Workshop on Nonparametric Statistics

Conference proceedings

Mieres, 16–18 July 2025

Universidad de Oviedo

Committees

Scientific Committee

- Jose Ameijeiras Alonso (Universidade de Santiago de Compostela)
- Arís Fanjul Hevia (Universidad de Oviedo)
- Eduardo García Portugués (Universidad Carlos III de Madrid)
- Andrea Meilán Vila (Universidad Carlos III de Madrid)

Organizing Committee

- Jose Ameijeiras Alonso (Universidade de Santiago de Compostela)
- Arís Fanjul Hevia (Universidad de Oviedo)
- Laura Freijeiro González (Universidad de Oviedo)
- Andrea Meilán Vila (Universidad Carlos III de Madrid)

Editors: Scientific Committee

Conference website: https://sites.google.com/view/wnpseio/

Contents

Schedule

All sessions are held in the "Salón de Grados" of the Escuela Politécnica de Mieres at Universidad de Oviedo.

Tuesday, 15 July

- 16:00 Check-in
- 21:00 Dinner

Wednesday, 16 July

- 08:00–09:00 Breakfast and registration
- 09:00–09:20 Welcome reception
- 09:20–11:20 **Short Course (Part I)** *Chair: Jose Ameijeiras Alonso*
  Measurement errors: from basic concepts to recent advances (**Ingrid Van Keilegom**)
- 11:20–12:00 Coffee Break
- 12:00–12:40 **Invited Talk I** *Chair: Arís Fanjul Hevia*
  Tests for almost stochastic dominance and goodness-of-fit (**Javier Cárcamo Urtiaga**)
- 12:40–13:40 **Contributed Session I** *Chair: Arís Fanjul Hevia*
  - Uncertainty estimation for random forests with metric data (**Diego Serrano Ortega**)
  - A new proposal of specification tests for the asynchronous functional concurrent model (**Laura Freijeiro González**)
  - On boxplots for functional data (**Stanislav Nagy**)
- 13:40–14:40 Lunch
- 14:40–15:20 **Invited Talk II** *Chair: Laura Freijeiro González*
  Sequential data depth computation with statistical guarantees (**Alicia Nieto Reyes**)
- 15:20–16:40 **Contributed Session II** *Chair: Laura Freijeiro González*
  - Optimal filament estimation: a geometrical approach (**Héctor González-Vázquez**)
  - Kernel-based density estimation in tests of symmetry (**Raúl Pérez-Fernández**)
  - Projection depth and L^r-type depths for fuzzy sets (**Luis González de la Fuente**)
  - Cluster analysis via concave density regions (**José E. Chacón**)
- 16:40–17:20 Coffee Break

- 17:20–18:00 **Invited Talk III** *Chair: Eduardo García Portugués*
  Comparing many populations using Gini covariance (**María Dolores Jiménez Gamero**)
- 18:00–19:00 **Contributed Session III** *Chair: Eduardo García Portugués*
  - Distributional convergence of empirical entropic optimal transport and applications (**Santiago Arenas Velilla**)
  - Quick quantile segmentation (**Luis Alberto Rodríguez Ramírez**)
  - Generalized distorted expectiles (**Ignacio Cascos**)
- 21:00 Dinner

Thursday, 17 July

- 08:00–09:00 Breakfast
- 09:00–11:00 **Short Course (Part II)** *Chair: Jose Ameijeiras Alonso*
  Measurement errors: from basic concepts to recent advances (**Ingrid Van Keilegom**)
- 11:00–11:40 Coffee Break
- 11:40–12:20 **Invited Talk IV** *Chair: Andrea Meilán Vila*
  On a Parzen–Rosenblatt type estimator for circular data (**Carlos Tenreiro**)
- 12:20–13:20 **Contributed Session IV** *Chair: Andrea Meilán Vila*
  - The k-sample problem with incomplete survival data (**Adrián Lago**)
  - Simulation study of information criteria for copula model selection (**Aibat Kossumov**)
  - Goodness-of-fit testing with survival data (**Jacobo de Uña Álvarez**)
- 13:20–14:20 Lunch
- 14:20–15:00 **Invited Talk V** *Chair: María Alonso Pena*
  Specification tests for statistical models with recent results under a correlation distance–kernel approach (**Wenceslao González Manteiga**)
- 15:00–16:20 **Contributed Session V** *Chair: María Alonso Pena*
  - Testing uniformity on the sphere: from pairs to m-tuples (**Alberto Fernández-de-Marcos**)
  - A test for the equality of an estimable parameter in many populations (**Marcos Romero**)
  - Testing regression model specification via kernel methods (**María Vidal-García**)
  - Empirical study of distance correlation tests for random fuzzy numbers (**Miriam Alonso de la Fuente**)
- 16:20–16:40 Closing Ceremony
- 17:30–00:00 **Social Activity I**
  - 18:15–19:30 Short Tour: Oviedo
  - 20:00–23:15 Gala Dinner: *Espicha* at Tierra Astur Águila (Colloto)

Friday, 18 July

- 08:20–09:20 Breakfast
- 09:20–17:30 **Social Activity II**: Descent of the Sella River

Wednesday 16

Measurement errors: from basic concepts to recent advances

16th July
9:20–11:20
17th July
9:00–11:00
Short Course

Ingrid Van Keilegom

KU Leuven

In the nonparametric literature on the classical measurement error problem, where the goal is to do inference on the distribution of a variable X observed with independent additive error U, it is often assumed that the error density of U is perfectly known. In applications this is often too restrictive, since in many cases one does not know much about the complexity and imperfect nature of the underlying data. In this short course we present approaches for non- and semiparametric inference on the distribution of X developed in papers that do not assume the density of U to be (fully) known. Both the case where additional data is available (e.g. in the form or replicated measurements), and the case without additional data will be considered. We will also consider a number of more specific topics in which measurement errors play a prominent role, like boundary estimation, survival analysis and directional data.

Keywords: Kernel density estimation; Measurement error; Polynomial basis functions.

Tests for almost stochastic dominance and goodness-of-fit

16th July
12:00–12:40
Invited Talk I

Javier Cárcamo

University of the Basque Country

Stochastic dominance rules are partial orders among random elements and, hence, not every pair of distributions can be ranked. To address this limitation, Leshno and Levy [5] introduced the notion of *almost stochastic dominance* (ASD), which refers to the situation where strict SD does not hold but the noncompliance is minor. The unfulfillment degree, called *violation ratio*, is quantified by a parameter $\epsilon \in [0, 1]$; see [4].

In this work we study the asymptotic behavior of the *minimal violation ratio* (MVR) and establish conditions under which its bootstrap estimator is strongly consistent. These results enable the development of consistent testing procedures for ASD.

We also extend the framework to construct *almost goodness-of-fit* (AGoF) tests, procedures to decide if a (parametric) model provides a good representation of the observed sample up to a predefined margin of error. This leads to a flexible notion of model validation and allows quantifying the gain of a proposed model over a constant (uninformative) benchmark.

Our approach relies on empirical process theory [6] and differentiability techniques [3]. The methodology and main results are presented in [1] and [2].

Keywords: Bootstrap consistency; Empirical processes; Model validation; Stochastic orders; Violation ratio.

Acknowledgements This research is supported by the Spanish MCyT grant PID2023-148081NB-I00.

Bibliography

[1] A. Baíllo, J. Cárcamo, C. Mora-Corral. Tests for Almost Stochastic Dominance. *Journal of Business & Economic Statistics* **43(2)** (2024) 338-350.

[2] A. Baíllo, J. Cárcamo. Almost goodness-of-fit tests (2024). arXiv:2410.20918.

[3] J. Cárcamo, A. Cuevas, L.A. Rodríguez. Directional differentiability for supremum-type functionals: Statistical applications. *Bernoulli* **26** (2020) 2146-2175.

[4] H. Levy. Almost stochastic dominance and efficient investment sets. *American Journal of Operations Research* **2(3)** (2012) 313-321.

[5] M. Leshno, H. Levy. Preferred by "all" and preferred by "most" decision makers: Almost stochastic dominance. *Management Science* **48(8)** (2002) 1074-1085.

[6] A.W. van der Vaart, J.A. Wellner, (2023). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Second edition. Springer.

Uncertainty estimation for random forests with metric data

16th July
12:40–13:00
Contributed
Session I

Diego Serrano and Eduardo García-Portugués

Universidad Carlos III de Madrid, Departamento de Estadística, Spain

Random forests have been recently generalized for variables taking values in general metric spaces. The most relevant methods in this sense are Fréchet Random Forests (FRF) [1] and Random Forest Weighted Local Constant Fréchet Regression (RFWLCFR) [2]. We present a new methodology to quantify the uncertainty in the prediction of FRF/RFWLCFR. The confidence regions are based on the out-of-bag errors obtained from a single forest training, allowing the use of the entire dataset for both prediction and uncertainty estimation. This results in more precise and computationally efficient estimates compared to generic split-conformal approaches. Asymptotic coverage theory is presented in four different scenarios for the type of coverage. The proposed prediction regions are illustrated through simulations using RFWLCFR in different metric spaces, such as the Euclidean space, unit sphere, unit hyperboloid, and the space of symmetric positive definite matrices. The methodology is also applied to create prediction regions for the estimated death location of sunspot groups using their birth locations.

Keywords: Fréchet regression; Confidence regions; Random forests; Metric space data analysis.

Acknowledgements The first and second authors acknowledge support from the project INFLUENTIA-CM-UC3M and "PIPF Programa Inteligencia Artificial", funded by Universidad Carlos III of Madrid. The second author also acknowledges support from "Convocatoria de la Universidad Carlos III de Madrid de Ayudas para la recualificación del sistema universitario español para 2021–2023", funded by Spain's Ministerio de Ciencia, Innovación y Universidades.

Bibliography

[1] R. Qiu, Z. Yu, R. Zhu. *Random forest weighted local Fréchet regression with random objects*. J. Mach. Learn. Res. **25** (2024) 1-69.

[2] L. Capitaine, J. Bigot, R. Thiébaut, R. Genuer. *Fréchet Random Forests for Metric Space Valued Regression with Non Euclidean Predictors*. J. Mach. Learn. Res. **25** (2024) 1-41.

A new proposal of specification tests for the asynchronous functional concurrent model

16th July
13:00–13:20
Contributed
Session I

Laura Freijeiro-González[1], Manuel Febrero-Bande[2,3] and Wenceslao González-Manteiga[2,3]

[1]Department of Statistics and Operational Research and Didactics of Mathematics; Universidad de Oviedo, Oviedo, Spain.

[2]CITMAga; Universidade de Santiago de Compostela, Santiago de Compostela, Spain.

[3]Department of Statistics, Mathematical Analysis and Optimization; Universidad de Santiago de Compostela, Santiago de Compostela, Spain.

Functional concurrent models are specific functional regression models where both response (Y) and covariates (X) are functions of the same argument t, and their relation is concurrent or point-by-point. Its asynchronous design (AFCM) enables the observed curves for $Y(t)$ and $X(t)$ to be measured at different time instants for each individual in a given sample.

In the existing literature, it is often assumed that a specific structure (such as linear or additive) exists between $Y(t)$ and $X(t)$ for estimating the regression model. However, this assumption can be quite restrictive and difficult to verify in practice. Additionally, estimation of the model can be challenging when multiple covariates are involved. To address this issue, we propose new specification tests for dimensionality reduction within the general AFCM formulation. To this end, we adapt the conditional distance covariance (CDC) ideas of [1]. We illustrate the behavior of our procedure in a real dataset.

Keywords: AFCM; Bandwidth selection; Covariates selection; Conditional dependence; CDC.

Acknowledgements This work was supported by the Consellería de Cultura, Educación e Ordenación Universitaria along with the Consellería de Economía, Emprego e Industria of the Xunta de Galicia under Project ED481A-2018/264; and MICIU/AEI/ 10.13039/501100011033 and the Competitive Reference Groups 2021–2024 (ED431C 2021/24) from the Xunta de Galicia through the ERDF under Project PID2020-116587GB-I00. We also acknowledge to the Centro de Supercomputación de Galicia (CESGA) for computational resources.

Bibliography

[1] Wang, X., Pan, W., Hu, W., Tian, Y., and Zhang, H. *Conditional distance correlation.* J. Am. Stat. Assoc. **110(512)** (2015) 1726–1734.

On boxplots for functional data

16th July
13:20–13:40
Contributed
Session I

Stanislav Nagy[1], Tomáš Mrkvička[2] and Antonio Elías[3]

[1]Charles University, Prague, Czech Republic; [2]University of South Bohemia, České Budějovice, Czech Republic; [3]University of Málaga, Málaga, Spain

The boxplot for functional data is a prominent visualization tool for datasets where each observation is a function [3]. Such boxplots are based on the notion of a functional depth, ranking the observations from the typical to the more extreme ones. While, in principle, any functional depth can be used to construct a boxplot, the most common depths used in practice are integrated depths, e.g., the popular modified band depth [1]. In this contribution we highlight that integrated depths do not result in well-defined boxplots. Instead, we argue that infimal depths [2] are the only functional depths that provide a valid construction of a functional boxplot [4].

Keywords: Functional Boxplot; Functional Data Analysis; Exploratory Data Analysis; Visualization.

Acknowledgements This work was supported by the Czech Science Foundation (project n. 24-10822S) and the ERC CZ grant LL2407 of the Ministry of Education, Youth and Sport of the Czech Republic.

Bibliography

[1] S. López-Pintado, J. Romo. *On the concept of depth for functional data.* J. Amer. Statist. Assoc. **104**(486) (2011) 718–734.

[2] K. Mosler. *Depth statistics.* Robustness and complex data structures (2013) 17–34.

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

[4] S. Nagy, T. Mrkvička, A. Elías. *Which depth to use to construct functional boxplots?* (2025) arXiv:2409.18603.

Sequential data depth computation with statistical guarantees

16th July
14:40–15:20
Invited Talk II

Alicia Nieto-Reyes[1], Felix Gnettner[2] and Claudia Kirch[2]

[1]Departamento de Matemáticas, Estadística y Computación, Facultad de Ciencias, Universidad de Cantabria, Spain; [2]Institut für Mathematische Stochastik, Fakultät für Mathematik, Otto-von-Guericke-Universität Magdeburg, Germany

seMCD is a new sequential Monte Carlo method [1] for efficiently computing both theoretical and empirical statistical depth functions in multivariate and functional spaces. Instead of returning a single estimate, seMCD outputs an interval (a "bucket") that contains the desired value with a high, user-specified probability. The algorithm uses ideas from sequential testing to guarantee performance with fewer samples than traditional Monte Carlo methods. It is particularly useful when depth depends on unknown distributions and is applicable to tasks like anomaly detection, classification, and central region estimation, offering reliable results with reduced computational cost.

Keywords: Statistical data depth; Sequentially implemented Monte Carlo; Anytime-valid probabilistic guarantees.

Acknowledgements This work was financially supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 314838170, GRK 2297 MathCoRe and by grant PID2022-139237NB-I00 funded by "ERDF A way of making Europe" and MCIN/AEI/10.13039/501100011033.

Bibliography

[1] F. Gnettner, C. Kirch, A. Nieto-Reyes. *seMCD: Sequentially implemented Monte Carlo depth computation with statistical guarantees.* To appear in ArXiv (2025).

Optimal filament estimation: a geometrical approach

16th July
15:20–15:40
Contributed
Session II

Héctor González-Vázquez, Beatriz Pateiro-López and Alberto Rodríguez-Casal

Department of Statistics, Mathematical Analysis and Optimization, University of Santiago de Compostela, Spain; Galician Centre for Mathematical Research and Technology (CITMAga), Santiago de Compostela, Spain

The term manifold learning encompasses different methods that allow for a non-linear dimension reduction of high-dimensional data, revealing their low-dimensional manifold structure. Filament estimation techniques aim to recover the one-dimensional structure when the manifold is a curve.

In this work, we propose a new filament estimator, in a certain model with compact noise, that achieves the minimax rate of convergence in Hausdorff distance (up to logarithmic factor), as shown by [1], when the ambient space is the plane. Our estimator builds upon the Euclidean Distance Transform (EDT) estimator introduced in [2], enhancing it by incorporating a shape restriction known as r-convexity. This generalization of convexity offers greater flexibility. More specifically, we propose to estimate the support of the sampling distribution using the r-convex hull of the sample, see [3]. We also present a real data application illustrating the estimation of tree stem cross-sections in forest inventory.

Keywords: Filament estimation; Manifold learning; Minimax rate; r-convex hull; Set estimation.

Acknowledgements The authors acknowledge support from the R&D project PID2020-116587GB-I00 granted by MICIU/AEI/10.13039/501100011033.

Bibliography

[1] C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, L. Wasserman. *Minimax manifold estimation.* J. Mach. Learn. Res. **13** (2012) 1263-1291.

[2] C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, L. Wasserman. *The geometry of nonparametric filament estimation.* J. Am. Stat. Assoc. **107** (2012) 788-799.

[3] A. Rodríguez-Casal. *Set estimation under convexity type assumptions.* Ann. Inst. H. Poincaré Prob. Statist. **43** (2007) 763-774.

Kernel-based density estimation in tests of symmetry

16th July
15:40–16:00
Contributed
Session II

Marina Iturrate-Bobes and Raúl Pérez-Fernández

Department of Statistics and O.R. and M.D., University of Oviedo, Spain

Symmetry is a fundamental and often desirable property in many statistical contexts. For instance, many tables for symmetric distributions —such as the normal distribution— are greatly simplified, as cumulative probabilities in the right tail can be easily derived from those in the left tail. As another example, symmetry is a key assumption that simplifies the understanding of statistical tests such as Wilcoxon's signed rank test.

The earliest efforts to quantify the lack of symmetry of a random variable using what is now known as a skewness coefficient are commonly attributed to Pearson in the 19th century, although a rigorous axiomatic framework for these coefficients was developed much later with notable contributions from van Zwet [4] and Oja [3]. Numerous skewness coefficients have been proposed, ranging from classical ones such as those by Fisher, Pearson, and Bowley to more modern alternatives like the medcouple. Every skewness coefficient has an associated sample version that, under certain conditions, follows an asymptotically normal distribution centered at the population value. In particular, for symmetric distributions, the asymptotic distribution is always centered at zero but the asymptotic variance varies from one symmetric distribution to another. This is an issue that has been faced by different authors when developing tests of symmetry. In order to circumvent such a problem, some authors have considered the bootstrap [2] and some others have opted for either fixing a reference distribution or estimating the asymptotic variance directly [1].

In the present work, we study the influence of using kernel-density estimation techniques in tests of symmetry based on different skewness coefficients and, in particular, present a comparative power analysis across different symmetric and asymmetric distributions for different choices of kernel and bandwidth. The results show that the choice of skewness coefficient plays a much bigger role than the choices of kernel and bandwidth.

Keywords: Test of symmetry; Skewnesss coefficient; Kernel density estimation.

Bibliography

[1] P. Cabilio, J. Masaro. *A simple test of symmetry about an unknown median.* Can. J. Stat. **24** (1996) 349–361.

[2] M. Iturrate-Bobes, R. Pérez-Fernández. *Tests of symmetry based on the bootstrap estimation of the asymptotic variance of a skewness coefficient.* Inf Sci. **646** (2023) 1–13.

[3] H. Oja. *On location, scale, skewness and kurtosis of univariate distributions.* Scand. J. Stat. **8** (1981) 154–168.

[4] W. R. van Zwet. *Convex Transformations of Random Variables.* Mathematisch Centrum, Amsterdam, 1964.

Projection depth and L^r-type depths for fuzzy sets

16th July
16:00–16:20
Contributed
Session II

Luis González-De La Fuente[1], Alicia Nieto-Reyes[1] and Pedro Terán[2]

[1]Departamento de Matemáticas, Estadística y Computación, Universidad de Cantabria; [2]Departamento de Estadística e Investigación Operativa y Didáctica de las Matemáticas, Universidad de Oviedo

Statistical depth functions have become a standard tool in nonparametric statistics, since they pretend to extend the order in the real line to higher dimensions and more complex spaces. Two instances of depth are the projection depth and the family of L^r-type depths proposed in [3]. We aim to extend those instances of depth to the fuzzy space. In [1] we extend those depth functions and prove that the projection depth for fuzzy sets fulfills the notions of depth proposed in [2]. This implies that the fuzzy projection depth performs well in ordering fuzzy sets with respect to fuzzy random variables. Furthermore, we illustrate the good empirical behavior of the projection depth with a real data example involving trapezoidal fuzzy sets. Moreover, we prove that the L^r-type depths exhibit good properties, even though they do not satisfy all the axioms required by the definition of depth in the fuzzy setting. Finally, since trapezoidal fuzzy sets can be represented as elements of the multivariate space, we support our proposals by empirically demonstrating the superiority of the fuzzy depths compared to the multivariate projection depth when applied to trapezoidal fuzzy sets.

Keywords: Statistical depth functions; Fuzzy sets; Projection depth; L^r-type depths; Fuzzy random variables.

Acknowledgements The authors are supported by grant PID2022-139237NB-I00 funded by MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe". Additionally, L. González was supported by the Spanish Ministerio de Ciencia, Innovación y Universidades grant MTM2017-86061-C2-2-P. P. Terán is also supported by the Ministerio de Ciencia, Innovación y Universidades grant PID2019- 104486GB-I00.

Bibliography

[1] L. González-De La Fuente, A. Nieto-Reyes, P. Terán. *Projection depth and L^r-type depths for fuzzy random variables.* Fuzzy Sets Syst. **487** (2024) 108991.

[2] L. González-De La Fuente, A. Nieto-Reyes, P. Terán. *Statistical depth for fuzzy sets.* Fuzzy Sets Syst. **443**(A) (2022) 58–86.

[3] Y. Zuo, R. Serfling. *General Notions of Statistical Depth Function.* Ann. Stat. **28**(2) (2000) 461–482.

Cluster analysis via concave density regions

16th July
16:20–16:40
Contributed
Session II

José E. Chacón

Departamento de Matemáticas, Universidad de Extremadura, Spain

Nonparametric density-based cluster analysis can be approached through various techniques. Methods based on highest density regions [3] rely solely on density values, defining clusters as the connected components of the region where the density exceeds a given threshold. However, such regions do not cover the entire sample space and, besides, discerning the whole cluster structure may require scanning across all possible density levels [4]. In contrast, defining clusters as domains of attraction of density modes [1], based on density gradients, enables a full partitioning of the sample space for sufficiently smooth distributions.

In this work, we explore the use of concave density regions as proxies for cluster cores [2]. This approach shares the goal of high-density regions in the sense that we are interested only in the most prominent region of the clusters, but with the advantage that no density level is required. Moreover, the definition of this novel type of regions can be based either on the density or, under stronger smoothness assumptions, on the density Hessian.

Keywords: Cluster analysis; Concave density regions; Highest density regions.

Acknowledgements This work has been supported by the MICINN grants PID2021-124051NB-I00 and PID2023-148081NB-I00.

Bibliography

[1] J. E. Chacón. *A population background for nonparametric density-based clustering.* Stat. Sci. **30** (2015) 518-532.

[2] J. E. Chacón, J. Fernández Serrano. *Bump hunting through density curvature features.* TEST **32** (2023) 1251-1275.

[3] R. J. Hyndman. *Computing and graphing highest density regions.* Am. Stat. **50** (1996) 120-126.

[4] W. Stuetzle. *Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample.* J. Classif. **20** (2003) 25-47.

Comparing many populations using Gini covariance

16th July
17:20–18:00
Invited Talk III

M. Dolores Jiménez-Gamero and M. Remedios Sillero-Denamiel

Universidad de Sevilla, Spain

Given k populations and assuming that independent samples are available from each of them, the problem of testing for the equality of the k populations is addressed. With this aim, an unbiased estimator of the Gini covariance is taken as test statistic. In contrast to the classical setting, where k is kept fixed and the sample size from each population increases without bound, here k is assumed to be large and the size of each sample can either remain bounded or increase with k. The asymptotic distribution of the test statistic is stated under the null hypothesis as well as under alternatives, which allows us to study the consistency of the test. Specifically, it is shown that the test statistic is asymptotically free distributed under the null hypothesis. The finite sample performance of the test based on the asymptotic null distribution and the comparison with existing tests are studied via simulation.

Keywords: k-sample problem; Energy distance; Gini correlation; Asymptotic power; Consistency.

Acknowledgements The research in this paper has been supported by grant PID2022-137818OB-I00 (Ministerio de Ciencia, Innovación y Universidades, Spain).

Distributional convergence of empirical entropic optimal transport and applications

16th July
18:00–18:20
Contributed
Session III

Santiago Arenas Velilla, Axel Munk and Luis Alberto Rodríguez Ramírez

Institute for Mathematical Stochastics, Georg August-Universität Göttingen, Germany

The statistical properties of entropic optimal transport have recently become of great interest as this quantity has been shown to be useful for complex data analysis, among others due to its fast computability. Prominent applications meanwhile include tasks in machine learning, computer vision and various subject matters, such as econometrics or particle physics. In cell biology, colocalization analysis based on entropic optimal transport (EOT) has been used as a measure for quantification of spatial proximity of different protein assemblies. Using properties of the entropic optimal transport plans, we derive an asymptotic weak convergence result for a large class of functionals of the EOT plan, in which the colocalization process is included. The proof is based on Hadamard differentiability and the extended delta method. As applications, we obtain uniform confidence bands for colocalization curves and bootstrap consistency. Our theory is supported by simulation studies and is illustrated by real world data analysis from mitochondrial protein colocalization.

Keywords: Limit Laws; Entropic Optimal Transport; Colocalization Analysis; Uniform Confidence Bands; Bootstrap.

Quick quantile segmentation

16th July
18:20–18:40
Contributed
Session III

Housen Li, Axel Munk and Luis Alberto Rodríguez Ramírez

Institut für Mathematische Stochastik (IMS), Georg August-University of Göttingen, Germany

Change point estimation is often formulated as a search for the maximum of a gain function describing improved fits when segmenting the data. Searching through all candidates requires a linear amount of evaluations, i.e., $\mathcal{O}(n)$, of the gain function for an interval with n observations. If each evaluation is computationally demanding, this can become infeasible. Instead, we propose optimistic search methods with logarithmic amount of evaluations exploiting the specific structure of the gain function.

Towards a solid understanding of our strategy, in this talk we provide asymptotic consistency results for robust versions of the gain function. Specifically, we apply empirical process theory to provide a general asymptotic result for the change point estimator. Additionally, exponential bounds for the optimistic search methods are given. Finally, the performance of these algorithms will be illustrated by simulation.

Keywords: Change Point; Optimistic Search; Empirical Process; Trimmed Wasserstein Distance.

Bibliography

[1] S. Kovács, H. Li, L. Haubner, A. Munk, P. Bühlmann, *Optimistic search: Change point estimation for large-scale data via adaptive logarithmic queries.* (2020) arXiv preprint arXiv:2010.10194.

[2] S. Bobkov, M. Ledoux, (2019). *One-dimensional empirical measures, order statistics, and Kantorovich transport distances.* American Mathematical Society (Vol. 261, No. 1259).

Generalized distorted expectiles

16th July
18:40–19:00
Contributed
Session III

Ignacio Cascos

Department of Statistics, Universidad Carlos III de Madrid (Spain)

For a univariate sample or probability distribution, its *expectiles* are the solution to an asymmetric least squares minimization problem. Among other applications, expectiles have been used to construct *coherent* (subadditive) *risk measures*, see [2], and in data analysis, to build the expectile data depth, see [3]. In order to robustify the procedures based on the expectile depth, a *distortion* has been applied to the probability distribution in [5], which requires the use of *Choquet integrals*. A similar construction was previously employed in [1] to define distorted risk measures. In this work, we propose a generalization of the family of risk measures introduced in [1], based on two distortion functions and an expectile-based construction applied to Choquet integrals. Taking advantage of some results in [4], we will show that, under some specific conditions on the distortions, the risk measures are coherent. Finally, we propose a general back-testing procedure for *validating* capital requirement models that are based on risk measures indexed by confidence levels, as it is the case of the Value-at-Risk, the Expected Shortfall, the expectiles or any other construction for which there exists an associated notion of data depth.

Keywords: Choquet integral; Coherent risk measure; Distortion; Expectile; Validation.

Bibliography

[1] A. Balbás, J. Garrido, S. Mayoral. *Properties of distortion risk measures*. Methodol. Comput. Appl. Probab. **11** (2009) 385-399.

[2] F. Bellini, B. Klar, A. Müller, E.R. Gianin. *Generalized quantiles as risk measures*. Insur. Math. Econ. **54** (2014) 41-48.

[3] I. Cascos, M. Ochoa. *Expectile depth: Theory and computation for bivariate datasets*, J. Multivar. Anal. **184** (2021) 104757.

[4] D. Denneberg, (1994). *Non-additive measure and integral*. Kluwer Academic Publishers, 1994.

[5] M. Ochoa, I. Cascos. *Data depth and multiple output regression, the distorted M-quantiles approach*, Math. **10** (2022) 3272.

Thursday 17

On a Parzen-Rosenblatt type estimator for circular data

17th July
11:40–12:00
Invited Talk IV

Carlos Tenreiro

CMUC, Department of Mathematics, University of Coimbra, Portugal

Given an independent and identically distributed sample of angles $X_1, \ldots, X_n \in [0, 2\pi[$ from some absolutely continuous random variable X with unknown probability density function f, the standard kernel density estimator is defined, for $\theta \in [0, 2\pi[$, by

$$\tilde{f}_n(\theta; g) = \frac{c_g(L)}{n} \sum_{i=1}^{n} L\Big(\frac{1 - \cos(\theta - X_i)}{g^2}\Big),$$

where $L : [0, \infty[\to \mathbb{R}$ is a bounded function satisfying some additional conditions, $g = g_n$ is a sequence of positive numbers such that $g_n \to 0$, as $n \to \infty$, and $c_g(L)$, depending on the kernel L and the bandwidth g, is chosen so that $\tilde{f}_n(\cdot; g)$ integrates to unity. In this talk we consider an alternative estimator of f which is close in spirit to the Parzen-Rosenblatt estimator for linear data. For $\theta \in [0, 2\pi[$, it is defined by

$$\hat{f}_n(\theta; h) = \frac{d_h(K)}{n} \sum_{i=1}^{n} K_h(\theta - X_i),$$

where $h = h_n > 0$ is a sequence of strictly positive real numbers converging to zero as n tends to infinity, K_h is a real-valued periodic function on $\mathbb{R}$, with period 2π, such that $K_h(\theta) = K(\theta/h)/h$, for $\theta \in [-\pi, \pi[$, with K a kernel on $\mathbb{R}$, that is, an integrable real-valued function on $\mathbb{R}$ with $\int_{\mathbb{R}} K(u)du > 0$, and $d_h(K)$ is a normalizing constant depending on the kernel K and the bandwidth h which is chosen so that $\hat{f}_n(\cdot; h)$ integrates to unity. The connection between the two estimators, the consistency of the Parzen-Rosenblatt type estimator, and the automatic selection of the bandwidth, are, among others, issues that will be addressed in this talk.

Keywords: Circular data; Kernel density estimation; Bandwidth selection.

Acknowledgements Research partially supported by the Centre for Mathematics of the University of Coimbra – UID/MAT/00324.

Bibliography

[1] C. Tenreiro. *Kernel density estimation for circular data: a Fourier series-based plug-in approach for bandwidth selection.* J. Nonparametr. Stat. **34** (2022) 377-406.

[2] C. Tenreiro. *A Parzen-Rosenblatt type density estimator for circular data: asymptotic and exact optimal bandwidths.* Comm. Statist. Theory Methods **53** (2024) 7436-7452.

[3] C. Tenreiro. *A note on a Parzen-Rosenblatt type density estimator for circular data.* In New Frontiers in Statistics and Data Science, (SPE2023, Guimarães, Portugal, October 11-14), L. Henriques-Rodrigues et al. (eds.), Springer Proceedings in Mathematics & Statistics **469** (2025) 1-11.

The k-sample problem with incomplete survival data

17th July
12:20–12:40
Contributed
Session IV

Adrián Lago, Juan Carlos Pardo-Fernández and Jacobo de Uña-Álvarez

Department of Statistics and Operations Research, CITMAga, Universidade de Vigo, Vigo, Spain

Truncation and censoring frequently arise in applied fields such as medicine, reliability, or economics, often leading to partial observation of individuals or their failure times. As a main consequence, these mechanisms make the sample unrepresentative of the underlying population. The main statistical techniques available to test for the equality of distributions in this context are mainly rank-based methods, which are not omnibus, or standard tests, which are not consistent either. In this talk, we introduce new procedures for testing the equality of k density functions in the presence of left truncation and possible right censoring. First, a smooth test based on a nonparametric estimator of the density functions is proposed, specifically adapted to truncated data. Second, natural extensions of the classical Kolmogorov–Smirnov and Cramér–von Mises tests are developed for the k-sample problem under truncation and censoring. For each proposed test, the asymptotic distribution under the null hypothesis is derived, and the consistency is established. Since the limiting distributions depend on unknown quantities, two bootstrap procedures are introduced to approximate the null distribution of the proposed tests. The method is validated via Monte Carlo. The performance of the proposed tests under the alternative hypothesis is addressed in a simulation study, where the log-rank test is also considered for comparison. Finally, the practical performance of the new proposals is illustrated through a real dataset on the unemployment duration of Galician women.

Keywords: Bootstrap; Censoring; K-sample problem; Log-rank test; Truncation.

Simulation Study of Information Criteria for Copula Model Selection

17th July
12:40–13:00
Contributed
Session IV

Aibat Kossumov

Department of Probability and Mathematical Statistics, Charles University, Czech Republic

One of the fundamental problems in dependence modeling is the selection of an appropriate parametric copula model. In [1], it was shown that an adaptation of the Akaike Information Criterion (AIC) does not exist for most commonly used copula models. As a possible alternative, the authors proposed the Copula Information Criterion (CIC) based on leave-one-out cross-validation. In [2], the AIC and CIC were compared, and only minor differences were observed. In the context of linear model selection, Jun Shao [3] demonstrated that the optimal selection procedure is leave-n_v-out cross-validation, where n_v is of the same order as the sample size n, i.e., $n_v/n \to 1$ as $n \to \infty$. This idea is adapted to the context of copula model selection. Its performance is compared with that of AIC and CIC.

Keywords: Cross-validation; Parametric copula model.

Bibliography

[1] S. Grønneberg, N. L. Hjort. *The copula information criteria.* Scand. J. Stat. **41** (2014) 436–459.

[2] L. A. Jordanger, D. Tjøstheim. *Model selection of copulas: AIC versus a cross validation copula information criterion.* Statist. Probab. Lett. **92** (2014) 249–255.

[3] J. Shao. *Linear model selection by cross-validation.* J. Amer. Statist. Assoc. **88** (1993) 486–494.

Goodness-of-fit testing with survival data

17th July
13:00–13:20
Contributed
Session IV

Jacobo de Uña-Álvarez[1] and Juan Carlos Escanciano[2]

[1]Universidade de Vigo; [2]Universidad Carlos III de Madrid

In this talk a new general strategy for goodness-of-fit testing with survival data will be introduced. The setting is that of testing for a parametric family of distribution functions when the data are deteriorated due to random censoring and/or random truncation. A key step is the characterization of the null hypothesis through a moment equation which involves the estimation of the observable distribution under both the null and the alternative. An omnibus test based on a maximum mean discrepancy principle will be proposed, and its theoretical properties will be presented. The finite sample performance of the proposed test will be investigated through simulations. Illustrative real data applications will be given.

Keywords: Censoring; Truncation.

Acknowledgements Work supported by the Grant PID2023-148811NB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU.

Specification tests for statistical models with recent results under a correlation distance-kernel approach

17th July
14:20–15:00
Invited Talk V

Wenceslao González-Manteiga

Department of Statistics, Mathematical Analysis and Optimization. University of Santiago de Compostela, Spain; Galician Centre for Mathematical Research and Technology (CITMAga), Santiago de Compostela, Spain

Specification tests in statistical models refer to the assertion or rejection of some assumption about the statistical model structure. These cover well studied topics as for example the goodness-of-fit for distributions, densities or regression models between others. The objectives, in general, can be different: Testing if one regression model belongs to some parametric family, the comparison of the structure of k populations, testing the significance of covariates in one high dimensional regression model... etc. In this talk we will revise the different methodologies developed in the last 30 years (of my academic life) with recent results in complex models with some different applications and with additional studies for the testing the goodness of fit under a correlation distance-kernel approach.

Keywords: Specification Tests; Goodness-of-Fit; Correlation Distance.

Testing uniformity on the sphere: from pairs to m-tuples

17th July
15:00–15:20
Contributed
Session V

Alberto Fernández-de-Marcos[1], Eduardo García-Portugués[1] and Thomas Verdebout[2]

[1]Department of Statistics, Universidad Carlos III de Madrid (Spain);
[2]Department of Mathematics, Université Libre de Bruxelles (Belgium)

When uniformity on the sphere is tested, most classical tests employ a V-statistic of second degree investigating the relation between pairwise observations to detect deviations from uniformity. In this work, we propose two new classes of uniformity tests based on U- and V-statistics of arbitrary degree m that capture the interaction between m-tuples of observations. The computation of the new class of V-statistics is shown to be manageable even for large m values, and we provide closed-form circular test statistics that extend classical tests for $m = 2$. We study the asymptotic null distribution for these m-tests, its usability in practice, and its asymptotic behavior under local alternatives. Furthermore, we investigate the impact of m on the rotational invariance of the proposed test statistics and introduce a modified construction that ensures this invariance for all $m > 2$, with closed expressions and known asymptotic distribution in the circular case. Through simulations, we show that tests with $m > 2$ lead to increases in power, compared to $m = 2$, for certain scenarios.

Keywords: Circular data; Spherical data; Uniformity tests.

Acknowledgements The first and second authors acknowledge support from PID2021-124051NB-I00, funded by MCIN/AEI/10.13039/50110001103 and by "ERDF A way of making Europe". The first author is supported by grant PIPF-2023/COM-30702, funded by Comunidad de Madrid. The second author also acknowledges support from "Convocatoria de la Universidad Carlos III de Madrid de Ayudas para la recualificación del sistema universitario español para 2021–2023", funded by Spain's Ministerio de Ciencia, Innovación y Universidades. The third author acknowledges support from a PDR and a Hubert-Curient grants from the Fonds National de la Recherche Scientifique (FNRS).

A test for the equality of a estimable parameter in many populations

17th July
15:20–15:40
Contributed
Session V

Marcos Romero-Madroñal, M. Remedios Sillero-Denamiel and M. Dolores Jiménez-Gamero

Department of Statistics and Operations Research, Universidad de Sevilla, Spain
Instituto de Matemáticas de la Universidad de Sevilla, Spain

The problem of comparing a parameter, estimable through U-statistics, across k populations is addressed. The kernel is assumed to have degree two. A statistic is proposed and shown to be distribution-free under the null hypothesis and certain alternatives for large k. This property under the null allows for the construction of a test, while the distribution under alternatives enables a power analysis. The finite performance of the test is numerically assessed through a simulation study.

Keywords: Testing; Degree-2 U-statistics; Consistency.

Acknowledgements This research has been financed by research project PID2022-137818OB-I00 (Ministerio de Ciencia, Innovación y Universidades, Spain).

Testing regression model specification via kernel methods

17th July
15:40–16:00
Contributed
Session V

María Vidal-García[1], Ingrid Van Keilegom[2], Rosa Crujeiras[1] and Wenceslao González-Manteiga[1]

[1]Department of Statistics, Mathematical Analysis and Optimization, CITMAga-Universidade de Santiago de Compostela, Spain; [2]Faculty of Economics and Business, Katholieke Universiteit Leuven, Belgium

In recent years, significant advances have been made in statistical methodologies based on the kernel approach. This approach, which originated as a tool for classification in the field of Machine Learning, has since become a valuable tool in Statistics, with many applications in specification testing [3] (including two-sample testing, independence testing, and goodness-of-fit tests).

This work introduces a kernel-based method for testing the goodness-of-fit of regression functions within scale-location models. Building on prior research, the test relies on the reformulation of the original contrast in terms of the distribution of the residuals. Previous strategies [4] employed Kolmogorov-Smirnov and Cramér-von Mises type statistics. The proposed method offers a novel alternative using the maximum mean discrepancy (MMD), a statistical distance derived from embedding observations in a Reproducing Kernel Hilbert Space. A theoretical review of MMD and the underlying assumptions is provided [1, 2], followed by a presentation of the new test and its properties.

The last part of the talk focuses on practical challenges. Since reliance on asymptotic distributions often yields suboptimal results, resampling techniques are used for implementation, particularly the residual and smooth residual bootstrap adapted to account for heteroskedasticity [5]. The performance of the proposed method is assessed via simulation and compared with previous alternatives in the literature.

Keywords: Maximum mean discrepancy; Specification test; Regression; Kernel mean embedding.

Acknowledgements This work is part of R&D project PID2020-116587GBI00 granted by MICIU/AEI/10.13039/501100011033.

Bibliography

[1] A. Berlinet, C. Thomas-Agnan, (2011). *Reproducing kernel Hilbert spaces in probability and statistics.* Springer Science & Business Media, 2011.

[2] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, G. R. Lanckriet. *Hilbert space embeddings and metrics on probability measures.* JMLR **11** (2010) 1517-1561.

[3] G. J. Székely, M. L. Rizzo. *The energy of data.* Annu. Rev. Stat. Appl. **4(1)** (2017) 447-479.

[4] I. Van Keilegom, W. González Manteiga, C. Sánchez Sellero. *Goodness-of-fit tests in parametric regression based on the estimation of the error distribution.* TEST **17** (2008) 401–415.

[5] N. Neumeyer. *Smooth residual bootstrap for empirical processes of non-parametric regression residuals.* Scand. J. Stat. **36(2)** (2009) 204–228.

Empirical study of distance correlation tests for random fuzzy numbers

17th July
16:00–16:20
Contributed
Session V

Miriam Alonso de la Fuente and Pedro Terán

Department of Statistics and Operations Research and Mathematics Didactics, University of Oviedo, Spain

Random elements taking values on the space of fuzzy numbers, that is, upper semicontinuous functions in the real line with values in the unit interval, are called random fuzzy numbers and allow us to capture both imprecision and randomness. Distance correlation statistics [5, 4] provide a way of testing independence of random vectors. We develop a way of applying these statistics to random fuzzy numbers by simulating random vectors, which are measurable as a consequence of [1], with exchangeable coordinates with finite variance in order to maintain the requirements of the second test. Next, we apply these methods to real data from [3] and perform a simulation study for random fuzzy numbers with trapezoidal shape using the method described in [2] to compare the power of both variants with different distributions, sample sizes and lengths of the random vectors generated. We observe that the dimension of the vectors has little influence over the power when compared to the sample size. This shows that we can generate vectors of low dimension, which has less computational cost.

Keywords: Random fuzzy numbers; Distance correlation; Random vectors; Independence.

Acknowledgements Research in this paper was partially funded by grants from Spain via the project PID2019-104486GB-I00 and the Principality of Asturias via the Severo Ochoa grant PA-21-PF-BP20-112. Their contribution is gratefully acknowledged.

Bibliography

[1] M. Alonso de la Fuente, P. Terán. *Joint measurability of mapings induced by a fuzzy random variable*. Fuzzy Sets Syst. **424** (2021) 92-104.

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

[3] SMIRE+CoDIRE `https://bellman.ciencias.uniovi.es/smire+codire/index.html`

[4] G. J. Székely, M. L. Rizzo. *The distance correlation t-test of independence in high dimension*. J. Multivar. Anal. **117** (2013) 193-213.

[5] G. J. Székely, M. L. Rizzo, N. K. Bakirov. *Measuring and testing dependence by correlation of distances*. Ann. Stat. **35** (2007) 2769-2794.

List of all attendees

1. Miriam Alonso de la Fuente (University of Oviedo)
2. María Alonso Pena (Universidade de Santiago de Compostela)
3. Jose Ameijeiras Alonso (Universidade de Santiago de Compostela)
4. Santiago Arenas Velilla (Georg August-University of Göttingen)
5. Diego Bolón (Université Libre de Bruxelles)
6. Francisco Branquinho (Universidade do Porto)
7. Catarina Campos Silva (University of Porto)
8. Javier Cárcamo Urtiaga (University of the Basque Country)
9. Ignacio Cascos (Universidad Carlos III de Madrid)
10. José E. Chacón (Universidad de Extremadura)
11. Rui Costa-Miranda (University of Porto)
12. Tomás R. Cotos-Yáñez (Universidade de Vigo)
13. Jacobo de Uña Álvarez (Universidade de Vigo)
14. Arís Fanjul Hevia (Universidad de Oviedo)
15. Alberto Fernández de Marcos (Universidad Carlos III de Madrid)
16. Laura Freijeiro González (Universidad de Oviedo)
17. Rita Gaio (University of Porto)
18. Jesus García García (Dartmouth College)
19. Eduardo García Portugués (Universidad Carlos III de Madrid)
20. Luis Angel García-Escudero (Universidad de Valladolid)
21. Matilde Gonçalves (Universidade do Porto)
22. Alberto González (Columbia University)
23. Luis González de la Fuente (Universidad de Cantabria)
24. Wenceslao González-Manteiga (Universidade de Santiago de Compostela)
25. Héctor González-Vázquez (Universidade de Santiago de Compostela)
26. María Dolores Jiménez Gamero (University of Seville)

27. Aibat Kossumov (Charles University)
28. Adrián Lago (Universidade de Vigo)
29. María Dolores Martínez Miranda (Universidad de Granada)
30. Agustín Mayo-Iscar (Universidad de Valladolid)
31. Andrea Meilán Vila (Universidad Carlos III de Madrid)
32. Stanislav Nagy (Charles University)
33. Alicia Nieto-Reyes (Universidad de Cantabria)
34. Juan Carlos Pardo-Fernández (Universidade de Vigo)
35. Ana Pérez González (Universidade de Vigo)
36. Raúl Pérez-Fernández (Universidad de Oviedo)
37. Sonia Pérez-Fernández (Universidad de Oviedo)
38. Luis Alberto Rodríguez Ramírez (Georg August-University of Göttingen)
39. David Rodríguez Vítores (Universidad de Valladolid)
40. Marcos Romero Madroñal (Universidad de Sevilla)
41. Diego Serrano Ortega (Universidad Carlos III de Madrid)
42. Germán Ernesto Silva Gómez (University of Oviedo - University of Granada)
43. Beatriz Sinova Fernández (Universidad de Oviedo)
44. Carlos Tenreiro (University of Coimbra)
45. Pedro Terán (Universidad de Oviedo)
46. Lucía Trapote (Universidad de Valladolid)
47. Ingrid Van Keilegom (KU Leuven)
48. María Vidal-García (Universidade de Santiago de Compostela)

Index of authors