demand and price using functional additive models. These models combine flexibility and the control of the dimensionality effects. Three approaches, taking information from functional covariates (one is endogenous), are considered. In the first case, the effect of the covariates on the response is linear (functional linear model), while in the other two proposals the predictor is the sum of smoothing functions applied to the covariates. When dealing with demand prediction, it is convenient to introduce the temperature and other weather variables as covariates. In the case of price, one may consider the wind power production and the forecast demand.

An application to the electrical dataset previously used in [1] allows to compare the accuracy of those functional additive models with the nonparametric and semi-functional partial linear models, among other prediction methods.

When dealing with forecasting, it is important to consider also prediction intervals and the prediction density, which help to understand the behaviour of the forecasts in a deeper way. For that purpose, algorithms for the construction of prediction intervals and prediction density associated with the functional additive models are proposed, using residual-based bootstrap methods. The obtained results, within the electrical dataset study, are compared with the prediction intervals obtained in [2].

Key words: Load and price; functional data; functional time series forecasting; additive model; prediction intervals.

Acknowledgements This research is partly supported by MTM2014-52876-R from Spanish Ministerio de Economía y Competitividad, from the Xunta de Galicia (Centro Singular de Investigación de Galicia accreditation ED431G/01 2016-2019 and Grupos de Referencia Competitiva ED431C2016-015) and the European Union (European Regional Development Fund - ERDF).

References

[1] Aneiros, G., Vilar, J., Raña, P. (2016). Short-term forecast of daily curves of electricity demand and price. *Electr. Power Energy Syst.*, **80**, pp. 96–108.

[2] Vilar, J., Raña, P., Aneiros, G. (2017). Prediction intervals for electricity demand and price using functional data. *Submitted.*

[1]Department of Mathematics, University of A Coruña, A Coruña, Spain. Email: paula.rana@udc.es
[2]Department of Mathematics, University of A Coruña, A Coruña, Spain. Email: juan.vilar@udc.es
[3]Department of Mathematics, University of A Coruña, A Coruña, Spain. Email: ganeiros@udc.es

SYSØR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

The deployment of automated vehicles: dedicated zones as a urban planning strategy

Lígia Conceição [1] , Gonçalo Correia [2] , José Pedro Tavares [3]

L. Conceição (PhD Student).- Lígia Conceição is a 3rd year PhD student from University of Porto (Portugal), under the MIT Portugal Program focused on Transportation Systems. She focuses her research on the deployment of automated vehicles in urban centers. The methodological approach regards optimisation, moreover mixed integer problems.

In the past two decades, an increased interest has been growing towards vehicle automation which brings great potential changes on mobility in urban centres. However, vehicle automation is not yet a reality which casts huge speculation of what will really happen when implemented in the near future. On the one hand, it may enhance the current mobility system, on the other hand, it might disrupt the current transportation paradigm in a way that is still difficult to foresee.

Since the deployment of fully automated vehicles cannot be realized instantaneously in all areas of a city, a transitional phase must be assumed to mitigate the changes to come. It is critical to devise policies in order to implement such technology to leverage the benefits that it may bring [1].

In urban environments, automated vehicles (AVs) are believed to begin as speedy last mile taxis by 2020 whereas automated taxis will only become a reality by 2028. The deployment of AVs in urban environments is expected to start with segregated lanes and then with dedicated lanes by 2020 and mixed with conventional vehicles (CVs) by 2028 [2].

The literature regarding automated vehicles is scarce and disperse. Most of the existent research covers the vehicle technical features and the forthcoming impacts in interurban traffic environments [3].

In order to address that gap, we want to support city planners by developing a strategy of integration for AVs into urban networks. Moreover, at a traffic level, a strategy of dedicated zones for automated vehicles. We develop a mathematical programming model whereby the aim is to minimize the congestion problem through dedicated links where only automated vehicles can drive. A traffic assignment approach is used where the minimization of the sum of travel times is part of the objective function. The number of automated vehicles is changed in function of a penetration rate. Each scenario is simulated and compared.

This research begins the discussion of how to help public authorities plan the deployment of such automated vehicles and bring improvement to traffic in cities.

Key words: Automated Vehicles; Urban Planning; Optimization

Acknowledgements The first author would like to thank the support by the Portuguese Foundation for Science and Technology (PD/BD/113760/2015) under the MIT Portugal Program. We also thank FICO (Xpress supplier) for establishing a research agreement with the Department of Transport and Planning at TU Delft.

References

[1] Milakis, D., van Arem, B., van Wee, B. (2017). Policy and society related implications of automated driving: A review of literature and directions for future research. *J. Intell. Transp. Syst.*, pp. 1–25. doi:10.1080/15472450.2017.1291351

[2] Zlocki, A. (2014). Automated Driving. *Transp. Res. Rec.*, **2416**, pp. 64–72. doi:10.3141/2416-08

[3] Correia, G., Milakis, D., Arem, Bart van Hoogendoorn, R. (2015). Vehicle automation for improving transport system performance: conceptual analysis, methods and impacts. *In: Bliemer, M.C.J. (Ed.), Handbook on Transport and Urban Planning in the Developed World*, pp. 498–516.

[1]CITTA, Department of Civil Engineering, Faculty of Engineering of the University of Porto, Porto, Portugal. Email: ligia.conceicao@fe.up.pt

[2]Department of Transport and Planning, Delft University of Technology, Delft, The Netherlands. Email: G.Correia@tudelft.nl

[3]CITTA, Department of Civil Engineering, Faculty of Engineering of the University of Porto, Porto, Portugal. Email: ptavares@fe.up.pt

SYSØR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Modelling protein structure evolution by toroidal diffusions

E. García-Portugués [1], Michael Sørensen [2], Kanti V. Mardia [3], Thomas Hamelryck [4], Michael Golden [5], Jotun Hein[5]

E. García-Portugués (Assistant Professor).- Eduardo García Portugués (León, 1987) is Assistant Professor at the Department of Statistics of Carlos III University of Madrid. He is PhD in Statistics and Operations Research (2014, European distinction, cum laude), MSc in Statistical Techniques (2012, extraordinary award), and BSc in Mathematics (2010, extraordinary award), all of them by the University of Santiago de Compostela. He was postdoc at the Dynamical Systems Interdisciplinary Network of the University of Copenhaguen and did research stays at the Université Catholique de Louvain, and the University of North Carolina at Chapel Hill. He is *Young Researchers Award* (2013) by the Galician Society for the Promotion of Statistics and Operations Research, and *Ramiro Melendreras Award* (2015) by the Spanish Society of Statistics and Operations Research. Up to date, he has published 9 papers, 7 as a first author (one of them as a sole author). His main research lines are focused on the development of nonparametric methods and diffusive models for directional data, as well as the software associated with their implementation. Personal web: `https://egarpor.github.io/`

We present a probabilistic model for pairs of related proteins that, through the use of novel diffusions on the torus, aims to provide new insights into the relationship between protein sequence and structure evolution.

Proteins are large biomolecules with associated complex three-dimensional structures that are hard to obtain experimentally and are often crucial for determining biological functionality. Mathematically, a protein $\mathbf{P}$ comprised of n amino acids can be effectively parametrized by three sequences of size n: $\mathbf{P} \equiv (\mathbf{A}, \mathbf{X}, \mathbf{S})$, where $\mathbf{A}$ is the sequence of *amino acids* labels, $\mathbf{X} = \{(\phi_i, \psi_i)\}_{i=1}^n$ encodes the backbone of the protein through *dihedral angle pairs*, and $\mathbf{S}$ represents the *secondary structure* labels giving the main structural motifs. This triple is exploited by our model, termed Evolutionary Torus Dynamic Bayesian Network (ETDBN, [3]), to build three continuous-time Markovian processes that induce time-dependent joint distributions for the pairs of amino acids, dihedral angles, and secondary structures of two different-size proteins $\mathbf{P}_a$ and $\mathbf{P}_b$. These joint distributions are coupled together via a hidden Markov model, in the spirit of the non-evolutionary TorusDBN [1], that accounts for aligned positions dependence and yields a distribution for $(\mathbf{P}_a, \mathbf{P}_b)$.

The most challenging part in the development of ETDBN is the construction of a diffusive process on the torus $\mathbb{T}^2 = [-\pi, \pi) \times [-\pi, \pi)$ that is well-founded, ergodic, time-reversible, and tractable. To that aim, we present a class of toroidal analogues of the celebrated Ornstein-Uhlenbeck process that are generated by well-known distributions in Directional Statistics [4]. Their likelihood function is a product of transition densities that are analytically untractable, but that can be computed by solving numerically a Fokker-Planck equation. We propose several *approximate likelihoods* that are computationally less demanding, among them a specific approximation to the transition density of the wrapped normal process [2], which is the one we employ in ETDBN due to its tractability and reliability.

We provide simulation studies and empirical benchmarks that show the adequate performance of the approximate likelihoods and of ETDBN. Finally, a case study shows how ETDBN is able to deliver new evolutionary insights.

Key words: Evolution; Directional Statistics; Probabilistic model; Protein structure; Stochastic differential equation.

Acknowledgements This work is part of the Dynamical Systems Interdisciplinary Network, University of Copenhagen, and was funded by the University of Copenhagen 2016 Excellence Programme for Interdisciplinary Research (UCPH2016-DSIN), and by project MTM2016-76969-P from the Spanish Ministry of Economy, Industry and Competitiveness, and European Regional Development Fund (ERDF).

References

[1] Boomsma, W., Mardia, K. V., Taylor, C. C., Ferkinghoff-Borg, J., Krogh, A., Hamelryck, T. (2008). A generative, probabilistic model of local protein structure. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, pp. 8932–8937.

[2] García-Portugués, E., Sørensen, M., Mardia, K. V., Hamelryck, T. (2017). Langevin diffusions on the torus: estimation and applications. *arXiv preprint*, arXiv:1705.00296.

[3] Golden, M., García-Portugués, E., Sørensen, M., Mardia, K. V., Hamelryck, T., Hein, J. (2017). A generative angular model of protein structure evolution. *Mol. Biol. Evol.*, **34**, pp. 2085–2100.

[4] Mardia, K. V., Jupp, P. E. (2000). *Directional Statistics*. John Wiley & Sons, Chichester 2000.

[1]Department of Statistics, Carlos III University of Madrid, Leganés, Spain. Email: edgarcia@est-econ.uc3m.es
[2]Department of Mathematical Sciences, University of Copenhagen, Copenhagen, Denmark. Email: michael@math.ku.dk
[3]School of Mathematics, University of Leeds, Leeds, United Kingdom. Email: K.V.Mardia@leeds.ac.uk
[4]Department of Biology, University of Copenhagen, Copenhagen, Denmark. Email: thamelry@binf.ku.dk
[5]Department of Statistics, University of Oxford, United Kingdom. Email: golden@stats.ox.ac.uk

SYSØR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Calmness modulus of the optimal value function

M. J. Gisbert Francés[1], M. J. Cánovas Cánovas*, J. Parra López*, F.J. Toledo Melero*

M. J. Gisbert Francés (PhD Student).- María Jesús Gisbert is a PhD student at Miguel Hernández University of Elche. She graduated from University of Alicante with a BSc in Mathematics. Before coming to Elche, she studied a MSc in Mathematical Engineering at Carlos III University of Madrid, with Statistics specialization. Her fields of interest are Functional Data Analysis, Optimization, Linear Programming and Variational Analysis. Her current research is focused on obtaining exact formulae for calmness and Lipschitz moduli of the optimal value function in a solvable linear problem.

The final goal of the paper presented in this talk consists in computing/estimating the calmness moduli from below and above of the optimal value function restricted to the set of solvable linear problems. Roughly speaking these moduli provide measures of the maximum rates of decrease and increase of the optimal value under perturbations of the data (provided that solvability is preserved). This research is developed in the framework of (finite) linear optimization problems under canonical perturbations; i.e., under simultaneous perturbations of the right-hand-side (RHS) of the constraints and the coefficients of the objective function. As a first step, part of the work is developed in the context of RHS perturbations only, where a specific formulation for the optimal value function is provided. This formulation constitutes the starting point in providing exact formulae/estimations for the corresponding calmness moduli from below and above. We point out the fact that all expressions for the aimed calmness moduli are conceptually tractable (implementable) as far as they are given exclusively in terms of the nominal data.

Key words: Calmness; Optimal Value; Linear Programming.

Acknowledgements This research has been partially supported by Grant MTM2014-59179-C2-2-P from MINECO, Spain and FEDER (EU).

References

[1] Gisbert, M. J., Cánovas, M. J., Parra, J., Toledo, F. J. (2017). Calmness of the optimal value in linear programming. *Submitted*.

[1]Center of Operations Research (CIO), Miguel Hernández University of Elche (UMH), Elche (Alicante), Spain. Email: mgisbert@umh.es, canovas@umh.es, parra@umh.es, javier.toledo@umh.es

SYSØR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Probabilistic methods for combining internal migration data

G. Vinué [1] , G. Abel [2] , D. Yildiz [3] , A. Wisniowski [4]

G. Vinué (Postdoctoral Researcher).- Guillermo Vinué received his PhD in Statistics and, currently, he is working as a postdoctoral researcher at Vienna Institute of Demography.

Movement of people is becoming a complex phenomenon. In order to fully understand the causes and consequences of population movements, and how they evolve over time, researchers and policy makers require timely consistent data. Traditionally, data are obtained from censuses, registers or surveys, which only give a brief picture of the migration activity. In addition, these sources provide estimates of movements with different qualities according to their data collection methods or sample sizes. This paper proposes a Bayesian probabilistic methodology to harmonize movement data from traditional sources. US internal migration data is used. In migration studies, using a probabilistic approach is very natural and common [1]. In particular, the Bayesian paradigm provides a formal framework for combining different data types and for dealing with inherent uncertainties in migration [2]. The methodology presented will be able to achieve the following goals: (i) To get estimates of the true migration flows over a range of timing criteria; (ii) To explicitly address all potential sources of data inadequacies; (iii) A better understanding of people's movement patterns beyond the confines of a single source.

Key words: Internal migration; Combining data; Bayesian model; US data.

Acknowledgements This work has been supported for the project STE-Projekts 0059 funded from the Jubiläumsfonds der Stadt Wien.

References

[1] Nowok, B., Willekens, F. (2011). A probabilistic framework for harmonisation of migration statistics. *Population, Space and Place*, **17**, pp. 521–533.

[2] Raymer, J., Wisniowski, A., Forster, J., Smith, P., Bijak, J. (2013). Integrated modeling of European migration. *Journal of the American Statistical Association*, **108**, pp. 801–819.

[1]Wittgenstein Centre for Demography and Global Human Capital (IIASA, VID/ÖAW, WU), Vienna Institute of Demography/Austrian Academy of Sciences, Austria. Email: guillermo.vinue.visus@oeaw.ac.at

[2]Wittgenstein Centre for Demography and Global Human Capital (IIASA, VID/ÖAW, WU), Vienna Institute of Demography/Austrian Academy of Sciences, Austria ; Asian Demographic Research Institute, Shanghai University, China. Email: guy.abel@oeaw.ac.at

[3]Wittgenstein Centre for Demography and Global Human Capital (IIASA, VID/ÖAW, WU), Vienna Institute of Demography/Austrian Academy of Sciences, Austria. Email: dilek.yildiz@oeaw.ac.at

[4]Cathie Marsh Institute for Social Research, School of Social Sciences, University of Manchester, United Kingdom. Email: a.wisniowski@manchester.ac.uk

SYSOR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Some novel approaches to portfolio optimization

M. Leal [1] , D. Ponce[1] , J. Puerto[1]

M. Leal (PhD Student).- Marina Leal is currently a PhD student at University of Sevilla. She received her BSc in Mathematics at University of Alicante and a MSc in Operational Research at University of Murcia. During her PhD, she has carried out research stays at Rotman School of Management of Toronto and at Free University of Brussels (ULB). Her research is based on the construction of optimization models in the context of complex network analysis that include random or uncertain aspects in the parameters that define the model and / or multiple optimization criteria.

Portfolio problems with given transaction cost has been extensively studied in the literature. We present a bilevel leader-follower portfolio selection model in which the bank has to decide on the transaction costs for investing in some securities, maximizing its benefits, and the investor has to choose his portfolio, minimizing the risk and ensuring a given expected profit. In order to minimize the risk of the investor different risk measures can be considered. This gives rise to general non linear bilevel problems in both levels. We model different bilevel versions of the problem (cooperative model, bank-leader model, ...), determine some properties of the models, provide Mixed Integer Linear Programming formulations for some cases and report some preliminary computational results.

Key words: Portfolio selection; bilevel optimization; mixed integer linear programming.

Acknowledgements This research has been funded by the Spanish Ministry of Science and Technology projects MTM2013-46962-C2-1-P and MTM2016-74983-C2-1-R.

References

[1] Mansini, R., Ogryczak, W., Speranza, M. G. (2003). On LP Solvable Models for Portfolio Selection. *Informatica*, **14**, pp. 37–62.

[2] Mansini, R., Ogryczak, W., Speranza, M. G. (2014). Twenty years of linear programming based on portfolio optimization. *European Journal of Operational Research*, **234**, pp. 518–535.

[1]Department of Statistics and O.R., University of Sevilla and IMUS, Sevilla, Spain. Emails: mleal3@us.es, ponce@us.es, puerto@us.es.

SYSØR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Robust budget optimization and forecasting under uncertainty in the media industry

V. Gallego Alcalá [1] , P. Angulo Ardoy [2] , P. Suárez García [3] , D. Gómez-Ullate Oteiza [4]

V. Gallego Alcalá (PhD Student).- Víctor Gallego Alcalá graduated with a double degree in Mathematics and Computer Science, from Complutense University of Madrid, and a MSc in Mathematical Engineering. He has just joined ICMAT-CSIC as a predoctoral researcher in the SPOR-Datalab group under the supervision of Profs. David Gómez-Ullate and David Ríos. His research focuses on topics such as Artificial Intelligence, Machine Learning and Bayesian Statistics, ranging from fundamental aspects to applications in areas such as online advertising or natural language processing.

We are faced with the problem of optimizing the allocation of the marketing budget of a firm. The dependence of sales on the investment levels on each possible media channel and other external factors is stochastic, it is not well understood and it is hard to transfer from one product to another. Thus, we must learn this dependence, but the data available is clearly insufficient for such a task, so we must take into account the risk associated with our decisions.

We model the dependence using a bayesian approach and perform robust optimization to find the optimal investment strategy for each possible risk preference, without assuming knowledge of the firm's utility function.

We present some possible models, their practical implementation, and their application to one real data set resulting from the activity of a large chain of restaurants spreaded across a wide region.

Key words: DLM, MCMC, risk analysis.

Acknowledgements The authors acknowledge support from the *Institute of Mathematical Sciences* via the ICMAT-Severo Ochoa Excellence Program SEV-2015-0554.

References

[1] Campagnoli, P., Petrone, S., Petris, G. (2009). *Dynamic Linear Models with R.* Springer-Verlag, New York.

[2] Box, G., Jenkins, G., Reinsel, G. (2008). *Time Series Analysis: Forecasting and Control.* 4th ed. Wiley.

[3] Terence C. M. (1990). *Time Series Techniques for Economists.* Cambridge University Press, Cambridge.

[1]SPOR-Datalab, Instituto de Ciencias Matemáticas, Madrid, Spain. Email: vicgalle@ucm.es
[2]SPOR-Datalab, Instituto de Ciencias Matemáticas, Madrid, Spain. Email: pablo.angulo@upm.es
[3]SPOR-Datalab, Instituto de Ciencias Matemáticas, Madrid, Spain. Email: pasuarez@fis.ucm.es
[4]SPOR-Datalab, Instituto de Ciencias Matemáticas, Madrid, Spain. Email: david.gomez-ullate@icmat.es

SYSØR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Risk measures on threshold exceedance structural indicators in spatiotemporal processes

J. L. Romero [1], A. E. Madrid [2] , J. M. Angulo [3]

J. L. Romero (PhD Student).- José L. Romero has a BSc in Mathematics, a BSc in Statistics, a MSc in Research Design and Applications to Psychology and Health Sciences, and a MSc in Applied Statistics, all of them at University of Granada. Currently, José Luis is a PhD Student in the Mathematical and Applied Statistics PhD program at University of Granada. His topics of current research interest are related to environmental risk assessment, conditional risk measures, extreme values, threshold exceedance probabilities, deformation of random fields, spatiotemporal processes, etc.

Risk assessment in real phenomena (e.g. in Actuarial Sciences and Finance, Environmental Sciences, Geophysics, etc.), where the spatial features of the definition domain have to be taken into account to assess possible hazard situations, is an increasing area of research interest. Different probabilistic and statistical aspects can be addressed in terms of random field models (see, for example, [1, 4]). So-called first-order indicators derived from structural characteristics of threshold exceedances of random fields describe the extremal behaviour in the spatiotemporal dynamics of these phenomena (see, for example, [2, 3, 5, 9]).

There exist a recent well-founded theory of risk measures, mainly motivated by areas of applications such as Finance and Insurance, with an increasing interest in many other areas because of its potential applicability (see [7, 8]). Quantile-based risk measures such as Value-at-Risk and Expected Shortfall have received special attention because of their direct interpretation and easy computational implementation.

In this work, an original approach based on conditional simulation for the analysis of risks at both global and local scales is introduced. In particular, quantile-based risk measures such as Value-at-Risk and Expected Shortfall are applied to different threshold exceedance indicators by means of their empirical distributions, thus allowing the construction of dynamic risk maps with meaningful information on risk assessment. Effects of local variability and dependence range of the underlying random field, varying reference thresholds, and varying confidence levels of the considered quantile-based risk measures, among other aspects, are discussed and illustrated by simulation under different scenarios with the application of this methodology. The compound cumulative distribution function (see, [6]) plays a key role, both formally and regarding the practical threshold specification, for these indicators.

Key words: conditional simulation; quantile-based risk measures; space-time random fields; threshold exceedance indicators.

Acknowledgements This work has been partially supported by grants MTM2012-32666 and MTM2015-70840-P of Spanish MINECO/FEDER, EU.

References

[1] Adler, R. J. (1981). *The Geometry of Random Fields*. Wiley, Chichester.

[2] Adler, R. J., Taylor, J. E. (2007). *Random Fields and Geometry*. Springer, New York.

[3] Angulo, J. M., Madrid, A. E. (2010). Structural analysis of spatiotemporal threshold exceedances. *Environmetrics*, **21**, pp. 415–438.

[4] Christakos, G. (1992). *Random Field Models in Earth Sciences*. Academic Press, San Diego.

[5] Christakos, G., Hristopulos, D. T. (1996). Stochastic indicators for waste site characterization. *Water Resources Research*, **32**, pp. 2563–2578.

[6] Craigmile, P. F., Cressie, N., Santner, T. J., Rao, Y. (2005). A loss function approach to identifying environmental exceedances. *Extremes*, **8**, pp. 143–159.

[7] Föllmer, H., Shied, A. (2016). *Stochastic Finance. An Introduction in Discrete Time*. Walter de Gruyter GmbH & Co. KG, Berlin/New York.

[8] Klüppelberg, C., Straub, D., Welpe, I. M. (eds) (2014). *Risk - A Multidisciplinary Introduction*. Springer, Berlin.

[9] Romero, J. L., Madrid, A. E., Angulo, J. M. (2017). Quantile-based spatiotemporal risk assessment of exceedances. *Submitted*.

[1]Department of Statistics and O.R., University of Granada, Granada, Spain. Email: jlrbejar@ugr.es
[2]Department of Sciences and Informatics, University Centre of Defence at the Spanish Air Force Academy, Murcia, Spain. Email: anae.madrid@cud.upct.es
[3]Department of Statistics and O.R., University of Granada, Granada, Spain. Email: jmangulo@ugr.es

SYSOR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

New mathematical optimization models for Air Traffic Flow Management

David García-Heredia [1], Elisenda Molina [2] , Antonio Alonso-Ayuso [3]

D. García-Heredia (PhD Student).- David García Heredia got his BSc in Engineering (2015) at University Rey Juan Carlos, obtaining the best promotion academic record. In 2017, he received his MSc in Mathematical Engineering at Carlos III University of Madrid, where he is currently a PhD student in the department of Statistics and Operational Research. His research interests are combinatorial and stochastic optimization problems; particularly, the part related with the development of methods and algorithms that can solve them.

Nowadays, different regions of the world, such as Europe or USA are facing the problem of a saturated air traffic demand; i.e.: the number of scheduled flights is sometimes larger than the capacity of the airspace and/or the airports. This produces the necessity of delaying or even cancelling some flights to ensure the safety of the system.

To mitigate this problem, different approaches exist. One of them is the Air Traffic Flow Management (ATFM), where the objective is to modify flight plans so delays and cancellations are done at minimum cost, while fulfilling capacity constraints at all time (see [1, 2, 3, 4] among others).

To that purpose, we propose a new 0-1 mathematical optimization model based on 4D trajectories (position and time) for each flight. This allows us to see the problem not as a binary optimization problem, but as a shortest path problem with capacity constraints. A property that produces a nice structure in the constraint matrix, as well as some interesting decomposition possibilities.

The two most important contributions of this new model are: 1) In the computational experience that we present we achieve really good times for real size problems; and 2) Compared with previous models, we obtain more realistic costs' representations, as well as a better control of the decisions involved, making our approach closer to reality.

Key words: ATFM; Combinatorial Optimization; 4D graph.

Acknowledgements This research was partially funded by projects MTM2015-63710-P from the Spanish Ministry of Economy and Competitiveness.

References

[1] Agustín, A., Alonso-Ayuso, A., Escudero, L. F., Pizarro, C. (2012). On air traffic flow management with rerouting. Part I: Deterministic case. *European Journal of Operational Research*, **219**, pp. 156–166.

[2] Bertsimas, D., Patterson, S. S. (1998). The air traffic flow management problem with enroute capacities. *Operations research*, **46**, pp. 406–422.

[3] Bertsimas, D., Lulli, G., Odoni, A. (2011). An integer optimization approach to large-scale air traffic flow management. *Operations research*, **59**, pp. 211–227.

[4] Balakrishnan, H., Chandran, B. G. (2014). *Optimal large-scale air traffic flow management*. MIT, Massachusetts, USA.

[1]Department of Statistics, University Carlos III de Madrid, Leganés (Madrid), Spain. Email: dgheredi@est-econ.uc3m.es
[2]Department of Statistics, University Carlos III de Madrid, Leganés (Madrid), Spain. Email: emolina@est-econ.uc3m.es
[3]Department of Statistics and O.R., University Rey Juan Carlos, Móstoles (Madrid), Spain. Email: antonio.alonso@urjc.es

SYSOR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

A lack-of-fit test for quantile regression models using logistic regression

M. Conde-Amboage [1], V. Patilea [2], C. Sánchez-Sellero [3]

M. Conde-Amboage (Postdoctoral Researcher).- Mercedes Conde Amboage is currently working as a Postdoctoral Researcher at the Research Centre for Operations Research and Business Statistics (ORSTAT), from the University of Leuven (Belgium).

Quantile regression is employed when the aim of study is centred on the estimation of the different positions (quantiles). This kind of regression allows a more detailed description of the behaviour of the response variable, adapts to situations under more general conditions of the error distribution and enjoys properties of robustness. Hereby it facilitates a more complete and robust analysis of the information. For all that, quantile regression is a very useful statistical technology for a large diversity of disciplines.

Quantile regression was introduced by [4] as a weighted absolute residuals fit which allows to extend some properties of classical least squares estimation to quantile regression estimates. Several classical statistical tools and procedures have been adapted to quantile regression scenario over the years. In this line, [3] is a good review about quantile regression.

A new lack-of-fit test for parametric quantile regression models will be presented. The test is based on interpreting the residuals from the quantile regression model fit as response values of a logistic regression, the predictors of the logistic regression being functions of the covariates of the quantile model. Then a correct quantile model implies the nullity of all the coefficients but the constant in the logistic model. Given this property, we use a lack-of-fit test in the logistic regression to check the quantile regression model following the ideas introduced by [1]. In the case of a multivariate quantile regression, to avoid working in very large dimension, we use predictors obtained as functions of univariate projections of the covariates from the quantile model. Finally, we look for a "least favourable" projection for the null hypothesis of the lack-of-fit test. Our test can detect general departures from the parametric quantile model. To approximate the critical values of the test, a wild bootstrap mechanism is used, similar to that proposed by [2]. A simulation study and an application to real data show the good properties of the new test versus other nonparametric tests available in the literature.

Key words: Quantile regression; Lack-of-fit testing; Logistic regression; Bootstraping.

Acknowledgements This study was supported by Projects MTM2013–41383–P (Spanish Ministry of Economy, Industry and Competitiveness) and MTM2016–76969–P (Spanish State Research Agency, AEI), both co–funded by the European Regional Development Fund (ERDF). Support from the IAP network StUDyS from the Belgian Science Policy is also acknowledged. V. Patilea acknowledges financial support from the research program *New Challenges for New Data* of LCL and Genes.

References

[1] Aerts, M., Claeskens, G. and Hart, J. D. (2000). Testing lack of fit in multiple regression. *Biometrika*, **87**, pp. 405–424.

[2] Feng, X., He, X. and Hu, J. (2011). Wild bootstrap for quantile regression. *Biometrika*, **98**, pp. 995–999.

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

[4] Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, **46**, pp. 33–50.

[1]Research Centre for Operations Research and Business Statistics (ORSTAT), University of Leuven, Leuven, Belgium. Email: mercedes.condeamboage@ukleuven.be

[2]Center of Research in Economics and Statistic (CREST), École Nationale de la Statistique et de l'Analyse de l'Information, Rennes, France. Email: Valentin.PATILEA@ensai.Fr

[3]Department of Statistics, Mathematical Analysis and Optimization, Universidade de Santiago de Compostela, Santiago de Compostela, Spain. Email: cesar.sanchez@usc.es

SYSOR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

New methods and results for the optimisation of solar power tower plants

T. Ashley [1] , Emilio Carrizosa [2], Enrique Fernández-Cara [3]

T. Ashley (PhD Student).- Thomas Ashley is a PhD student in Mathematics at University of Sevilla (Spain), researching into the optimisation of the renewable energy technology Solar Tower Power Plants. Thomas has an MSc in Advanced Mathematics from University of Exeter (UK) and has previously worked for six years applying mathematics within the Defense engineering industry.

Research into renewable energy sources has continued to increase in recent years, and in particular the research and application of solar energy systems. Concentrated Solar Power (CSP) used by a Solar Power Tower (SPT) plant is one technology that continues to be a promising research topic for advancement.

CSP is a method of solar energy collection, where the energy from the Sun is concentrated by a field of heliostat onto a central receiver. In Solar Power Tower (SPT) plants, this receiver is mounted atop a tower, and the resultant thermal load is used to drive a steam generator. This allows high temperatures to be achieved and is an increasingly investigated method into renewable energy production, see [1, 2].

We model the incident radiation in the system using a Gaussian distribution, as in [3], and utilise integer programming techniques in order to optimise the energy reaching the receiver surface. We optimise the aiming strategy for the heliostat field [4], where we constrain the system to produce a homogenous distribution on the receiver with a maximum limit to prevent damage. Inclement weather effects are taken into account, and an aiming strategy formed when cloud movement may be predicted.

We also consider the cleaning strategy for the heliostat field, where two integer programming problems are formed; the allocation of heliostats to cleaning periods, and the route optimisation for each period.

Further work in this project will consider other programming methods for the optimisation of SPT plants, as well as the application of thermal modelling in the system and how this affects the aiming strategies already developed.

Key words: Solar Thermal Power; Aiming Strategy; Integer Programming

Acknowledgements This research is being supported by the Spanish Government through the research project PCIN-2015-108 and is being conducted within the group MTM2015-65915-R at the University of Seville.

References

[1] Barberena, J. G., Mutuberria Larrayoz, A., Sánchez, M., Bernardos, A. (2016). State-of-the-art of Heliostat Field Layout Algorithms and their Comparison. *Energy Procedia*, **93**, pp. 31–38.

[2] Besarati, S. M., Yogi Goswami, D. (2014). A computationally efficient method for the design of the heliostat field for solar power tower plant. *Renewable Energy*, **69**, pp. 226–232.

[3] Carrizosa, E., Domínguez-Bravo, C., Fernández-Cara, E., Quero, M. (2015). A heuristic method for simultaneous tower and pattern-free field optimization on solar power systems. *Computers & Operations Research*, **57**, pp. 109–122.

[4] Ashley, T., Carrizosa, E., Fernández-Cara, E. (2017). Optimisation of aiming strategies in Solar Power Tower plants. *Energy*, **137C**, pp. 285–291.

[1]Instituto de Matemáticas Universidad de Sevilla (IMUS), Seville, Spain. Email: tashley@us.es
[2]Instituto de Matemáticas Universidad de Sevilla (IMUS), Seville, Spain. Email: ecarrizosa@us.es
[3]Dep. EDAN and IMUS, Universidad de Sevilla, Spain. Email: cara@us.es

SYSØR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Analysing biological rhythms using order restricted inference.

Y. Larriba [1] , C. Rueda [2], M.A. Fernández [3]

Y. Larriba (PhD Student).- Yolanda Larriba is a PhD student at the Department of Statistics and Operational Research of University of Valladolid. Her main research concerns the development of statistical methodologies and software to analyse circular data models and their applications in biology, from the perspective of the order restricted inference. She completed a MSc in Mathematical Research on July 2015 at University of Valladolid with the project entitled: *A new method for detection of cycling circadian genes using order restricted inference*. Before that, she obtained a BSc in Statistics on July 2014 and a BSc in Mathematics on July 2012, both at University of Valladolid.

Many biological processes, such as cell cycle, circadian clock or blood pressure, are governed by oscillatory systems consisting of numerous components that exhibit periodic patterns over time. Modelling these rhythms is a challenge in literature since usually the sampling density is low, the number of periods is generally two and the underlying signals adopt a wide range of temporal patterns, see [1]. Several authors proposed parametric functions of time, such as the sinusoidal function, to model these signals. However, these parametric functions might be too rigid for data derived from cell-cycle or circadian clock.

These signals usually have a unique peak at time point U and a unique trough at time point L within each period, so that they monotonically increase up to U (when $L > U$) and then decrease up to L; before increasing again. It is clear that the shape of these signals can be entirely described in the euclidean space by mathematical inequalities among their components. The main novelty of this work is the definition of circular signals using restrictions to model common signal shapes in biology. We will give a definition that allows us to state equivalent signal formulations both in the euclidean and in the circular spaces. This formulation is crucial to set and interpret rhythmicity issues easily.

Additionally to the definition of circular signals, this work proposes a novel general methodology to analyse rhythmicity which is based on order restricted inference (ORI). Specifically, it includes an efficient algorithm to compute the restricted maximum likelihood estimate (RMLE) under circular constraints, rhythmicity tests based on likelihood ratio test (LRT), a procedure to compute confidence intervals for the landmarks L and U as well as an approach to estimate sampling order when time sampling is unknown. The obtained results are compared with classical methods in literature to analyse rhythmicity both in simulations and in real data bases.

Key words: Order Restricted Inference; Rhythmicity; Circular Space; Circadian Genes; Oscillatory Systems.

References

[1] Larriba, Y., Rueda, C., Fernández, M. A. and Peddada, S. D. (2016). Order restricted inference for oscillatory systems for detecting rhythmic signals. *Nucleic Acids Res.*, **44**, doi: 10.1093/nar/gkw771.

[1]Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Valladolid, Spain. Email: yolanda.larriba@uva.es
[2]Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Valladolid, Spain. Email: cristina.rueda@uva.es
[3]Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Valladolid, Spain. Email: miguelaf@eio.uva.es

SYSOR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Improving interpretability in linear regression

A.V. Olivares Nadal [1] , Emilio Carrizosa Priego [2] , Pepa Ramírez Cobo [3]

Alba V. Olivares Nadal (Assistant Professor).- Born and raised in Sevilla, Alba V. Olivares Nadal obtained a PhD in Mathematics at University of Sevilla. She has frequently visited the London Business School, and she has worked at University of Edinburgh as a Postdoctoral Research Associate in Optimization. She has also been employed by University of Cádiz, and currently, she is working as a Lecturer at University Pablo de Olavide (Sevilla). His research is based on addressing trending problems by merging optimization techniques with statistical concepts, leading to methods that outperform the classic approaches and bridge theoretical mathematics with real life problems. Since one of these problems is that real life is actually limited, if not short, she has to share my time with my other passion aside from researching: travelling.

A plethora of real world data involve multiple features interacting between each other. As a consequence, one of the most common problems in real life is trying to predict a variable by making use of attributes that are deterministic or easier to access. A widely studied tool to achieve this are the linear regression models $\mathbf{Y} = \beta_0 + \boldsymbol{\beta}\mathbf{X} + \mathbf{a}$, where $\mathbf{Y} = (y_1, ..., y_K)'$ contains the K realizations of the stochastic variable to be predicted, $\mathbf{X} \in \mathbb{R}^{K \times N}$ contains the observations of the attributes $X^1, ..., X^N$ that influence on $\mathbf{Y}$, and $\mathbf{a} \in \mathbb{R}^K$ denote the error terms. Estimating the coefficients $\beta_0, \beta_1, ..., \beta_N$ yielding a good fit to the data set may lead to a highly dense solution; i.e., a high number of non-zero coefficients is likely to be obtained, which may make the model difficult to interpret.

A more interpretable output may be obtained not only by requiring a sparse output, but also paying attention to highly correlated predictors. This is a critical issue, since almost linear relationships amongst the features $X^1, ..., X^N$ increase the variance of the estimated coefficients $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_N$, allowing for misleading interpretations.

With the aim to enhance the interpretability of linear regression models, we propose to use tools based on Mathematical Optimization. In particular, we will model constraints and add them to an optimization problem expressing some estimation procedure as OLS or Lasso [3]. These constraints will express the desire to

(i) attain sparsity, by upper bounding the number of non-zero coefficients and/or avoiding spurious coefficients so that only significant features are represented in the model

(ii) avoid misleading interpretations yielded by collinearity, by forcing the sign of the estimated coefficients to be consistent with the sign of the correlations between predictors or incorporating the information gathered by principal components.

The so-obtained constrained regression problems will become Mixed Integer Non-Linear Programs (MINLP). Although not frequently used amongst the statistical community, MINLP are becoming an useful tool to address statistical problems in a tractable and versatile manner. In fact, supported by the recent improvements in computational times that MINLPs have enjoyed, MINLPs have recently been used to tackle sparse models in linear regression [1, 2].

The numerical experiments carried out on real and simulated datasets suggest that embedding constraints that model (i) and/or (ii) into the estimation procedures for $\boldsymbol{\beta}$ may help to improve the sparsity and interpretability of the solutions with competitive predictive quality.

Key words: Linear regression; Variable selection; Sparsity; Cardinality constraint; Multicollinearity; Mixed Integer Non Linear Programming

Acknowledgements This research is supported by projects MTM2015-65915 (Ministerio de Economía y Competitividad, Spain), P11-FQM-7603 and FQM-329 (Junta de Andalucía), all with EU ERD Funds.

References

[1] Bertsimas, D., King, A. (2015). OR forum: An algorithmic approach to linear regression. *Operations Research.*

[2] Bertsimas, D., King, A., Mazumder, R. (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics*, **44**, pp. 813–852.

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

[1]Department of Statistics and O.R., University of Cádiz, Cádiz, Spain. Email: alba.olivares@uca.es
[2]Department of Statistics and O.R., University of Sevilla, Sevilla, Spain. Email: ecarrizosa@us.es
[3]Department of Statistics and O.R., University of Cádiz, Cádiz, Spain. Email: pepa.ramirez@uca.es

SYSØRM 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

A multiple criteria decision aiding method for nominal classification

A.S. Costa[1], J.R. Figueira [2] , J. Borbinha [3]

A. S. Costa (PhD Student).- Ana Sara Costa is currently a PhD student in Engineering and Management at Instituto Superior Técnico of University of Lisboa, and a researcher at Center for Management Studies (CEG-IST) and Instituto de Engenharia de Sistemas e Computadores - Investigação e Desenvolvimento (INESC-ID). She holds a MSc degree in Biomedical Engineering and a Post-graduation in Project Management. Her research interests include the development and application of multiple criteria decision aiding methods to classification problems.

Multiple Criteria Decision Aiding (MCDA) methods are considered a suitable tool for classifying actions (alternatives or options) into different categories (classes or groups). In this work, we propose a new MCDA method for nominal classification problems. They are encountered in a wide range of areas, such as genetics, medicine, psychology, education and training, business and environment. A multiple criteria nominal classification problem exists when the categories are pre-defined and no order exists among them. Addressing a problem of this kind consists of assigning each action, which is assessed according to multiple criteria, to the different non-ordered categories. The proposed method, Cat-SD (CAtegorization by Similarity-Dissimilarity), is based on the concepts of similarity and dissimilarity. We propose a way of modeling similarity and dissimilarity between two actions. Each category is characterized by a set of reference actions (the most representative actions of the category), and the interaction between some pairs of criteria is possible. Application of the Cat-SD method should follow a decision-aiding constructive approach, which means that an interactive process between the analyst and the decision-maker should be followed during application of the method. The proposed method fulfills a certain number of structural requirements (fundamental properties). We present these fundamental properties of the method and provide their proofs. Some potential applications are introduced, and a numerical example is presented to illustrate the way in which the proposed method can be applied. Robustness concerns are also considered in our work.

Key words: Multiple Criteria Decision Aiding; Decision Support Systems; Nominal classification.

Acknowledgements This work was supported by national funds through Fundação para a Ci'encia e a Tecnologia (FCT) with reference UID/CEC/50021/2013. The first author acknowledges financial support from Universidade de Lisboa, Instituto Superior Técnico, and CEG-IST (Ph.D. Fellowship), and financial support from Associação Portuguesa de Investigação Operacional (APDIO) to participate in SYSORM 2017.

[1]CEG-IST, INESC-ID, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal. Email: anasaracosta@tecnico.ulisboa.pt
[2]CEG-IST, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal. Email: figueira@tecnico.ulisboa.pt
[3]INESC-ID, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal. Email: jlb@tecnico.ulisboa.pt

SYSOR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Numerical methods for optimal mixture experiments

I. García-Camacha Gutiérrez * , R. Martín Martín [1] , B. Torsney [2]

I. García-Camacha Gutiérrez (Assistant Professor).- Irene García-Camacha Gutiérrez is an Assistant Professor in the Department of Mathematics, Statistics and Operational Research Area, at the University of Castilla-La Mancha, where she has been member since 2012. She has a Mathematics and Statistics Degree (BSc) by the University of Salamanca and a Mathematics and Physics Master (MSc) by the Universities of Castilla-La Mancha and Granada. Irene completed her PhD at University of Castilla-La Mancha in 2017. Her research interests lie in the area of Optimal Experimental Design (OED), mixture experiments and optimization algorithms.

Many experiments are formed by mixing two or more components or ingredients. Applications of mixture problems can be found in several areas including the chemical, pharmaceutical and biological sciences, among others [1]. Their main purpose is to identify the composition of different blends which optimally describe the characteristic-response of certain products. Mixture design background has received a classical approach. However there is a growing interest in Optimal Experimental Design (OED) nowadays. This is not only due to resource optimization considerations but also due to its flexibility in handling non-standard conditions such as design space constraints where traditional designs are inappropriate. Analytical results can be only obtained in systematic examples and under strict assumptions, definitively far from realistic problems. Heuristic solutions provided by numerical techniques are currently the only viable option to tackle such problems. Algorithmic techniques for computing optimal designs continue to be a need in the optimal experimental design field due to the current-state-of-the-art methods. The increasing interest in finding optimal experimental conditions demands new methods for more complex frameworks arising in realistic situations.

Two efficient algorithms are proposed in this work for identifying exact $D-$optimal designs in mixture experiments. The first one is based on a Multiplicative Algorithm (MA) [2]. It consists of an update rule of probability measures and its convergence has been extensively studied for approximate design theory. However, the application of this methodology is not straightforward in mixture settings. In this work, we provide a new approach of the MA using a special class of designs known as exchangeable designs [3]. The idea of these designs is to generate candidate points in the mixture designs using permutations of a fixed set of component values. In this work, this class of designs are called *permutation mixture experimental designs* (PMED). Under this context, the use of MA takes advantage of exploiting the general equivalence theorem. On the other hand, Genetic Algorithms (GAs) are a flexible class of stochastic optimization methods which are easy to implement [4]. The nature of mixture experiments requires special conditions on the operators and even more if there are experimental limitations on the proportions. An efficient GA is shown to be an heuristic alternative which is also valid in constrained mixture problems. It is strongly demanded in real applications since there are experimental limitations or ingredient availability considerations.

One of the most important challenges in the experimental field is to provide efficient designs in order to set-up the trials. Throughout the new proposed methods, we provide practitioners with highly efficient optimal designs in comparison with the existing ones in order to carry out their experiments related to mixing laws for fluid viscosity or drug formulations.

Key words: Optimal Experimental Design; Mixture Experiments; Multiplicative Algorithm; Genetic Algorithm.

Acknowledgements The authors have been sponsored by Ministerio de Ciencia e Innovacion (MMTM2013-47879-C2-1-P, MTM2016-80539-C2-1-R)

References

[1] Cornell, J. A. (2002). *Experiments with Mixtures*. Wiley, New York 2002.

[2] Silvey, S. D., Titterington, D. M., Torsney, B. (1978). An algorithm for optimal designs on a finite design space. *Communications in Statistic*, **14A**, pp. 1379–1389.

[3] Draper, N. R., Pukelsheim, F. (1999). Kiefer ordering of simplex designs for first- and second-degree mixture models. *Journal of Statistical Planning and Inference*, **79**, pp. 325–348.

[4] Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. MIT Press, Cambridge 1975.

[1]Department of Mathematics, University of Castilla-La Mancha, Avda. Carlos III s/n, 45071, Toledo, Spain. Email: irene.garciacamacha@uclm.es, raul.mmartin@uclm.es

[2]School of Mathematics & Statistics, University of Glasgow, Glasgow G128QW, UK. Email: Bernard.Torsney@glasgow.ac.uk

SYSOR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

A hybrid multiple criteria approach to improve supplier relationship management: A PROMETHEE and MAUT comparison

M. Segura Maroto [1] , C. Maroto Álvarez[1]

M. Segura Maroto (Assistant Professor).- Marina Segura is an Assistant Professor in the Department of Applied Statistics and Operations Research and Quality at UPV. She received her PhD in Statistics and Optimization (2015). MSc in Data Analysis, Process Improvement and Decision Support Engineering (2011) and Master in Business Administration and Management (2010), all of them from UPV. Her main research topics are Multiple Criteria and Group Decision Making, Supplier Segmentation and Ecosystem Services Assessment. She obtained a grant of the Ministry of Education for developing her PhD thesis (2012-2015). In addition, she did four research stays, two at the Centre for Ecology, Society and Biosecurity (UK, 2012 and 2015) and two at University of Strathclyde, Department of Management Science (UK, 2014 and 2015). She is also co-author of the teaching book *Teaching book: Operations Research in Business Administration and Management.*

As supported by many conceptual models, supplier management has nowadays become a strategic function in companies with growing importance. Nevertheless, companies are focused on products, without appropriating tools to inform relationships with suppliers. The objective of this research is to propose and validate a decision support system which allows qualifying, selecting and segmenting suppliers based on multiple criteria and group decision making approach, also comparing the strengths and weaknesses of outranking and value function methods for this purpose.

We have proposed a hybrid method which integrates Analytic Hierarchy Process (AHP) with Preference Ranking Organization Method for Enrichment Evaluation (PROMETHEE) and Multi-Attribute Utility Theory (MAUT). AHP is useful to define the hierarchy with evaluation criteria, grouping them into relevant dimensions, as well as to elicit the weights, which represent the point of view of several departments of the company.

After defining the criteria to qualify suppliers for each product, a general two-phase procedure has been developed and implemented in a big manufacturing company, which works for several sectors such as nutrition and chemicals. First, an assessment of products has carried out based on criteria classified in two dimensions, critical and strategic. Safety and environmental regulations, number of suppliers, delivery time and provisioning factor are critical criteria, mainly related to the market. The strategic criteria include internal features, such as contact with the final product, stopping the production of the factory, impact on the image of the company and the purchase volume. Second, the procedure evaluates suppliers also based on critical and strategic criteria, which takes into account the product indicators obtained in the first phase in addition to others, such as delays and several types of risks.

PROMETHEE and MAUT are used to carry out both evaluations of products and suppliers, deriving critical and strategic indicators, useful to supplier selection and segmentation. AHP is used to elicit the weights of criteria. The indicators obtained are represented in a graph, where axes are the critical and strategic scores. Dividing this graph into four areas, the suppliers can be segmented according to their performance. The appropriate relationships between the company and less critical and strategic suppliers should be guided by market price. On the contrary, the very critical and very strategic suppliers should be partners of the company. When suppliers maintain a strategic character, but they are not critical, one suitable strategy for the company to follow is to sign long term contracts. Finally, when a supplier is very critical but little strategic, the company should remove this supplier from its portfolio.

The results have shown the flexibility and robustness of PROMETHEE as an outranking method for supplier segmentation. MAUT is also flexible, but as compensatory method it has less discriminant power among suppliers than PROMETHEE. On the contrary, MAUT is easier to understand by the company staff. We propose a system with both methods in order to provide more advanced capabilities. Finally, this research is the first supplier segmentation applied to a real industry, which integrates historical data and expert knowledge in order to inform decision making [1].

Key words: Supplier segmentation; PROMETHEE; AHP; MAUT; Group decision making.

Acknowledgements Ministry of Education (Marina Segura, scholarship of Training Plan of University Teaching) and Ministry of Economy (Ref. ECO2011- 27369) has supported this research. The authors also thank the purchasing department for providing real data.

References

[1] Segura, M., Maroto, C. (2017). A multiple criteria supplier segmentation using outranking and value function methods. *Expert Syst. Appl.*, **69**, pp. 87–100. doi:10.1016/j.eswa.2016.10.031

[1]Dept. of Applied Statistics and OR, and Quality, Universitat Politécnica de Valencia, Valencia, Spain. Email: masema@upvnet.upv.es

SYSØR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Time series in function spaces: autoregressive Hilbertian processes

Javier Álvarez Liébana [1] and M. Dolores Ruiz Medina [2]

J. Álvarez Liébana (PhD Student).- Javier Álvarez Liébana was born in Madrid (1989). In 2013, he was graduated in BSc in Maths by Complutense University (Madrid, Spain). He studied at University of Bologna, as an Erasmus student during the period 2011-2012. In 2014, he was postgraduated in MSc in Mathematical Engineering, from Complutense University. He has been employed in different insurance and consultancy companies. Since 2014, he is working on his PhD thesis (3rd year) at University of Granada, supervised by M. Dolores Ruiz Medina. His main goal is the formulation of new outcomes on the estimation and prediction of functional time series, where Hilbert or Banach spaces are regarded. Its application to the forecasting of meteorological, genetic or financial data is also being achieved as well. Currently, he is developing his work at Université Pierre et Marie Curie (Paris, France), under the supervision of Denis Bosq. His research interests are wide, such as big data, biostatistics, clustering, finance, functional data, Geodesy, machine learning and time series.

Since the beginning, time series analysis has played a key role in the statistical analysis of temporally correlated data. Because of its high flexibility, it is well-known that prediction draw from time series framework has been crucial in a wide range of applications, such as stock market analysis, quality control, biological processes and sales forecasting, among others. Due to the huge computing advances, data began to be gathered with an increasingly temporal resolution level, such that the values are collected with a hourly, even minute-to-minute, frequency. In this high-dimensional framework, stored observations are now considered as points coming from a continuous process valued in a function space.

This communication is aimed at introducing how data with an autoregressive structure can be predicted by using an autoregressive process of order one, taking values in Hilbert spaces (ARH(1) processes). To provide to the listener a comprehensive idea about the great potential of this Hilbert-valued time series framework, we also address the main estimation and prediction results existing in the current literature, in the context above-referred (see [1, 3, 4, 5, 6, 7]). The simulation study pays attention to the forecasting of functional Ornstein-Uhlenbeck processes (see [2, 8, 9]), such that the results derived allow to predict the curve representing the interest rate over a temporal interval, in a consistent way.

A real-data application on the forecasting of the zero-coupon yield curve (also known as term structure), which shape is read as the main measure of future expectations and the assessment of monetary policies, will be also illustrated. Namely, this doubt product is used as a benchmark to compare with other debt instruments in the secondary market.

Key words: ARH processes; functional time series review; Hilbert-Schmidt autocorrelation operator; Ornstein-Uhlenbeck process; strongly-consistent estimator; yield curve

Acknowledgements Work supported by project MTM2015-71839-P (co-funded by Feder funds), of the DGI, MINECO, Spain.

References

[1] Álvarez-Liébana, J. (2017). A review and comparative study on functional time series techniques. Submitted.

[2] Álvarez-Liébana, J., Bosq, D., Ruiz-Medina, M. D. (2016). Consistency of the plug-in functional predictor of the Ornstein-Uhlenbeck process in Hilbert and Banach spaces. *Statist. Probab. Lett.*, **117**, 12–22.

[3] Álvarez-Liébana, J., Bosq, D., Ruiz-Medina, M. D. (2017). Asymptotic properties of a componentwise ARH(1) plug-in predictor. *J. Multivariate Anal.*, **155**, 12–34.

[4] Antoniadis, A., Sapatinas, T. (2003). Wavelet methods for continuous-time prediction using Hilbert-valued autoregressive processes. *J. Multivariate Anal.*, **87**, 133–158.

[5] Bosq, D. (2000). *Linear Processes in Function Spaces.* Springer, New York.

[6] Guillas, S. (2001). Rates of convergence of autocorrelation estimates for autoregressive Hilbertian processes. *Statist. Probab. Lett.*, **55**, 281–291.

[7] Ruiz-Medina, M. D., Álvarez-Liébana, J. (2017). Consistent diagonal componentwise ARH(1) prediction. Submitted.

[8] Uhlenbeck, G. E., Ornstein, L. S. (1930). On the theory of Brownian motion. *Phys. Rev.*, **36**, 823–841.

[9] Wang, M. C., Uhlenbeck, G. E. (1945). On the theory of Brownian motion II. *Rev. Modern Phys.*, **17**, 323–342.

[1] PhD Student (3rd year). Department of Statistics and O.R., University of Granada, Granada, Spain. Email: javialvaliebana@ugr.es

[2] Full Professor. Department of Statistics and O.R., University of Granada, Granada, Spain. Email: mruiz@ugr.es

SYSØR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Minimum density power divergence estimators for one-shot device model

E. Castilla [1]

E. Castilla (PhD Student).- Elena Castilla was born in 1993, in Madrid (Spain). She did her undergraduate studies in Mathematics and Statistics at Complutense University of Madrid (UCM). When she finished (2015), she did a MSc about Statistics and Computation at the same university, jointly with Politécnica University of Madrid (UPM). In 2016, she started my PhD thesis under the supervision of Leandro Pardo and Nirian Martín. Her research interest include robustness, logistic regression, one-shot devices and divergence measures.

The reliability of a product can be defined as the probability that the device does not fail when used. While engineers asses reliability by repeatedly testing the device and observing its failure rate, this is a challenging approach for "One-shot devices", that can only be used once and after use the device is either destroyed or must be rebuilt. Some examples of one-shot devices are nuclear weapons, space shuttles or fuses. In survival analysis,these data are called "current status data". For example, in animal carcinogenicity experiments, one observes whether a tumour occurs at the examination time for each subject.

We shall assume that the failure times of devices follow an exponential distribution, depending on an adjusting controllabe factor such as temperature (see [1]). A new family of estimators, the minimum density power divergence estimators (MDPDEs, introduced in [2]) are developed as a natural extension of the maximum likelihood estimator (MLE) for the parameters of the one-shot device model under exponential distribution. We also develop a new family of test statistics, Z-type test statistics based on MDPDEs, for testing the corresopnding testing parameters. A simulation study shows how some MDPDEs have a better behaviour than the MLE in relation to robustness ([3]).

Key words: One-shot devices; Minimum density power divergence estimator, Exponential distribution; Robustness.

References

[1] Balakrishnan, N., Ling, M. H. (2012). EM algorithm for one-shot device testing under the exponential distribution. *Computational Statistics & Data Analysis*, **56**, pp. 502–509.

[2] Basu, A., Harris, I. R., Hjort, N. L., Jones, M. C. (1998). Robust and efficient estimation by minimizing a density power divergence. *Biometrika*, **85**, pp. 549–559.

[3] Balakrishnan, N., Castilla, E., Martín, N., Pardo, L. (2017). Robust Estimators and Test-Statistics for One-Shot Device Testing Under the Exponential Distribution. *arXiv:1704.07865*

[1]Department of Statistics and O.R., Complutense University of Madrid, Madrid, Spain. Email: elecasti@ucm.es

SYSØR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Adversarial Classification: An Adversarial Risk Analysis Approach

R. Naveiro Flores [1] , A. Redondo [2] , A. Adebanji [3] , F. Ruggeri [4] , D. Ríos Insua [5]

R. Naveiro Flores (PhD Student).- Born in the lovely city of León, Roi Naveiro Flores studied theoretical physics in Salamanca and Complutense University of Madrid. Recently, he has joined the amazing world of applied maths, such as statistics, operational research and machine learning. Currently, he is developing his PhD thesis at ICMAT-CSIC, working on how to modify machine learning sin order to be secure when intelligent adversaries that try to cheat them are present.

A pioneering attempt to modify machine learning classification algorithms to deal with situations in which an adversary, that actively manipulates the data to fool the classifier; is present, is denominated Adversarial Classification, see [1]. The authors describe the problem as a game between a classifier (C) and an adversary (A). The classifier aims at finding an optimal classification strategy against the adversary's optimal attacking strategy. Computing Nash equilibria in this general game becomes too complex. Therefore, the authors propose a simplified version in which C first assumes the data is untainted and computes her optimal classifier; then, A deploys an optimal attack against this classifier; subsequently, C implements the optimal classifier against this attack, etc. As pointed out by the authors, a very strong assumption is made: all parameters of both players are known to each other. Although standard in game theory, this common knowledge assumption is actually unrealistic in real life scenarios of adversarial classification. Despite the inoperability produced by this assumption, most approaches expanding and developing the framework by Dalvi *et al* have been unable to overcome such strong assumption. See for instance [2], [3], [4]. In this paper, we outline a new framework for adversarial classification, based on Adversarial Risk Analysis (ARA), described in [5], where no assumptions of common knowledge are made. ARA provides one-sided prescriptive support to a DM, maximizing her subjective expected utility, treating the adversaries' decisions as random variables. It models the adversaries' decision-making problems and, under assumptions about their rationality, such as them being expected utility maximizers, tries to assess their probabilities and utilities. However, the uncertainty about the adversaries' probabilities and utilities is propagated into their decisions, leading to random optimal adversarial decisions which provide the required distributions over the adversaries' decisions. Here we shall build Adversarial Classification Risk Analysis (ACRA), an ARA adversarial classification approach on top of the pioneering framework in [1].

Key words: Statistical classification; adversarial classification; adversarial risk analysis; game theory; spam detection.

Acknowledgements R.N.F acknowledges the Spanish Ministry of Education for the FPU Ph.D. scholarship. The work of D.R.I is supported by the Spanish Ministry of Economy and Innovation program MTM2014-56949-C3-1-R and the AXA-ICMAT Chair on Adversarial Risk Analysis. This work has also been partially supported by the Spanish Ministry of Economy and Competitiveness through the "Severo Ochoa" Program for Centers of Excellence in R&D (SEV-2015-0554) and project MTM2015-72907-EXP.

References

[1] Dalvi, N., Domingos, P., Sanghai, S., Verma, D. (2004). Adversarial classification. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 99–108.

[2] Lowd, D., Meek, C. (2005). Adversarial learning. *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 641–647.

[3] Kantarcioglu, M., Xi, B., Clifton, C. (2011). Classifier evaluation and attribute selection against active adversaries. *Data Mining and Knowledge Discovery*, **22**, pp. 291–335.

[4] Zhou, Y., Kantarcioglu, M., Thuraisingham, B., Xi, B. (2012). Adversarial support vector machine learning. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1059–1067.

[5] Ríos Insua, D., Rios, J., Banks, D. (2009). *Adversarial risk analysis. Journal of the American Statistical Association*, **104**, pp. 841–854.

[1]Institute of Mathematical Sciences (ICMAT-CSIC), Madrid, Spain. Email: roi.naveiro@icmat.es
[2]Institute of Mathematical Sciences (ICMAT-CSIC), Madrid, Spain. Email: alberto.redondo@icmat.es
[3]KNUST, Ghana. Email: aoadebanji.cos@knust.edu.gh
[4]CNR-IMATI, Milano, Italy. Email: fabrizio@mi.imati.cnr.it
[5]Institute of Mathematical Sciences (ICMAT-CSIC), Madrid, Spain. Email: david.rios@icmat.es

SYSØR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Prediction bands for functional data based on depth measures

A. Elías Fernández [1], R. Jiménez Recaredo [2]

A. Elías Fernández (PhD Student).- Antonio Elías is a Phd student in Statistics at Carlos III Unviersity of Madrid. He completed a postgraduate program in Business and Quantitative Methods after a Degree in Economics. The topic of his PhD research is related to non parametric functional data methods and depth measures.

Suppose a sample of random functions $\{Y_1, \ldots, Y_n\}$ with values on $C(I)$, $I = [a, b]$ and a function Y_{short} that it is observed in a domain such that $I_{short} \subseteq I$. In this work, we deal with the problem of extending the partially observed function Y_{short} to the unobserved domain.

We propose a non parametric methodology for solving this problem motivated by the work of [1]. The novelty of our approach relies on the selection of subsamples that make the function to predict a deep datum in the range of observation. The central regions delimited by the deepest curves of the selected subsamples provide tight bands that wrap not only in the observed part but even in the unobserved domain, preserving also its shape.

The involved subsampling problem is dealt by algorithms specially designed to be used in conjunction with two different tools for computing and visualizing central regions for functional data. Following [3], the first algorithm is based on Tukey's depth and the first two robust principal components scores. This two dimensional feature space allows us to find neighbourhoods or subsamples that surround the curve to predict in a natural way. In contrast, the second algorithm attemps to solve the problem in the functional space by applying functional depth measures and following the functional boxplot of [2].

We present two case studies for putting into practice our methodology. First, we tackle the problem of forecasting the Spanish electricity demand during the last three months of 2016 with a data sample of daily functions from 2014. On other hand, we propose an exercise of missing interval imputation with a data set of Spanish daily temperatures measured in 53 weather stations along the country.

The performance of both algorithms is similar for samples where dimensional reduction does not lead to considerable loss of information. Furthermore, the methodology is easy to adapt to a wide range of processes and the proposed algorithms could be considered with other depth measures or distances.

Key words: depth measures; central regions; functional boxplot; Delaunay triangulation.

Acknowledgements Supported by the Spanish Ministry of Education, Culture and Sport under grant FPU15/00625 and partially supported by the Spanish Ministry of Economic and Competitiveness under grant ECO2015-66593-P.

References

[1] Sugihara, G., May, R. M. (1990). Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, **344**, pp. 734–741.

[2] Sun, Y., Genton, M. G. (2011). Functional boxplots. *J. Comput. Graph. Stat.*, **20**, pp. 316–334.

[3] Hyndman, R. J., Shang, H. L. (2010). Rainbow plots, bagplots, and boxplots for functional data. *J. Comput. Graph. Stat.*, **19**, pp. 29–45.

[1]Department of Statistics, University Carlos III of Madrid, Madrid, Spain. Email: aelias@est-econ.uc3m.es
[2]Department of Statistics, University Carlos III of Madrid, Madrid, Spain. Email: rjjimene@est-econ.uc3m.es

SYSOR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

New heuristic approaches for the Probabilistic p-Center problem

Maria Albareda-Sambola [1] , Luisa I. Martínez-Merino [2] , Antonio M. Rodríguez-Chía[2]

L. I. Martínez-Merino (PhD Student).- Luisa I. Martínez Merino is a PhD student at University of Cádiz. She studied BSc in Mathematics in this university and after that, she studied MSc in Mathematics. Currently, she is working on the Operational Research field. Namely, her thesis is focused on location problems under uncertainty and classification methods. With respect to location problems under uncertainty, she worked with the p-center problem considering uncertainty through a stochastic programming perspective and she is now working on a generalization of covering problems considering multi-periods and uncertainty. Regarding the classification methods, Luisa is also working on new enhancements in Support Vector Machines.

Among the variety of discrete location models, this work focuses on a generalization of the p-Center Problem (pCP). Given a set of sites and the matrix of distances between them, the objective of the pCP is to locate p centers out of n sites minimizing the maximum distance between each site and its closest center, see [1].

A drawback of this kind of models is that, in some cases, focusing on the worst served customer may yield solutions with high average service costs. This work presents a stochastic pCP variant that aims at smoothing this loss of spatial efficiency, trying to keep the centers close to where they are needed and considering how likely each customer is to require the service. This variant is called probabilistic p-center problem (PpCP), see [3].

Given a set of sites $N = \{1, \dots, n\}$, a matrix of distances among these sites, $d_{ij} \geq 0, i, j \in N$, and a vector of demand probabilities $q = (q_i)_{i \in N}$, let ξ be an n-dimensional random vector with marginal Bernoulli distributions defined by q, $\xi_i \sim B(q_i)$ modeling the requests for service of the different sites ($\xi_i = 1$ iif site i requests being served in a given realization). Then, the PpCP is defined as:

$$\min_{\substack{S \subset N \\ |S|=p}} \mathbb{E}_\xi \big(d(N(\xi), S)\big) \tag{1}$$

where for any two subsets $A, B \subseteq N$ the distance $d(A, B)$ is defined as $\max_{i \in A} \min_{j \in B} d_{ij}$ and, for a given realization of the random vector ξ, $N(\xi) = \{j \in N : \xi_j = 1\}$.

We explore not only three formulations for the $PpCP$ but also two heuristic approaches. For the formulations, an ordered objective function has been considered [4]. Moreover, we analyze two heuristics: the Variable Neighborhood Search (VNS) and the Sample Average Approximation (SAA).

The first heuristic approach considered, VNS, has been successfully applied to the Discrete Ordered Median Problem (DOMP), see [2]. The major difference between the PpCP and the DOMP is that, in the case of the PpCP, given a solution S, sorting the service distances associated with the different sites, $\underline{d}_i^S$, is necessary to keep track of what site is associated with each ordered distance, since the values of their weights are not fixed as in the DOMP, but depend on the specific sites with larger associated distances.

On the other hand, Sample Average Approximation (SAA) is a very practical tool in the context of stochastic programming, [5]. The method consists in solving a sequence of PpCP instances, each restricted to a sample of possible scenarios (subsets of customers with demand) each. The optimal value of the stochastic program can be estimated through the average of the optimal values to these instances.

Key words: Discrete location; p-center; demand uncertainty;

Acknowledgements This research has been partially supported by Ministerio de Economía y Competitividad under grants MTM2012-36163- C06-05, MTM2013-46962-C02-02, MTM2015-63779-R, MTM2016- 74983-C2-2-R and by FEDER-Junta de Andalucía under grant FQM 05849.

References

[1] Calik, H., Labbé, M., Yaman, H. (2015). p-Center problems. In Laporte, G., Nickel, S. and Saldanha da Gama, F. editors, *Locat. Sci.*, chapter 4, pp. 79–92.

[2] Dominguez-Marín, P., Nickel, S., Mladenović N. (2005). Heuristic procedures for solving the discrete ordered median problem. *Ann. Oper. Res.*, **136**, pp. 145–173.

[3] Martínez-Merino, L. I., Albareda-Sambola, M., Rodríguez-Chía, A. M. (2017). The probabilistic p-center problem: Planning service for potential customers. *Eur. J. Oper. Res.*, **262**, pp. 509–520.

[4] Nickel, S., Puerto, J. *Facility location: a unified approach.* Springer, 2005.

[5] Kleywegt, A. J., Shapiro, A., Hommem de Mello, T. (2001). The sample average approximation method for stochastic discrete optimization. *SIAM J. Optim.*, **12**, pp. 479–502.

[1]Department of Statistics and O.R., Universitat Politècnica de Catalunya, Terrassa, Spain. Email: maria.albareda@upc.edu
[2]Department of Statistics and O.R., University of Cádiz, Cádiz, Spain. Email: luisa.martinez@uca.es, antonio.rodriguezchia@uca.es

SYSØR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Application of Multi-Objective Constrained Optimisation to Minimise the Expected Loss in Credit Risk

Francisco-Javier Granados-Ortiz [1]

F. J. Granados-Ortiz (Postdoctoral Researcher).- Francisco Javier Granados Ortiz joined in 2013 the University of Greenwich (London, UK) for 3 years as Marie Curie Early Stage Researcher in AeroTraNet2. His role was to address the issues of analysing and handling data when understanding how uncertainties are propagated in the design life cycle of computational simulations of jet flows. Since 2016, he is currently applying and developing his data science skills at the Spanish bank Grupo Cajamar, in order to improve predictive modelling, sampling and optimisation in credit risk portfolios.

Credit Risk is an increasingly important analysis in banks, thanks to the application of advanced predictive modelling procedures and theory to measure the probability of default and risk exposure [1, 2, 3]. Banks have access to large datasets that can be used for modelling purposes, producing quick mathematical evaluations of new and old customers applying for *i. e.* loans or credit cards. However, it does happen often in risk methodologies that the model outcome is incomplete, not taking into account the minimisation of the expected loss in accordance with the quantity of the potential obligor.

In this work, basics on the credit risk modelling workflow will be given. Once these models are available, Multi-Objective Optimisation by means of Genetic Algorithms [4] is applied to several credit risk models subject to some constraints. This is aimed to demonstrate the relevance of this approach in order to minimise the expected loss in SMEs Portfolio.

Key words: Predictive Modelling; Optimisation; Credit Risk; Data Science; Banking

Acknowledgements The credit risk framework has been improved over recent years by all the risk Methodology Unit staff at Grupo Cajamar.

References

[1] Bluhm, C., Overbeck, L., Wagner, C. (2016). *Introduction to Credit Risk Modeling.*, Crc Press.

[2] Hosmer Jr, D. W., Lemeshow, S., Sturdivant, R. X. (2013). *Applied Logistic Regression*. **398**. John Wiley & Sons.

[3] Engelmann, B., Rauhmeier, R. (2006). *The Basel II risk parameters: estimation, validation, and stress testing*. Springer Science & Business Media.

[4] Deb, K., Pratap, A., Agarwal, S., Meyarivan, T. A. M. T. (2002). *A fast and elitist multiobjective genetic algorithm: NSGA-II*. IEEE transactions on evolutionary computation, **6**(2), pp. 182–197.

[1] Data Scientist at Risk Methodology Unit in Banco de Crédito Cooperativo Grupo Cajamar. Email: frangranados@live.com / franciscojaviergranados@bcc.es

SYSOR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Minimum Density Power Divergence Estimators in Loglinear Models with multinomial sampling

A. Calviño[1]

A. Calviño (Assistant Professor).- Aida Calviño received the BSc degree in Statistics with honors from Complutense University of Madrid in 2010, and the MSc degree from University of Santiago de Compostela in 2011. Additionally, she got the PhD degree at University of Cantabria, in 2013 (*International PhD Abertis prize*). She is an Assistant Professor at Complutense University of Madrid, where she is currently doing research in the *Divergence-based Inference Procedures* group. Her research interests include statistics, optimization methods and their applications.

Categorical data analysis is an essential tool when the data are nominal. Even when the data are ordinal, it sometimes makes sense to categorize them into a discrete number, $k > 1$, of classes.

Let $Y_1, Y_2, ..., Y_n$ be a sample of size $n \geq 1$, with realizations from $\mathcal{X} = \{1, ..., k\}$ i.i.d. according to a probability distribution $\boldsymbol{p}(\boldsymbol{\theta}_0)$. This distribution is assumed to be unknown, but belonging to a known family, $\mathcal{P} = \left\{\boldsymbol{p}(\boldsymbol{\theta}) = (p_1(\boldsymbol{\theta}), ..., p_k(\boldsymbol{\theta}))^T : \boldsymbol{\theta} \in \Theta\right\}$, of distributions on $\mathcal{X}$ with $\Theta \subseteq \mathbb{R}^t$, $t < k-1$. Thus, the true value $\boldsymbol{\theta}_0$ of parameter $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^t$ is fixed but unknown. We denote $\widehat{\boldsymbol{p}} = (\widehat{p}_1, ..., \widehat{p}_k)^T$, with $\widehat{p}_j = \frac{N_j}{n}$ and $N_j = \sum_{i=1}^n I_{\{j\}}(Y_i)$, $j = 1, ..., k$, where $I_A(.)$ is the indicator function of A. The statistic $(N_1, ..., N_k)$ is sufficient for the statistical model above and is multinomially distributed.

A class of models often used in this context is the family of loglinear models: $p_u(\boldsymbol{\theta}) = \exp\left(\boldsymbol{w}_u^T\boldsymbol{\theta}\right) / \sum_{v=1}^k \exp\left(\boldsymbol{w}_v^T\boldsymbol{\theta}\right)$; $u = 1, ..., k$. This is the model we consider for the theoretical results in this talk.

Most statistical software use the maximum likelihood method to obtain the estimates (MLE) $\widehat{\boldsymbol{\theta}}$ for the parameter vector $\boldsymbol{\theta}$. However, although it is well-known that it is a BAN (Best Asymptotically Normal) estimator, it is also well-known that the MLE is not a robust estimator in general and in particular in loglinear models. For that reason, in this talk, we aim at introducing *robust inferential procedures* for estimating and testing that will be not sensitive to outliers. Note that, in the context of contingency tables, we deal with outlying cells rather than individual outlying observations contributing to cell counts.

Based on the fact that the MLE can be equivalently defined as the minimization of the Kullback-Leibler divergence, we shall introduce the minimum *Density Power Divergence* (DPD) estimator for loglinear models. The minimum density power divergence estimators (MDPDE) were defined by [1] as a robust and efficient estimation tool. They have been used in different statistical problems (see, for instance, [2]) showing very good robust behavior for both point estimates and test statistics based on it.

The minimum density power-divergence estimator of $\boldsymbol{\theta}$ for the loglinear model previously stated is given by $\widehat{\theta}_{(\beta)} \equiv \arg\min_{\theta\in\Theta} d_\beta(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\theta}))$, where $d_\beta(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\theta})) \equiv \sum_{i=1}^k \left\{p_i(\boldsymbol{\theta})^{1+\beta} - \left(1+\frac{1}{\beta}\right) p_i(\boldsymbol{\theta})^\beta \widehat{p}_i + \frac{1}{\beta}\widehat{p}_i^{\beta+1}\right\}$, for $\beta > 0$, and, for $\beta = 0, d_0(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\theta})) = d_{Kullback}(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\theta}))$.

Estimators based on divergence measures, as well as test statistics based also on them, are not new in loglinear models (see, for instance, [3]). However, in many of those papers ϕ-divergence measures were used. As far as we know, this is the first time that DPD measures are used in loglinear models.

In this talk we will show how to obtain the MDPDEs, as well as Wald-type test statistics based on them for solving the classical test statistics appearing in loglinear models. Additionally, we will study its robust properties both theoretically and through a simulation study. Finally, some numerical examples will be presented to illustrate the proposed methods.

Key words: logistic regression; robustness; density power divergence; Wald-type test statistics.

Acknowledgements The author is indebted to Prof. L. Pardo and N. Martin for their insightful comments.

References

[1] Basu, A., Harris, I. R., Hjort, N. L., Jones, M. C. (1998). Robust and efficient estimation by minimizing a density power divergence. *Biometrika*, **85**, pp. 549–559.

[2] Basu, A., Mandal, A., Martin, N., Pardo, L. (2016). Generalized Wald-type tests based on minimum density power divergence estimators. *Statistics*, **5**, pp. 1–26.

[3] Cressie, N., Pardo, L. (2000). Minimum $\phi-$divergence estimator and hierarchical testing in loglinear models. *Statistica Sinica*, **10**, pp. 867–884.

[1]Department of Statistics and O.R. III, Complutense University of Madrid, 28040 Madrid, Spain. Email: aida.calvino@ucm.es

SYSOR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Statistical methods to improve estimates obtained with probabilistic and non-probabilistic samples

Ramón Ferri-García [1] , María del Mar Rueda[1]

R. Ferri-García (PhD Student).- Ramón Ferri-García received the BSc degree in Statistics and the MSc degree in Data Science and Computer Engineering from University of Granada (Spain), in 2014 and 2015, respectively. He is currently a PhD candidate at the Department of Statistics and Operations Research of University of Granada. His research interests include survey sampling in finite populations, survey methodology and design, simulation, predictive modelling and its applications in sampling theory.

The development of new technologies in the last decades has made a considerable impact on survey methods worldwide. New forms of questionnaire administration, such as online or smartphone surveys, are replacing more traditional forms such as face-to-face or landline telephone surveys. The advantages of new methods are numerous: critical reduction of costs and a wider range of possibilities for questionnaires can be mentioned as some of them. However, they also raise serious issues on non-sampling errors. Particularly, online and smartphone surveys often suffer from bias caused by non-probabilistic sampling methods, lack of coverage because of the absence of sampling frames, and non-response issues.

In recent years, there has been an increasing development of techniques to deal with the mentioned issues. Regarding non-probabilistic samples, calibration weighting [1] and Propensity Score Adjustment (PSA) [3] have been proposed in order to reduce the impact of such sampling. Current research proves that both methods and its combination can be effective to reduce the bias but at the cost of increasing the variance of the estimators. To deal with coverage issues, literature suggests the use of multiple frames [2] to combine the coverage of several frames to reach an adequate efficiency for estimates. Since its initial development, literature has mainly considered the dual frame situation, while the case with more than two frames has mostly been ignored. Non-response treatment has been widely studied as it concerns all kind of surveys, and the proposed methods focus on reweighting [5] and missing data imputaton [4].

The aim of the present research is to develop the existing techniques to deal with the mentioned issues in online and smartphone surveys, which involves adapting the different techniques to the context where these surveys are applied. In this sense, the research is focused on the study of the bias produced by self-selected samples and the calibration procedures used for its removal, as well as its software implementation. In addition, the extension of multiple frames theory to the case where three or more frames are available is considered, with an analysis of its implications on nexus sampling in social networks. Other objectives include the development of estimation techniques to overcome non-response bias with new survey methods, and the comparison between smartphone surveys and other survey methods in terms of efficiency.

Research findings have proved so far that the combination of PSA with calibration weighting efficiently reduces the bias of non-probabilistic sampling when the self-selection mechanism is exogenous to the target variable. It has been also found that the increase in variance gets reduced as the reference sample increases in size, and it remains stable even if population estimates for auxiliary information totals are used instead of actual population totals, as long as these estimates are unbiased.

Key words: online surveys; bias; non-probabilistic sampling; multiple frames; non-response.

References

[1] Deville, J. C., Sarndal, C. E. (1992). Calibration Estimators in Survey Sampling. *J Am Stat Assoc*, **87**, pp. 376–382.

[2] Hartley, H. O. (1962). Multiple frame surveys. In *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 203–206.

[3] Lee, S. (2006). Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys. *J Off Stat*, **22**, pp. 329–349.

[4] Rubin, D. B. (1996). Multiple imputation after 18+ years. *J Am Stat Assoc*, **91**, pp. 473–489.

[5] Särndal, C. E., Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Wiley, New York 2005.

[1]Department of Statistics and O.R., University of Granada, Granada, Spain. Email: rferri@ugr.es, mrueda@ugr.es

SYSOR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Black-box optimization of expensive functions using Bayesian optimization with a novel batch acquisition algorithm

C. Domínguez Bravo [1] , J. Lozano [2]

C. Domínguez-Bravo (Postdoctoral Researcher).- Since February 2016, Carmen-Ana Domínguez-Bravo is working as a Postdoctoral Fellow at BCAM-Basque Center for Applied Mathematics- in the Machine Learning line. She obtained her PhD in January 2016 at the Institute of Mathematics of University of Sevilla. During her PhD studies, she developed heuristic methods to address an optimization problem coming from the solar industry. This work was done in collaboration with a Spanish solar company called *Abengoa Solar*. She has studied an MSc and BSc in Mathematics at University of Sevilla, with one Erasmus-studio year at Pierre et Marie Curie University (Paris VI).

The problem of optimizing black-box expensive functions appears in many practical situations. Bayesian optimization methodology has been successfully applied to solve them using Gaussian processes as surrogate models.

Following this method, the original minimization problem is transformed into an iterative adaptive process where the points to be observed with the black-box function are obtained by maximizing an auxiliary function called acquisition function. Of course, this auxiliary optimization problem can be solved easily applying standard techniques.

Different single–point acquisition functions appear in the literature considering a different trade-off between exploration (select points with high uncertainty) and exploitation (select points with high expected value). Nowadays, the technology development allows in some cases to perform evaluations of the expensive function in parallel at the same cost as a single evaluation. Therefore, there is a need of developing batch acquisition criteria to take advantage of this parallelism.

In this presentation we will introduce a novel batch acquisition algorithm which selects the batch of points in two separated steps. Firstly, the Pareto–optimal set of the bi– objective problem defined by maximizing the variance and minimizing the mean of the model is approximated by means of an evolutionary multi-objective algorithm. Then, the batch is extracted from the approximated Pareto set as the set of points that maximize the mutual information with respect to the black–box function. Some variants of the algorithm will be also presented.

We compare our strategies with state-of-the-art approaches in benchmark functions, obtaining better results. For a brief bibliography on the topic see: [1, 2, 3, 4]

Key words: black–box expensive optimization; Bayesian optimization; batch acquisition function; multi–objective optimization.

Acknowledgements This research has been mainly supported by the Spanish Ministry of Economy and Competitiveness MINECO (BCAM Severo Ochoa excellence accreditation SEV-2013-0323) and the Basque Government (BERC 2014-2017 program and ELKARTEK).

References

[1] Brochu, E., Cora, V. M., de Freitas, N. (2010). *A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning.*

[2] Contal, E. (2016). *Statistical learning approaches for global optimization.* Université Paris-Saclay.

[3] Krause, A., Singh, A., Guestrin, C. (2008). Near-Optimal Sensor Placements in Gaussian Processes: Theory, Efficient Algorithms and Empirical Studies. *Journal of Machine Learning Research*, **9**, pp. 235–284.

[4] Rasmussen, C. E. (2006). *Gaussian Processes Covariance Functions and Classification.* Max Planck Institute for Biological Cybernetics.

[1]BCAM–Basque Center for Applied Mathematics, Bilbao, Spain. Email: cdominguez@bcamath.org
[2]BCAM & University of the Basque Country UPV/EHU, Donostia, Spain. Email: ja.lozano@ehu.eus

SYSØR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Nonparametric techniques for assessing the number of modes

J. Ameijeiras–Alonso [1] , R. M. Crujeiras[1] , A. Rodríguez–Casal[1]

J. Ameijeiras-Alonso (PhD Student).- José Ameijeiras Alonso is a PhD student at the Department of Statistics, Mathematical Analysis and Optimization of Santiago de Compostela University. His main research concerns the development of statistical methodologies for determining the number of modes and whether the underlying distribution presents symmetry or not, for linear and circular data. His main background is in nonparametric techniques, but he is also interested on the Le Cam methodology and on the Hidden Markov Models. During the last years, his main focus of application was related with the study of wildfires.

Probability density functions fully describe the behaviour of random variables, although in many cases, a complete characterization is not necessary and the main interest is to determine which variable values are most likely. For continuous random variables, this procedure is just the identification of the density modes (local maxima of the probability density function). There are several alternatives for detecting modes in scalar variables. While there are various exploratory tools, these do not provide a formal statistical hypothesis test. Within this context, it is worth mentioning the critical bandwidth and also those tests based on the excess mass or on the dip. However, the poor calibration in the multimodal cases provides unsatisfactory results in practice. A review on the different methods for determining the number of modes and the design of an appropriate calibration procedure in practice based on the excess mass and an application of this new method are the objective of this work.

Key words: Exploratory Tools; Kernel Density Estimation; Multimodality; Nonparametric; Testing Procedure

Acknowledgements The authors gratefully acknowledge the support of Projects MTM2016–76969–P (Spanish State Research Agency, AEI) and MTM2013–41383–P (Spanish Ministry of Economy, Industry and Competitiveness), both co–funded by the European Regional Development Fund (ERDF), IAP network from Belgian Science Policy. Work of J. Ameijeiras-Alonso has been supported by the PhD Grant BES–2014–071006 from the Spanish Ministry of Economy, Industry and Competitiveness.

[1]Department of Statistics, Mathematical Analysis and Optimization, Universidade de Santiago de Compostela, A Coruña, Spain. Email: jose.ameijeiras@usc.es, alberto.rodriguez.casal@usc.esrosa.crujeiras@usc.es

SYSOR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

New valid inequalities for a class of p-hub location problems

Mercedes Landete [1] , Ángel Corberán [2] , Francisco Saldanha-da-Gama [3] , Juanjo Peiró [4]

J. Peiró (Assistant Professor).- Juanjo Peiró has been recently appointed as an Assistant Professor at the Department of Statistics and Operations Research at University of Valencia (Spain). Previously to this, he studied Business Administration, and MSc and PhD degrees in Operational Research. His research interests involve Operational Research and Management Science in general and Combinatorial Optimization in particular, with special focus in network design, routing and facility location problems, all of them tackled with metaheuristics and exact algorithms that are based on computational geometry.

p-hub location problems in transportation networks are $\mathcal{NP}$-hard combinatorial optimization problems with many industrial applications. In the $r-$allocation variant [7], three optimization subproblems are involved: a service facility location problem, an assignment problem, and a routing problem.

In this work we focus on finding new valid inequalities for this variant. Some of them have been adapted from inequalities proposed for related problems ([1], [2], [3], [4], [5]) while other inequalities are new contributions.

The intersection of many of them defines a *set packing* polyhedron [6], which has an associated *conflict graph* that we have studied in order to generate new valid inequalities for the problem, especially those of the *clique* and *odd-hole* classes.

Computational results will be provided showing that the new inequalities help in strengthen the linear relaxation of the original formulation.

Key words: hub location; non-stop services; clique facets; odd holes; set packing problem.

Acknowledgements This work was supported by the Spanish Ministerio de EconomÃa y Competividad and Fondo Europeo de Desarrollo Regional (MINECO/FEDER) (project TIN-2015-65460-C02-01, project MTM-2015-68097, and PhD. grant BES-2013-064245), and by the Generalitat Valenciana (project Prometeo 2013/049). Francisco Saldanha-da-Gama was supported by the Portuguese Science Foundation (FCT—Fundação para a Ciência e Tecnologia) under the project UID/MAT/04561/2013 (CMAF-CIO/FCUL). All this support is gratefully acknowledged.

References

[1] Baumgartner, S. (2003) *Polyhedral analysis of hub center problems*. Diploma thesis, Universitat Kaiserslautern.

[2] García, S., Landete, M., Marín, A. (2012). New formulation and a branch-and-cut algorithm for the multiple allocation p-hub median problem. *European Journal of Operational Research*, **220**, pp. 48–57.

[3] Marín, A. (2005). Uncapacitated euclidean hub location: Strengthened formulation, new facets and a relax-and-cut algorithm. *Journal of Global Optimization* **33**, pp. 393–422.

[4] Marín, A. (2005). Formulating and solving splittable capacitated multiple allocation hub location problems. *Computers & Operations Research*, **32**, pp. 3093–3109.

[5] Marín, A., Cánovas, L., Landete, M. (2006). New formulations for the uncapacitated multiple allocation hub location problem. *European Journal of Operational Research*, **172**, pp. 274–292.

[6] Padberg, M. W. (1973). On the facial structure of set packing polyhedra. *Mathematical Programming*, **5**, pp. 199–215.

[7] Yaman, H. (2011). Allocation strategies in hub networks. *European Journal of Operational Research*, **211**, pp. 442–451.

[1]Departamento de Estadística, Matemáticas e Informática. Instituto Centro de Investigación Operativa. Universidad Miguel Hernández. Elche, Spain. Email: landete@umh.es

[2]Departament d'Estadística i Investigació Operativa. Universitat de Valencia. Valencia, Spain. Email: angel.corberan@uv.es

[3]Departamento de Estatística e Investigação Operacional. Centro de Matemática, Aplicações Fundamentais e Investigação Operacional. Universidade de Lisboa. Lisboa, Portugal. Email: fsgama@ciencias.ulisboa.pt

[4]Departament d'Estadística i Investigació Operativa. Universitat de Valencia. ValÃšncia, Spain. Email: juanjo.peiro@uv.es

SYSOR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Functional Surface Classification Using Neigbourhood Information with Applications to Facial Shapes

P. Montero Manso [1]

P. Montero Manso (PhD Student).- Pablo Montero Manso is a PhD Student at University of A Coruña. His research interests include both Supervised and Unsupervised Statistical Learning, Time Series Analysis and Big Data. He is exploring the use of distances/dissimilarities to approach this problems in complex objects such as time series, shapes or functional data.

Statistical analysis of shapes has attracted a great amount of research [3], mostly focused on 2-dimensional shapes due to their simplicity. The analysis of 3-dimensional shapes (surfaces) has received less attention due to the high cost of surface information acquisition devices, such as stereo cameras and laser scanners. The dramatic reduction on the cost of these devices and the availability of smartphones with 3D cameras forecasts a wealth of surface data, increasing the need and interest of analysis techniques for this kind of data.

One of the most studied problems regarding surfaces is the analysis of human faces and is therefore a good testbed for new methods in shape analysis. Applications range from medicine to marketing, such as [4], where significant differences are found in face shape between patients with a certain genetic disorder and the control group, or in [5], where face shape information is proposed for modeling consumer behavior.

Traditional approaches to shape analysis are based on landmarks, i.e. a human annotates the positions of important features such as the position of the centers of the eyes or the corners of the lips in the case of facial shapes, and then standard multivariate analysis is performed. The landmark-based approach has several drawbacks, such as the requirement of human intervention, preventing fully automated systems to be developed. The landmark acquisition may not even be possible because of noisy acquisition, such as occlusion due to the pose of the object. Additionally, a huge amount of information is being discarded in the process. On the other hand, the methods proposed to overcome these issues lack in interpretability.

Recent advances in surface [1] and general shape analysis [2] from the functional data analysis point of view show promising results in supervised analysis tasks such as classification. In this work, we extend the landmark-free method proposed in [2] for shape classification. The approach works by sampling points inside a given surface on a first step and then the neighbourhood of each sampled point is sampled again on a second step. Both sampling schemes are data-driven in order to increase classifaction acuracy and reduce computing time. The surface is then interpreted as the multivariate distribution function generating these sampled points and compared to other surfaces using a distance between distributions on a two step process. We use face shape data as the focus of our experiments.

Key words: Shape; Surface; Classification; Functional Data; Distance.

Acknowledgements This work is supported by the Spanish Ministerio de EconomÃa y Competitividad grant MTM2014-52876-R and Xunta de Galicia ED431C 2016-015.

References

[1] Álvarez-Liébana, J., Ruiz-Medina, M. (2015). Functional statistical classification of non-linear dynamics and random surfaces roughness in control systems. *International Journal of Mathematical Models and Methods in Applied Sciences*, **9**, pp. 1–20.

[2] Berrendero, J. R., Cuevas, A., Pateiro-López, B. (2016). Shape classification based on interpoint distance distributions. *J. Multivariate Anal.*, **146**, pp. 237–247.

[3] Dryden, I. L. Mardia, K. V. (2016). *Statistical Shape Analysis: With Applications in R.* John Wiley & Sons, New York.

[4] Prasad, S., Katina, S., Hennessy, R. J., Murphy, K. C., Bowman, A. W., Waddington, J. L. (2015). Craniofacial dysmorphology in 22q11. 2 deletion syndrome by 3d laser surface imaging and geometric morphometrics: illuminating the developmental relationship to risk for psychosis. *American Journal of Medical Genetics Part A*, **167**, pp. 529–536.

[5] Zhong, H., Xiao, J. (2015). Big data analytics on customer behaviors with kinect sensor network. *International Journal of Human Computer Interaction*, **6**, p. 36.

[1]MODES Group, Department of Mathematics, University of A Coruña, Spain. Email: p.montero.manso@udc.es

SYSOR 2017
1ST SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Visualizing dynamic data: A Mathematical Optimization approach

E. Carrizosa[1] , V. Guerrero[2] , D. Romero Morales[3]

V. Guerrero (Assistant Professor).- Vanesa Guerrero was born in Guadalcanal, Sevilla (1989). She received her BSc (2012), MSc (2013) and PhD (2017) in Mathematics at University of Sevilla. She is currently working as an Assistant Professor at Carlos III University of Madrid (UC3M). Her research aims to enhance interpretability in complex datasets by combining methods of Statistical Data Analysis and Mathematical Optimization. In particular, she has worked in developing sparse models for Principal Components Analysis and in developing new visualization frameworks which faithfully preserve the inherent data structure.

The usefulness of Information Visualization lies with its power to improve interpretability of the unknown phenomena described by raw data to aid decision making. In particular, datasets involving time-varying frequency distributions and proximity relations are the ones studied in this work. In order to visualize this structured data, we develop a visualization framework which extends the standard Multidimensional Scaling and has a global optimization model at its heart. Difference of Convex functions and Nonconvex Quadratic Binary Optimization techniques are combined as a solution approach. Our methodology is illustrated using a dynamic linguistic real-world dataset.

Key words: Visualization; Dynamic data; Multidimensional Scaling; Difference of Convex Algorithm

References

[1] Carrizosa, E., Guerrero, V., Romero Morales, D. (2017) Visualizing data as objects by DC (difference of convex) optimization. *Mathematical Programming* (in press). `https://doi.org/10.1007/s10107-017-1156-1`.

[2] Carrizosa, E., Guerrero, V., Romero Morales, D. (2017) Visualization of complex dynamic datasets by means of Mathematical Optimization. *Submitted*.

[3] Carrizosa, E., Guerrero, V., Hardt, D., Romero Morales, D. (2017) On Building Online Visualization Maps for News Data. *Submitted*.

[1]Department of Statistics and Operations Research, Universidad de Sevilla, Sevilla, Spain. Email: ecarrizosa@us.es
[2]Department of Statistics, Universidad Carlos III de Madrid, Madrid, Spain. Email: vanesa.guerrero@uc3m.es
[3]Department of Economics, Copenhagen Business School, Frederiksberg, Denmark. Email: drm.eco@cbs.dk