Doc-Course on Data Science

Institute of Mathematics of the University of Granada (IMAG)

Institute of Mathematics of the University of Seville (IMUS)

September 17 - November 14, 2023

Workshop proceedings

IMAG
INSTITUTO DE MATEMÁTICAS
Universidad de Granada

imus
instituto de matemáticas
universidad de sevilla

INSTITUTO ANDALUZ DE MATEMÁTICAS

UNIVERSIDAD
DE GRANADA

UNIVERSIDAD
DE SEVILLA
1505

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN

Esta actividad es parte de los proyectos de I+D+i:
> PID2020-114594GB-C21, financiado por MCIN/AEI/10.13039/501100011033
> IMAG-María de Maeztu grant CEX2020-001105-M/AEI/10.13039/501100011033

Contents

Presentation

Due to the large amount of data generated by new technologies, Data Science is essential not only for enhancing the performance of companies but also for improving the lives of citizens of today's society in all their facets. This is why the training of researchers in advanced and novel Statistics, Operation Research and Data Science methodologies is really important to improve decision-making based on the available data.

The International Doctoral Course on Data Science (DCDS23) was held at the Institutes of Mathematics of Granada (IMAG) and Seville (IMUS) during two months, September 17 to November 14, 2023.

The Doc-Course activity was aimed at PhD/Master students and consisted of a training period (a 4-week course), visits by researchers and professors to the IMAG/IMUS facilities throughout the period, a supervised research period for participating students (4 weeks) and an international closing workshop in which researchers, lecturers, supervised students and tutors participate.

Eighteen students who were competitively selected among all applicants, participated in this Doc-Course. All accommodation and living expenses of the selected students were paid by the organization. More detailed information in the DCDS23 web https://www.imus.us.es/congresos/DCDS23/

The International Doc-Course Closing Workshop took place from November 13 to 14, 2023 at IMAG. It consisted of the presentation of the research works carried out by the students in the period of supervision. In addition, 2 plenary talks by US/UGR professors were given. All students presented their research work very successfully, demonstrating a very good assimilation of the courses and high capacity to delve deeper into them and develop applications of interest in very diverse areas.

DCDS 2023 has been for all of us, students and professors, a very pleasant and enriching experience that has allowed us to establish personal and professional relationships that will last forever.

Ana M. Aguilera and Miguel A. Pozo (Organizing Committee)

Committees

Organizing Committee

Ana M. Aguilera - Universidad de Granada - IMAG, Spain

Miguel A. Pozo Universidad de Sevilla - IMUS, Spain

Scientific Committee and Editors

Ana M. Aguilera - Universidad de Granada - IMAG, Spain

Miguel A. Pozo Universidad de Sevilla - IMUS, Spain

Schedule

Monday, November 13

9.30 - 10.00	*Opening Session*
10.00 - 11.00	*Plenary session 1*
	Speaker: María Alonso-Pena
	Chair: Ana Aguilera
11.00 - 11.30	*Coffee break*
11.30 - 13.30	*Session 1*
	Chair: Francesca Bellissimo
13.30 - 16.00	*Lunch break*
16.00 - 17.00	*Session 2 (Part 1)*
	Chair: Alberto Torrejón
17.00 - 17.30	*Coffee break*
17.30 - 18.30	*Session 2 (Part 2)*
	Chair: Alberto Torrejón

Tuesday, November 14

10.00 - 11.00	*Plenary session 2*
	Speaker: Luisa I. Martínez-Merino
	Chair: Miguel A. Pozo
11.00 - 11.30	*Coffee break*
11.30 - 13.30	*Session 3*
	Chair: Sergio García-Carrión

Plenary session 1 (Monday 10:00)

Contents

Going beyond classical: circular data, regression and mode estimation

María Alonso-Pena [*] Rosa M. Crujeiras [†]

María Alonso-Pena (Postdoctoral researcher) María Alonso-Pena is a postdoctoral researcher at the Institute of Mathematics of the University of Granada. Before that, she was a postdoctoral researcher at the Research Center for Operations Research and Statistics of KU Leuven, Belgium. María obtained her PhD in statistics from the Universities of Santiago de Compostela, Vigo and A Coruña in 2022. Before that, she obtained a master´s degree in statistics from the same universities and a bachelor degree in mathematics from the University of Santiago de Compostela. Her main research interests are nonparametric methods for density and regression estimation and statistical inference for directional variables. In addition, she is a member of the cohort of the Young Academy of the European Mathematical Society.

When analyzing data, one often employs statistical learning methods that assume that the data are supported on a Euclidean space. Although this assumption is frequently true, there are some cases in which observations are naturally defined on other manifolds. A really simple example is observations that correspond to angles or directions, or that are simply periodic, which are supported on the unit circumference. While the circumference seems like a straightforward manifold, the lack of Euclidean support complicates the use of classical data science methods (see [1]).

An interesting problem when dealing with directional data is the study of regression models, where three different scenarios arise depending on the nature of the variables: circular response and real-valued covariate, real-valued response and circular covariate and the case where both variables are directional. Depending on the setting, the observations might *live* on the surface of either a cylinder or a torus. Due to this, it is necessary to formulate new regression models and estimation methods to obtain statistical techniques which respect the periodicity of the variables, starting by the definition of the regression function, usually considered as the conditional mean (see, for example, [2]).

Another important regard is the fact that the mean direction is not always defined for directional random variables. This implies that the classical definition of regression function, the conditional mean direction, is also not always defined, and thus the estimation of this quantity might not be always appropriate. Instead, other conditional location measures can be targeted, such as the modes, whose definition matches the one in the real line.

In this talk, based on [3], we will introduce some basic concepts about circular data, motivating why it is necessary to employ different methods for this kind of observations through the use of a real data example. The three circular regression settings will be also introduced, presenting both parametric and nonparametric estimation methods. Lastly, we will present the modal regression approach for circular variables, and study some of its properties.

Keywords: Circular data; Directional variables; Kernel regression; Modal regression; Nonparametric statistics.

Acknowledgements This work was supported by grant PID2020-116587GB-I00 funded by MCIN/ AEI/10.13039/501100011033 and the Competitive Reference Groups 2021/2024 (ED431C 2021/24) from the Xunta de Galicia through the ERDF. Research of M. Alonso-Pena was supported by the Xunta de Galicia through the grant ED481A-2019/139 from the Consellería de Educación, Universidade e Formación Profesional.

References

[1] Mardia, K.V. and Jupp, P.E. (2000). *Directional Statistics*. Wiley.

[2] Di Marzio, M., Panzera, A., and Taylor, C.C. (2013). Non-parametric regression for circular responses. *Scand. J. Stat.*, **40**, pp. 238–255.

[3] Alonso-Pena, M. and Crujeiras, R.M. (2023). Analyzing animal escape data with circular nonparametric multimodal regression. *Ann. Appl. Stat.*, **17**, pp. 130–152.

[*] Institute of Mathematics, University of Granada, Spain. Email: mariaalonso.pena@usc.es

[†] CITMAga, Universidade of Santiago de Compostela, Spain. Email: rosa.crujeiras@usc.es

Session 1 (Monday 11:30)

Contents

Population prediction using functional data with spatial dependence

Damián Pérez [*] Carmen Aguilera [†] Ana Aguilera [‡] María Durbán [§]

Damián Pérez Asensio (Master's student) Damián Pérez Asensio is a Master´s student at the Polytechnic University of Valencia. He has studied a double degree in Telecommunication Technologies and Services Engineering and Business Administration and Management. Ever since he discovered the world of Operations Research through a subject in his Business Administration and Management degree, he has been interested in continuing his training in this field. In fact, when he did his Final Degree Project, he decided to work on the optimisation of health resources, resulting in a work entitled "Algorithms to solve a real problem of allocation of health centres". Damián is now studying a Master's Degree in Data Analysis, Process Improvement and Decision Making Engineering carried out by the Department of Applied Statistics and Operations Research and Quality at the Polytechnic University of Valencia.

Health is an important part of welfare state services and therefore of public policy. This is reflected in the amount of money allocated each year in the state budget to this sector. However, this does not seem to be sufficient to alleviate the heavy burden of care in health centres.

In view of the ageing of the population [1], the pressure on hospitals is going to increase, since, as a general rule, as age increases, so does the number of consultations. Moreover, in the wake of the COVID-19 pandemic, the centres have experienced an unforeseen increase in the burden of care, resulting in a perception that the quality of the healthcare system has deteriorated.

Despite the social changes, both in terms of place of residence and age of patients, the delimitations of health departments have remained the same over the years. A new allocation of the population to the departments using optimisation methods could be a solution. In this way, the burden of care could be balanced between the health centres already built, making more efficient use of current resources.

However, before carrying out a new allocation, it is interesting to predict how the population will evolve in these places, since, as it seems logical, the creation of new boundaries for the health departments would imply an administrative change that would have to be maintained over time.

In this work we are going to use a penalised functional spatial regression model, PFSRM, [2] to predict the population that will live in each of the localities of the Province of Valencia next year based on the observed population from 1998 to 2022, which can be extrapolated to any territory. The application of this methodology is relevant because it takes into account both the number of inhabitants over the years and the latitude and longitude of each of the localities, that is a prediction is obtained considering space-time.

Keywords: Functional Data Analysis; Spatial dependence; P-spline penalty; Prediction; Population.

Acknowledgements Work supported by Doc-Course on Data Science, DCDS23, organised by the Institute of Mathematics of the University of Granada (IMAG) and the Institute of Mathematics of the University of Seville (IMUS).

References

[1] Generalitat Valenciana. Proyecciones de población - Portal estadístico de la Generalitat Valenciana. *Portal Estadístico de la Generalitat Valenciana, https://pegv.gva.es/es/temas/demografiaypoblacion/poblacion/proyeccionesdepoblacion*

[2] Aguilera-Morillo, M.C., Durbán, M. and Aguilera, A.M. (2017) .Prediction of functional data with spatial dependence: a penalized approach. *Stoch Environ Res Risk Assess* **31**, 7–22

[*] Polytechnic University of Valencia, Spain. Email: dapeas@upv.edu.es

[†] Department of Applied Statistics and Operations Research and Quality, Polytechnic University of Valencia, Spain. Email: mdagumor@eio.upv.es

[‡] Department of Statistics and Operations Research, University of Granada, Spain. Email: aaguiler@ugr.es

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

Methods for integrating probability and non-probability survey data

Jorge L. Rueda Sánchez [*] Beatriz Cobo Rodríguez [†]

J.L. Rueda Sánchez (PhD Student) Jorge Luis Rueda Sánchez is a PhD student in the Mathematical and Applied Statistics program at the University of Granada. His main interest is Sampling Theory. He obtained the degree in Statistics from the University of Granada in 2022 and the official Master in Applied Statistics from the University of Granada in 2023. During these years he has been awarded with different prizes such as the Extraordinary Prize of the Degree in Statistics (academic year 2021/2022) by the University of Granada or the Recognition of Excellence in Academic Performance (2022 edition) by the University of Granada. He currently participates in the research group "FQM-365: Diseño y análisis estadístico de encuestas por muestreo (DAE)", in the research project of the national plan "Técnicas modernas para reducción de sesgos en las estimaciones. Aplicación al estudio de adicciones" and has a PhD grant, with code PRE2022-103200, from the F.P.I. grants awarded by the Spanish National Research Plan for the completion of the doctorate.

In the last decade, national statistical agencies have started to consider alternative methods to conduct their surveys [1], due to: drastic decrease in response rates of probability surveys, which causes an increase in costs [3]; the need to obtain updated data quickly after the COVID-19 crisis; and the emergence of new data sources from new technologies such as the Internet or large volumes of data (also called Big Data). These alternative methods using such innovative data sources are the so-called non-probability surveys, since the probability of inclusion of each individual is not known with certainty or is zero, i.e. the selection mechanism is not random.

However, this type of survey has a great defect due to its non-probability nature, and this is the great intrinsic bias of its estimates [2]. Therefore, probability surveys are still the most accurate and reliable alternative, but in some cases it can be very useful to integrate both types of data. In other words, using a sample with a non-probability design would complement a classic probability survey, especially when such a survey cannot achieve a recommendable sample size or when a part of the population is underrepresented.

In this work we will develop a new methodology to combine probability and non-probability survey samples, using Machine Learning techniques, in order to achieve more accurate estimates of the population of interest. To test the performance of this method we will perform a simulation study with many other techniques already widely studied, seeking to keep the bias and the RMSE as small as possible. We will check how this new methodology produces the best results and we will confirm with a real application that using non-probability data in a complementary way improves the estimates of the classical probability survey.

Keywords: Sampling, non-probability survey, online surveys, machine learning, Kernel Weighting Method.

Acknowledgements Work supported by project PID2019-106861RB-I00 and PhD grant PRE2022-103200 from the F.P.I. grants awarded by the Spanish National Research Plan.

References

[1] Beaumont, J. F., & Rao, J. N. K. (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? Surv. *Stat, 83*, 11-22.

[2] Elliott, M. and Haviland, A. (2007). Use of a Web-Based Convenience Sample to Supplement a Probability Sample. *Survey Methodology, 33.* 211-215.

[3] Kennedy, C., Hartig, H. (2019). Response rates in telephone surveys have resumed their decline. Pew Research Center. https://www.pewresearch.org/fact-tank/2019/02/27/response-rates-in-telephone-surveys-have-resumed-their-decline/

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

[†] Department of Quantitative Methods for Business and Economics, University of Granada, Spain. Email: beacr@ugr.es

Functional principal components analysis of the S&P500 and IBEX35: a comparison during the COVID-19 period

Pierdomenico Duttilo [*] Ana M. Aguilera [†] Christian Acal [‡]

P. Duttilo (PhD student) Pierdomenico Duttilo completed his bachelor's and master's degree in Economics at the University "G. d'Annunzio" of Chieti-Pescara. He is currently a third year PhD student in Human Science with curriculum in Economics and Statistics at the University "G. d'Annunzio" of Chieti-Pescara, under the academic mentor-ship of Prof. Stefano Antonio Gattone. His research is focused on financial statistics and mixture models. He is particularly interested in mixture models able to capture the heavy-tailed behaviour of financial assets returns.

This work aims to analyse the monthly returns of the S&P500 and IBEX35 during the COVID-19 period via functional data analysis (FDA) methods. Specifically, the most significant components of the two indices are explored by means of functional principal component analysis (FPCA).

Firstly, the daily closing prices of the S&P500 and IBEX35 stocks have been collected from yahoo.com. The S&P500 is one of the most important US stock index which consists of 503 stocks, while the IBEX35 consists of 35 stocks. The time period taken for the study is from January 2020 to December 2020. Monthly returns were calculated with the natural log difference approach.

Secondly, the raw data are converted into functional data by using the B-splines basis functions [1, 2]. The function $x(t)$ is represented as a linear expansion of K B-splines basis functions

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t), \tag{1}$$

where c_k are the basis coefficients and ϕ_k are the basis of functions. As a result, each curve represents a stock. Next, the relevant statistical features and the FPCA are carried out [3].

Results show that the first principal component accounts most of the variability and represents the COVID-19 outbreak. On the contrary, the second and third component represent the two recovery peak after the outbreak of the pandemic. Moreover, the score plots highlight important clusters of economic sectors. For example, the health care sector show a positive performance during the COVID-19 outbreak.

Future developments of this work consist of studying the theoretical aspects of the Karhunen-Loève decomposition when the scores are fitted with mixtures of generalized normal distribution [4] in order to characterize the distribution of the stochastic process $X(t)$ that generates the curves

$$X(t) = \widehat{\mu}(t) + \sum_{j=1}^{J} z_j w_j(t), \tag{2}$$

where $\widehat{\mu}(t)$ is the sample mean function, z_j are the scores and $w_j(t)$ are the weight functions.

Keywords: Functional Data Analysis; Functional Principal Component Analysis; Financial Data; Karhunen-Loève Decomposition; Mixtures of Generalized Normal Distribution.

References

[1] Ramsay, J.O. and Silverman B.W. (2005). *Functional Data Analysis.* Springer, New York 2005.

[2] Wang, Z., Sun, Y., and Li, P. (2014). Functional Principal Components Analysis of Shanghai Stock Exchange 50 Index. *Discrete Dynamics in Nature and Society*, **2014**, 1–7.

[3] Greenacre, M., Groenen, P.J.F., Hastie, T. et al. (2022). Principal component analysis. *Nat Rev Methods Primers*, **2**, 100, 1–24.

[4] Wen, L., Qiu, Y., Wang, M., Yin, J., P., Chen (2022). Numerical characteristics and parameter estimation of finite mixed generalized normal distribution. *Communications in Statistics - Simulation and Computation* **51**, 3596–3620.

[*] University "G. d'Annunzio" of Chieti-Pescara, Italy. Email: pierdomenico.duttilo@unich.it

[†] Department of Statistics and Operation Research and IMAG, University of Granada, Spain. Email: aaguiler@ugr.es

[‡] Department of Statistics and Operation Research and IMAG, University of Granada, Spain. Email: chracal@ugr.es

Analyzing approaches to community detection in networks

Sebastián V. Taboh [*] Miguel A. Pozo [†]

S.V. Taboh Sebastián Víctor Taboh has been pursuing his Ph.D. in computer science since 2020 at the University of Buenos Aires under the supervision of Paula Zabala and Isabel Méndez-Díaz. His main research interests encompass mathematical programming, algorithms, graph theory and computational complexity theory. He earned his "Licenciatura" in computer science, a comprehensive 6-year degree combining a B.S. and an M. Sc., from the University of Buenos Aires, Faculty of Exact and Natural Sciences in 2019. Additionally, he obtained his bachelor's degree in data science from the same university in 2021.

Graphs are abstract structures that have been studied for centuries, proving highly useful for modeling the relationships between different entities. They serve as a fundamental tool in various fields, allowing us to represent and analyze connections between elements in a structured way. One of the many contexts in which they naturally emerged was the study of relationships between people. For instance, each person could be represented by a vertex, and edges would connect vertices if the corresponding individuals had a certain relationship, such as being friends in a social network. As classification has always been a human need, the problem of grouping entities based on different criteria also arose, giving rise to the problem of community detection. Given a graph, this problem involves assigning each vertex to a set of communities, with each community being a non-empty set of vertices. The objective is to achieve a high relationship index between vertices within the same community and a low relationship index between vertices that do not belong to a common community. This broad formulation of the problem gives rise to various specific variants depending on different components arising from the real-world situation being studied. On the other hand, the relationship metric, known as the "modularity function", can be defined in many different ways, and its utility may also depend on the real-world data on which it is applied.

While the topic of community detection was already under study in the second half of the 20th century [1], it garnered increased attention from the scientific community at the beginning of this century, with numerous papers published following the seminal works of Girvan and Newman [2, 3]. In this work, we conduct a comprehensive review of the literature, identifying the most studied variants and the methods developed to address them. A particularly noteworthy observation is the limited number of works that delve into these issues using mathematical programming, despite its proven effectiveness in solving combinatorial optimization problems. Subsequently, we delve into a detailed exploration of a recent work employing mathematical programming [4], focusing on potential methodological enhancements for both problem-solving and the analysis of proposed methods' effectiveness. Finally, we present our conclusions and outline future research directions that we will pursue shortly.

Keywords: Complex Networks; Community Detection; Modularity Functions; Mathematical Programming.

Acknowledgements Work supported by Institute of Mathematics of the University of Granada (IMAG), Institute of Mathematics of the University of Seville (IMUS) and by the PhD grant PICT-2017-1826.

References

[1] Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, **33(4)**. doi: 10.1086/jar.33.4.3629752

[2] Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.*, **99(12)**, 7821–7826. doi: 10.1073/pnas.122653799

[3] Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, **69(2)**, 026113. doi: 10.1103/PhysRevE.69.026113

[4] Benati, S., Puerto, J., Rodríguez-Chía, A. M., & Temprano, F. (2022). A mathematical programming approach to overlapping community detection. *Physica A*, **602**, 127628. doi: 10.1016/j.physa.2022.127628

[*] Department of Computer Science, University of Buenos Aires, Argentina. Email: staboh@dc.uba.ar

[†] Department of Statistics and Operations Research, University of Seville, Spain. Email: miguelpozo@us.es

Minimal second order cone reformulation for SVM with ℓ_p-norm

Miguel Martínez Antón* Víctor Blanco†

Abstract In this work, we analyze minimal second order cone (SOC) reformulations of SVM with ℓ_p-norms. We provide a procedure to construct a minimal representation by means of second order cones in case p is rational. The construction is based on the identification of the cones with a graph, the mediated graph. Then, we develop a mixed integer linear optimization formulation to obtain the optimal mediated graph in terms of its number of nodes, and then, the minimal SOC reformulation.

The management of the conic structure of the feasible region of SVM, when you generalize the problem to an ℓ_p-norm, is crucial at the time of obtaining a good classifier by means of the SVM solution using an off-the-shell solver such as Gurobi, CPLEX, FICO, etc. The conic structure that arises in this problem is given by the following definition of p-order cone

$$\mathcal{K}_p^{d+1} = \{(\mathbf{x}, z) \in \mathbb{R}^d \times \mathbb{R}_+ : \|\mathbf{x}\|_p \leq z\}.$$

We recall SVM consists on finding $(\mathbf{w}, b)$ witch minimize $\|\mathbf{w}\|_q$ under the constraints

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 \geq 0, \forall i = 1, \dots, n$$

Then we can write it as a standard q-order cone program problem minimizing z under the constraints

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 \geq 0, \forall i = 1, \dots, n \tag{1}$$

$$(\mathbf{w}, z) \in \mathcal{K}_q^{d+1} \tag{2}$$

where q is the conjugate of p, i.e., $\frac{1}{p} + \frac{1}{q} = 1$.

The aim of this work is to give a reformulation of 2 by means of 3-dimensional rotated second order cones defined in the following way

$$\hat{\mathcal{K}}_2^3 = \{(x, y, z) \in \mathbb{R}^2 \times \mathbb{R}_+ : xy \leq z^2\}.$$

that is equivalent to the regular second order cone $\mathcal{K}_2^{2+1}$, such that be minimal in terms of the number of constraints and auxiliary variables.

The problem of finding the minimal second order cone reformulation is equivalent to the problem of finding the minimal cardinality of q-mediated graph. Let $q = \frac{r}{s}$ in its irreducible form, we define a q-mediated graph as a digraph $G = (\mathcal{A}_q \cup X, A)$ with $\mathcal{A}_q = \{0, r\}$ and $X \subset \mathbb{R}$ such that:

1. $s \in X$, and for each $x \in X$ there exist $y \neq z \in \mathcal{A}_q \cup X$ such that $x = \frac{1}{2}(y + z)$.
2. $A = \Big\{(x, y) : \exists z \in (\mathcal{A}_q \cup X) \backslash \{y\}$ such that $x = \frac{1}{2}(y + z)\Big\}$.

Let $k = \lceil log_2(r) \rceil$, the solution of this problem provides a list of auxiliary variables $\{t_1, \dots, t_{k-1}\}$ and a set of projections $\pi_j : \mathbb{R}^{2d+k} \to \mathbb{R}^3$ for $j = 1, \dots, k$; that get reformulate the SVM as minimize z under the constraints

$$\begin{aligned} &y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 \geq 0, \forall i = 1, \dots, n \\ &s_1 + \cdots + s_d \leq z \\ &\pi_j(\mathbf{w}, z, \mathbf{s}, \mathbf{t}) \in \hat{\mathcal{K}}_2^3, , \forall j = 1, \dots, k. \end{aligned}$$

Keywords: Second order cone; p-Order cone; Conic programming; Supervised Classification; Mediated graph.

Acknowledgements This research has been partially supported by Spanish Ministerio de Ciencia e Innovación, AEI/FEDER grant number PID 2020 - 114594GBC21, RED2022-134149-T (Thematic Network on Location Science and Related Problems), Junta de Andalucía projects B-FQM-322-UGR20 and AT 21_00032, and the IMAG-Maria de Maeztu grant CEX2020-001105-M /AEI /10.13039/501100011033.

References

[1] V. Blanco, J. Puerto, and S. ElHaj-BenAl (2014). *Revisiting several problems and algorithms in continuous location with τ-norms.* Computational Optimization and Applications, 58 (2014), pp. 563–595.

[2] V. Blanco, J. Puerto, and A. Rodriguez-Chia (2020). *On ℓ_p-Support Vector Machines and Multidimensional Kernels.* JMLR, 21(14):1-29, 2020.

*Institute of Mathematics of University of Granada, IMAG, University of Granada. Email: mmanton@ugr.es
†Dpto. Quantitative Methods for Economics & Bussines, University of Granada. Email: vblanco@ugr.es

Propensity Score Adjustment to reduce selection bias

Francesca Bellissimo [*] Maria del Mar Rueda [†] Beatriz Cobo Rodríguez [‡]

F. Bellissimo Francesca Bellissimo graduated in Mathematics in 2022 at the University of Calabria. She's currently a master student in Data Science for Business Administration at the Department of Economics, Statistics and Finance, University of Calabria.

The development of communication technologies and the need to obtain information quickly and with relatively low costs have led to the spread of non-probability surveys, potentially conducted online by volunteers. This can cause problem of undercoverage of the target population. Specifically, when participants are a subset of the population of interest that systematically differ from individuals not taking part in the survey, the results are affected by selection bias. The effects of it can be reduced using *Propensity Score Adjustment* (PSA).

In order to give each unit a certain weight and build a reweighted estimator, it is necessary to identify a measure of how much each unit is likely to participate in the survey, based on some covariates. For a generic unit i of the target population, propensity score is defined as

$$\pi_{vi} = Pr(R_i = 1|\boldsymbol{X}_i),$$

where R_i takes value 1 if the unit belongs to the non-probability sample and $\boldsymbol{X}_i$ are a set of covariates.

Propensity scores are usually estimated through logistic regression, which turns out to be a robust technique for the estimation [1], but it is possible to do it using Machine Learning algorithms. In this study, *Gradient Boosting Machine* and *Neural Networks* have been used.

Once the propensity scores have been estimated, it is possible to build a new weight system in several ways. In this study, Valliant weights, Schonlau and Couper weights, Lee and Valliant weights, Valliant and Dever weights have been used.

All the methods have been applied to a non-probability dataset, containing socio-demographic variables (province, gender, level of education, etc.) and some related to the perception of work performance during the pandemic.

Keywords: Propensity Score Adjustment; Non-probability survey; Gradient Boosting Machine; Neural Networks.

References

[1] Ferri-García, R., & Rueda, M. D. M. (2020). Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. *PloS One*, **15(4)**, e0231500.

[2] Castro-Martín, L., Rueda, M. D. M., & Ferri-García, R. (2020). Estimating general parameters from non-probability surveys using propensity score adjustment. *Mathematics*, **8(11)**, 2096.

[*] Department of Economics, Statistics and Finance, University of Calabria

[†] Department of Statistics and Operational Research, University of Granada

[‡] Department of Quantitative Methods for Economics and Business, University of Granada

Session 2 (Monday 16:00)

Contents

On the Hierarchical Districting Problem

Andreana Ferraioli [*] Federico Perea [†]

Andreana Ferraioli is a PhD student in the course of Technology, Innovation and Management at the Universita degli studi di Bergamo. Her main interest is Location Science with particular reference to Districting. She earned her bachelor's and master degree in Management Engineering from the Università degli studi di Napoli Federico II.

The Hierarchical Districting Problem (HDP) is the problem of partitioning a set of Territorial Units (TUs) in $|K|$ different district plans, with K representing a generic set of hierarchically organized services (or levels).

The meaning of the hierarchy depends on the specific application context. It can reflect the progressive complexity of the offered services, as, for example, in the case of health care networks (hierarchically organized in a Hub & Spoke fashion), or the level of expertise/responsibility in the provision of some services, e.g., the administrative organization of some territories (municipalities, provinces, regions, etc.).

The objective is the integrated definition of a districting plan for each level $k \in K$, where each districting plan represents the partitioning of the TUs into a predefined number of districts, say $p_k, \quad k \in K$, meeting some planning criteria. The latter usually include integrity, balancing, compactness, and contiguity. Integrity means that each TU belongs to only one district. Balancing expresses the need for districts of similar size w.r.t. some activity measures associated with the TUs. Contiguity implies that the devised districting plan does not include enclaves, which also ensures that there is always a path connecting two TUs belonging to the same district that does not cross any other district. Finally, compactness is a topological property requiring that districts do not have elongated shapes. We assume compactness to be represented by a surrogate measure expressed by the allocation costs of the TUs to a "special" TU of the district, referred to as the center or representative of the district. In practice, it is a p-median dispersion function.

Different variants of such problems have been formulated to deal with various types of hierarchies to account for two main practical attributes of the hierarchy, i.e., nestedness and coherency. Borrowing some ideas from the related literature on hierarchical facility location problems [1], in the "districting" terminology, we say that the hierarchy is nested if the location of a higher-level center can be chosen only among the lower-level ones. Besides, we say the hierarchy is coherent if the higher-level districts are built by aggregating the lower-level ones. In other words, two TUs belonging to the same lower-level district will belong to the same higher-level one. Clearly, both these characteristics may be simultaneously required.

A further extension is the case of the so-called "partial coherence", where the coherency between two consecutive districting plan is ensured to a minimum desirable degree (percentage), which is a user-defined parameter. Such problems may be applied to the spatial organization of service/distribution networks, where the presence of a hierarchical system of facilities/competence areas (that is, the districts) emerges as relevant.

Some (initial) possible ideas for devising heuristic approaches for the problem are discussed.

Keywords: Location Science; Districting; Hierarchical Facility Location; Heuristics.

Acknowledgements Work supported by project Doc-Course on Data Science 2023.

References

[1] Şahin, G., & Süral, H. (2007). A review of hierarchical facility location models. *Computers & Operations Research*, **34(8)**, pp. 2310-2331.

[2] Serra, D., & Revelle, C. (1992). The pq-median problem: location and districting of hierarchical facilities. Part ii: heuristic solution methods.

[*] Department of Management, Information and Production Engineering, Università degli studi di Bergamo, Italy. Email: andreana.ferraioli@unibg.it

[†] Department of Applied Mathematics II, Universidad de Sevilla, Spain. Email: perea@us.es

A co-evolutionary metaheuristic algorithm for a pricing hub location bilevel problem

Carlos Corpus [*] Víctor Blanco Izquierdo [†] José-Fernando Camacho-Vallejo [‡]

Carlos Corpus Carlos Corpus holds a Master's degree in Science and is a final-year Ph.D. student in Science with a specialization in Mathematics in the Centro de Investigación Ciencias Fisíco Matemáticas en la Universidad Autonóma de Nuevo León. His research area is focused on the development of metaheuristics and the classification of solutions for bilevel programming problems.

In recent years, metaheuristic algorithms have gained a prominent position in problem-solving across various fields. Their ability to discover high-quality solutions within a reasonable time frame has rendered them excellent tools for addressing the need for efficient and quality solutions.

One of the primary opportunities for the use of metaheuristic algorithms, which has seen significant impact in recent years, is in bilevel programming. These bilevel programming problems have the peculiarity that one of the constraints in the mathematical optimization problem is another mathematical optimization problem.[1].

$$\begin{aligned} \max_{x\in X, y\in Y} \quad & F(x,y) \\ \text{s.t} \quad & G(x,y) \leq 0 \\ & y \in \arg\min_{y\in Y}\{f(x,y)\,|\,g(x,y)\leq 0\} \end{aligned}$$

These problems, which involve the simultaneous optimization of two levels of decision-making, pose unique computational challenges due to the interdependence between the two levels. Therefore, developing effective solution schemes is of utmost importance.

This work discusses the implementation of a novel coevolutionary algorithm. This algorithm leverages the combination of multiple evolving populations, incorporating characteristics from one population into another. To illustrate its effectiveness, the study addresses the bilevel tree of hubs location problem with pricing. This problem is unique in that the decision-maker not only determines which hubs to open and connect in a tree-like structure but also sets the prices users must pay for using this network. The algorithm is presented in detail, and previous results are provided to demonstrate its effectiveness.

Keywords: Co-evolutionary algorithm, metaheuristics, bilevel optimization,

Acknowledgements Work supported by CONACyT PhD grant 958882 and Facultad de Ciencias Físico Matemáticas, Universidad Autonoma de Nuevo León.

References

[1] Camacho-Vallejo, José-Fernando and Corpus, Carlos and Villegas, Juan G (2023). Metaheuristics for bilevel optimization: A comprehensive review. *Computers & Operations Reseach* Elsevier, 2023

[2] Contreras, Ivan and Fernández, Elena and Marín, Alfredo (2010). The tree of hubs location problem *European Journal of Operational Research*, **2**, pp. 390–400.

[*] Centro de Investigación Ciencias Físico Matemáticas, Universidad Autonoma de Nuevo Léon, Nuevo León, México. Email: carlos.corpuscrd@uanl.edu.mx

[†] Departamento de Métodos Cuantitativos para la Economía y la Empresa, Universidad de Granada, España. Email: vblanco@ugr.es

[‡] Tecnologico de Monterrey, Escuela de Ingeniería y Ciencias, México. Email: fernando.camacho@tec.mx

The Storyboard Problem

Víctor Blanco [*] Gabriel González [†] Justo Puerto [‡]

Gabriel González (Phd student) In this work we analyze a network design problem that arises in naval designs that we call the Storyboard Problem. The problem consists of locating a given set of devices that allows to provide service from a given finite set of pipelines and cables to a set of final sources. The technical requirements for the connections and location of devices is given by the "storyboard", a forest-based graph that states how many devices are needed and which is the hierarchical structure of the final network, as well as some technical requirements for the links as capacities, separation between devices, etc. We provide different mathematical optimization based methodologies to solve the problem and heuristic approaches that allow to obtain good quality solutions in reasonable CPU time for real size instances

The work addresses the intricate problem of determining cable routes in the design of naval structures, a critical aspect of ensuring the functionality and safety of complex systems on ships. This challenge involves strategically placing cables and pipelines while considering factors such as obstacles, incompatible cable interactions, and safety distances. Related works in this topic are [1] and [2]. Designers are provided with a reference diagram known as the "storyboard", which outlines the topological structure that cables must follow to connect the main cable trunk to various elements in need of service. The storyboard is represented as a set of trees, with branching nodes on main pipelines and leaves representing final service points.

The cable routing problem involves determining the positions of branching nodes, intermediate nodes, and routes while adhering to the storyboard's structure. The main pipelines consist of segments defined by their endnodes, and the final service points and intermediate nodes are specified. Furthermore, intermediate nodes must be located within predefined regions in three-dimensional space. A safety distance constraint is imposed to ensure that nodes from different pipelines do not interfere with one another.

This work presents the Minimal Storyboard-restricted Cable Routing Problem (MSBP), aiming to find optimal solutions for positioning root nodes, intermediate nodes, and cable routes while minimizing the spatial footprint of the cables. Solving this problem is crucial for efficient naval structure design and operation.

In this work we propose a mathematical optimization based approach to solve the MSBP. For obtaining a suitable mathematical optimization formulation of the problem, as usual in other type of routing problems in continuous domains, we adequately discretize the space.

Keywords: Pipeline routing, Cable routing, Network design, Matheuristics, Naval engineering

References

[1] V. Blanco, G. González, Y. Hinojosa, D. Ponce, M. A. Pozo, J. Puerto (2022). *Network flow based approaches for the pipelines routing problem in naval design.*, Omega, 2022.

[2] V. Blanco, G. González, Y. Hinojosa, D. Ponce, M. A. Pozo, J. Puerto (2023). *The pipelines and cable trays location problem in naval design.*, Ocean Engineering, 2023.

[*] University of Granada. Email: vblanco@ugr.es

[†] University of Granada. Email: ggdominguez@ugr.es

[‡] University of Sevilla. Email: puerto@us.es

Analysis of Ambulance Demand in the Province of Valencia Using Spatial Point Processes

Esther Rodenas Moreno [*] Federico Perea Rojas-Marcos [†]

E. Rodenas Moreno (Master's Student) Esther Rodenas Moreno is a master's student at the Polytechnic University of Valencia. Her main interests are Data Analysis, Spatial Statistics, Econometrics and Operations Research. She obtained a Double Bachelor's Degree in Telecommunications Engineering and Business Administration and Management from the Polytechnic University of Valencia. After a one-year training programme abroad, she returned in 2022 to complete her studies with a Master's Degree in Data Analysis, Process Improvement and Decision Support Engineering.

Providing a swift response when attending emergency calls that require emergency medical vehicles (EMV) is crucial for the well-being of people requiring assistance. However, the EMV fleet management and dynamic deployment decision process is quite challenging. The project iReves[1] aims to develop smart tools that aid fleet management decision-makers in best-locating EMVs. In this case, it may be helpful to have a prediction of the EMV demand to improve the quality of the service by responding to more emergencies in less amount of time.

The available dataset contains more than 98.000 emergency calls where EMVs were deployed in the Province of Valencia for 2019. Furthermore, it contains the coordinates (longitude and latitude) of the spatial position where the emergency occurred, the main focus of this research project. The aim is to analyse how these calls are spatially arranged in the bounded spatial region W. This dataset type is considered a spatial point pattern [1] and it will be treated as such.

After an exploratory analysis is performed [1] using the statistical software R [2], it is found that the data is clustered and does not follow a complete spatial randomness (CSR) point process. Therefore, a series of models are proposed to study the dependency of the location and possible covariates, evaluating them with the pseudo R^2. After better understanding the dataset, the objective is to improve the obtained results by researching further possible processes and/or covariates.

Keywords: Spatial statistics; EMV Demand; Point Process Analysis; Clustering; Spatial Point Patterns; R Software.

Acknowledgements Work supervised by Federico Perea Rojas-Marcos during the Doc-Course On Data Science 2023 at the University of Sevilla and University of Granada. This is also part of a Master's Degree Thesis supervised by Mª Fulgencia Villa Juliá, Victor Sanchez Anguix and Yulia Karpova Krylova.

References

[1] Baddeley, A., Rubak, E. H., and Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R.* CRC Press.

[2] Roger S. Bivand, E. J. P. u. V. G.-R. (2008). *Applied Spatial Data Analysis with R.* Springer New York.

[*] Department of Applied Statistics and Operational Research, and Quality, Polytechnic University of Valencia, Valencia, Spain. Email: esromo@ade.upv.es

[†] Department of Applied Mathematics, University of Sevilla, Sevilla, Spain. Email: perea@us.es

[1] https://ireves.webs.upv.es/

Generative models of vascular structures: Deep Learning vs Constrained Optimization approach

Paula Feldman [*] Angel M. González-Rueda [†]

Paula Feldman Biomedical Engineer from Universidad Nacional de Tucumán. Currently PhD candidate at Universidad Torcuato Di Tella with a scholarship from the National Reaserch Council (CONICET).

Accurate 3D models of blood vessels are increasingly required for several purposes in Medicine and Science. These meshes are typically generated using either image segmentation or synthetic methods. Despite significant advances in vessel segmentation, reconstructing thin features accurately from medical images remains challenging, which explains the scarcity of curated datasets. As a result, several methods have been developed to adequately synthesize blood vessel geometry.

Within the existing literature on generating vascular 3D models, we identified two primary types of algorithms: fractal-based, and space-filling algorithms. Fractal-based algorithms use a set of fixed rules that include different branching parameters, such as the ratio of asymmetry in arterial bifurcations and the relationship between the diameter of the vessel and the flow. On the other hand, space-filling algorithms allow the blood vessels to grow into a specific perfusion volume while aligning with hemodynamic laws and constraints. A particular class of such space-filling algorithms is known as Constrained Constructive Optimisation (CCO). Although these *model-based* methods provide some degree of control and variation in the structures produced, they often fail to capture the diversity of real anatomical data.

In recent years, deep neural networks led to the development of powerful generative models, such as Generative Adversarial Networks, and Diffusion Models, which produced groundbreaking performance in many applications, ranging from image and video synthesis to molecular design. These advances have inspired the creation of novel network architectures to model 3D shapes using voxel representations, point clouds, signed distance functions, and polygonal meshes. In particular, More recently, a GAN model was proposed. It is capable of generating coronary artery anatomies. However, this model is limited to generating single-channel blood vessels and it does not support the generation of more complex, tree-like vessel topologies.

VesselVAE [2] proposes a novel *data-driven* framework based on a Recursive variational Neural Network (RvNN), that has been applied in various contexts, including natural language, shape semantics modeling, and document layout generation. In contrast to previous data-driven methods, the recursive network fully exploits the hierarchical organization of the vessel and learns a low-dimensional manifold encoding branch connectivity along with geometry features describing the target surface. Once trained, the VesselVAE latent space is sampled to generate new vessel geometries. In contrast, Georg et al [1] presents a CCO framework for the construction of vascular systems based on optimality principles of theoretical physiology. Specifically, given the position and flow distribution of endpoints of a vascular system, optimization is driven by intravascular volume minimization with constraints derived from physiological principles. The goal of this talk is to make a review of these two methodologies.

Keywords: Vascular 3D model; Generative modeling; Neural Networks; Constrained Constructive Optimization.

References

[1] Georg, Manfred; Preusser, Tobias; and Hahn, Horst K., "Global Constructive Optimization of Vascular Systems" Report Number: WUCSE-2010-11 (2010). All Computer Science and Engineering Research.

[2] Feldman, P., Fainstein, M., Siless, V., Delrieux, C., & Iarussi, E. (2023, October). VesselVAE: Recursive Variational Autoencoders for 3D Blood Vessel Synthesis. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 67-76). Cham: Springer Nature Switzerland.

[*] Universidad Torcuato Di Tella, Buenos Aires, Argentina. Email: paulafeldman@conicet.gov.ar

[†] Universidade de Santiago de Compostela, Santiago de Compostela, España.

New Advances in the Stackelberg Minimum Spanning Tree Problem

Miguel A. Pozo [*] Justo Puerto [†] Alberto Torrejón [‡]

Alberto Torrejón (PhD Student). Alberto Torrejón Valenzuela is a PhD student at the University of Seville in the Department of Statistics and Operations Research, and also a member in formation at Instituto de Matemáticas de la Universidad de Sevilla. His mains interests are discrete and combinatorial optimization problems. He received the double degree from University of Seville in 2020 in Mathematics and Statistics and the master's degree in Mathematics from the University of Seville in 2021.

Let G be a given a graph whose edge set is partitioned into a set of red edges and a set of blue edges, and assume that red edges are weighted and contain a spanning tree of G. Then, the Stackelberg Minimum Spanning Tree Game (StackMST) consists in pricing (i.e., weighting) the blue edges in such a way that the total weight of the blue edges selected in a Minimum Spanning Tree (MST) of the resulting graph is maximized. In the context of data science and decision theory, StackMST, together with other bi-level pricing problems, has great application in the analysis and modelling of competitiveness in spatial data.

The StackMST, see [1], can be seen as a bilevel optimization problem where the second level is a MSTP and where the objective functions are bilinear at both levels. The StackMST can be formally defined as follows. Let $G = (V, E)$ be a given graph whose edge set E is partitioned into a set B of blue edges (controlled by the leader) and a set R of red edges, and assume that red edges are weighted and contain at least one spanning tree of G, thus, $|R| \geq |V| - 1$. A positive cost c_e is associated to each red edge $e \in R$ and a positive price T_e ($T = [T_1, ..., T_{|E|}]$) has to be determined for each blue edge $e \in B$. We denote by $x = [x_1, ..., x_{|E|}]$ the design variables used to describe the STP polytope $\mathcal{T}$. Then, StackMST consists in determining T_e for each $e \in B$ in such a way that the total weight of the blue edges selected in a minimum spanning tree of the resulting graph is maximized.

$$\text{StackMST}: \max_{T\geq 0} \sum_{e\in B} T_e x_e$$

$$\text{s.t. } x = \text{argmin}_{x\in\mathcal{T}}\{\sum_{e\in B} T_e x_e + \sum_{e\in R} c_e x_e\}.$$

In this talk we present different enhancements to the state-of-the-art for the StackMST (see in [2]) based on the properties of the Minimum Spanning Tree Problem and the bilevel optimization. Also, a Benders decomposition for the StackMST and several enhancements for the heuristic approach to the problem are presented

Keywords: Minimum Spanning Tree, Bilevel optimization, Stackelberg game.

Acknowledgements This research is supported by the Plan Estatal 2017-2020 Generación Conocimiento - Proyectos I+D+i project *"Optimization on data science and network design problems: Large scale network models meet optimization and data science tools"* (PID2020-114594GB-C21) by Ministerio de Ciencia e Innovación. This support is gratefully acknowledged.

References

[1] J. Cardinal; E.D. Demaine; S. Fiorini; G. Joret; S. Langerman; I. Newman; O. Weimann (2011). The Stackelberg minimum spanning tree game. Algorithmica, 59(2), 129-144.

[2] Labbé, M.; Pozo, M.A.; Puerto, J. (2019) Computational comparisons of different formulations for the Stackelberg minimum spanning tree game. International Transactions in Operational Research, 28, 48–69.

[*] Department of Statistics and Operational Research. Faculty of Mathematics, University of Seville, Spain. Institute of Mathematics of the University of Seville, Spain. Email: miguelpozo@us.es

[†] Department of Statistics and Operational Research. Faculty of Mathematics, University of Seville, Spain. Institute of Mathematics of the University of Seville, Spain. Email: puerto@us.es

[‡] Department of Statistics and Operational Research. Faculty of Mathematics, University of Seville, Spain. Institute of Mathematics of the University of Seville, Spain. Email: atorrejon@us.es

Plenary session 2 - (Tuesday 10:00)

Contents

How to improve classic SVM? Recent insights

Luisa I. Martínez Merino [*]

Luisa I. Martínez-Merino (Assistant professor) Luisa I. Martínez-Merino works in the department of Statistics and Operations Research of Universidad de Cádiz as assistant professor. She earned her PhD in Mathematics from Universidad de Cádiz in 2018. The focus of her PhD were different location models under uncertainty and classification methods. For some years, she has worked as a postdoctoral fellow and lecturer until she obtained her current position. Her main interests are optimization models in the field of location, networks and supervised classification.

Keywords: Support Vector Machines; Ramp-loss; conditional constraints; feature selection.

The support vector machine (SVM) is a mathematical programming approach which has become widely used among supervised classification methods. Given a training sample with the feature values and class (-1 or 1) of its objects, the main goal of SVM is to provide a separator hyperplane that optimally classifies the sample in two classes. The classic SVM, introduced in [4], seeks for the maximization of the margin between the supporting hyperplanes associated with each class and the minimization of the deviations related to the misclassified objects.

Since the introduction of this classic model, many authors have extended it, trying to improve the classification performance of this approach. Particularly, in this talk, we will analyze the SVM with Ramp-Loss (RL-SVM) introduced in [3]. RL-SVM modifies the penalization that misclassified objects present in the objective function. One advantage of this model is that it can produce robust classifiers when applied to data with outliers.

Conditional constraints appearing in RL-SVM is the focus of research in some recent papers. Particularly, in [1], several techniques are developed in order to improve the time performance of the formulation. These techniques are based on three strategies for obtaining tightened values of the big M parameters which are present in the model. Two of them require solving a sequence of continuous problems, while the third uses the Lagrangian relaxation to tighten the bounds.

SVM with Ramp-Loss and feature selection (RL-FS-SVM) is another extension of classic SVM recently introduced in [2]. RL-FS-SVM is a robust model since it deals with outliers detection and feature selection at the same time. For this model, some techniques are proposed to tighten some M parameters appearing in this model. In addition, a heuristic approach is developed with the goal of obtaining an appropriate feasible solution for this model in reduced times.

References

[1] Baldomero-Naranjo, M., Martínez-Merino, L.I., 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, 84–100.

[2] Baldomero-Naranjo, M., Martínez-Merino, L.I., 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.

[3] Brooks, J. P. (2011). Support vector machines with the ramp loss and the hard margin loss. *Operations Research*, 59 (2), 467–479.

[4] Cortes, C. , Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20 (3), 273–297.

[*] Department of Statistics and Operations Research, Universidad de Cádiz, Cádiz, Spain. Email: luisa.martinez@uca.es

Session 3 (Tuesday 11:30)

Contents

A brief application of Multi Omics Factor Analysis for Parkinson patient data

Pablo P. Jurado-Bascón * Ramón Ferri-García †

Pablo P. Jurado (Phd student) Pablo Jurado is a Phd student at the University of Granada. His main interest is bioinformatics, specifically omics sciences. He received his B.S.C. from University of Granada in 2022 in statistics. He earned his M.S.C. in applied statistic also from the university of Granada in 2023, in which he participated in the bioinformatics unit at GENYO Center for Genomics and Oncological Research. He is now enrolled in the biostatistics doctoral program in statistics under the supervision of Pedro Carmona Sáez. He is a member of the Statistical Bioinformatics & Systems Biomedicine research group.

In this presentation we will be covering an interesting approach to the PCA applied to omics sciences. Omics sciences are a collection of branches of biology, such as genomics, proteomics, metabolomics, phenomics, transcriptomics, etc. These branches aim for the collective characterization and quantification of pools of biological molecules that translate into the structure, function and dynamics of an organism.

In particular, the analysed data comes from the *Accelerating Medicines Partnership* [1] program with data for Parkinson´s Disease. The data consist on Proteomics and Transcriptomics of different cohorts of patients, from which we will focus only in two batches with a total size of 319 patients, 2826 RNA-Transcripts selected for the transcriptomic data after QC filters and 1472 proteins, classified by cardiometabolic, neurology, oncology and inflammation. The data was collected from samples of plasma and cerebrospinal fluid from Parkinson´s Disease patients and control individuals.

The data was analysed from the multi-omics point of view, which means the use of tools that are capable of include data from multiple omics; in this case we used Multi Omics Factor Analisys (MOFA) [2] via their R package *MOFA2* available in the bioconductor repository. The master equation of this tool is:

$$Y^m = Z(W^m)^T + \epsilon^m; \quad \forall m = 1, \dots, M$$

where Y^m denotes the m-th matrix out of M data matrices of different omics that share most of the sample subjects, W^m denotes weights matrices for Y^m and the final factors (components), Z denotes the final representation of the original data in global terms and ϵ^m denotes the m-th error term matrix associated to the m-th data matrix.

The multiple input matrices is not the only difference with the common PCA or FA, as it also uses a Bayesian framework for "disentangling" the sources of variance. In other words, it has two regularization layers for applying sparsity to the model parameters via specific prior distribution to said parameters, possibly rendering a more relevant and interpretable model.

MOFA is a promising tool to extract the common direction in which biological patterns function, letting us check which genes, proteins, base-pair, transcription factors and so on act in concert in fundamental physiological processes.

Keywords: Omics; Multi Omics; Parkinson disease; PCA; Transcriptomics; Proteomics

Acknowledgements This project has received funding from grant PID2020-119032RB-I00 supported by MCIN/AEI/10.13039/501100011033

References

[1] Hirotaka Iwaki, Hampton L Leonard, Mary B Makarious, Matt Bookman, Barry Landin, David Vismer, Bradford Casey, J Raphael Gibbs, Dena G Hernandez, Cornelis Blauwendraat, et al. Accelerating medicines partnership: Parkinson's disease. genetic resource. *Movement Disorders*, 36(8):1795–1804, 2021

[2] R. Argelaguet, D. Arnol, D. Bredikhin, Y. Deloro, B. Velten, J.C. Marioni, and O. Stegle. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21(1):111, 2020.

[*] Department of Statistics and Operations Research, University of Granada. Center for Genomics and Oncological Research. Email: pjurado@ugr.es

[†] Department of Statistics and Operations Research, University of Granada Email: rferri@ugr.es

Gait pattern identification via functional analysis of variance

Helena Ortiz Alcalá [*] Christian J. Acal González [†] Ana M. Aguilera del Pino [‡]

H. Ortiz (doctoral student) Helena Ortiz Alcalá is a grant holder in her first year of PhD at the Department of Statistics and O.R. from the University of Granada (UGR). She received her B.S.C. in Statistics and her M.S.C. in Applied Statistics at UGR. Currently, she teaches subjects in the Statistics Degree at UGR, specifically, Bayesian Statistics. Her areas of interest cover Stochastic Modeling and Forecasting, and data driven statistics. In particular, her Doctoral Thesis aims to introduce the Bayesian Inference to the field of Functional Data Analysis.

Biomechanics data are usually curves that represent the human movement when subjects are submitted to multiple conditions. Gait analysis has been studied traditionally via discrete measures, which causes a great loss of information. However, functional data analysis uses whole curves and trajectory information, revealing the true nature of movement. This fact allows to model and forecast this kind of data in a more complete way. The main goal of this document is detecting the possible differences in gait patterns in kids between the age of 8 and 11 years old while moving to school using different types of book-bags and/or weight. For that purpose, we use the data available from the experimental study of the biomechanics laboratories of the Sport and Health Institute of the University of Granada (IMUDS). In particular, the angle of each joint (ankle, foot progress, hip, knee, pelvis and thorax) was measured for each axis (X,Y and Z, except for foot progress which only registered in one axis) and experimental condition (walking, backpack and trolley) with 10,15 and 20% of each subject's weight. Therefore we obtained 53 discrete curves for each joint, axis and condition. Taking into account the functional nature of the data, we assume an expansion of each curve in terms of a functional basis and then fit the smoothed basic coefficients in order to obtain the functional form of the data. By means of Functional Analysis of Variance (FANOVA) we want to study if there are significance differences on the mean curves according to the type of satchel and the weight contained. From a theoretical viewpoint, it has been proved that two-way FANOVA for repeated measures (FANOVA-RM) is equivalent to a two-way multivariate ANOVA-RM (MANOVA-RM) for the multivariate response of functional variable, in this case, the angle of the joints. This MANOVA has been solved through mixed multivariate model (MMM) over the basic coefficients and the first q principal components, which account for at least 95 % of the variability. As the MMM assumes multivariate normality and sphericity, we use a permutation test when any of theses assumptions is not fulfilled. All this methodology has been computationally implemented on R, a free software environment for statistical computing and graphics. In particular, we made use of fda package (reconstruction of data and Functional Principal Component Analysis) and several functions implemented by the supervisors of this manuscript for in the two-way MANOVA-RM through MMM and permutation test. Finally, FANOVA-RM methodology was applied to contrast the differences in rotation angle over each joint, axis and experimental condition (weight) differentiating by sex.

Keywords: Functional data analysis; Multivariate analysis of variance; Principal component analysis; Repeated measures; Biomechanics..

References

[1] Ramsay J.O. and Silverman B.W. (2005). *Functional Data Analysis.* Springer Science and Business Media.

[2] Zhang J.T.(2014). *Analysis of Variance for Functional Data.* CRC Press.

[3] Acal C. and Aguilera A.M. (2022). Basis expansion approaches for functional analysis of variance with repeated measures. *Advances in Data Analysis and Classification*, **17**, pp. 291–321.

[4] Aguilera A.M.,Acal C., Aguilera-Morillo M.C., Jiménez-Molinos F. and Roldán J.(2021). Homogeneity problem for basis expansion of functional data with applications to resistive memories. *Mathematics and Computers in Simulation*, **186**, pp. 41–51.

[*] Department of Statistics and O.R. and IMAG, Uniersity of Granada, Spain. Email: hortiz@ugr.es

[†] Department of Statistics and O.R. and IMAG, University of Granada, Spain. Email: chracal@ugr.es

[‡] Department of Statistics and O.R., University of Granada, Spain. Email: aaguiler@ugr.es

Integrating Stringing and Functional Logistic Regression for Gene Expression Data Analysis

Laura Antequera Pérez [*] Ana M. Aguilera [†] Manuel Escabias Machuca [‡]

L. A. P. Laura Antequera Pérez holds a Bachelor's degree in Mathematics from the University of Granada. Her passion for data analysis led her to pursue a Master's degree in Data Science and Computer Engineering at the same institution. With a strong academic foundation, she embarked on a journey into the world of data-driven solutions.

Currently, she is a dedicated professional working as part of a project-based team at the University of Granada. Her research and analysis focus on Bioinformatics, where she specialize in examining gene expression and microbiota data. Her primary objective is to identify and select significant variables that play a crucial role in predicting diseases based on patient samples.

Gene expression data is characterized by a multitude of variables, specifically genes, while the number of available samples remains consistently limited. This inherent limitation presents a significant challenge in the identification of the most pertinent and predictive genes for determining a patient's cancer status. As a result, this study primarily aims to not only enhance prediction accuracy but also to derive meaningful interpretations of the identified significant genes. To achieve this, we employ a recent technique known as "stringing" [3] which facilitates the reorganization of genes based on their similarity, thereby constructing a functional dataset that represents gene expression profiles as curves.

$$\{Z_i\left(s\right) : s \in S, i = 1, \ldots, n\}$$

$$z_i\left(s\right) = \sum_{j=1}^{p} a_{ij}\phi_j\left(s\right), \qquad i = 1, \ldots, n,$$

where the< $\phi_1, \ldots, \phi_p$ > are B-spline functions and the basis coefficients are estimated by least squares.

Subsequently, we utilize a functional logistic regression model to evaluate performance and interpret the parameter function. In parallel, a traditional methodology is also employed to facilitate a comprehensive comparative analysis. This research seeks to leverage both approaches for a comprehensive analysis of gene expression data, aiming for improved predictive accuracy and meaningful gene interpretation.

Keywords: Functional data analysis; Gene expression; High-dimensional predictors; Feature Selection; Prediction.

References

[1] Grossman, Robert L and Heath, Allison P and Ferretti, Vincent and Varmus, Harold E and Lowy, Douglas R and Kibbe, Warren A and Staudt, Louis M (2016). *Toward a shared vision for cancer genomic data.* N. Engl. J. Med.

[2] TCGA Research Network. `https://www.cancer.gov/tcga`.

[3] Kun Chen, Kehui Chen, Hans-Georg Muller and Jane-Ling Wang (2010). *Stringing High Dimensional Data for Functional Analysis*

[*] University of Granada. Email: lantequera@ugr.es

[†] Department of Statistics and Operation Research and IMAG, University of Granada, Spain. Email: aaguiler@ugr.es

[‡] Department of Statistics and Operation Research and IMAG, University of Granada, Spain. Email: escabias@ugr.es

A Mathematical Programming Approach to Sparse Canonical Correlation Analysis

Carlos Valverde [*] Justo Puerto [†]

C. Valverde I am a PhD student at the University of Seville. My main interests are location and routing problems using conic programming. I received my degree from University of Seville in 2016 in Mathematics. I earned my master's degree in mathematics from the University of Seville in 2017.

Recent developments in the interplay between Operational Research and Statistics allowed us to exploit advances in Mixed-Integer Optimisation (MIO) solvers to improve the quality of statistical analysis. In this work, we tackle Canonical Correlation Analysis (CCA), a dimensionality reduction method that jointly summarises multiple data sources while retaining their dependency structure. We propose a new technique for encoding sparsity in CCA by means of a mathematical programming formulation that allows one to obtain an exact solution using readily available solvers (such as Gurobi) or design solution algorithmic procedures based on it. Finally, we evaluate the performance of alternative solution strategies presented on multiple datasets from the literature. The results of the extensive comparison study highlight that the proposed approach is capable of finding the optimal correlation or finding good quality solutions, better than those provided by other conventional methods.

Keywords: Data science; Canonical Correlation Analysis; Mixed-Integer Optimisation; Sparsity

Acknowledgements This research has been partially supported by the Agencia Estatal de Investigación (AEI) and the European Regional Development Fund (ERDF): PID2020-114594GB-C21; Andalusia Regional Government: projects P18-FR-1422; Fundación BBVA: NetmeetData project (Ayudas Fundación BBVA a equipos de investigación científica 2019) and Sapienza University of Rome grant number RM11916B7F962975, RM1221816BB0FEE3 and Ministerio de Universidades grant number FPU2018-04591.

[*] Department of Statistical Sciences and Operational Research, University of Seville, Spain. Email: cvalverdeus.es

[†] Department of Statistical Sciences and Operational Research, University of Seville, Spain. Email: puertous.es

The Ordered Travelling Salesman Problem: A Comparison Of Formulations

Francisco Temprano [*] Justo Puerto [†]

F. Temprano F. Temprano is a PhD student at the University of Seville. His main interests are Combinatorial and Network design problems using mathematical programming. He received the degree from University of Seville in 2020 in Mathematics and the master's degree in mathematics from the University of Seville in 2021.

The OTSP is to find a Hamiltonian cycle in a network with minimum weighted cost. The costs of the edges of the cycle have to be sorted in decreasing order before multiplying them times the given weights. When all weights are equal and positive, the TSP is obtained as a particular case. When the unique non-null weight (positive) is the first one, the bottleneck TSP is obtained. Positive weights in the first positions and negative weights in the last positions will provide us with balanced cycles. Many other particular cases of interest exist, and the aim of the problem is to study all of them as a whole. In this preliminary work we study and compare different integer programming formulations of reduced size for the OTSP, paying special attention to the sorting part of the models.

Keywords: Routing; Combinatorial; Location Theory; Ordered Median Problems.

Acknowledgements This research has been partially supported by the Agencia Estatal de Investigación (AEI) and the European Regional Development Fund (ERDF): PID2020-114594GB-C21; Andalusia Regional Government: projects P18-FR-1422; Fundación BBVA: NetmeetData project (Ayudas Fundación BBVA a equipos de investigación científica 2019) and Sapienza University of Rome grant number RM11916B7F962975, RM1221816BB0FEE3 and Ministerio de Universidades grant number FPU2018-04591.

[*] Department of Statistical Sciences and Operational Research, University of Seville, Spain. Email: ftgarcia@us.es

[†] Department of Statistical Sciences and Operational Research, University of Seville, Spain. Email: puerto@us.es

Functional data analysis for monitoring and fault diagnosis of batch processes

Sergio García-Carrión * Ana M. Aguilera † M. Carmen Aguilera-Morillo ‡ Alberto Ferrer §

Sergio García-Carrión is a PhD candidate in the Statistics and Optimization Doctoral Program at Universidad Politecnica de Valencia, in the Multivariate Statistical Engineering Researh Group (MSEG). His main interest focuses on industrial process optimization, specially through (retrospective) Design of Experiments and latent variables-based multivariate statistical techniques. He holds a bachelor's and master's degree in Chemical Engineering, and a Black Belt in Six Sigma from the Universidad Politecnica de Valencia.

Batch processes operate for a finite period, from the starting point to the ending point, exhibiting a time-varying behavior. Data collected from batch processes have a unique structure. They are arranged in a three-way (3D) array (batches x variables x time) containing the time-varying trajectory measurements for all process variables throughout the entire batch duration. Consequently, batch data analysis differs significantly from the analysis of traditional continuous processes, given the distinct nature of the data.

There are several different approaches for analyzing batch data. One widely studied and adopted approach [1] consists of using Principal Component Analysis (PCA) [2] models. They are bilinear models, so the three-way array must be unfolded in two dimensions. After the unfolding, a regular PCA is performed in the usual way. Two complementary multivariate control charts can be constructed from the Hotelling's T^2 and the SPE statistics, which allow the detection of batches exhibiting potentially abnormal behavior. Additionally, by using contribution plots, the contribution of the original process variables to the potential out-of-control batch can be obtained.

On the other hand, one approach of growing interest in batch data analysis and monitoring involves Functional Data Analysis (FDA) [3], where Multivariate Functional Principal Component Analysis (MFPCA) [4] could play a crucial role. This arises from the fact that the time-varying trajectories measured for the process variables in the different batches can be considered functional data. Here, multivariate control charts based on the Hotelling's T^2 and the SPE statistics, and contribution plots can also be developed and exploited for monitoring and fault diagnosis purposes.

A comparative study was carried out to assess the performance of both approaches. The database employed comes from a Sequencing Batch Reactor (SBR) used for wastewater treatment. It is composed of 72 batches and 5 process variables measured along 340 time instances.

Keywords: Multivariate Functional Principal Component Analysis; Principal Component Analysis; Batch Process Monitoring; Multivariate Control Charts; Fault Diagnosis.

Acknowledgements Sergio García-Carrión acknowledges the funding from the *Programa de Ayudas de Investigación y Desarrollo (PAID-01-20) de la Universitat Politècnica de València.*

References

[1] Nomikos, P.; MacGregor, J.(1995). Multivariate SPC Charts for Monitoring Batch Processes. *Technometrics*, **37:1**, pp. 41–59.

[2] Wold, S.; Esbensen, K.; Geladi, P. (1987). Principal component analysis. *Chemom. Intell. Lab. Syst.*, **2:1-3**, pp. 37–52.

[3] Ramsay, J.O.; Silverman, B.W. (2000). *Functional Data Analysis.* Springer-Verlag, New York 2005.

[4] Happ, C.; Greven, S. (2018). Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains. *J. Am. Stat. Assoc.*, **113:522**, pp. 649–659.

[*] Department of Applied Statistics and Operational Research, and Quality, Universidad Politécnica de Valencia, Valencia, Spain. Email: sergarc6@eio.upv.es

[†] Department of Statistics and Operations Research, Universidad de Granada, Granada, Spain. Email: aaguiler@ugr.es

[‡] Department of Applied Statistics and Operational Research, and Quality, Universidad Politécnica de Valencia, Valencia, Spain. Email: mdagumor@eio.upv.es

[§] Department of Applied Statistics and Operational Research, and Quality, Universidad Politécnica de Valencia, Valencia, Spain. Email: aferrer@eio.upv.es