SMDTA 2022 - ATHENS 7-10/6/2022

Stochastic and Statistical Data Analysis Models and Methods

Virtual Session

Chair: Ilia Vonta, National Technical Univ. of Athens

1. **Elisavet Beki**, A. Karagrigoriou and K. Ntotsis
   A Novel Dimensionality Reduction Approach through Partial Least Squares Method
2. **Konstantinos N. Makris** and Ilia Vonta
   An introduction to the indices μ^{MKN} and to the mathematical measure V^{MKN}
3. **Christina Parpoula** and Alex Karagrigoriou
   The key role of change-point analysis in public health decision-making
4. **Milan Stehlik**
   On tissue discrimination for cancer research

Stochastic and Statistical Data Analysis Models and Methods

On-Site Session

Chair: Alex Karagrigoriou, Univ. of the Aegean

1. **Christos Meselidis,** Alex Karagrigoriou and Takis Papaioannou
   Robust Statistical Inference based on Dual Measures
2. Vlad S. Barbu and **Thomas Gkelsinis**
   A class of Homogeneity tests for independent and dependent data with asymmetrically important transitions
3. **Chrysoula Kroustalli**, Alex Karagrigoriou, Andreas Makrides
   Stochastic Processes and Reliability Analysis: Theoretical Issues and Applications
4. **Ioannis Mavrogiannis**, A. Karagrigoriou, G. Papasotiriou, I. Vonta
   Detecting Exponentiality via the catastrophe and conspiracy principles
5. Fatiha Mokhtari, Chafiâa Ayhar, Saâdia Rahmani, **Vlad Stefan Barbu**
   Asymptotic properties of kernel estimators for continuous-time semi-Markov processes

A Novel Dimensionality Reduction Approach through Partial Least Squares Method

Elisavet Beki[1], Kimon Ntotsis[2] and Alex Karagrigoriou[3]

[1] Department of Statistics and Actuarial-Financial Mathematics, School of Sciences, University of the Aegean, Samos, Greece
(E-mail: sasm20006@aegean.gr)
[2] Department of Statistics and Actuarial-Financial Mathematics, School of Sciences, University of the Aegean, Samos, Greece
(E-mail: kntotsis@aegean.gr)
[3] Department of Statistics and Actuarial-Financial Mathematics, School of Sciences, University of the Aegean, Samos, Greece
(E-mail: alex.karagrigoriou@aegean.gr)

High-dimensional data sets give researchers from different fields, the ability to answer various scientific questions. However, their commonly complicated structure and other features, like multicollinearity and noise accumulation, set their handling with classic methods insufficient. New computational and statistical techniques are required to conduct analysis, which can result in reliable conclusions. Such a method is considered to be Partial Least Squares. It is a method with many applications in a wide spectrum of fields, that serves various statistical purposes (such as regression, classification, etc.) and though, till recent years, was unfamiliar to a major part of statisticians. The aim of this master thesis is to investigate the prospects of Partial Least Squares Method in both univariate and multivariate level as a reliable method for the analysis of high-dimensional data. The theoretical basis of this dimensionality reduction method is presented along with Principal Component Analysis, to further compare their performance in linear regression problems.

Keywords: Dimensionality Reduction, Partial Least Squares Regression, Principal Component Regression, Model Selection.

An introduction to the indices $\mu^{[MKN]}$ and to the mathematical measure $V^{[MKN]}$

Konstantinos N. Makris[1] and Ilia Vonta[2]

[1] National Technical University of Athens, School of Applied Mathematical and Physical Sciences, Zografou Campus, 15780 Zografou, Greece (E-mail: constantinosmakris@yahoo.gr)

[2] National Technical University of Athens, School of Applied Mathematical and Physical Sciences, Zografou Campus, 15780 Zografou, Greece (E-mail: vonta@math.ntua.gr)

In this work two methods are presented for the comparison of the time points where the maximum or minimum values of N data sets appear. More specifically, an attempt is made to investigate whether the maximum values of N time series appear at the same or adjacent time points among them or on the contrary, the maximum values of some time series occur at adjacent time points with the corresponding minimum values of other time series. Two methods are considered, firstly mathematical measures are used which are calculated based solely on the time points and secondly, novel indices are defined which are based on the actual data of the time series.

Keywords: Time series, Matrix $V^{[MKN]}$, Index, Matrix $[\mu]^{[MKN]}$.

The key role of change-point analysis in public health decision-making

Christina Parpoula[1] and Alex Karagrigoriou[2]

[1] Department of Psychology, Panteion University of Social and Political Sciences, 17671 Athens, Greece (E-mail: chparpoula@panteion.gr)
[2] Lab of Statistics and Data Analysis, Department of Statistics and Actuarial-Financial Mathematics, University of the Aegean, 83200 Karlovasi, Samos, Greece (E-mail: alex.karagrigoriou@aegean.gr)

Surveillance is a core public health activity that provides information vital for the protection and promotion of health. In particular, surveillance is critical for detecting disease outbreaks rapidly and for guiding interventions to effectively control epidemics. Considerable research has been directed towards early detection of the start of the epidemic, in order to initiate a timely response, and rarely the focus has been given on the whole signal and/or the end of the epidemic. However, identifying the full-time course of an epidemic (i.e., beginning and ending dates of a disease outbreak) is useful for two key reasons. First, detecting the full extent of past outbreaks in surveillance data can improve future outbreak identification. Second, identifying the end of the epidemic helps public health officials decide when response activities can cease and determine whether new cases are part of a known or a new outbreak. Toward this end, this study aims at evaluating the ability of change-point analysis methods to locate the whole disease outbreak signal. Depending on the underlying model used to solve the change-point problem, we compare the performance of some state of the art parametric, nonparametric and Bayesian change-point model approaches with those considered as "gold standard" methods. The empirical and simulation-based results highlight the key role of change-point analysis in outbreak detection. The derived findings support that change-point analysis is a useful analytic tool that can be extensively used to understand disease development, evaluate the design of new strategies of prevention and control of the disease, and thus steer public health decision-making processes.

Keywords: Change-point analysis, Decision-making, Public health, Signal, Outbreak detection.

On tissue discrimination for cancer research

Milan Stehlík

Department of Applied Statistics, Johannes Kepler University Linz, Linz, Austria & Institute of Statistics, Univesidad de Valparaiso, Chile (E-mail: Milan.Stehlik@jku.at)

When we consider fractal-based cancer diagnostic, many times a statistical procedure to assess the fractal dimension is needed. [1] discussed planar tissue preparations in mice which has a remarkably consistent scaling exponents (fractal dimensions) for tumor vasculature even among tumor lines that have quite different vascular densities and growth characteristics. In previous investigations, it has been shown that the texture of mammary tissue, as seen at low magnification, may be characterized quantitatively in terms of stereology (see [3] and references therein). In [4], the images of the mammary cases were reexamined. We will construct a statistical test, which is able to distinguish between the two groups and decide for a possibly new image if it belongs to masthopathic group or not (see [5]). We will address some important inverse problems related to extreme process estimation and scaling. Scaling may lead to a range of p-values and powers, which constitutes an inverse problem. We will discuss these issues in the context of our recent results (see e.g. [2]). During the talk we will discuss several issues which bring light into both fractal-based cancer modelling and general stochastic geometry models.

References

[1] Baish J. W. and Jain R. K. (2000). Fractals and cancer. Cancer Research, 60, 3683-3688.

[2] Filus J., Filus L. and Stehlík M. (2009). Pseudoexponential modelling of cancer diagnostic testing. Biometrie und Medizinische Informatik, Greifswalder Seminarberichte Heft 15, 41-54.

[3] Mattfeldt T. (2003). Classification of binary spatial textures using stochastic geometry, nonlinear deterministic analysis and artificial neural networks. Int. J. Pattern Recogn. Artif. Intel. 17, 275--300.

[4] Mrkvička T. and Mattfeldt, T. (2011). Testing histological images of mammary tissues on compatibility with the boolean model of random sets, Image Analysis and Stereology, 30:11-18.

[5] Stehlík M., Mrkvička T., Filus J. and Filus L. (2012). Recent development on testing in cancer risk: a fractal and stochastic geometry, Journal of Reliability and Statistical Studies, 5, 83-95.

Robust Statistical Inference based on Dual Measures

Christos Meselidis[1], Alex Karagrigoriou[1] and Takis Papaioannou[3]

[1] Lab of Statistics and Data Analysis, Dept. of Statistics and Actuarial-Financial Mathematics, Univ. of the Aegean, Greece (E-mail: Meselidis; alex.karagrigoriou@aegean.gr)

[2] Dept. of Statistics and Insurance Science, University of Piraeus, Greece (E-mail: tak-pap@unipi.gr)

Divergence measures have vast applicability in statistical inference. Not only can they be used for estimation purposes but also for the construction of tests of fit. In this work the focus is placed on contaminated data, thus in order to derive robust procedures for estimation and testing we exploit the (Φ, α)-power divergence family [1] for multinomial populations, which is a general class of measures. In particular, for estimating the unknown parameter we propose the minimum (Φ, α)-power divergence estimator which afterwards is used in the new dual divergence-test statistic [2], for the problem of goodness-of-fit, that involves two indices α_1 and α_2. The values of these indices play an important role in the robustness of the estimator and the stability of the test statistic in the case of contaminated data. All the aforementioned notions are presented through an extensive simulation study where the level of closeness of the contamination distribution to the postulated one varies in a wide range. The proposed estimator and test statistic are compared with the classical ones that can be derived through the Cressie and Read family of divergence measures [3]. The comparison is based on the mean square error of the estimator, whereas regarding the test statistic the size and the power of the test are taken into account.

Keywords: Divergence Measures, Robustness, Multivariate Data, Goodness-of-fit Tests.

References:

1. K. Mattheou and A. Karagrigoriou. A new family of divergence measures for tests of fit. *Aust. & New Zeal. J. of Stat.*, 52, 187-200, 2010.

2. C. Meselidis and A. Karagrigoriou. Statistical inference for multinomial populations based on a double index family of test statistics. *J. of Stat. Computation and Simulation*, 90(10): 1773-1792, 2020.

3. N. Cressie and T. R. C. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society*, 5: 440-454, 1984.

A class of Homogeneity tests for independent and dependent data with asymmetrically important transitions

Vlad S. Barbu[1] and Thomas Gkelsinis[2]

[1] Laboratory of Mathematics Raphaël Salem, University of Rouen-Normandy, France (E-mail: barbu@univ-rouen.fr)
[2] Laboratory of Mathematics Raphaël Salem, University of Rouen-Normandy, France (E-mail: thomas.gkelsinis@univ-rouen.fr)

In this paper we present a class of Homogeneity tests for independent data and its extension to Markov Chains of general order (dependent case) with transitions of different importance. The underlying mechanism of these tests is based on the family of weighted φ-divergences which is a generalization of the usual φ-divergence. The appropriate asymptotic theory is presented according with useful simulations for the power and size of the proposed tests.

Keywords: Homogeneity, Markov Chains, Divergence Measures, CWKL divergence

Stochastic Processes and Reliability Analysis: Theoretical Issues and Applications

Chrysoula Kroustalli, Alex Karagrigoriou and Andreas Makrides

Department of Statistics and Actuarial-Financial Mathematics, School of Sciences, University of the Aegean, Samos, Greece
(E-mail: sasm20018; alex.karagrigoriou; amakridis@aegean.gr)

In Reliability Analysis the main focus is on stochastic processes and in particular Semi-Markov processes since they allow for general lifetime distributions. In this work, we aim our attention at modified distributions, especially on Modified Weibull Poisson, for which we provide their properties as well as the regions for which we can approximate them by simpler distributions that are closed under minimum and we establish the expressions for parameter estimation. Simulation results show the performance of the proposed estimators.

Keywords: Multi-state system, semi-Markov processes, H-class of distributions, Modified Weibull distribution, Gompertz distribution, parameter estimation.

.

Detecting Exponentiality via the catastrophe & conspiracy principles

I. Mavrogiannis[1], A. Karagrigoriou[1], G. Papasotiriou[2] and I. Vonta[2]

[1] Lab of Statistics and Data Analysis, Dept. of Statistics and Actuarial-Financial Mathematics, Univ. of the Aegean, Greece (E-mail: sasm20029@sas.aegean.gr; alex.karagrigoriou@aegean.gr)

[2] National Technical University of Athens, School of Applied Mathematical and Physical Sciences, Zografou Campus, Greece (E-mail: georgiapap22@yahoo.gr; vonta@math.ntua.gr)

This work is filling up the gap in the literature regarding the verification of the log-concavity property which is a widely studied topic due to the fact that it provides desirable estimating properties. At the same time, log-concavity together with log-convexity are vital in reliability, engineering and stochastic modeling for distinguishing between an exponential, a light-tailed and a heavy tailed distribution. In this work we propose an exponentiality test of fit to be used for distinguishing between exponential and log-convex or long-concave distributions. The proposed test statistic is based on the conspiracy and catastrophe principles through which a characterization for the (tail part of the) exponential distribution is established. The details of the formulation of the test are provided, an extended simulated study which shows the performance of the proposed test statistic is given, and some concluding remarks are stated.

Keywords: Exponentiality test · characterization · log-concavity · log-convexity· goodness of fit test · catastrophe principle · conspiracy principle.

Asymptotic properties of kernel estimators for continuous-time semi-Markov processes

Fatiha Mokhtari[1], Chafiâa Ayhar[1], Saâdia Rahmani[1] & Vlad S. Barbu[2]

[1]Laboratory of Stochastic Models, Statistics and Applications, University of Saida–Doctor Moulay Taher, Saïda, Algeria

[2]Laboratory of Mathematics Raphaël Salem, University of Rouen-Normandy, France (E-mail: barbu@univ-rouen.fr)

In this paper we present asymptotic properties of kernel estimators for continuous-time semi-Markov processes.

Keywords: Kernel estimator, semi-Markov process.