SYSORM 2022

sysorm.umh.es

3RD SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING (SYSORM 22) **Conference proceedings**

Elche, 21st-23rd of September 2022

UNIVERSITAS Miguel Hernández RESEARCH INSTITUTE

Seio
Sociedad de Estadística e Investigación Operativa

Contents

Committees

Editors: Organizing Committee

Organizing Committee

M. Carmen Aguilera Morillo (**Chair**) - Universitat Politècnica de València

Javier Álvarez Liébana - Universidad Complutense de Madrid

Elena Castilla González - Universidad Rey Juan Carlos

Eduardo García Portugués - Universidad Carlos III de Madrid

Vanesa Guerrero Lozano - Universidad Carlos III de Madrid

Harold Hernández Roig - Universidad Carlos III de Madrid

Juan Carlos Laria de la Cruz - Universidad Carlos III de Madrid

Álvaro Méndez Civieta - Universidad Carlos III de Madrid

Carmen Minuesa Abril - Universidad de Extremadura

Beatriz Sinova Fernández - Universidad de Oviedo

Scientific Committee

Elena Castilla González - Universidad Rey Juan Carlos

Eduardo García Portugués - Universidad Carlos III de Madrid

Vanesa Guerrero Lozano - Universidad Carlos III de Madrid

Carmen Minuesa Abril - Universidad de Extremadura

Beatriz Sinova Fernández - Universidad de Oviedo

Local Committee

Juan Aparicio Baeza - Universidad Miguel Hernández

Lidia Ortiz Henarejos - Universidad Miguel Hernández

José Luis Sainz-Pardo Auñón - Universidad Miguel Hernández

Laura Antón Sánchez - Universidad Miguel Hernández

Sixto Alonso Mateu - Universidad Miguel Hernández

Daniel Valero Carreras - Universidad Miguel Hernández

Víctor Javier España Roch - Universidad Miguel Hernández

ORGANIZER

SPONSORS

Schedule

Wednesday, September 21

9.00 - 9.30	Opening Session
9:30 - 10:30	Plenary session 1:
	Claudia D'Ambrosio
10.30 - 11.00	Coffee break
11.00 - 12.20	Session 1
12.20 - 12.35	Short Break
12:35 - 13.35	Session 2
13.35 - 15.50	Lunch
15.50 - 17.10	Session 3

Thursday, September 22

9:30 - 10:30	Plenary session 2:
	Rubén Ruiz
10.30 - 11.00	Coffee break
11.00 - 12.20	Session 1
12.20 - 12.35	Short Break
12:35 - 13.35	Session 2
13.35 - 16.00	Lunch
16.00 - 17.00	Session 3

Friday, September 23

9:30 - 10:30	Plenary session 3:
	Ricardo Cao
10.30 - 11.00	Coffee break
11.00 - 12.20	Session 1
12.20 - 12.35	Short Break
12:35 - 13.35	Session 2
13.35 - 16.00	Lunch
16.00 - 17.00	Plenary session 4:
	Frederic Ferraty
17.00 - 17.15	Short break
17.15 - 18.00	Closing session

Plenary speakers

Speakers

Urban air mobility: Models and algorithms for tactical deconfliction

Claudia D'Ambrosio * Mercedes Pelegrín † Rémi Delmas ‡ Youssef Hamadi §

C. D'Ambrosio (Directeure de Recherche CNRS) Claudia D'Ambrosio is a research director at CNRS (France) and an adjunct professor at École Polytechnique (France). She is the head of the International Academic and Research Chair "Integrated Urban Mobility". She holds a Computer Science Engineering Master Degree and a Ph.D. in Operations Research from University of Bologna (Italy). Her research speciality is mathematical optimization, with a special focus on mixed integer nonlinear programming. During her whole carrier, she was involved both in theoretical and applied research projects. She was awarded the EURO Doctoral Dissertation Award for her Ph.D. thesis on "Application-oriented Mixed Integer Non-Linear Programming" and the 2nd award "Prix Robert Faure" (3 candidates are awarded every 3 years) granted by ROADEF society.

In this talk, we focus on how Mathematical Optimization (MO) could help build the bricks to develop Urban Air Mobility (UAM). In particular, we focus on passengers' transportation via eVTOLs, i.e., electric flying vehicles that will allow exploiting the sky to help smooth ground traffic in densely populated areas. Clearly, one of the biggest challenges in UAM is to ensure safety. From an operational viewpoint, the flights planning can be highly affected by different kinds of disruptions, which have to be solved at the tactical deconfliction level. Inspired by the classical aircraft deconfliction (see, for example, [1, 2, 3]), we propose a MO formulation based on a mathematical definition of vehicles separation, specialized in the UAM context. The deconfliction is based on speed changes or delayed takeoff, when possible. To the best of our knowledge, our MO model is the first that considers the whole set of conflicts at the same time. In the computational study, we, thus, compare our approach against a variant considering only pairwise conflicts, on three sets of realistic scenarios. More details can be found in Pelegrín et. al [4].

Keywords: Air traffic management; Reoptimization; Mixed integer linear programming.

Acknowledgements Supported by the Chair "Integrated Urban Mobility", backed by L'X - École Polytechnique and La Fondation de l'École Polytechnique. The Partners of the Chair shall not under any circumstances accept any liability for the content of this publication, for which the author shall be solely liable.

References

[1] Pelegrín, M. and D'Ambrosio, C. (accepted). Aircraft deconfliction via Mathematical Programming: Review and insights. *Transportation Science*.

[2] Cerulli, M., Pelegrín, M., Cafieri, S., D'Ambrosio, C., and Rey, D. (accepted). Aircraft deconfliction. *Encyclopedia of Optimization*.

[3] Cerulli, M., D'Ambrosio, C., Liberti, L., and Pelegrín, M. (2021). Detecting and solving aircraft conflicts using bilevel programming. *Journal of Global Optimization*, 81:529–557. doi:10.1007/s10898-021-00997-1.

[4] Pelegrín, M., D'Ambrosio, C., Delmas, R., and Hamadi, Y. (2021). Urban air mobility: From complex tactical conflict resolution to network design and fairness insights. https://hal.archives-ouvertes.fr/hal-03299573.

*LIX - CNRS, École Polytechnique, Institut Polytechnique de Paris, France. Email: dambrosio@lix.polytechnique.fr.
†LIX - CNRS, École Polytechnique, Institut Polytechnique de Paris, France. Email: pelegrin@lix.polytechnique.fr.
‡Integrated Urban Mobility chair, France.
§Integrated Urban Mobility chair, France.

Optimizing Amazon Elastic Compute Cloud. How Operations Research improves instance placement

Rubén Ruiz *

Rubén Ruiz (Professor) is currently a Principal Applied Scientist at Amazon Web Services (AWS) and Professor (on leave of absence) of Statistics and Operations Research at the Polytechnic University of Valencia, Spain. He is co-author of more than 100 papers in International Journals and has participated in presentations of more than 200 papers in national and international conferences. He is editor of Elsevier's Operations Research Perspectives (ORP) and co-editor of the European Journal of Industrial Engineering (EJIE), both journals listed in the Journal Citation Reports (JCR). He is also associate editor of other important journals like TOP as well as member of the editorial boards of several journals, most notably European Journal of Operational Research and Computers and Operations Research. His research interests revolve around cloud computing optimization, scheduling and logistic problems in real environments.

Cloud computing is a paradigm where users employ computing resources through Internet in a pay-per-use or pay-as-you-go basis. Resources typically entail data storage (cloud storage) and computing power. Instead of purchasing and operating expensive servers, users only pay for what they use on-demand, significantly reducing capital expenses. Other important problems are also removed from the equation as are for example maintaining (updates, patching, security) resources or hiring personnel for such operations.

Amazon Web Services (AWS) is the world's most comprehensive and broadly adopted cloud platform, offering over 200 fully featured services from data centers globally. Millions of customers –including the fastest-growing startups, largest enterprises, and leading government agencies– are using AWS to lower costs, become more agile, and innovate faster. AWS has the largest and most dynamic community, with millions of active customers and tens of thousands of partners globally. Customers across virtually every industry and of every size are running every imaginable use case on AWS.

One big advantage of cloud computing services like AWS is elasticity. When running computing resources on-premise, users have to either provision servers for peak demand, which results in lower utilization of resources off-peak and large capital expenses, or provide a lower than ideal service with limited resources during peaks. With AWS, users can elastically provision resources as they are needed, just paying for whatever resources are used at any moment.

One of the main services in AWS is Amazon Elastic Compute Cloud (Amazon EC2) where users can rent virtual servers on-demand in a reliable, scalable, secure and optimized way. Amazon EC2 has millions of users around the globe. Companies like Netflix, Twitch, LinkedIn, Meta (Facebook) or BBC use EC2 everyday.

Some interesting problems arise within EC2. Servers can run different types of virtual machines and within any server, several virtual machines might be running at any given time. Servers come in different sizes and capacities and Amazon EC2 offers more than 500 different types of virtual machines (called "instances"). Servers have a myriad of limited resources, like RAM, number of cores or CPUs, Graphics cards (GPUs), disks, network capacities, etc. Placing instances at servers randomly results in fragmentation, where servers are not completely full and incoming instances of large size might not fit. Other problems arise as well since the best servers for some specific instances might not be available if a non-optimized placement is carried out. To counter these problems, more servers have to be provisioned, which means higher operational costs that have to be translated to clients. The interested reader may have identified a potential for optimization. Indeed, the instance placement problem is a form of a generalized bin packing problem or, more specifically, a multi-dimensional vector bin packing problem where bins are heterogeneous. Of course, there are many additional constraints and implications, resulting in a rich problem of gargantuan size that is very challenging to optimize.

In this talk we will introduce AWS and Amazon EC2 with some interesting use cases. We will later define some optimization problems in instance placement and sketch some of the methods we are employing for their resolution. The talk will not disclose proprietary details or specifics that might disclose intellectual property, but will have enough depth so as to grasp how much Operations Research can bring to cloud computing.

Keywords: Cloud computing; virtual machine placement; multi-dimensional heterogeneous vector bin packing.

*Principal Applied Scientist. Amazon Web Services. EC2 Placement Quality. Email: rruizg@amazon.es

Nonparametric inference for big-but-biased data

Ricardo Cao * Laura Borrajo †

R. Cao Ricardo Cao is Professor of Statistics and Operations Research at the Universidade da Coruña (Spain). His main interests are Nonparametric Statistics, Bootstrap, Survival Analysis, Functional Data Analysis, Big Data Statistical Analysis, Dependent Data, Empirical Likelihood, Credit Risk and Statistical Methods in Genomics, Neuroscience, Epidemiology and Weed Science. He received his Bachelor Degree in Mathematics in 1988 and Doctor Degree in Mathematics in 1990 both from Universidade de Santiago de Compostela (Spain). Ricardo Cao has been the (Co-)Editor-in-Chief of the journals TEST (2009-2012), Computational Statistics (2016-2018) and Journal of Nonparametric Statistics (2019-2021).

It is often believed that in a Big Data context, given the large amount of data available, the data reflect precisely the underlying population. However, the data are often strongly biased due to the procedure used for obtaining them. In order to reduce the significant bias that may appear in Big Data (Big-but-Biased Data, B3D), different testing methods for bias detection are used and completely nonparametric estimation methods for bias correction are proposed for large-sized but possibly biased samples.

Nonparametric estimators for the mean of a transformation of the random variable of interest are considered when the bias mechanism is known ([1]). When ignoring the biasing weight function, two different setups are proposed. In Setup 1 ([3]) a small-sized simple random sample of the real population is assumed to be additionally observed, while in Setup 2 it is assumed that a twice biased sample of small size is observed. The asymptotic properties of the proposed estimators are extensively studied under suitable limit conditions on the small and the large sample sizes and standard and non-standard asymptotic conditions on the two bandwidths. The performance of the proposed nonparametric estimators is compared with the classical estimators based on the two samples involved in each setup through Monte Carlo simulation studies. Simulation results show that the new mean estimators outperform the classical empirical means for suitable choices of the two smoothing parameters involved. The influence of these smoothing parameters on the performance of the final estimators is also studied, exhibiting a striking limit behaviour of their optimal values. In addition, bootstrap bandwidth selection methods for each nonparametric mean estimator are introduced.

Finally, the proposed techniques are applied to several real data sets from different areas, including business intelligence ([3]) and air quality ([2]).

Keywords: Bandwidth selection; Big data; Kernel density estimation; Large sample size; Sampling bias.

Acknowledgements Laura Borrajo research was sponsored by the Xunta de Galicia predoctoral grant (with reference ED481A-2016/367) for the universities of the Galician University System, public research organizations in Galicia and other entities of the Galician R&D&I System, whose funding comes from the European Social Fund (ESF) in 80% and in the remaining 20% from the General Secretary of Universities, belonging to the Ministry of Culture, Education and University Management of the Xunta de Galicia. Her research was also found by a Vodafone Campus Lab grant. Both authors acknowledge partial support by MINECO Grants MTM2017 -82724-R and PID2020-113578RB-I00, by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and ED431C-2020-14, Centro Singular de Investigación de Galicia ED431G/ 01 and Centro de Investigación del Sistema Universitario de Galicia ED431G 2019/ 01), all of them through the European Regional Development Fund (ERDF).

References

[1] Cao, R. and Borrajo, L. (2018). Nonparametric mean estimation for big-but-biased data. In E. Gil, E. Gil, J. Gil, and M. Á. Gil (Eds.), *The Mathematics of the Uncertain* (pp. 55–65). Springer International Publishing. doi:10.1007/978-3-319-73848-2_5

[2] Borrajo, L. and Cao, R. (2020). Big-but-biased data analytics for air quality. *Electronics*, 9:1551. doi:10.3390/electronics9091551.

[3] Borrajo, L. and Cao, R. (2021). Nonparametric estimation for big-but-biased data. *TEST*, 30:861–883. doi:10.1007/s11749-020-00749-5.

*Research Group on Modeling, Optimization and Statistical Inference (MODES), Department of Mathematics, Center for Information and Communication Technologies (CITIC), Universidade da Coruña, Spain. Email: rcao@udc.es.

†Research Group on Modeling, Optimization and Statistical Inference (MODES), Department of Mathematics, Center for Information and Communication Technologies (CITIC), Universidade da Coruña, Spain. Email: laura.borrajo@udc.es.

Scalar-on-function local linear regression and beyond

Frederic Ferraty * Stanislav Nagy †

F. Ferraty Frédéric Ferraty is Professor of Statistics at the Department of Computer Science and Mathematics of the University Toulouse Jean Jaurès, France. He is also a permanent member of the Toulouse Mathematics Institute. His research focuses on statistical modelling of high-dimensional data. In particular, he is one of the world leaders in functional data analysis (FDA) with more than 70 references and 1500 citations according to "MathSciNet".

Local linear regression is one of the most popular nonparametric regression method when the predictor is a finite-dimensional covariate. It is well known that the local linear regression outperforms the usual kernel estimator and the literature dealing with this topic is huge. To our knowledge (and surprisingly) there are only two papers extending the local linear regression to the situation when one considers a functional predictor. Problem: the theoretical developments of one of these works is approximative where in the second one, the authors require strong assumptions with respect to the distribution of the functional predictor. Even if the infinite-dimensional feature of the predictor makes challenging the asymptotics in the functional local linear regression, it is clear that this topic is still underdeveloped. So this talk aims to bring a relevant response by proposing new theoretical developments. As a by-product, we also provide the asymptotics for the Fréchet derivative of the functional linear regression operator. On simulated datasets we illustrate good finite sample properties of the proposed methods. On a real data example of a single-functional index model we indicate how the functional derivative of the regression operator provides an original, fast, and widely applicable estimation method.

Keywords: Functional data; Functional derivative of regression operator; Functional index model; Local linear regression; Scalar-on-function model.

References

[1] Ferraty, F. and Nagy, S. (2022). Scalar-on-function local linear regression and beyond. *Biometrika*, 109(2):439–455. doi:10.1093/biomet/asab027.

*Toulouse Mathematics Institute, University of Toulouse, France. Email: ferraty@math.univ-toulouse.fr.
†Department of Probability and Mathematical Statistics, Charles University, Czech Republic. Email: nagy@karlin.mff.cuni.cz.

Session 1 (Wednesday 11:00)

Session talks

Weighted Bayesian regression for selection bias

Hieu Cao * M. Remedios Sillero-Denamiel † Simon Wilson ‡

M. Remedios Sillero-Denamiel (Postdoctoral Researcher) M. Remedios Sillero-Denamiel studied the BSc and MSc in Mathematics at the University of Seville. In 2017, she was awarded a scholarship to pursue a PhD within the Department of Statistics and Operations Research of University of Seville, and she obtained the PhD degree in Mathematics in July 2021. Since September 2021, she has been employed as a postdoctoral researcher at Trinity College Dublin. Her research bridges the disciplines of Operations Research and Statistics to develop novel computational methods for the extraction of knowledge from current datasets describing timely and relevant real-world situations.

In the regression setting, it is typically assumed that training and test sets follow similar distributions, but that is not always true, as is the case with the sky surveys of galaxies where faint ones are not observed in favour of brighter ones. In addition, when data follow complicated non-Gaussian distributions, the full conditional density has to be estimated to properly quantify the uncertainty in the predictions [1]. In this work, we develop a Bayesian approach [2] to estimate the conditional density under selection bias. Concretely, a non-linear relationship between the response variable and the predictors is considered, and the distance between a sample and a distribution is taking into account when building the regression model to improve the performance results for data that we are interested in predicting. Thus, we present a Bayesian framework for supervised learning that accounts for differences in training and test data, and can handle highly asymmetric and multimodal distributions. This methodology has a direct application in galaxies' redshift (z) estimation, which is an important measure to explain the structure and evolution of the universe [1, 3]. Finally, empirical studies concerning the new method will be shown.

Keywords: Non-linear regression models; Selection bias problem; Bayesian inference; Photo-z estimation

Acknowledgements This research is supported by the research project 12/RC/2289_P2 (Insight Centre for Data Analytics (Ireland)). This support is gratefully acknowledged.

References

[1] Izbicki R., Lee, A.B. and Freeman, P.E. (2017). Photo-z estimation: An example of nonparametric conditional density estimation under selection bias. *Annals of Applied Statistics*, 11(2):698–724. doi:10.1214/16-AOAS1013.

[2] Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *Bayesian Data Analysis* (1st ed.). Chapman and Hall/CRC. doi:10.1201/9780429258411.

[3] Almosallam, I. A., Jarvis, M. J. and Roberts, S. J. (2016). GPz: Non-stationary sparse Gaussian processes for heteroscedastic uncertainty estimation in photometric redshifts. *Monthly Notices of the Royal Astronomical Society*, 462(1):726–739. doi:10.1093/mnras/stw1618.

*School of Computer Science & Statistics, Trinity College Dublin, Ireland. Email: caok@tcd.ie.
†School of Computer Science & Statistics, Trinity College Dublin, Ireland. Email: sillerom@tcd.ie.
‡School of Computer Science & Statistics, Trinity College Dublin, Ireland. Email: swilson@tcd.ie.

On the kernel trick for high-dimensional two-sample problems

Javier Cárcamo * Antonio Cuevas † Luis Alberto Rodríguez ‡

L.A. Rodríguez (PhD candidate) Luis Alberto Rodríguez Ramírez is a PhD candidate at Autonomous University of Madrid. His main interests are empirical process theory and functional data analysis. He received his BSc from Autonomous University of Madrid in 2017 in mathematics. He earned his MSc in mathematics and applications from Autonomous University of Madrid one year later. In 2018, he enrolled in the doctoral program in mathematics at Autonomous University of Madrid under the supervision of Javier Cárcamo and Antonio Cuevas.

Two-sample tests aim to decide whether or not it can be accepted that two random elements have the same distribution, using the information provided by two independent samples from such distributions. This problem is omnipresent in practice on account of their applicability to a great variety of situations, ranging from biomedicine to quality control. Since the classical Student's t-tests or rank-based (Mann–Whitney, Wilcoxon,...) procedures, the subject has received an almost permanent attention from the statistical community. In this work we focus on two-sample tests valid, under broad assumptions, for general settings in which the available data are observations drawn from two random elements X and Y taking values in some separable Hilbert space $\mathcal{X}$; so, that the two-sample problem would be within the framework of Functional Data Analysis (FDA).

Many important methods, including goodness of fit and homogeneity tests, are based upon the use of an appropriate probability distance or discrepancy metric. Probability distances between probability measures allow the practitioner to measure the dissimilarity between two random quantities. Therefore, the estimation of a suitable distance helps detect (significant) differences between two populations. Some well-known, classic examples of such distances are the Kolmogorov distance, that leads to the the popular Kolmogorov–Smirnov statistic, and L^2-based discrepancy metric, leading to Cramér–von Mises or Anderson–Darling statistics. However, this classical methods have a low performance when dealing with high-dimensional data and cannot be applied in general non-Euclidean data (e.g., in FDA problems).

In [1], a proposal for the two-sample test was given based upon a different mathematical approach: kernel distance. These authors show that kernel-based procedures perform better than the classical ones when dimension grows. Unfortunately, this methodology introduce some new restrictions, such that a limitation on sample sizes of the homogeneity problem and data-driven kernel selection. These constraints have been replicated since then in the literature (see for instance [2]). In this talk we introduce an uniform version of this discrepancy metric between probability measures in order to remove them: the Supremum Kernel Distance (SKD). We will show the key ideas of our extension of the kernel distance as well as an asymptotic results on the plug-in estimators of this novel metric. Additionally, some empirical results on the performance of the associated test will be outlined.

Keywords: Functional data analysis; Homogeneity test; Kernel distance; Probability metrics; Two-sample problem.

Acknowledgements Work supported by project PID2019-109387GB-I00. The third author would also like to thank Arthur Gretton for his patient and disponibility and Bojan Mihaljevic for his invaluable support.

References

[1] Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. J. (2007). A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems* 513–520. doi:https://doi.org/10.48550/arXiv.0805.2368

[2] Zhang, J. T., Guo, J., and Zhou, B. (2022). Testing equality of several distributions in separable metric spaces: A maximum mean discrepancy based approach. *Journal of Econometrics*. doi:https://doi.org/10.1016/j.jeconom.2022.03.007

*Department of Mathematics, Universidad del País Vasco, Spain. Email: javier.carcamo@ehu.eus.
†Department of Mathematics, Universidad Autónoma de Madrid, Spain. Email: antonio.cuevas@uam.es.
‡Department of Mathematics, Universidad Autónoma de Madrid, Spain. Email: luisalberto.rodriguez@uam.es.

Computing upper bounds for the maximum chi-square index through a combinatorial relaxation

Davide Duma * Stefano Gualandi † Federico Malucelli ‡

D. Duma (Postdoctoral researcher) Davide Duma is a researcher (tenure-track) at the University of Pavia, where he is currently involved in research projects with the Joint Research Center in Siviglia and Fedegari Autoclavi SpA on real-world applications of optimization methods. His main interests are operations research applications for health care management and combinatorial optimization. He received his BSc from University of Salento in 2011 in Mathematics and Computer Science. He earned his MSc in Mathematics from the University of Turin in 2014. Then, he enrolled in the doctoral program in Computer Science at the University of Turin with a thesis project on online optimization for health services management under the supervision of Prof. Roberto Aringhieri.

Consider a set of observations consisting of measures on two variables. Given a set of marginal frequencies $\mathbf{a} = (a_1, \ldots, a_m)$ and $\mathbf{b} = (b_1, \ldots, b_n)$, a statistical test of independence of the two variables is the maximum Pearson's chi-square (χ^2) index, defined as a Quadratic Transportation Problem (QTP) in [1]. Such an optimization problem is derived from the Transportation Problem, where the objective function is quadratic in the flow variables. The objective is to find the joint distribution x_{ij} (contingency table) that maximizes the χ^2 index, that can be formulated as follows:

$$
\begin{aligned}
\chi^2 := \max_{\mathbf{x}} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{1}{a_i b_j} x_{ij}^2 & \\
\text{s.t.} \quad & \sum_{j=1}^{n} x_{ij} = a_i & \forall i \\
& \sum_{i=1}^{m} x_{ij} = b_j & \forall j \\
& 0 \le x_{ij} \le \min\{a_i, b_j\} & \forall i, j
\end{aligned}
$$

Since $a_i, b_j > 0$, the problem consists in maximizing a convex function over a convex set, which is a difficult problem, since the transportation polytope has a large number of extreme points. In the literature, the solutions approaches are mainly based on Lagrangean relaxations [2], except in [1], where three combinatorial optimization heuristics are proposed and evaluated experimentally.

In this work, we introduce a combinatorial relaxation of the QTP and we exploit the structure of the problem, providing a decomposition method, consisting of $m + n$ quadratic optimization subproblems. Then, we present a combinatorial algoritm for each subproblem, which is solved through a further decomposition that allows us to deal with further linear subproblems reduced to the well-known 0–1 knapsack formulation.

Finally, we report a computational analysis performed on several tests using as benchmarks a set of random instances similar to those proposed in [1] in order to compare the upper bounds (UBs). Results show the effectiveness of the proposed relaxation in finding tighter UBs for rectangular contingency tables, contrary to those found in [1], which often coincide with the theoretical UBs used in the statistical index Cramér's V.

Keywords: Chi-square index; Combinatorial optimization; Convex optimization; Quadratic programming; Transportation problem.

Acknowledgements Work supported by the Italian Ministry of Education, University and Research (MIUR) under the funding "Dipartimenti di Eccellenza" (law 232 of 2016).

References

[1] Kalantari, B., Lari, I., Rizzi, A., and Simeone, B. (1993). Sharp bounds for the maximum of the chi-square index in a class of contingency tables with given marginals. *Computational Statistics & Data Analysis*, 16(1):19–34. doi:10.1016/0167-9473(93)90242-L.

[2] Adlakha, V. and Kowalski, K. (2013). On the quadratic transportation problem. *Open Journal of Optimization*, 2(3): 89–94. doi:10.4236/ojop.2013.23012.

*Department of Mathematics, University of Pavia, Italy. Email: davide.duma@unipv.it.
†Department of Mathematics, University of Pavia, Italy. Email: stefano.gualandi@unipv.it.
‡Department of Electronics and Information, Politectico di Milano, Italy. Email: federico.malucelli@polimi.it.

On new omnibus tests of uniformity on the hypersphere

Alberto Fernández-de-Marcos * Eduardo García-Portugués †

A. Fernández-de-Marcos (PhD candidate) Alberto Fernández de Marcos is a part-time PhD candidate in the Mathematical Engineering program at Carlos III University of Madrid, where he earned his Master in Statistics for Data Science. He is now focused on the development of one- and two-sample tests in directional statistics. He received his BSc in Aerospace Engineering from the Technical University of Madrid in 2018. After working as data scientist in several companies, he now works as a secondary education teacher in mathematics and technology.

We propose two new uniformity tests for data on the hypersphere, $\mathbb{S}^d := \{\mathbf{x} \in \mathbb{R}^{d+1} : \|\mathbf{x}\| = 1\}$, $d \geq 1$. This testing problem is arguably the most fundamental analysis when dealing with data on $\mathbb{S}^d$. It is also an auxiliary tool for other related testing problems for data on $\mathbb{S}^d$ and $\mathbb{R}^{d+1}$ (e.g., goodness-of-fit or spherical symmetry).

Given an independent and identically distributed sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$, the test statistics are defined as

$$T_{n,k} := \frac{1}{n} \sum_{i,j=1}^{n} \psi_k(\theta_{ij}) - n\mathbb{E}_{\mathcal{H}_0}[\psi_k(\theta_{12})], \tag{1}$$

where ψ_k is the kernel of the test statistic and $\theta_{ij} := \cos^{-1}(\mathbf{X}_i'\mathbf{X}_j)$. The tests constructed from (1) are seen to belong to the so-called "Sobolev class" of uniformity tests.

The first test uses the kernel $\psi_1(\theta) := \exp(\kappa \cos\theta)$, $\kappa \geq 1$. It can be interpreted as a "smooth maximum" statistic, since it is connected to the "LogSumExp" function. It can also be regarded as an interpolation test between Cai & Jiang's [1] maximum statistic test and Rayleigh's test, both arising as limiting cases when $\kappa \to \infty$ and $\kappa \to 0$.

The second test uses the Poisson-like kernel $\psi_2(\theta) := (1 - \rho^2)/((1 - 2\rho\cos\theta + \rho^2)^{(d+1)/2})$, $0 \leq \rho < 1$. This test allows an extension of Pycke's [3] circular-only test to the hypersphere. It is also related to Rayleigh's test when $\rho \to 0$.

Leveraging closed-form expressions for Gegenbauer polynomials, we obtain the null asymptotic distributions of the two presented test statistics and their explicit powers against local alternatives, some of them specially tailored to the tests. The asymptotic distributions are seen to be effective and usable in practice, as the convergence is relatively fast.

A simulation study compares the powers of the new tests with others in the Sobolev class, showing that the new tests present higher powers when facing challenging multimodal alternatives to uniformity. The simulation study also explores the optimal choice of ρ and κ for different alternatives to uniformity.

As in [2], we also study transformations *à la* Stephens [4] for stabilizing the exact-n null distributions of the two test statistics about the asymptotic distribution, thus facilitating the application of the test in practice.

A real data application in astronomy illustrates the tests.

Keywords: Directional data; Nonparametric statistics; Uniformity; Sobolev tests.

Acknowledgements The authors acknowledge support from PGC2018-097284-B-100, funded by MCIN/AEI/10.13039/-501100011033 and by "ERDF A way of making Europe", and from the Community of Madrid through the framework of the multi-year agreement with Carlos III University of Madrid in its line of action "Excelencia para el Profesorado Universitario" (EPUC3M13).

References

[1] Cai, T. and Jiang, T. (2012). Phase transition in limiting distributions of coherence of high-dimensional random matrices. *Journal of Multivariate Analysis*, 107:24–39. doi:10.1016/j.jmva.2011.11.008.

[2] Fernández-de-Marcos, A. and García-Portugués, E. (2021). Data-driven stabilizations of goodness-of-fit tests. arXiv:2112.01808.

[3] Pycke, J.-R. (2010). Some tests for uniformity of circular distributions powerful against multimodal alternatives. *The Canadian Journal of Statistics*, 37(1):80–96. doi:10.1002/cjs.10048.

[4] Stephens, M. A. (1970). Use of the Kolmogorov-Smirnov, Cramér-von Mises and related statistics without extensive tables. *Journal of the Royal Statistical Society, Series B (Methodological)*, 32(1):115–122. doi:10.1111/j.2517-6161.1970.tb00821.x.

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

Session 2 (Wednesday 12:35)

Session talks

Feature selection and outliers detection using support vector machines

Marta Baldomero-Naranjo * Luisa I. Martínez-Merino † Antonio M. Rodríguez Chía ‡

M. Baldomero-Naranjo (Assistant Professor) Marta is Assistant Professor at Universidad Complutense de Madrid. Her main interests are Supervised Classification and Localization Theory. She received her BSc in Mathematics from Universidad de Cádiz in 2016. She earned her MSc in Mathematics (2017) and her PhD in Mathematics (2021) from Universidades de Almería, Cádiz, Granada, Jaén, and Málaga. Her thesis was entitled *Supervised Classification and Network Location Problems via Mathematical Optimization* and was supervised by Antonio M. Rodríguez Chía.

We focus our attention on the study of support vector machines (SVM) models. Since their introduction, SVM have been deeply analyzed in the literature. Although classical SVM models have high predictive power in comparison with other state-of-the-art classifying methods, some drawbacks also arise from their use. In this talk, we focus our attention in two of them: influence of outliers and feature selection.

First, we consider various models of support vector machines with ramp loss, these being an efficient and robust tool to limit the influence of outliers in the classifier. The exact solution approaches for the resulting optimization problem are of high demand for large datasets. Hence, the goal is to develop algorithms that provide efficient methodologies to exactly solve these optimization problems. These approaches are based on three strategies for obtaining tightened values of the big M parameters included in the formulation of the problem. Two of them require solving a sequence of continuous problems, while the third uses the Lagrangian relaxation to tighten the bounds. The proposed resolution methods are valid for the ℓ_1-norm and ℓ_2-norm ramp loss formulations. They are tested and compared with existing solution methods in simulated and real-life datasets, showing the efficiency of the developed methodology. More details can be found in [1].

In the second part of the talk, we propose a robust classification model, based on support vector machines, which simultaneously deals with outliers detection and feature selection. The classifier is built considering the ramp loss margin error and it includes a budget constraint to limit the number of selected features. The search of this classifier is modeled using a mixed-integer formulation with big M parameters. Two different approaches (exact and heuristic) are proposed to solve the model. The heuristic approach is validated by comparing the quality of the solutions provided by this approach with the exact approach. In addition, the classifiers obtained with the heuristic method are tested and compared with existing SVM-based models to show their efficiency. See [2] for further details.

Keywords: Supervised classification; Support vector machine; Outliers detection; Feature selection; Mixed integer programming.

Acknowledgements Work supported by projects MTM2016-74983-C2-2-R, FEDER-UCA18-106895, P18-FR-1422 and PhD grant UCA/REC01VI/2017. The authors would also like to thank Fundación BBVA for their support in the project NetmeetData.

References

[1] Baldomero-Naranjo, M., Martínez-Merino, L.I. and Rodríguez Chía, A.M. (2020). Tightening big Ms in integer programming formulations for support vector machines with ramp loss. *European Journal of Operational Research*, 286(1):84–100. doi:10.1016/j.ejor.2020.03.023.

[2] Baldomero-Naranjo, M., Martínez-Merino, L.I. and Rodríguez Chía, A.M. (2021). A robust SVM-based approach with feature selection and outliers detection for classification problems. *Expert Systems with Applications*, 178:115017. doi:10.1016/j.eswa.2021.115017.

*Departamento de Estadística y Ciencia de los Datos, Universidad Complutense de Madrid, España. Email: martbald@ucm.es.
†Departamento de Estadística e Investigación Operativa, Universidad de Cádiz, España. Email: luisa.martinez@uca.es.
‡Departamento de Estadística e Investigación Operativa, Universidad de Cádiz, España. Email: antonio.rodriguezchia@uca.es.

Solving mixed-integer programs with warm-starting constraint generation methods via machine learning tools

Asunción Jiménez-Cordero * Juan Miguel Morales † Salvador Pineda ‡

A. Jiménez-Cordero (Lecturer) Asunción Jiménez-Cordero received the Mathematics degree from the University of Seville (Seville, Spain) in 2013, and a Ph.D. degree in Mathematics also from the University of Seville, in 2019. She is currently a lecturer in the Department of Statistics and Operations Research at the University of Málaga in Spain. Her research interests are in the fields of mathematical programming; optimization; machine learning; data-driven approaches, and power systems applications.

Mixed Integer Linear Programs (MILP) are well known to be NP-hard problems in general, [4]. Even though pure optimization-based methods, such as constraint generation, are guaranteed to provide an optimal solution if enough time is given, their use in online applications is still a great challenge due to their usual excessive time requirements. To alleviate their computational burden, some machine learning techniques have been proposed in the literature [1, 2] using the information provided by previously solved MILP instances. Unfortunately, these techniques report a non-negligible percentage of infeasible or suboptimal instances.

By linking mathematical optimization and machine learning, this paper proposes a novel approach that speeds up the traditional constraint generation method, [3], preserving feasibility and optimality guarantees. In particular, we first identify offline the so-called invariant constraint set of past MILP instances. We then train, also offline, a machine learning method (any strategy can be chosen) to learn an invariant constraint set as a function of the problem parameters of each instance. Next, we predict online an invariant constraint set of the new unseen MILP application and use it to initialize the constraint generation method. This warm-started strategy significantly reduces the number of iterations to reach optimality, and therefore, the computational burden to solve online each MILP problem is significantly reduced. Very importantly, the proposed methodology inherits the feasibility and optimality guarantees of the traditional constraint generation method. The computational performance of the proposed approach is quantified through synthetic and real-life MILP applications.

Keywords: Mixed integer linear programming; Machine learning; Constraint generation; Warm-start; Feasibility and optimality guarantees.

Acknowledgements This work was supported in part by the Spanish Ministry of Science and Innovation through project PID2020-115460GB-I00, in part by the European Research Council (ERC) under the EU Horizon 2020 research and innovation program (grant agreement No. 755705) in part, by the Junta de Andalucía (JA), the Universidad de Málaga (UMA), and the European Regional Development Fund (FEDER) through the research projects P20_00153 and UMA2018-FEDERJA-001, and in part by the Research Program for Young Talented Researchers of the University of Málaga under Project B1-2020-15. The authors thankfully acknowledge the computer resources, technical expertise, and assistance provided by the SCBI (Supercomputing and Bioinformatics) center of the University of Málaga.

References

[1] Bengio, Y., Lodi, A., and Prouvost, A. (2021). Machine learning for combinatorial optimization: A methodological tour d'horizon. *European Journal of Operational Research*, 290:405–421. doi:10.1016/j.ejor.2020.07.063.

[2] Gambella, C., Ghaddar, B., and Naoum-Sawaya, J. (2021). Optimization problems for machine learning: A survey. *European Journal of Operational Research*, 290:807–828. doi:10.1016/j.ejor.2020.08.045.

[3] Minoux, M. (1989). Networks synthesis and optimum network design problems: Models, solution methods and applications. *Networks*, 19:313–360. doi:10.1002/net.3230190305.

[4] Wolsey, L. A. (2008). *Mixed integer programming.* Wiley Encyclopedia of Computer Science and Engineering (pp. 1–10). American Cancer Society. doi:10.1002/9780470050118.ecse244.

*OASYS group, Department of Statistics and Operations Research, University of Málaga, Spain. Email: asuncionjc@uma.es.
†OASYS group, Department of Applied Mathematics, University of Málaga, Spain. Email: juanmi82mg@gmail.com.
‡OASYS group, Department of Electrical Engineering, University of Málaga, Spain. Email: spinedamorente@gmail.com.

Robust inference for non-destructive one-shot devices testing under the step-stress model and Weibull lifetime distributions

María Jaenada *

M. Jaenada (PhD candidate) María Jaenada received the BSc degrees in mathematics and mathematics and statistics at Universidad Complutense de Madrid, where she later obtained a MSc in computational statistics. She is currently working toward the doctoral degree in mathematical engineering, statistics, and operations research with the Department of Statistics and Operational Research at the Complutense University of Madrid thanks to a FPU grant, under the supervision of Prof. Leandro Pardo. Her research interests include reliability analysis and efficiency and robustness for high dimensional generalized regression models.

One-shot devices analysis involves an extreme case of interval censoring. Some kind of one-shot devices do not get destroyed when tested, and so can continue within the experiment, providing extra information for inference. Further, one-shot devices usually have large mean lifetimes under working conditions, and so their lifetime distributions must be estimated via accelerated life tests (ALTs) by running the tests at varying and higher stress levels than working conditions.

Step-stress tests allow the experimenter to increase the stress levels at pre-fixed times gradually during the lifetesting experiment. The cumulative exposure model is commonly assumed for step-stress models, relating the lifetime distribution of units at one stress level to the lifetime distributions at preceding stress levels.

Classical estimators for one-shot devices under the step-stress ALT model are based on the maximum likelihood estimator (MLE), and so they enjoy nice efficiency properties but lack of robustness. Robust divergence-based inference for non-destructive one-shot devices was first studied in [1] for the case of exponential lifetime distributions.

In this work we develop robust estimators and Wald-type test statistics based on the density power divergence (DPD) for testing composite null hypothesis for non-destructive one-shot devices under the step-stress ALTs with Weibull lifetime distributions, as developped in [2]. Moreover, we examine theoretically and empirically the asymptotic and robustness properties of the estimators and test statistics, as well as prediction accuracy of different lifetime characteristics such as reliability, distribution quantiles and mean lifetime of the devices.

Keywords: Accelerated lifetests; One-shot devices; Robustness; Weibull lifetime distributions.

Acknowledgements This work was supported by the Spanish Grants PGC2018-095194-B-100 and FPU/018240.

References

[1] Balakrishnan, N., Castilla, E., Jaenada, M. and Pardo, L. (2022). Robust inference for non-destructive one-shot device testing under step-stress model with exponential lifetimes. arXiv Preprint. arXiv:2204.11560.

[2] Balakrishnan, N., Jaenada, M. and Pardo, L. (2022). Non-destructive one-shot devices testing under the step-stress model and Weibull lifetimes. Preprint.

*Interdisciplinary Mathematics Institute. Department of Statistics and O.R., Complutense University of Madrid, Spain. Email: mjaenada@ucm.es.

Session 3 (Wednesday 15:50)

Session talks

Decomposition methods for large-scale routing problems

Alberto Santini* Michael Schneider† Thibaut Vidal‡ Daniele Vigo§

A. Santini (**post-doctoral researcher**) is a Marie Skłodowska-Curie Fellow "EUTOPIA-SIF" at ESSEC Business School (France). He is also an affiliated professor at the Institute of Advanced Studies of CY Cergy Paris Université (France), the Barcelona School of Economics (Spain) and the Barcelona Graduate School of Mathematics (Spain). Before starting his fellowship, he was a tenure-track Assistant Professor at Universitat Pompeu Fabra (Spain) and a post-doctoral researcher at RWTH Aachen University (Germany). He received his PhD in 2017 from the University of Bologna (Italy). He works in Combinatorial Optimisation, using both exact and heuristic methods and with applications in areas spanning from logistics to production planning and from graph problems to urban design.

Decomposition techniques are an important component of modern heuristics for large instances of vehicle routing problems (VRPs). The current literature lacks a characterisation of decomposition strategies and a systematic investigation of their impact when integrated into state-of-the-art heuristics. We aim to fill this gap: we discuss the main characteristics of decomposition techniques in vehicle routing heuristics, highlight their strengths and weaknesses, and derive a set of desirable properties. Through an extensive numerical campaign, we investigate the impact of decompositions within two algorithms for the capacitated vehicle routing problem: Adaptive Large Neighbourhood Search [1] and Hybrid Genetic Search [2]. We evaluate the quality of popular decomposition techniques from the literature and propose new strategies that outperform existing methods. The best performing decomposition creates subproblems from existing routes in a high-quality solution, after clustering the routes using tools from unsupervised machine learning. Our results also confirm the validity of the desirable properties identified in the analysis of the literature.

Keywords: Vehicle routing; Heuristics; Problem decomposition; Genetic algorithms; Adaptive large neighbourhood search.

Acknowledgements

The authors are grateful to Stefan Ropke for sharing the source code of the ALNS heuristic. Alberto Santini was partially funded by MICINN (Spain) through the programme Juan de la Cierva Formación, the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement 945380, ESSEC Business School (France) through the Visiting Professor programme. The research of Daniele Vigo has been supported by Ministero dell'Istruzione, dell'Università e della Ricerca, Italy under grant PRIN 2015JJLC3E_002, and by USAF under grant FA9550-17-1-0234. The research of Thibaut Vidal in Brazil has been supported by CAPES, by CNPq under grant 308528/2018-2, and by FAPERJ under grant E-26/202.790/2019.

References

[1] Ropke, S. and Pisinger, D. (2006) An adaptive large neighborhood search heuristic for the pickup and delivery problem with time windows. *Transportation Science*, 40(4):455–472. doi:10.1287/trsc.1050.0135.

[2] Vidal, T., Crainic, T.G., Gendreau, M. and Prins, C. (2013) A hybrid genetic algorithm with adaptive diversity management for a large class of vehicle routing problems with time-windows. *Computers & Operations Research*, 40(1):475–489. doi:10.1016/j.cor.2012.07.018.

*ESSEC Business School, France; Institute of Advanced Studies, CY Cergy Paris Université, France. Email: santini@essec.edu.
†Deutsche Post Chair for the Optimisation of Distribution Networks, RWTH Aachen University, Germany. Email: schneider@dpo.rwth-aachen.de.
‡CIRRELT and Scale AI Chair in Data-Driven Supply Chains, Canada. Email: thibaut.vidal@cirrelt.ca.
§Alma Mater Studiorum University of Bologna, Italy. Email: daniele.vigo@unibo.it.

New model and solution methods for a routing problem with compartmentalized trucks and trailers

Laura Davila-Pena [*‡§] David R. Penas [†] Balbina Casas-Méndez [*‡]
Maria Antónia Carravilla [§] José Fernando Oliveira [¶]

L. Davila-Pena (PhD student) Laura Davila Pena is a PhD candidate at the University of Santiago de Compostela. Her main interests include Operations Research problems such as Vehicle Routing, Game Theory, or Sequencing. She received her BSc from University of Santiago de Compostela in 2017 in mathematics, and earned her master's degree in statistical techniques from the University of Santiago de Compostela in 2019. During the second year of her master's degree, she enrolled in the doctoral program in statistics and operations research at the same university, under the supervision of Ignacio García Jurado and Balbina Casas Méndez.

Vehicle routing problems admit different variants depending on the clients' needs. One of them is the truck and trailer routing problem, TTRP, where a fleet of trucks and trailers serves a set of customers such that when the trailer is not able to reach a customer, they are attended only by the truck [1].

This work proposes a novel mixed-integer linear programming approach to combine the TTRP with product compartmentalization, which we call the multi-compartment truck and trailer routing problem (MC-TTRP). The combination of these two features is motivated by the needs of a Spanish agricultural cooperative that produces feed for cattle [3].

We present two heuristic algorithms for the MC-TTRP: an iterated tabu search (ITS) and an adaptive large neighbourhood search (ALNS). Both proposals consist of two stages: the first phase iteratively builds an initial solution, based on the savings method of Clarke and Wright, and then the second phase aims to refine the solution. We carried out a computational study on new 21 test problems adapted from those in preexisting literature. The results obtained prove the effectiveness of our proposals. In particular, the ITS outperforms previous approaches for some truck and trailer routing problem instances [2]. Furthermore, an application of the proposed model and heuristics is demonstrated in the field of agricultural logistics by comparing the obtained results through different approaches.

Keywords: Truck and trailer routing problem; Compartmentalized vehicles; Heuristics; Logistics.

Acknowledgements Laura Davila-Pena's research was funded by the Ministry of Education, Culture and Sports of Spain (contract FPU17/02126). David R. Penas' research was funded by the Xunta de Galicia (post-doctoral contract ED481B-2019-010). Work also supported by the ERDF (MINECO/AEI grant MTM2017-87197-C3-3-P) and by the Xunta de Galicia (Competitive Reference Group ED431C 2017/38 and ED431C 2021/24).

References

[1] Chao, I-M. (2002). A tabu search method for the truck and trailer routing problem. *Computers & Operations Research*, 29(1):33–51. doi:10.1016/S0305-0548(00)00056-3.

[2] Davila-Pena, L., R. Penas, D., and Casas-Méndez, B. (2021). A new two-phase heuristic for a problem of food distribution with compartmentalized trucks and trailers. *International Transactions in Operational Research*. doi:10.1111/itor.13071.

[3] Guitián de Frutos, R. M. and Casas-Méndez, B. (2019). Routing problems in agricultural cooperatives: a model for optimization of transport vehicle logistics. *IMA Journal of Management Mathematics*, 30(4):387–412. doi:10.1093/imaman/dpy010.

[*]CITMAga, Spain. Email: lauradavila.pena@usc.es.
[†]Computational Biology Lab, MBG-CSIC, Spain. Email: david.rodriguez.penas@csic.es.
[‡]Department of Statistics, Mathematical Analysis and Optimization, Universidade de Santiago de Compostela, Spain. Email: balbina.casas.mendez@usc.es.
[§]INESC TEC, Faculty of Engineering, University of Porto, Portugal. Email: mac@fe.up.pt.
[¶]INESC TEC, Faculty of Engineering, University of Porto, Portugal. Email: jfo@fe.up.pt.

Mixture cure model estimators of the probability of default in credit risk

Rebeca Peláez * Ingrid Van Keilegom † Ricardo Cao ‡ Juan Manuel Vilar §

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

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

The probability of default conditional on the credit scoring, $PD(t|x)$, can be written as a transformation of the conditional survival function of the variable "time to default", $S(t|x)$:

$$PD(t|x) = 1 - \frac{S(t+b|x)}{S(t|x)}$$

This property is used to propose new PD estimators. Throughout the study of a set of credits, the default is not observed for all of them and the variable "time to default" is censored. As a consequence, censored data and survival analysis are widely used. However, the time to default does not only face a problem of right censoring. Some customers never default, that is, no matter how long you observe such individuals, they will never experience the event of interest. This work discusses techniques for estimating the probability of default (PD) based on cure models, which are the survival models that take into account the existence of a group of cured individuals who are not susceptible to default. A nonparametric survival estimator proposed by [1] and [2] is considered and transformed to obtain an estimator of the probability of default (NPCM estimator). Its behaviour is compared by simulation with Beran's PD estimator based on the generalised limit-product estimator of the conditional survival function and parametric methods based on cure models such as the proportional hazards and the accelerated failure time methods.

The results obtained show that the NPCM estimator provides good estimations of PD and reduces the error committed by the parametric alternatives. Beran's PD estimator is competitive with the NPCM estimator in most scenarios.

The asymptotic properties of the NPCM PD estimator are analysed: an almost sure representation of the estimator and asymptotic expressions of the bias and variance are obtained, as well as its asymptotic normality. Finally, to illustrate the use of Beran's and NPCM estimators, a statistical analysis of German bank loans is carried out.

Keywords: Censored data; Kernel method; Nonparametric estimation; Survival analysis.

Acknowledgements This research has been supported by MICINN Grant PID2020-113578RB-100, by the Xunta de Galicia (Grupo de Referencia Competitiva ED431C-2020-14 and Centro Singular de Investigación de Galicia ED431G 2019/01), all of them through the ERDF and by the European Research Council (2016-2022, Horizon 2020 / ERC grant agreement No. 694409). RP was sponsored by inMOTION Programme of grants for pre-doctoral stays Inditex-UDC 2021.

References

[1] López-Cheda, A., Cao, R. and Jácome, M. A. (2017). Nonparametric latency estimation for mixture cure models. *TEST*, 26(2):353-376. doi:https://doi.org/10.1007/s11749-016-0515-1.

[2] López-Cheda, A., Cao, R., Jácome, M. A. and Van Keilegom, I. (2017). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Computational Statistics & Data Analysis*, 105:144-165. doi:https://doi.org/10.1016/j.csda.2016.08.002.

*Research Group MODES, Department of Mathematics, CITIC, University of A Coruña, Spain. Email: rebeca.pelaez@udc.es.
†Research Centre for Operations Research and Statistics (ORSTAT), KU Leuven, Belgium. Email: ingrid.vankeilegom@kuleuven.be.
‡Research Group MODES, Department of Mathematics, CITIC, University of A Coruña, Spain. Email: ricardo.cao@udc.es.
§Research Group MODES, Department of Mathematics, CITIC, University of A Coruña, Spain. Email: juan.vilar@udc.es.

Mixture cure models in the presence of vector and functional covariates. A dimension reduction approach

Beatriz Piñeiro-Lamas * Ricardo Cao † Ana López-Cheda ‡

B. Piñeiro-Lamas (PhD student) Beatriz Piñeiro-Lamas is a predoctoral researcher at Universidade da Coruña. Her main interest is survival analysis; in particular, cure models. She received her BSc from Universidade de Santiago de Compostela in 2018 in Mathematics. She earned her master's degrees in Statistical Techniques and Biomedical Research from the same university in 2020.

Standard survival models assume that, in the absence of censoring, all individuals will experience the event of interest. However, sometimes this is not realistic. For example, if we consider cancer patients being treated and the event is the appearance of an adverse effect, there will be patients that will never experience it. Those who will never develop this health condition will be considered as cured. To incorporate this cure fraction, classical survival analysis has been extended to cure models. In particular, mixture cure models allow to estimate the probability of being cured and the survival function for the uncured subjects. In the literature, nonparametric estimation of both functions is limited to continuous univariate covariates ([1], [2]). We fill this important gap by considering both vector and functional covariates and proposing a single-index model for dimension reduction. This approach has been studied in the presence of censoring ([3]), but not in the presence of cure. The methodology is applied to a cardiotoxicity dataset from the University Hospital of A Coruña.

Keywords: Cardiotoxicity; Censored Data; Kernel Estimator; Survival Analysis.

Acknowledgements This research has been supported by MINECO grant MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2020-14 and Centro de Investigación del Sistema Universitario de Galicia ED431G 2019/01), all of them through the ERDF. The first author acknowledges financial support from Axudas Predoutorais da Xunta de Galicia, with reference ED481A-2020/290, and from Centro de Investigación en Tecnoloxías da Información e das Comunicacións (CITIC) of the University of A Coruña, funded by Xunta de Galicia and the European Union (ERDFGalicia 2014-2020 Program), by grant ED431G 2019/01.

References

[1] López-Cheda, A., Cao, R., Jácome, M. A., and Van Keilegom, I. (2017a). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Computational Statistics & Data Analysis*, 105:144–165. doi:10.1016/j.csda.2016.08.002.

[2] López-Cheda, A., Jácome, M. A., and Cao, R. (2017b). Nonparametric latency estimation for mixture cure models. *Test*, 26:353–376. doi:10.1007/s11749-016-0515-1.

[3] Strzalkowska-Kominiak, E., and Cao, R. (2013). Maximum likelihood estimation for conditional distribution single-index models under censoring. *Multivariate Analysis*, 114:74–96. doi:10.1016/j.jmva.2012.07.012.

*Grupo MODES, CITIC, Departamento de Matemáticas, Universidade da Coruña, Spain. Email: b.pineiro.lamas@udc.es.
†Grupo MODES, CITIC, Departamento de Matemáticas, Universidade da Coruña, Spain. Email: ricardo.cao@udc.es.
‡Grupo MODES, CITIC, Departamento de Matemáticas, Universidade da Coruña, Spain. Email: ana.lopez.cheda@udc.es.

Session 4 (Thursday 11.00)

Session talks

How statistics can aid a chemist: the case of multivariate curve resolution

Raffaele Vitale*

R. Vitale (Associate Professor) Raffaele Vitale is Associate Professor at the University of Lille (France). He received his MSc in Analytical Chemistry at the University of Rome "La Sapienza" (Italy) and his PhD in Statistics and Optimisation at the Technical University of Valencia (Spain). Currently, his work is mainly focused on the development and application of multivariate statistical approaches for the analysis of hyperspectral and optical microscopy images.

In chemistry, the term *spectroscopy* refers to an ensemble of instrumental approaches that permits to characterise the inherent structure and composition of complex samples based on how light interacts with them. Measurements resulting from platforms of this specific type are generally high-dimensional and multivariate and can be usually regarded as the sum of signal contributions deriving from the individual *pure* components (*e.g.*, atoms of a particular element, compounds with a common molecular backbone, *etc.*) that underlie such samples. In order to get relevant insights into their nature, though, these individual contributions need to be somehow unravelled by solving an *inverse* resolution or unmixing problem. Although there exists a plethora of diverse strategies to successfully tackle this task, practitioners frequently resort to least squares methodologies, like the well-known Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS [1]). MCR-ALS is a soft-modelling technique that performs the resolution of multi-component evolving chemical systems according to the low-rank bilinear model in the following equation:

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \tag{1}$$

where $\mathbf{X}$ ($N \times J$) is a matrix containing a series of N J-dimensional spectroscopic (or spectral) recordings, $\mathbf{C}$ ($N \times A$) and $\mathbf{S}^{\mathrm{T}}$ ($A \times J$) carry the pure profiles of the aforementioned individual signal contributions associated to the data variation along the rows and columns of $\mathbf{X}$, respectively, and $\mathbf{E}$ denotes the residuals array, that is to say the portion of $\mathbf{X}$ not *explained* at the user-defined rank, A. MCR-ALS iteratively estimates $\mathbf{C}$ and $\mathbf{S}^{\mathrm{T}}$ by minimising the squared Euclidean norm of $\mathbf{E}$, *i.e.*, $\left\|\mathbf{X} - \mathbf{C}\mathbf{S}^{\mathrm{T}}\right\|^2$, throughout a sequence of alternating least squares projections of the rows and columns of $\mathbf{X}$, executed under appropriate constraints (for example, non-negativity) to somehow restrict the final MCR-ALS solution[1] and obtain a factorisation of $\mathbf{X}$ that is meaningful from a physicochemical point of view.

In this presentation, a comprehensive overview of the spectral unmixing framework, of the operating principles of MCR-ALS and of the practical implications resulting from its utilisation will be given. A distinctive focus will be put on the analogy between the algorithmic procedure behind MCR-ALS and multivariate regression which has lately enabled the formalisation of a deleterious effect that might jeopardise the quality of the MCR-ALS output and that is related to the leverage properties of the various data objects at hand [2]. Solutions to overcome this effect, borrowed from the robust statistics domain (for instance, object weighting), will also be discussed.

Keywords: Spectroscopic data; Inverse problems; Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS); Leverage; Object weighting.

Acknowledgements The author acknowledges funding from the project "ANR-21-CE29-0007" (Agence Nationale de la Recherche).

References

[1] Tauler, R., Smilde, A. K. and Kowalski, B. (1995). Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *Journal of Chemometrics*, 9(1):31–58. doi:10.1002/cem.1180090105.

[2] Vitale, R. and Ruckebusch, C. (2022). On the black hole effect in bilinear curve resolution based on least squares ChemRxiv:10.26434/chemrxiv-2022-34g1j-v2.

*Univ. Lille, CNRS, LASIRE (UMR 8516), Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, France. Email: raffaele.vitale@univ-lille.fr.

[1] Given the fact that MCR-ALS components are not necessarily orthogonal, the decomposition of $\mathbf{X}$ is usually non-unique or *ambiguous*, *i.e.*, different combinations of $\mathbf{C}$ and $\mathbf{S}^{\mathrm{T}}$ may result in an identical data fit.

ADMM-based unit and time decomposition for price arbitrage by cooperative price-maker electricity storage units

Miguel F. Anjos * James R. Cruise † Albert Solà Vilalta ‡

A. Solà Vilalta (PhD candidate) Albert Solà Vilalta is a PhD candidate in Optimization and Operations Research at the University of Edinburgh. His main research interests are the decarbonisation challenges faced by modern societies, with a focus on power systems and energy storage. More broadly, he is interested in any problem with an optimization component, and the mathematics behind it. He received a BSc in Mathematics from the University of Barcelona in 2015, and a MSc in Mathematics from the University of Bonn in 2017.

Decarbonization via the integration of renewables poses significant challenges for electric power systems, but also creates new market opportunities. Electric energy storage can take advantage of these opportunities while providing flexibility to power systems that can help address these challenges. We propose a solution method for the optimal control of multiple price-maker electric energy storage units that cooperate to maximize their total profit from price arbitrage [2].

The proposed method can tackle the nonlinearity introduced by the price-maker assumption. The main novelty of the proposed method is the combination of a decomposition by unit and a decomposition in time. The decomposition by unit is based on the Alternating Direction Method of Multipliers (ADMM) and breaks the problem into several one-unit subproblems. Every subproblem is solved using an efficient algorithm for one-unit problems from the literature [3] that exploits an on the fly decomposition in time, and this results in a time decomposition for the whole solution method.

Our computational experiments show very promising performance in terms of accuracy and computational time. In particular, compared to the approach in [1], we observed two orders of magnitude computational time reduction with very small accuracy losses ($< 0.1\%$) when solving 2-unit instances. This allows to consistently solve instances with more than two storage units, and suggests that computational time scales linearly with the number of storage units.

Keywords: ADMM; Control; Energy storage; Horizons; Price-maker.

Acknowledgements The authors thank Fraser Daly, Chris Dent, Jonathan Eckstein, Jean Lasserre, Seva Shneer and Stan Zachary for very helpful discussions and Jalal Kazempour for delivering a course that inspired part of this research. Albert Solà Vilalta was supported by The Maxwell Institute Graduate School in Analysis and its Applications, a Centre for Doctoral Training funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016508/01), the Scottish Funding Council, Heriot-Watt University and the University of Edinburgh.

References

[1] Anjos, M. F., Cruise, J. R., and Solà Vilalta, A. (2020). Control of two energy storage units with market impact: Lagrangian approach and horizons. *Proceedings of the 2020 International Conference on Probabilistic Methods Applied to Power Systems* (Liege), 1-6. doi:10.1109/PMAPS47429.2020.9183690.

[2] Anjos, M. F., Cruise, J. R., and Solà Vilalta, A. (2021). ADMM-based unit and time decomposition for price arbitrage by cooperative price-maker electricity storage units. Preprint available at: http://www.optimization-online.org/DB_HTML/2021/10/8644.html.

[3] Cruise, J. R., Flatley, L., Gibbens, R. J., and Zachary S. (2019). Control of energy storage with market impact: Lagrangian approach and horizons. *Operations Research* 67(1):1–9. doi:10.1287/opre.2018.1761.

*Maxwell Institute for Mathematical Sciences, School of Mathematics, University of Edinburgh, United Kingdom. Email: anjos@stanfordalumni.org.

†Maxwell Institute for Mathematical Sciences, Department of Actuarial Mathematics and Statistics, Heriot-Watt University, United Kingdom. Email: r.cruise@hw.ac.uk

‡Maxwell Institute for Mathematical Sciences, School of Mathematics, University of Edinburgh, United Kingdom. Email: albert.sola@ed.ac.uk.

Interpreting an image classification model using superpixels

Cristian Pachón-García * Pedro Delicado † Verónica Vilaplana ‡

C. Pachón García (PhD candidate) Cristian Pachón-García is a PhD student at Universitat Politècnica de Catalunya. He holds a BSc in Mathematics and a MSc in Statistics and Operations Research. He has 10 years of professional experience as a Data Scientist, where he has developed machine learning models for fraud detection. Part of his PhD project has been focused on multivariate analysis, concretely on the dimensionality reduction techniques for large datasets.

Interpretability is one of the hottest topics currently in the field of machine learning. The number of publications in this field has grown recently, since the complexity of the machine learning models has increased, in part due to the emergence of deep learning models.

For an image classification problem, the goal of a local interpretability method is to identify which parts of a certain image are the most relevant for the model to make a prediction. Although there are many local interpretability methods, like LIME [4], there are few methods that explain an image model classification globally, as far as we know.

Some methods depend on the underlying machine learning model. This is the case of Grad-CAM [2] or saliency maps [3], which are designed to explain convolutional neural networks (CNN). As soon as CNN are no longer used to classify images, these methods will loose traction. That is why it is desirable to have interpretability methods that do not depend on the machine learning model. These type of methods are called model-agnostic.

In this work we develop a global agnostic interpretability method for an image classification model. The goal of the method is to find important visual concepts encoded in superpixels (group of contiguous pixels as homogeneous as possible).

The interpretability method requires a test dataset of images, $\mathcal{D}$, and an image classification model, $f(\mathbf{I})$. For a given image $\mathbf{I}_j \in \mathcal{D}$, it is segmented into a set of superpixels by means of an unsupervised segmentation algorithm such as Quick shift [1]. The resulting number of superpixels is m_j. For a given superpixel $s \in \mathbf{I}_j$, p features are extracted.

The method applies the previous procedure for each image Therefore, a matrix $\mathbf{X}$ of size $m \times p$ is obtained, where m is the total number of superpixels ($m = \sum_j m_j$). Using a clustering algorithm over $\mathbf{X}$, ℓ clusters are obtained, $C_1, \ldots, C_\ell$.

Finally, the importance of each cluster $C_r, r \in \{1, \ldots, \ell\}$, is measured. The algorithm ranks each cluster according to its importance. Since they contain superpixels, we can inspect the visual concept they encode.

Keywords: Interpretable machine learning; Explainable artificial intelligence; Deep learning; Superpixel.

Acknowledgements This research is supported by the Spanish Agencia Estatal de Investigación grant PID2020-116294GB-I00.

References

[1] Vedaldi A., and Soatto S. (2008). Quick shift and kernel methods for mode seeking. *European Conference on Computer Vision.*

[2] Selvaraju R. R., Cogswell M., Das A., Vedantam R., Parikh D., and Batra D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International conference on computer vision.* doi:10.1109/ICCV.2017.74.

[3] Simonyan K., Vedaldi A., and Zisserman A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. *International conference on learning representations.*

[4] Ribeiro-Marco T., Singh S., and Guestrin C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *International conference on knowledge discovery and data mining.* doi:10.1145/2939672.2939778.

*Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Spain. Email: cristian.pachon@upc.edu.
†Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Spain. Email: pedro.delicado@upc.edu.
‡Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, Spain. Email: veronica.vilaplana@upc.edu.

Fetal growth models and its application to examine the effect of polluting environmental substances

Ana García Burgos * Beatriz González Alzaga † María José Jiménez Asensio ‡
Marina Lacasaña Navarro § Nuria Rico Castro ¶ Desirée Romero Molina ‖

A. García Burgos (PhD candidate) Ana García Burgos is a PhD Student at the University of Granada. She is currently studying the doctoral program in Applied Mathematics and Statistics. She received the BSc in Mathematics from University of Granada in 2019. In 2021, she studied the MSc in Teaching Compulsory Secondary Education and Baccalaureate, Vocational Training and Language Teaching. Then, she earned her MSc in Applied Statistics in 2021. After completing her academic degree, she started working as an interim substitute professor in the Department of Statistics and Operations Research to the present day.

The aim of this work is to examine the effect of contaminating substances and other factors on mothers and their children. This study is part of the GENEIDA project, carried out in Almería.

As part of the study, ultrasound biometry was performed on a sample of 800 pregnant women. The echographic information is measured at different instants of time. In addition, the number of ultrasounds is different for each mother. These facts pose a problem for the conduct of certain investigations. The objective is to create a model that predicts the ultrasound measurements at the same instant of time for each mother. In particular, following [1], at 12, 20 and 34 weeks of gestation.

Initially, we tried to use the generalized linear model, but some of its hypotheses failed. In particular, there are these two facts:

- Heteroskedasticity of the residuals:

$$Var[\epsilon_i] = \sigma_i^2, \quad \forall i = 1, \ldots, n.$$

- Autocorrelation:

$$\mathrm{Cov}[\epsilon_i, \epsilon_j] \neq 0 \quad \forall i, j \quad i \neq j.$$

where ϵ_i is the residual in observation i. Therefore, the generalized least squares model has been used to obtain the predictions, since this model is less restrictive in terms of assumptions.

Following [1], a model has been built to represent the growth over time of ultrasound biometric measurements, performing a stepwise inclusion method taking into account the Akaike criterion. For this, different confounding variables observed in each mother have been selected, which were considered in their linear version and in their quadratic version. After obtaining the optimal model, an autocorrelation structure and a variance structure were included to model the assumptions of heteroskedasticity and autocorrelation of the residuals.

After obtaining the predictions using the optimal model, it is possible to study the effect of different factors, such as environmental pollutants, on fetal growth. In particular, as an example of application of the results obtained with the growth curves, we will focus on the analysis of the effect of dialkyl phosphates.

Keywords: Ultrasound measurements; Imputation; Growth curve; Generalized least squares linear model; Dialkyl phosphates.

References

[1] Iñiguez, C., Esplugues, A., Sunyer, J., Basterrechea, M., Fernández-Somoano, A., Costa, O., Estarlich, M., Aguilera, I., Lertxundi, A., Tardón, A., Guxens, M., Murcia, M., Lopez-Espinosa, M. J., Ballester, F., and INMA Project (2016). Prenatal Exposure to NO_2 and Ultrasound Measures of Fetal Growth in the Spanish INMA Cohort. *Environmental Health Perspectives*, 124(2):235–242. doi:10.1289/ehp.1409423.

*Department of Statistics and Operational Research, University of Granada, Spain. Email: agburgos@ugr.es.
†Andalusian School of Public Health, Spain. Email: beatriz.gonzalez.easp@juntadeandalucia.es.
‡Andalusian School of Public Health, Spain. Email: mariajoseases@hotmail.com.
§Andalusian School of Public Health, España. Email: marina.lacasana.easp@juntadeandalucia.es.
¶Department of Statistics and Operational Research, University of Granada, Spain. Email: nrico@ugr.es.
‖Department of Statistics and Operational Research, University of Granada, Spain. Email: deromero@ugr.es.

Session 5 (Thursday 12.35)

Session talks

Multicriteria sorting algorithm based on PROMETHEE's net flows: An application to supplier segmentation

Iván Felipe Barrera * Marina Segura † Concepción Maroto *

I.F. Barrera is a PhD candidate at the Polytechnic University of Valencia. His main interests are the application of multi-criteria techniques for the supply chain management. He received his BSc from the Autonomous University of the West (Colombia) in 2014 in Industrial Engineering. After four years in the purchasing department of a financial institution, he studied a MSc in Data Analysis, Process Improvement, and Decision Making Engineering. Finally, at the end of 2020 he enrolled in the doctoral program in Statistics and Optimisation, which he is currently studying.

Supplier management research has grown significantly in recent decades and an extensive literature addresses supplier selection from different perspectives. Due to the numerous evaluation criteria, Multiple Criteria Decision-Making provides a useful approach to deal with supplier evaluation. Sorting is a type of classification problem, where the suppliers are assigned to predefined groups characterised by reference profiles and the groups are defined in an ordinal way.

According to the state of the art there has been an increasing number of methodological developments related to multicriteria sorting issues (especially in education, project evaluation and risk assessment). The most cited PROMETHEE-based algorithm is PROMSORT, which was developed for supplier evaluation and was validated with synthetic data [1]. We have proposed a sorting method to classify suppliers into ordered categories and its validation in real contexts. The algorithm is referred to as Global and Local search of Net Flows (GLNF) sorting, as it is based on global and local searches of the net flow, the main concept of the PROMETHEE II method. The results obtained are compared with PROMSORT by using real supplier evaluation data from a multinational manufacturing company. A sensitivity analysis of the parameters of both methods is also carried out.

GLNF allows sorting the suppliers into ordered classes by applying PROMETHEE in an initial global search, followed by intra- and inter-category local searches. Global search provides an initial classification according to net flow. The sign of the net flows defines the allocations of intra- and inter-category local searches. The quality of the assignments has been measured with the quality index by Rosenfeld De Smet [2] and the new index proposed for sorting quality based on the extension of the silhouette (SILS) concept of data mining.

The results have shown that, for both quality indicators, the GLNF sorting method can achieve equal or better supplier assignments than those obtained with PROMSORT. GLNF is a robust method for multicriteria sorting, classifies all suppliers and requires less information from the decision-maker. In contrast, PROMSORT requires an additional parameter to be defined and may leave some alternatives unclassified. The sensitivity analysis shows that GLNF sorting is a robust algorithm, as modifications in the parameters of preference functions do not affect the results obtained significantly.

Finally, both contributions, the GLNF sorting method and the SILS quality index, can be applied to classify alternatives into ordered groups in other multicriteria problems related to supply chain management.

Keywords: Supplier Segmentation; Multicriteria Sorting; Local Search; PROMETHEE; Supply Chain Management.

References

[1] Araz, C., and Ozkarahan, I. (2007). Supplier evaluation and management system for strategic sourcing based on a new multicriteria sorting procedure. *International Journal of Production Economics*, 106:585–606. doi:10.1016/j.ijpe.2006.08.008.

[2] Rosenfeld, J., and De Smet, Y. (2020). An extension of PROMETHEE to hierarchical multicriteria clustering. *International Journal of Multicriteria Decision Making*, 8(2):133–150. doi:10.1504/IJMCDM.2019.106911.

*Department of Applied Statistics and Operations Research and Quality, Universitat Politècnica de València, Spain. Emails: ivbarji@posgrado.upv.es, cmaroto@eio.upv.es.
†Department of Financial and Actuarial Economics Statistics, Universidad Complutense de Madrid, Spain. Email: marina.segura@ucm.es.

Multivariate functional ordering based on indexes. An application to clustering

Belén Pulido * Alba M. Franco-Pereira † Rosa E. Lillo ‡

B. Pulido (Predoctoral researcher) Belén Pulido is a PhD candidate at the UC3M-Santander Big Data Institute (Universidad Carlos III de Madrid). Her research is related to Functional Data Analysis and Big Data. She received her BSc in Mathematics from Universidad de Málaga in 2019, and her MSc in Big Data Analytics from Universidad Carlos III de Madrid in 2020. After one year and a half working as data scientist in a consultant company, she enrolled in the doctoral program in Mathematical Engineering at Universidad Carlos III de Madrid, where she is currently working.

When dealing with functional data, the problem of ordering functions arises naturally. In the literature there are different possibilities for ordering functions. Two well-known approaches consist of measuring the "centrality" or the "extremality" of functions. The concept of depth for functional data covers the first approach, while the epigraph and the hypograph indexes allow to order from top to bottom in a figurative sense, since there is no natural order in the space of functions.

In many applications, it is usual to find databases with functional data that contain more than one variable to describe a problem as a whole. In these cases, it is necessary to work with multivariate functional data in order to study a common behaviour between the observed variables.

The work [3] proposes a methodology for clustering univariate functional data based on the epigraph and the hypograph indexes. The clustering problem is addressed by applying the epigraph and the hypograph indexes to a functional dataset and thereby, converting it from a functional data problem into a multivariate problem, also considering the indices associated with the first and second derivatives of the data. Once the multivariate dataset is obtained, the techniques that have been fully studied in the literature for clustering multivariate data can be applied.

The work [2] adapts the concept of functional depth for multivariate datasets. Here, we propose an extension of the epigraph and the hypograph indexes, firstly proposed in [1], to the multivariate functional context. The extension is based on geometrically considering the graph of a function in several dimensions. In this sense, the new definitions are based on the contention, for each of the components, of the graph of each curve in that component in the epigraph or the hypograph of a given curve.

There are few contributions in the literature for clustering functional data, but when taking into consideration multivariate functional data, the number of contributions decrease drastically. Thus, the latter mentioned methodology can be applied in the multivariate functional context through the use of the epigraph and the hypograph indexes defined for multivariate functional data. Finally, this process is illustrated through different datasets.

Keywords: Multivariate functional data; Epigraph; Hypograph; Clustering.

Acknowledgements This work has been partially supported by Ministerio de Ciencia e Innovación, Gobierno de España, grant numbers PID2019-104901RB-I00, PID2019-104681RB-I00 and PTA2020-018802-I.

References

[1] Franco-Pereira, A. M., Lillo, R. E. and Romo, J. (2011). Extremality for functional data. In *Recent advances in functional data analysis and related topics*. Springer. 131–134. doi:10.1007/978-3-7908-2736-$1_2$0.

[2] López-Pintado, S., Sun, Y., Lin, J. K. and Genton, M. G. (2014). Simplicial band depth for multivariate functional data. *Advances in Data Analysis and Classification*, 8(3): 321–338. doi:10.1007/s11634-014-0166-6.

[3] Pulido, B., Franco-Pereira, A. M. and Lillo, R. E. (2021). Functional clustering via multivariate clustering. arXiv:2108.00217.

*UC3M-Santander Big Data Institute, Universidad Carlos III de Madrid, Spain. Email: belen.pulido@uc3m.es.
†Department of Statistics and O.R., Universidad Complutense de Madrid, Spain. Email: albfranc@ucm.es.
‡Department of Statistics and UC3M-Santander Big Data Institute, Universidad Carlos III de Madrid, Spain. Email: lillo@est-econ.uc3m.es.

Solving fuzzy multi-objective shortest path problems by ranking approximate Pareto sets

Irene Mariñas-Collado* Susana Montes* Agustina Bouchet*

I. Mariñas-Collado (Assistant Professor) received her BSc in Statistics from the University of Salamanca in 2013 and earned her PhD in Statistics from the University of Glasgow in 2017. She completed her postdoc at Universidad de Salamanca at the end of 2019. She is currently *Profesora Ayudante Doctora* (Assistant Professor) in the Department of Statistics and Operational Research and Didactics of Mathematics at the University of Oviedo and a member of the research group "Modeling of Uncertainty and Imprecision in Decision Theory" (UNIMODE). Her research lines also include optimal design of experiments and statistical modeling of evolutionary processes.

One of the most studied classical optimisation problems is the *Shortest Path Problem* (SPP). It is well-known that the SPP consists in finding a path between two vertices (or nodes) in a network so that the sum of its edges' weights is minimised. As the number of graphs and data used grows with each new application, the development of new algorithms and speedup strategies to handle SP challenges remains an active research topic. The value of the path is normally measured in terms of a single attribute (cost, duration, time, risk, etc.) defined in each edge of the input graph. However, in many cases, a single attribute is insufficient to define the preference between routes. As a result, *Multi-Objective Shortest Path* (MOSP) problems arise, in which various attributes are defined on the edges and thus on the paths. This scenario leads to solution sets that can be exponentially sized relative to the input size of the problem. [**?**]. One of the most popular methods for solving MOSP problems is the construction of approximate Pareto sets. In general terms, a set of Pareto paths includes all incomparable 'best' paths [**?**]. Moreover, in traditional SPPs, there is exact information about the parameters of the problem. However, real-world environments require dealing with uncertainty and so fuzzy notions can be used [**?**]. In Fuzzy SPP, the classic costs between nodes are replaced by fuzzy numbers. Despite their potential, fuzzy shortest path problems have received less attention.

The main aim of this work is to find a way to provide an unique solution to the MOSP problem, rather than a set of optimal paths, while dealing with the fuzzy costs. The proposed approach is based on existing techniques that can be combined. This involves: 1) Map the fuzzy costs into crisp numbers; 2) Find an approximate Pareto set of optimal solutions; 3) Use ranking methods to establish preferences between the optimal solutions.

Ranking methods are usually studied in the field of social choice theory in which several voters express their preferences over a set of alternatives in the form of rankings. The novelty lies in applying these ranking methods to the MOSP problem, where each criterion can be considered a voter, and so all the paths in the set can be ordered according to each criterion. Then, the rankings can be aggregated so that an ordered list of the possible paths in the Pareto set is obtained. Moreover, with this approach, the different criteria can be given more or less weight in the voting process, based on different preferences.

Keywords: Shortest path; Multicriteria; MOSP; Fuzzy; Ranking methods

Acknowledgements This research has been partially supported by Spanish MINECO projects PGC2018-098623-B-I00 (I.M-C. and S.M.) and TIN2017-87600-P (A.B.).

References

[1] Tarapata, Z.M. (2007) Selected multicriteria shortest path problems: An analysis of complexity, models and adaptation of standard algorithms. *International Journal of Applied Mathematics & Computer Science* 17(2): 269–287.

[2] Hanusse, N., Ilcinkas, D., Lentz, A. (2020) Framing algorithms for approximate multicriteria shortest paths. *20th Symposium on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems (ATMOS 2020)* 85.

[3] Dubois, D.J. (1980) *Fuzzy sets and systems: theory and applications* (Vol. 144) Academic press.

*Department of Statistics and Operation Research and Mathematics Didactics, Universidad de Oviedo, Spain. Email: marinasirene@uniovi.es.

Session 6 (Thursday 16.00)

Session talks

An extension of PLS to quantile regression

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

A. Méndez Civieta (Postdoctoral researcher) Álvaro Méndez is a postdoctoral researcher at the UC3M-Santander Big Data institute. His main lines of research are centered in high dimensional problems, regression, and dimensionality reduction techniques, with extensions to quantile regression, and his work has been published in top conferences and journals. He received his BSc degree in Mathematics from University of Oviedo in 2015, and his MSc degree in Big Data analytics from University Carlos III de Madrid in 2016. He also earned a PhD in mathematical engineering working at the statistics department of University Carlos III de Madrid under the supervision of Rosa E. Lillo and M. Carmen Aguilera-Morillo in 2022.

This work presents the fast partial quantile regression (fPQR) methodology [2], a new methodology suitable for solving high dimensional, possibly colinear regression problems where the response can be multivariate. This type of problems can usually be found in fields such as chemometrics, econometrics, genetics etc. The fPQR can be seen as an extension of partial least squares (PLS) [1] to the quantile regression framework, as it shares many of the PLS nice properties. First, it is a dimension reduction technique suitable for multicollinear or high dimensional data. Second, the new scores obtained by the algorithm are uncorrelated. Third, it maximizes a quantile covariance between predictor and response. However, a common problem with PLS is that it is an iterative process based on least squares, which implies that it provides mean based estimates and is extremely sensitive to the presence of outliers, skewness or heteroscedasticity.

The fPQR benefits from its connection to quantile regression to solve these limitations. It is a robust methodology, suitable for dealing with outliers or heteroscedastic data. The fPQR can obtain an estimate of the median as a robust alternative to the mean predictions provived by PLS, but it also can obtain an estimate of any other quantile of interest of the response matrix conditional to the predictors. The estimation of the lower quantiles thus becomes a way of obtaining predictions for a worst-case scenario, while the upper quantiles provide predictions of a best-case scenario, obtaining a complete view of the distribution of the response.

Opposed to the traditional covariance, there is not a unique definition of what a quantile covariance should be. For this reason in this work three different alternatives for this metric are studied through a series of synthetic datasets, and an efficient implementation of fPQR is derived, which is already available as an open-source Python package (https://pypi.org/project/fpqr/).

Keywords: Partial-least-squares; Quantile-regression; Dimension-reduction; Outliers; Robust.

Acknowledgements This research was partially supported by research grants and projects PID2020-113961GB-I00 and PID2019-104901RB-I00 from the Spanish Agencia Estatal de Investigación.

References

[1] Wold, H. (1973). *Nonlinear Iterative Partial Least Squares (NIPALS) Modelling: Some Current Developments.* Multivariate Analysis III. Elsevier Science Ltd.

[2] Mendez-Civieta, A., Aguilera-Morillo, M. C. and Lillo, R. E. (2022). Fast partial quantile regression. *Chemometrics and Intelligent Laboratory Systems*, 223. doi:10.1016/j.chemolab.2022.104533.

*UC3M-Santander Big Data Institute, Universidad Carlos III de Madrid, Spain. Email: alvaro.mendez@uc3m.es.
†Department of Applied Statistics and Operational Research, and Quality, Universitat Politècnica de València, Spain. Email: mdagumor@eio.upv.es.
‡UC3M-Santander Big Data Institute, Department of Statistics, Universidad Carlos III de Madrid, Spain, Email: lillo@est-econ.uc3m.es.

A functional PLS algorithm based on the penalized rank-one approximation of the data

Harold A. Hernández-Roig [*] M. Carmen Aguilera-Morillo [†] Eleonora Arnone [‡]
Rosa E. Lillo [§] Laura M. Sangalli [¶]

H.A. Hernández-Roig (PhD candidate) Harold A. Hernández-Roig is a PhD student at the Department of Statistics, Universidad Carlos III de Madrid. He holds a BSc degree in Mathematics from Universidad de La Habana, 2015, and a MSc in Mathematical Engineering from Universidad Carlos III de Madrid, 2019. His research focuses on functional data analysis, with emphasis on the representation of high-dimensional data as functional, regression models, and principal components techniques.

This work proposes a penalized rank-one approximation to partial least squares (PLS) regression in the context of functional data. The focus is on random fields $\mathcal{X}$ taking values in $L^2(\mathcal{D}) = \{f : \mathcal{D} \mapsto \mathbb{R}, \int_{\mathcal{D}} f^2 < \infty\}$, a space that is equipped with the usual inner product $\langle f, g\rangle = \int_{\mathcal{D}} fg$ and norm $\|f\|^2 = \langle f, f\rangle$, for any $f, g \in L^2(\mathcal{D})$. The random field is assumed to have finite second moment and square-integrable covariance function. In the context of functional regression, the predictor $\mathcal{X}$ has an associated response $Y \in \mathbb{R}^\ell$, $\ell \geq 1$.

The domain $\mathcal{D}$ can be an interval in the real line, a two-dimensional (2D) planar domain, or a Riemannian manifold. An example of the latter case is the cerebral cortex, which can be seen as a 2D surface embedded in a three-dimensional space. The flexibility of $\mathcal{D}$ is one of the advantages of our formulation, as all the literature on functional PLS deals with Euclidean domains that are usually intervals in $\mathbb{R}$.

A common situation in all the functional data analysis problems is the impossibility to have realizations $\mathcal{X}(\mathbf{p})$ at any point $\mathbf{p} \in \mathcal{D}$ of the domain. Usually, authors rely on a presmoothing step that can involve penalties and a representation of the data in terms of basis functions (see [1] and references therein). Another advantage of our approach is that it bypasses this presmoothing step and estimates the PLS components through a rank-one approximation of the raw data. Also, it adds a roughness penalty to control the smoothness of the PLS weights. This solution is similar to the regularized principal components analysis over 2D manifolds developed in [2].

The performance of our method is compared with the penalized functional PLS (based on a basis expansion) presented in [1]. Particularly, the comparison takes into account model accuracy, dimension reduction, and the ability to reconstruct the coefficient function when the data is generated artificially. The study also includes an application in which the predictors are the near-infrared (NIR) spectra of gasoline samples and the responses are their octane numbers.

In summary, the results show similar behavior in terms of accuracy, but a higher dimension reduction when our method is deployed. Furthermore, in the simulation study, our estimates are closer to the true coefficient function.

Keywords: Functional data analysis; Regression; Partial least squares; Roughness penalties.

Acknowledgements Work supported by the UC3M Own Research Program and project ID2019-104901RB-I00 financed by MCIN/AEI/10.13039/501100011033.

References

[1] Aguilera, A. M., Aguilera-Morillo, M. C. and Preda, C. (2016). Penalized versions of functional PLS regression. *Chemometrics and Intelligent Laboratory Systems,* 154:80–92. doi:10.1016/j.chemolab.2016.03.013.

[2] Lila, E., Aston, J. A. and Sangalli, L. M. (2016). Smooth principal component analysis over two-dimensional manifolds with an application to neuroimaging. *Annals of Applied Statistics*, 10(4):1854–1879. doi:10.1214/16-AOAS975.

[*]Department of Statistics, Universidad Carlos III de Madrid, Spain. Email: haroldantonio.hernandez@uc3m.es.
[†]Department of Applied Statistics and Operational Research, and Quality, Universitat Politècnica de València, Spain. Email: mdagumor@eio.upv.es.
[‡]Department of Statistical Sciences, University of Padova, Italy. Email: eleonora.arnone@unipd.it.
[§]Department of Statistics, Universidad Carlos III de Madrid, Spain. Email: rosaelvira.lillo@uc3m.es.
[¶]Department of Mathematics, Politecnico di Milano, Italy. Email: laura.sangalli@polimi.it.

On a conic optimization approach to estimate smooth hypersurfaces using $P-$splines and shape constraints

Manuel Navarro-García * Vanesa Guerrero † María Durban ‡

M. Navarro-García (PhD candidate) Manuel Navarro García is a PhD candidate enrolled in the Mathematical Engineering program at the University Carlos III of Madrid along with the startup Komorebi AI. He earned a double degree in Mathematics and Physics at the University Complutense of Madrid and he received his master degree in Statistics for Data Science at the University Carlos III of Madrid. His main interests are non-parametric regression and mathematical optimization.

The complexity of the data generated at present in many diverse areas makes necessary the development of innovative methodologies which are able to incorporate human knowledge to, for instance, enhance interpretability and avoid misleading out-of-range predictions. In this work, we address the problem of estimating smooth functions in a regression task for data lying on large grids, and where the data fit needs to satisfy requirements about their sign and shape.

We assume that the smooth function to be estimated is defined through a reduced-rank basis ($B-$splines) and fitted via a penalized splines approach ($P-$splines [1]). When dealing with simple regression, necessary and sufficient conditions on the sign of the smooth curve to be estimated are proposed and embedded into the fitting procedure as hard constraints, yielding a conic optimization model. This characterization can be adapted to enforce sign constraints on higher order derivatives of the curve, meaning that our methodology can deal with requirements concerning the monotonicity and the curvature of the curve. Our approach for multiple regression arises as a non-trivial extension of the one-dimensional shape constrained framework, imposing the constraints over a finite set of curves which belong to the hypersurface. Furthermore, previous results are generalized for the first time to out-of-sample prediction, either forward and backward.

In summary, the contributions of this work are fourfold. First, a mathematical optimization formulation for the estimation of non-negative $P-$splines is proposed. Second, the problem of constrained out-of-sample prediction using $P-$splines is successfully addressed for the first time. Third, our approach extends the non-negativity constraint in the constrained smoothing and prediction approaches to other requirements, such as monotonicity and curvature, and to the case in which multiple constrained functions of the same predictor and response need to be estimated simultaneously for different groups of observations. The fourth and last contribution is the development of an open source Python library, `cpsplines`, which contains the implementations of all the methodologies developed in this work.

The methodologies presented in this work are illustrated using simulated instances and data about the evolution of the COVID-19 pandemic and about mortality rates for different age groups.

Keywords: Data science; Penalized splines; Conic optimization; Smoothing; Prediction.

Acknowledgements The research of Manuel Navarro-García has been financed by the research project IND2020/TIC-17526 (Comunidad de Madrid) and PID2019-104901RB-I00 (funded by MCIN/AEI/10.13039/501100011033). This support is gratefully acknowledged.

References

[1] Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with $B-$splines and penalties. *Statistical Science*, 11(2): 89–121. doi:10.1214/ss/1038425655.

*Department of Statistics, University Carlos III of Madrid, Spain & Komorebi AI Technologies, Madrid, Spain. Email: mannavar@est-econ.uc3m.es.
†Department of Statistics, University Carlos III of Madrid, Spain. Email: vanesa.guerrero@uc3m.es.
‡Department of Statistics, University Carlos III of Madrid, Spain. Email: mdurban@est-econ.uc3m.es.

Session 7 (Friday 11:00)

Session talks

Solving the premarshalling problem under limited crane time in the constraint programming paradigm

Celia Jiménez-Piqueras * Rubén Ruiz † Consuelo Parreño-Torres ‡ Ramon Alvarez-Valdes §

C. Jiménez Piqueras (PhD candidate) Celia Jiménez Piqueras is a PhD student at Universitat Politècnica de València. Her main interests are Linear Programming and Constraint Programming. She received her BSc from Universitat de València in 2017 in Mathematics. She earned her master's degree in Statistics and Operations Research from Universitat Politècnica de Catalunya and Universitat de Barcelona in 2019. In 2020, she enrolled in the doctoral program in Statistics and Optimization at Universitat Politècnica de València. Currently, her research focuses on container relocation problems at port terminals.

The majority of world trade in goods is carried by sea, and there is a great difficulty in operating port terminals efficiently. In that context, the study of optimization problems is crucial in dealing with demand peaks, reducing shipping costs, and minimizing the environmental impact.

The premarshalling problem is defined in the port yard, an area where containers are temporarily stored. Containers are organized in stacks, so the container at the top is the only one the crane can reach. When a container has to be retrieved, every container above it has to be relocated first, which is very time-consuming. Given a set of adjacent stacks, called a bay, the premarshalling problem aims to find the shortest sequence of crane movements that produces a bay configuration where every container is accessible at its retrieval time. Hence, after performing the premarshalling process, every container in the bay can be retrieved without any additional relocation at the moment it is required.

The premarshalling process is performed when the cranes are not used in other operations. However, the time to complete the premarshalling may exceed the period the crane is free. In this case, it would be advantageous to find a solution that, even if the bay is not fully ordered in the final layout, at least those containers which have to leave the bay first are accessible. This partial premarshalling is also very interesting if the retrieval times of the containers that will be retrieved at the latter part of the process are unknown.

The classical formulation of the problem does not provide, in general, a good solution in case the premarshalling has to be partially performed. Also, it does not take into account the time spent by crane movements. In this work, we propose a constraint programming formulation for a more realistic version of the premarshalling problem that considers the time needed for each movement and provides the best partial premarshalling under a given limited crane time.

The premarshalling problem with the objective of minimizing crane times is introduced by [1]. Regarding the partial premarshalling, to the best of our knowledge, the closest proposal is the pre-processing phase in the stochastic container relocation problem approach developed by [2].

Keywords: Logistics; Container terminal optimization; Premarshalling problem; Constraint programming.

Acknowledgements This study has been partially supported by the Spanish Ministry of Science and Innovation under predoctoral grant PRE2019-087706 and the project "OPTEP-Port Terminal Operations Optimization" (No. RTI2018-094940-B-I00) financed with FEDER funds.

References

[1] Parreño-Torres, C., Alvarez-Valdes, R., Ruiz, R., and Tierney, K. (2020). Minimizing crane times in pre-marshalling problems. *Transportation Research Part E: Logistics and Transportation Review*, 137:101917. doi:10.1016/j.tre.2020.101917.

[2] Zweers, B.G., Bhulai, S., and van der Mei, R.D. (2020). Pre-processing a container yard under limited available time. *Computers & Operations Research*, 123:105045. doi:10.1016/j.cor.2020.105045.

*Grupo de Sistemas de Optimización Aplicada, Instituto Tecnológico de Informática, Universitat Politècnica de València, Spain. Email: cejipi@upvnet.upv.es.

†Grupo de Sistemas de Optimización Aplicada, Instituto Tecnológico de Informática, Universitat Politècnica de València, Spain. Email: rruiz@eio.upv.es.

‡Department of Statistics and Operations Research, Universitat de València, Spain. Email: consuelo.parreno@uv.es.

§Department of Statistics and Operations Research, Universitat de València, Spain. Email: ramon.alvarez@uv.es.

A new nonparametric approach for the latency: an application to the financial field

Ana López-Cheda * M. Amalia Jácome † Yingwey Peng ‡

A. López-Cheda (Beatriz Galindo fellow) Ana López-Cheda is a distinguished researcher at the University of A Coruña (UDC), hired with a Beatriz Galindo contract in the field of Mathematics. Her research work has been, mainly, associated with survival analysis and its application to medical databases. She received her bachelor's degree from UDC in 2012 in Computer Sciences. She earned her master's degree in Statistical Techniques from the UDC in 2014. After that, she enrolled in the doctoral program in Statistics and Operational Research at UDC, under the supervision of professor R. Cao and professor M. A. Jácome. Her thesis defense was held in 2018.

Nonparametric estimation methods for the cure rate and the distribution of the failure time of uncured subjects with covariates for censored survival data have attracted much attention in the last few years. To model the effects of covariates on the distribution of the failure time of uncured subjects (latency), existing works assume that the cure rate (incidence) is a constant or depends on the same covariate as the distribution of uncured subjects. A new nonparametric estimator for the distribution of uncured subjects that relaxes the assumption used in the existing works in the literature is proposed. The methodology further extends the approaches by [1], [2] and [3] to allow different covariates in the cure rate and latency parts. Specifically, the method is based on the EM algorithm, which is readily available for mixture cure models, for fitting the mixture cure model with different covariates in the cure rate and latency parts. Furthermore, an important step in the nonparametric mixture cure model is to determine the bandwidths for both the incidence and the latency estimators. A bootstrap procedure to determine the optimal bandwidth in the proposed method is presented.

In addition, the finite sample performance of the proposed estimator is assessed and compared with existing methods in an extensive simulation study. Finally, the nonparametric estimation method is employed to model the effects of some covariates on the time to bankruptcy among commercial banks insured by the Federal Deposit Insurance Corporation in the United States in 2006 – 2016.

Keywords: Bootstrap; Censored data; EM algorithm; Mixture cure models; Survival analysis.

Acknowledgements ALC was sponsored by the BEATRIZ GALINDO JUNIOR Spanish Grant from MICINN (Ministerio de Ciencia e Innovación) with code BGP18/00154. ALC and MAJ acknowledge partial support by the MICINN Grant PID2020-113578RB-I00 and partial support of Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2020-14). ALC and MJ wish to acknowledge the support received from the Centro de Investigación de Galicia "CITIC", funded by Xunta de Galicia and the European Union European Regional Development Fund (ERDF)- Galicia 2014-2020 Program, by grant ED431G 2019/01.

References

[1] Xu, J. and Peng, Y. (2014) Nonparametric cure rate estimation with covariates. *Canadian Journal of Statistics*, 42:1–17. doi:10.1002/cjs.11197.

[2] López-Cheda, A., Cao, R., Jácome, M. A. and Van Keilegom, I. (2017). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models *Computational Statistics and Data Analysis*, 105:144–165. doi:10.1016/j.csda.2016.08.002.

[3] López-Cheda, A., Jácome, M. A. and Cao, R. (2017). Nonparametric latency estimation for mixture cure models. *TEST*, 26:353–376. doi:10.1007/s11749-016-0515-1.

*MODES, CITIC, Departament of Mathematics, University of A Coruña, Spain. Email: ana.lopez.cheda@udc.es.
†MODES, CITIC, Departament of Mathematics, University of A Coruña, Spain. Email: maria.amalia.jacome@udc.es.
‡Departments of Public Health Sciences and Mathematics and Statistics, and Cancer Care and Epidemiology Division of Cancer Research Institute, Queen's University, Canada. Email: pengp@queensu.ca.

The length constrained rural postman problem with a fleet of drones

James F. Campbell * Ángel Corberán † Isaac Plana ‡ J. M. Sanchis § Paula Segura ¶

P. Segura (PhD candidate) Paula Segura Martínez is a PhD candidate at the Universitat Politècnica de València. Her main interests are combinatorial optimization and linear programming focused on routing problems. She received her degree in mathematics from the Universidad de Alicante in 2017. She earned her master in advanced mathematics from the Universidad de Murcia in 2018. Currently, she is doing her PhD at the Universitat Politècnica de València under the direction of Ángel Corberán, Isaac Plana, and José María Sanchis.

Classical arc routing problems (ARPs) consist of finding a tour, or a set of tours, with total minimum cost traversing (and servicing) a set of links (arc or edges), called required links, of a graph. Well–known ARPs are the Chinese postman problem (CPP) and the rural postman problem (RPP), where a single vehicle has to traverse all or some of the links of the graph, respectively. In classical ARPs, the streets to be cleaned, roads where snow must be removed, or pipelines to be inspected, for example, are represented by edges or arcs of a network that ignore the line shape (although not its true cost or distance) since the vehicles have to traverse an arc from one endpoint to the other one. Further, the vehicles in these classical ARPs are not permitted to travel off the network. Aerial drones may be used to replace ground vehicles because of their reduced costs, higher speeds and/or safety improvements.

The length constrained K–drones rural postman problem (LC K–DRPP) is a continuous optimization problem where a set of curved or straight lines of a network have to be traversed, in order to be serviced, by a fleet of homogeneous drones, with total minimum cost. Since the range and endurance of drones is limited, we consider here that the length of each route is constrained to a given limit L. Unlike ground based vehicles that are limited to follow local ground infrastructure such as roadways or paths, aerial drones allow more direct travel and can easily fly across areas without roads, over bodies of water, and so on. Therefore, drones are not restricted to travel on the network, and they can enter and exit a line through any of its points, servicing only a portion of that line [1]. To deal with this problem, LC K–DRPP instances are digitized by approximating each line by a polygonal chain with a finite number of points and allowing drones to enter and exit each line only at these points. Thus, an instance of the Length Constrained K–vehicles Rural Postman Problem (LC K–RPP) is obtained. This is a discrete arc routing problem, and therefore can be solved with combinatorial optimization techniques. However, when the number of points in each polygonal chain is very large, the LC K–RPP instance can be so large that it is very difficult to solve, even for heuristic algorithms.

We present here the main procedures we have developed for solving the problem: a matheuristic algorithm ([2]), a formulation for the LC K–RPP, summaryzing the polyhedral study of the set of solutions of a relaxed formulation with some families of facet–inducing inequalities, and a branch–and–cut algorithm that incorporates the separation of these inequalities. Extensive computational experiments to assess the performance of the algorithms will be discussed.

Keywords: Rural postman problem; Drones; Heuristic; B&C.

Acknowledgements Work supported by the Spanish Ministerio de Ciencia, Innovación y Universidades (MICIU) and Fondo Social Europeo (FSE) through project PGC2018–099428–B–I00.

References

[1] Campbell, J.F., Corberán, Á., Plana, I., and Sanchis, J.M. (2018). Drone Arc Routing Problems. *Networks*, 72:543–559. doi:10.1002/net.21858.

[2] Campbell, J.F., Corberán, Á., Plana, I., Sanchis, J.M., and Segura, P. (2021). Solving the length constrained K–drones rural postman problem. *European Journal of Operational Research*, 292(1):60–72. doi:10.1016/j.ejor.2020.10.035.

*Supply Chain and Analytics Department, University of Missouri–St. Louis, USA. Email: campbell@umsl.edu.
†Department d'Estadística i Investigació Operativa, Universitat de València, Spain. Email: angel.corberan@uv.es.
‡Departmento de Matemáticas Para la Economía y la Empresa, Universitat de València, Spain. Email: isaac.plana@uv.es.
§Departmento de Matemática Aplicada, Universidad Politécnica de Valencia, Spain. Email: jmsanchis@mat.upv.es.
¶Departmento de Matemática Aplicada, Universidad Politécnica de Valencia, Spain. Email: psegmar@upvnet.upv.es.

Comparisons of VaR and CoVaR in terms of the value of the conditional variable

Patricia Ortega-Jiménez * Franco Pellerey † Miguel A. Sordo ‡ Alfonso Suárez-Llorens §

P. Ortega-Jiménez (PhD candidate) Patricia Ortega-Jiménez is a PhD student, currently working at University of Cádiz. She earned both her BSc and MSc in mathematics in the University of Cádiz in 2016 and 2018, respectively. After one year working as an statistician in the research group "Observatorio del Dolor", specialized in chronic pain, she enrolled the doctoral program in mathematics at the University of Cádiz, supervised by Miguel A. Sordo and Alfonso Suárez-Llorens. Her main interests are stochastic orders and copula modeling.

Let us consider a random vector (X, Y) with copula C. Recall that the copula models the inner dependence of the vector, independently from the marginal behavior of the components. Given a risk level $v \in [0, 1]$, in order to study the risk of one of the components, Y, the most extended risk measure is the Value at Risk, $VaR_v(Y) = F_Y^{-1}(v)$, which represents the maximum expected loss. However, the $VaR_v(Y)$ measures the risk of the single institution, not taking into account the system as a whole or the interactions with other risks. In [1] it was adopted a dependence-adjusted version of the Value at Risk, the *Co*-Value at risk, $CoVaR_{v,u}(Y|X)$, which takes into account systemic risk and stands for $VaR_v(Y|X = VaR_u(X))$ for the risk levels $v \in [0, 1]$ and $u \in [0, 1]$. Our goal is to find the values of the institution X that lead to the $CoVaR$ being greater than the VaR of Y. For these values of X, the capital risk based on the VaR can not be enough to face financial losses and the $CoVaR$ should be considered.

We compare these two measures in terms of the risk-level of the conditional variable, u. If $v \in (0, 1)$ is fixed, then there always exist values of u for which $CoVaR_{v,u}(Y|X)$ lies above $VaR_v(Y)$. If C has continuous partial derivatives, and Y is stochastically increasing in X ($Y \uparrow_{SI} X$), which reflects positive dependence, there exists a unique cut point u_v such that $CoVaR_{v,u_v}(Y|X) \geq VaR_v(Y)$ if and only if $u \geq u_v$. This value u_v does not depend on the marginal distributions but only on the dependence structure of the vector. Also, if there exists an upper bound $u^* \in (0, 1)$ such that $u_v \leq u^*$ for all $v \in (0, 1)$, then, for all $u \geq u^*$ and all $v \in (0, 1)$, $CoVaR_{v,u}(Y|X)$ lies above $VaR_v(Y)$ and, as seen in [2], $Y \leq_{ST} Y|X = x$ for all $x \geq F_X^{-1}(u^*)$. Sufficient conditions are provided, in terms of the copula, under which the bound exists.

Several examples of copulas with bounded and unbounded cut points are analyzed. The analytical expression of such a value u^* is provided for many of the bivariate copula families. In order to illustrate the results, we study a real numerical example, analyzing the vector of the log returns of two companies from IBEX35: International Consolidated Airlines Group S.A. and Meliá Hotels International S.A. In order to decide whether we should base the capital risk of Meliá Hotels in the VaR or in the $CoVaR$, we study the cut points of $CoVaR$ and VaR in terms of the airlines group, obtaining the values of u_v and u^* for this particular case.

Keywords: CoVaR; Copula; Stochastic comparison.

Acknowledgements The authors acknowledge University of Cádiz for the PhD grant (call 2018) linked to the project MTM2017-89577-P financed by Ministerio de Economía y Competitividad of Spain.

References

[1] Adrian, T. and Brunnermeier, M. K. (2016). CoVaR. *American Economic Review, American Economic Association*, 106(7):1705–1741. doi:10.1257/aer.20120555.

[2] Navarro, J. and Sordo, M. A. (2018). Stochastic comparisons and bounds for conditional distributions by using copula properties. *Dependence Modeling*, 6(1):156–177. doi:10.1515/demo-2018-0010.

*Department of Statistics and Operation Research, University of Cádiz, Spain. Email: patricia.ortega@uca.es.
†Department of Mathematical Sciences, Politecnico di Torino, Turin, Italy. Email: franco.pellerey@polito.it.
‡Department of Statistics and Operation Research, University of Cádiz, Spain. Email: mangel.sordo@uca.es.
§Department of Statistics and Operation Research, University of Cádiz, Spain. Email: alfonso.suarez@uca.es.

Session 8 (Friday 12:35)

Session talks

Machine learning and optimization: an approach for real-world discrete problems

Alessandro Baldo* Matteo Boffa† Lorenzo Cascioli‡ Edoardo Fadda § Chiara Lanza¶ Arianna Ravera‖

E. Fadda (Tenured Assistant Professor) received the Laurea degree in mathematical engineering and the PhD degree in information technology and system engineering from the Politecnico di Torino, in 2014 and 2018, respectively. Now he is Tenured Assistant Professor with the Dipartimento di Scienze Matematiche, Politecnico di Torino His main interests are optimization under uncertainty, machine learning and their applications.

Optimization problems with binary decision variables appear over a wide range of real world applications and are often computationally expensive to solve. Therefore, the need for an heuristic to quickly provide good solutions. In this setting, we propose a new and easy to apply framework using Machine Learning (ML) to tackle linear optimization problems characterized by binary variables.

The idea of the framework is to apply classification techniques to assign to a large set of binary variables the value 0 or 1 (thus reducing the dimension of the problem). In order to do so, each binary variable is associated with a set of features that can be easily computed (e.g. its solution in the continuous relaxation of the problem, its reduced costs, and other problem-specific information). More specifically, we exploit the power of a supervised ML algorithm which assigns to each binary variable of an instance the probability to assume value 0 or 1 in the optimal solution. From these estimated probabilities, the variables whose result is most certain are fixed. Contrarily, the most uncertain ones are further processed by the exact solver, which however benefits from the previous reduction of the search space. It follows that the complexity of the last run can be modified by simply changing the percentage of variables which are fixed by the classifier. This eventually provides a two-fold tool, which can either focus on short running times or on excellent objective function results.

The aforementioned classification algorithm is trained on a supervised manner by creating an ad-hoc data-set which collects the information gather from the solutions of several small instances of the problem. In particular, each record of the data-set contains the variable features and the optimal value computed by an exact solver (0 or 1 since the variable are binary). With this interpretation, we are able to cast the optimization problem into a classification one, allowing the use of several machine learning algorithms such as neural networks, support vector machines, and many others.

We applied the proposed framework to several variants of the the knapsack problem inspired from real-world applications. These include the robust knapsack that considers cost uncertainty, the quadratic knapsack that takes into account the mutual benefits associated with tuples of items, the polynomial knapsack which generalizes the quadratic knapsack to benefits among subsets of items of any cardinality and to the polynomial robust knapsack problem which extends the polynomial problem to the robust setting [1]. The computational experiments prove that the heuristic is able to produce close-to-optimal results in a short computation time. Thus, potentially opening up new perspectives in the application of operations research to real-world problems.

Keywords: Heuristics; Machine learning; Discrete optimization problems.

References

[1] Baldo, A., Boffa, M., Cascioli L., Fadda E., Lanza C., and Ravera A. (2022). The Polynomial Robust Knapsack Problem. *Preprint*.

*ISIRES, Italy. Email: alessandro.baldo@isires.org.
†Department of Electronics and Telecommunications, Italy. Email: matteo.boffa@polito.it.
‡ISIRES, Italy. Email: lorenzo.cascioli@isires.org.
§Department of Mathematics, Politecnico di Torino, Italy. Email: edoardo.fadda@polito.it.
¶ISIRES, Italy. Email: chiara.lanza@isires.org.
‖ISIRES, Italy. Email: arianna.ravera@isires.org.

Robust multivariate control chart based on shrinkage for individual observations

Elisa Cabana * Rosa E. Lillo †

E. Cabana (Postdoctoral researcher) is a postdoctoral researcher at the IMDEA Networks Institute in Madrid, Spain. Her main interests are robust methods for data analysis, machine learning and artificial intelligence. She received her BSc in Mathematics from University of Havana in 2012. She earned her MSc and PhD degree in Mathematical Engineering from the University Carlos III of Madrid, Spain, in 2015 and 2019, respectively.

Statistical process control charts are the most popular tool for monitoring production quality in the industry since they allow to search for abnormal or out-of-control behavior. In the univariate case, the classical approach is to use control charts that study the behavior of the mean and the variability of the process at the same time. These are known as the Shewhart control charts. These control charts allow to define a quality control process that consists of two distinct phases. In Phase I, historical observations are used to create the control charts for retrospectively testing whether the process is in control detecting abnormal behavior. By removing the out-of-control data, the parameters of the in-control process can be estimated. Once this is accomplished, in Phase II, future samples obtained during the manufacturing process can be monitored. Nowadays, it is very common that the dataset available is characterized by more than one variable. In this case, monitoring the variables one by one using the univariate control charts can be misleading and the variability due to their relationship would not be taken into account. Therefore, multivariate statistical process control techniques are more appropriate. Harold Hotelling introduced the first approach within this area in his pioneer paper in 1947 (see [1]). It was basically an extension of the Shewhart control charts to the multivariate case.

The problem divides into two cases: (i) when there are groups or subsamples in the data, and (ii) when the data consists of individual observations. In this paper, we focus on the latter case. The classical approach suffers from the disadvantage of using the classical sample estimators which are sensitive to the presence of outliers. This motivates our proposed method, a robust multivariate quality control technique for individual observations, based on the robust reweighted shrinkage estimators. These estimators were introduced in previous work for robust multivariate outlier detection [2] and robust regression [3], and prove to have a good performance. A simulation study is done to check the performance and compare the proposed method with the classical Hotelling approach, and a robust alternative based on the reweighted minimum covariance determinant estimator. The simulations include diverse scenarios, for example, the outlier-free and outliers-present cases, and also scenarios with independent or correlated variables. Several choices for the parameters are explored (sample size, dimension, level of contamination, etc.). The results show the appropriateness of the method even when the dimension or the Phase I contamination are high, with both independent and correlated variables, showing additional advantages in terms of computational efficiency. The approach is illustrated with two real data-set examples from production processes.

Keywords: Robust data analysis; Mahalanobis distance; Multivariate outliers; Quality control; Shrinkage.

Acknowledgements Work supported by Project PID2019-104901RB-I00 from Ministerio de Ciencia e Innovacion.

References

[1] Hotelling, H. (1947). Multivariate quality control. *Techniques of Statistical Analysis.* McGraw-Hill, New York.

[2] Cabana, E., Lillo, R.E. and Laniado, H. (2021). Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators. *Statistical Papers*, 62(4), 1583–1609. doi:10.1007/s00362-019-01148-1.

[3] Cabana, E., Lillo, R.E. and Laniado, H. (2020). Robust regression based on shrinkage with application to Living Environment Deprivation. *Stochastic Environmental Research and Risk Assessment*, 34(2), 293–310. doi:10.1007/s00477-020-01774-4.

*IMDEA Networks Institute and UC3M-Santander Big Data Institute, University Carlos III of Madrid, Spain. Email: elisa.cabana@imdea.org

†Department of Statistics and UC3M-Santander Big Data Institute, University Carlos III of Madrid, Spain. Email: lillo@est-econ.uc3m.es

An alternative representation of neural networks using polynomials: NN2Poly

Pablo Morala * Jenny Alexandra Cifuentes † Rosa E. Lillo ‡ Iñaki Ucar §

P. Morala (PhD Candidate) Pablo Morala is a PhD candidate in Mathematical Engineering at Universidad Carlos III de Madrid, Spain. He received in 2019 a Double BSc degree in Mathematics and Physics at Universidad de Oviedo and in 2020 a MSc in Statistics and Data Science at Universidad Carlos III de Madrid. His main interests are machine learning modeling and its interpretabilty, specially neural networks.

Neural networks have been established as one of the most used machine learning tools, specially with the developments of deep learning. However, their opacity and black box nature still posses an open problem and is an active field of research. Furthermore, determining the appropriate size of a neural network is a procedure that, until now, has been carried out on a trial-and-error basis. In this context, there is an increasing interest in finding an overlapping area between neural networks and more traditional statistical methods, which could be used as alternative representations of the former and help overcome these issues.

In this context, the NN2Poly method presented here aims to obtain an alternative representation of a neural network by means of polynomial regressions. To do so, the weights from a given trained neural network are used to obtain the coefficients of a polynomial (or several polynomials in a classification setting) that obtains almost identical predictions as the original neural network. The initial setting of NN2Poly was settled in [1], where this problem was solved for single hidden layer neural networks. To obtain the explicit formula for the polynomial coefficients, Taylor expansion is applied at each activation function and then several combinatorial properties are used to find the coefficient associated to each combination of variables. The order of the final polynomial will be determined by the chosen degree at which the Taylor expansion is truncated.

The generalization of NN2Poly to arbitrarily deep feed forward neural networks (multilayered perceptrons) is presented in [2], where the same Taylor expansion idea is used but the combinatorial properties needed to obtain the final polynomial increase in complexity and an interesting problem about multiset partitions arises. Related with this, some computational limitations appear in practice, as the number of needed combinations grows rapidly with the order of the polynomial, and this order increases with each hidden layer. Furthermore, some restrictions need to be imposed on the hidden layers weights to ensure that the Taylor expansion remains valid. The performance of the method is empirically tested via simulations and real data examples.

This alternative representation can help interpret the neural network output, as each polynomial coefficient is associated with a combination of variables and the total number of parameters is significantly reduced. Additionally, using this classical statistical model to represent neural networks can be useful to study their properties employing a different framework with a distinct and novel perspective.

Keywords: Neural networks; Interpretabilty; Deep learning; Polynomial regression.

Acknowledgements This research is part of the I+D+i projects PID2019-104901RB-I00 and PID2019-106811GB-C32, funded by MCIN/AEI/10.13039/501100011033/.

References

[1] Morala, P., Cifuentes, J. A., Lillo, R. E. and Ucar, I. (2021). Towards a mathematical framework to inform neural network modelling via polynomial regression. *Neural Networks*, 142:57–72. doi:10.1016/j.neunet.2021.04.036.

[2] Morala, P., Cifuentes, J. A., Lillo, R. E. and Ucar, I. (2021). NN2Poly: A polynomial representation for deep feed-forward artificial neural networks. arXiv:2112.11397.

*Department of Statistics, Universidad Carlos III de Madrid, Spain. Email: pablo.morala@uc3m.es.
†ICADE, Department of Quantitative Methods, Faculty of Economics and Business Administration, Universidad Pontificia Comillas, Spain. Email: jennyalexandra.cifuentes@uc3m.es.
‡Department of Statistics and uc3m-Santander Big Data Institute, Universidad Carlos III de Madrid, Spain. Email: lillo@est-econ.uc3m.es.
§uc3m-Santander Big Data Institute, Universidad Carlos III de Madrid, Spain. Email: inaki.ucar@uc3m.es.

3RD SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Elche, 21st-23rd of September 2022