PROGRAMME AND ABSTRACTS

24th International Conference on

Computational Statistics (COMPSTAT 2022)

http://www.compstat2022.org

Plesso Belmeloro, University of Bologna, Italy
23-26 August 2022

CSDA & EcoSta Workshop on

Statistical Data Science (SDS 2022)

http://www.compstat2022.org/SatelliteWorkshop.php

Department of Statistical Sciences "Paolo Fortunati", University of Bologna, Italy
26-28 August 2022

ELSEVIER

ISBN: 978-90-73592-40-7

COMPSTAT 2022 Scientific Program Committee:

Ex-officio:
COMPSTAT 2022 organiser and chairperson of the SPC: Alessandra Luati and Maria Brigida Ferraro.
Past COMPSTAT organiser: Cristian Gatu.
Next COMPSTAT organiser: Erricos Kontoghiorghes.
Incoming IASC-ERS Chairman: Cristian Gatu.

Members:
Peter Filzmoser, Christian Hennig, Tsung-I Lin, Martina Mittlboeck, Domingo Morales and Miguel de Carvalho.

Consultative Members:
Representative of the IFCS: Berthold Lausen.
Representative of the ARS of IASC: Philip Yu.
Representative of the LARS of IASC: David Fernando Munoz Negron.
Representative of CMStatistics: Ana Colubi.

Local Organizing Committee:
Alessandra Amendola, Enea Bongiorno, Fabrizio Durante, Marzia Freo and Paolo Giordani.

SDS 2022 Scientific Program Committee:

Members:
Elvezio Ronchetti, Ivan Kojadinovic, Bertrand Clarke, Xinyuan Song, Michele Guindani, Peter Winker, Chenlei Leng, Stefano Castruccio, Taps Maiti, Igor Pruenster, Hans-Georg Mueller, Juan Romo, Cheng Yong Tang and Jane-Ling Wang.

Organizers:
Ana Colubi, Erricos Kontoghiorghes, M. Brigida Ferraro, Marzia Freo and Alessandra Luati.

Dear Colleagues and Friends,

We wish to warmly welcome you to Bologna for the 24th International Conference on Computational Statistics (COMPSTAT 2022) and the CSDA & EcoSta Workshop on Statistical Data Science (SDS 2022). After two years of postponements due to the pandemic, we are especially grateful to all those who have kept re-organizing their plans and agendas to join us for these events, either in person or virtually. For many of us, this will be the first opportunity to network in person since 2019. It will be, thus, a special occasion, and we have endeavoured to make it memorable.

These events are locally organized mainly by members of the University of Bologna and The Sapienza University of Rome, assisted by renowned international researchers. The COMPSTAT is an initiative of the European Regional Section of the International Association for Statistical Computing (IASC-ERS), a society of the International Statistical Institute (ISI). COMPSTAT is one of the most prestigious world conferences in Computational Statistics, regularly attracting hundreds of researchers and practitioners.

The first COMPSTAT conference took place in Vienna in 1974, and the last edition took place in Iasi, Romania, in 2018. It has gained a reputation as an ideal forum for presenting top-quality theoretical and applied work, promoting interdisciplinary research and establishing contacts amongst researchers with common interests.

Keynote lectures are addressed by Prof. Igor Pruenster, Bocconi University, Italy, Prof. Jean-Michel Zakoian, ENSAE, France, and Prof. Holger Dette, Ruhr-University of Bochum, Germany.

More than 550 submissions have been received for COMPSTAT, and about 450 have been retained for presentation at the conference. The conference programme has 50 contributed sessions, 8 invited sessions, 3 keynote talks, 54 organized sessions and 2 tutorials. There are approximately 520 participants. For the first time, the conference will be hybrid, and all the sessions will be live-streamed so that participants can attend online the full conference.

The CSDA & EcoSta Workshop on Statistical Data Science has about 65 participants and 50 talks. SDS keynote lectures are addressed by Prof. Geoffrey McLachlan, University of Queensland, Australia, Prof. Peter Rousseeuw, KU Leuven, Belgium and Prof. Patrick J. Wolfe, Purdue University, United States.

The organization would like to thank the authors, referees and all participants of COMPSTAT 2022 who contributed to the success of the conference. Our gratitude to sponsors, the scientific programme committee, session organizers, local hosts, the city of Bologna, and many volunteers who have contributed substantially to the conference. We acknowledge their work and support.

The forthcoming COMPSTAT conference, which has been affected by the postponement, will take place in an odd year as an exception. The COMPSTAT 2023 organizers invite you to participate in London, UK, 22-25 August 2023. We wish the best of success to Erricos Kontoghiorghes, Chair of the 25th COMPSTAT edition.

Alessandra Luati and Maria Brigida Ferraro.

Contents

An extension of Weighted Quantile Sum (WQS) regression is proposed which estimates the double effect of a mixture of chemicals on a health outcome in the same model through the inclusion of two indices, one in the positive and one in the negative direction, with the introduction of a penalization term. To evaluate the performance of this new model in terms of the estimation of the regression parameters and the weights we performed both a simulation study and a real case study where we assessed the effects of nutrients on obesity among adults. The results showed good performance of the method in estimating both the regression parameter and the weights associated with the single elements when the penalized term was set equal to the magnitude of the AIC of the unpenalized WQS regression. The two indices further helped to give a better estimate of the parameters (Positive direction Median Error (PME): 0.017; Negative direction Median Error (NME): -0.023) compared to the standard WQS (PME: -0.141; NME: 0.078). In the case study, WQS with two indices was able to find a significant effect of nutrients on obesity in both directions identifying caffeine and magnesium as the main actors in the positive and negative association respectively. We introduce an extension of the WQS regression that showed how to improve the accuracy of the parameter estimates when considering a mixture of elements that can have both a protective and a harmful effect on the outcome

C0407: **Divide and conquer approaches for nonparametric regression and variable selection**
Presenter: **Sapuni Chandrasena**, University of Toledo, United States
Co-authors: Rong Liu

The rapid emergence of massive data with increasing size requests new statistical methods, especially in the fields of nonparametric regression, which is flexible but usually computationally intensive. To overcome the limitations of computing and storage, various distributed frameworks for statistical estimation and inference have been proposed. We study the statistical efficiency and asymptotic properties of the spline estimation for generalized additive models using the divide-and-conquer (DAC) approach. We also provide a variable selection method based on the majority voting procedure. The simulation study strongly supports the asymptotic theory and shows that the DAC approach is much more computational expedient without losing much accuracy.

CO051 Room Aula G IASC-ARS SESSION: COMPUTATIONS FOR CATEGORICAL DATA (VIRTUAL) Chair: Yuichi Mori

C0292: **The Cressie-Read divergence statistic and correspondence analysis; a unifying approach with possible extensions**
Presenter: **Rosaria Lombardo**, University of Campania, Italy
Co-authors: Eric Beh

In the correspondence analysis literature, the foundations of visually and numerically summarising the association between two categorical variables rest with Pearson's chi-squared statistic. Not only is this statistic extremely popular and versatile, but it also yields some very useful visual and numerical properties. More recently, ties have been established that show the role that the Freeman-Tukey statistic plays in correspondence analysis and confirmed the advantages of the Hellinger distance that have long been advocated. Both Pearson's and the Freeman-Tukey statistics are special cases of the Cressie-Read divergence statistic, as are the Cressie-Read statistic, the likelihood ratio statistic and their modified versions. Therefore, correspondence analysis will be explored where the association, and the resulting low-dimensional correspondence plot, have at its foundation this divergence statistic. By doing so, the properties of correspondence analysis are described for any special case of the Cressie-Read divergences statistic which includes the Hellinger distance decomposition (HDD) method and log-ratio analysis (LRA). Some extensions to this method will also be discussed including its role in multiple and multi-way correspondence analysis.

C0369: **A multiple correspondence analysis for aggregated symbolic data**
Presenter: **Junji Nakano**, Chuo University, Japan
Co-authors: Nobuo Shimizu, Yoshikazu Yamamoto

When we have a huge amount of data, we sometimes are interested in comparing meaningful groups of data, not individual observations. Aggregated symbolic data (ASD) expresses a group of observations that have continuous and categorical variables by using up to second moments of variables. ASD for a group of data is equivalent to the set of means, variances, and correlations for continuous variables, Burt matrix for categorical variables, and means of a continuous variable against one value of a categorical variable. As ASD with many categorical variables is still complicated, it is preferable to have simple measures of location and dispersion for a categorical variable, and a measure of the correlation between two categorical and/or continuous variables. We propose such measures by extending multiple correspondence analysis to ASD. They are compared with other measures, for example, correlation measures based on the polychoric correlation coefficient.

C0489: **A general framework for implementing distance measures for categorical variables**
Presenter: **Michel van de Velden**, Erasmus University Rotterdam, Netherlands
Co-authors: Alfonso Iodice D Enza, Angelos Markos, Carlo Cavicchia

In many statistical methods, distance plays an important role. For instance, data visualization, classification and clustering methods require quantification of distances among objects. How to define such distance depends on the nature of the data and/or problem at hand. For the distance between numerical variables, in particular in multivariate contexts, there exist many definitions that depend on the actual observed differences between values. It is worth underlining that often it is necessary to rescale the variables before computing the distances. Many distance functions exist for numerical variables. For categorical data, defining a distance is even more complex as the nature of such data prohibits straightforward arithmetic operations. Specific measures, therefore, need to be introduced that can be used to describe or study the structure and/or relationships in the categorical data. We introduce a general framework that allows an efficient and transparent implementation of the distance between categorical variables. We show that several existing distances (for example distance measures that incorporate association among variables) can be incorporated into the framework. Moreover, our framework quite naturally leads to the introduction of new distance formulations as well.

C0544: **cGAPdb: A matrix visualization database for categorical data sets**
Presenter: **Chun-houh Chen**, Academia Sinica, Taiwan
Co-authors: Shao-An Chen, Chiun-How Kao, Sheau-Hue Shieh, Han-Ming Wu

cGAPdb is a graphical database for categorical data sets for public use. The major type of visualization provided in this database is matrix visualization with the cGAP (Categorical Generalized Association Plots) environment. Most of the categorical data sets from the UCI Machine Learning Repository are included in this graphical database. All elements of a cGAP display such as (homals analysis, data matrix, proximity matrix for variables and samples, seriation method, etc.) are provided for each data set for users to browse and download. Additional categorical data sets other than those from the UCI Repository have also been collected in cGAPdb. A cGAP working place is available in cGAPdb for users to upload their own data sets for creating cGAP matrix visualization displays.

CO037 Room Aula C CLUSTERING METHODS AND COPULA FUNCTION Chair: F Marta L Di Lascio

C0508: **Copula-based clustering of dependent variables with application to flood risks**
Presenter: **Roberta Pappada**, University of Trieste, Italy
Co-authors: Fabrizio Durante, Sebastian Fuchs

In recent years, copula-based measures of association have been exploited to develop clustering methods that can take into account the dependence structure characterizing the underlying data generating process, e.g., when the data objects to cluster are time series. Motivated by the interest in clustering flood data, which are characterized by a number of physical variables (such as flood peak and volume) and collected at specific

geographical sites, some dissimilarity measures are proposed to cluster continuous random variables. Such measures are rank-based, hence depend on the copula of the involved random variables and assign the smallest value to two subsets of random variables that are pairwise comonotonic. Two different notions of multivariate comonotonicity for pairs of random vectors are investigated, which correspond to the strongest version of comonotonicity and a weaker notion called π-comonotonicity. The proposed dissimilarities are embedded into a hierarchical clustering procedure, with the final aim to detect clusters that account for the comovements of random variables. An application to the analysis of flood risks concerning data collected in the Po river basin is presented, along with the results from different simulated scenarios.

C0392: **Copula-based non-metric unfolding**
Presenter: **Marta Nai Ruscone**, Universita degli Studi di Genova, Italy
Co-authors: Antonio Dambrosio, Daniel Fernandez
A multidimensional unfolding technique that is not prone to degenerate solutions and is based on multidimensional scaling of a complete data matrix is proposed. We adopt the strategy of augmenting the data matrix, trying to build a complete dissimilarity matrix, by using Copulas-based association measures among rankings (the individuals), and between rankings and objects (namely, a rank-order representation of the objects through tied rankings). The proposed technique leads to an acceptable recovery of given preference structures. Applications on real datasets show that our procedure returns non-degenerate unfolding solutions.

C0218: **Mixtures with a prior on the number of components and the telescoping sampler**
Presenter: **Gertraud Malsiner-Walli**, WU Vienna University of Economics and Business, Austria
Co-authors: Sylvia Fruhwirth-Schnatter, Bettina Gruen
Within a Bayesian framework, a comprehensive investigation of the model class of mixtures of finite mixtures (MFMs) where a prior on the number of components is specified is performed. This model class has applications in model-based clustering as well as for semi-parametric density approximation, but requires suitable prior specifications and inference methods to exploit its full potential. We contribute to the Bayesian analysis of MFMs by (1) considering static and dynamic MFMs where the Dirichlet parameter of the component weights either is fixed or depends on the number of components, (2) proposing a flexible prior distribution class for the number of components, (3) characterizing the implicit prior on the number of clusters as well as partitions by deriving computationally feasible formulas, (4) linking MFMs to Bayesian non-parametric mixtures, and (5) finally proposing a novel sampling scheme for MFMs called the telescoping sampler which allows Bayesian inference for mixtures with arbitrary component distributions. The telescoping sampler explicitly samples the number of components, but otherwise requires only the usual MCMC steps for estimating a finite mixture model. The ease of its application using different component distributions is demonstrated on real data sets.

C0425: **Clustering Italian regions on the basis of bivariate income and consumption distributions**
Presenter: **Francesca Condino**, University of Calabria, Italy
Co-authors: Antonio Irpino, Rosanna Verde
In an economic framework, modeling income and consumption characteristics simultaneously can be of considerable relevance. Moreover, it could be of interest to identify homogeneous regions in a country in terms of economic behaviour. With this aim, we propose to jointly model income and consumption data through the copula approach and use the obtained bivariate density functions as descriptors of regions for clustering analysis purposes. In particular, considering data from the Survey on Households Income and Wealth (SHIW) by the Bank of Italy, the bivariate distributions at the regional level are obtained. The Jensen-Shannon divergence can be usefully employed to measure the discrepancies across density functions, as it allows us to take into account marginal and copula effects. The Italian regions are then partitioned in clusters by using a dynamic clustering algorithm, a non-hierarchical iterative algorithm, based on the optimization of an adequacy criterion that measures the fit between clusters and their prototypes. It can be shown that the divergence of all considered objects can be decomposed into two quantities, one relating to the heterogeneity present in the clusters and the other reflecting the discrepancy across clusters, according to Huygens' theorem.

CO148 Room Aula D EARLY CAREER ADVICE FOR STATISTICIANS IN THE COMPUTATIONAL SCIENCES Chair: Thomas Yee

C0295: **Different flavors of publishing computational work**
Presenter: **Ursula Laa**, BOKU University, Austria
The publication of work on computational methods comes in different flavors: from publishing software through a repository such as CRAN (with accompanying documentation), all the way to describing new concepts and approaches theoretically in a journal article. However, most of the time the ideal solution is somewhere in the middle: we both share the software and its documentation, and also describe the details in an associated research paper. We will present a broad overview of different types of journals relevant to computational statistics, and the accompanying expectations in terms of presentation and availability of software. This will be illustrated with examples from my own experience.

C0298: **Software development and statistical research: Some reflections**
Presenter: **Thomas Yee**, University of Auckland, New Zealand
Statisticians developing new methodology are obliged to provide software implementing the work as it facilitates its use and promotes reproducible research. However, writing good quality software takes much time, and this could be spent writing more papers. With fewer journal publications in general, academics pursuing this line of output are disadvantaged from those traditionally involved in publishing only. Some thoughts on this topic are shared. It is aimed more toward early-career researchers, however people of all ages should find something to identify with.

C0315: **The role of communities of practice for career development in computational statistics**
Presenter: **Laura Vana**, TU Wien, Austria
The aim is to reflect on how modern communities of practice, with a focus on meetup groups such as useR, R Ladies, PyLadies, can be leveraged by early career statisticians in the computational sciences to enhance their expertise, network and gain visibility. Finally, we will provide an overview of such modern communities in the area of computational statistics and data science as well as provide some guidelines on how to build and maintain impactful, safe and inclusive communities.

C0409: **What is the best programming language for computational sciences: No need to choose, be a polyglot**
Presenter: **Michele La Rocca**, University of Salerno, Italy
Early in their careers, a common question for students and data scientists is which programming language is best to learn. The question is somewhat misleading: every programming language has its strengths and weaknesses. Often, R and Python are compared with conclusions that, in some cases, point towards Python in others towards R. However, the correct answer to the question is not R *or* Python, but R *and* Python. Besides R and Python, Julia is receiving more and more attention from the data science community, again with significant strengths and some weaknesses. Especially at the beginning of their careers, computational scientists should be multilingual and learn complementary programming languages to cover future needs in different fields of application and career perspectives. The knowledge of any programming language is exposed to some degree of obsolescence. At the beginning of a career, the focus should not be on coding but rather on programming, especially on programming paradigms (OOP, functional programming, etc.) that have a higher degree of resilience.

CO035 Room Aula F RECENT DEVELOPMENTS OF VARIATIONAL APPROXIMATIONS Chair: Mauro Bernardi