82nd International Scientific Conference of the University of Latvia 2024

Mathematical Statistics and Data Science

Book of Abstracts

February 29, 2024

Department of Mathematics
University of Latvia

Contents

Organizational Network Analysis and Better People Analytics Using Machine Learning Analysis

ĒRGLIS Aldis

Faculty of Business, Management and Economics
University of Latvia
Aspazijas boulevard 5, Riga, LV-1050, Latvia
E-mail: aldis.erglis@lu.lv

Quantitative analysis of the organisation's characteristics is increasingly demanding in today's environment. The research direction that provides solutions in this field is ONA (Organization Network analysis). One of the areas is employee characteristics and interactions that contribute to achieving specific organizational objectives. For example, calculations that can determine the impact of communication intensity on customer service quality. Or quantitative indicators showing a certain level of creativity within an organization. This presentation will look at mathematical and machine learning methods for graph analysis, which allow the calculation of personnel characteristics based on communication fact data. We'll look at calculating and interpreting multiple metrics on data from a real organization. We will look at metrics like influencer (impact person) and ideation (person with potentially new ideas). Although the calculations are based on well-known graph algorithms, using social and management sciences techniques is essential in interpreting the results.

Advancing Markowitz: Asset Allocation Forest

TETEREVA Anastasija

Erasmus School of Economics
Erasmus University Rotterdam
Oudlaan 50 3062 PA, Rotterdam, Netherlands
E-mail: tetereva@ese.eur.nl

We propose a novel Asset Allocation Forest (AAF) model that combines the well-established machine learning tool with the conventional portfolio optimization method. The determination of locally optimal portfolio weights, which dynamically respond to market conditions, effectively captures market regimes, structural breaks and smooth transitions in a data-driven manner. The introduced model consistently outperforms established benchmarks, including the Hidden Markov Model (HMM), for a multi-asset portfolio composed of equities, bonds, credit, high yield, and commodities, even when trading costs are taken into account. By constructing accumulated local effects (ALE) plots, we find evidence of "flight-to-safety", indicating a strategic shift from riskier assets to less volatile bonds during periods of increased market turbulence. Furthermore, our model shows a pronounced preference for bonds in inflationary periods, demonstrating its adaptability to different economic conditions.

Change-Point Detection for Dependent Data

AŅISKEVIČA Svetlana[1], VALEINIS Jānis[2] and ALKSNIS Reinis[2]

[1] *Latvian Environment, Geology and Meteorology Centre*
Maskavas street 165, Riga, LV-1019, Latvia
E-mail: svetlana.aniskevica@lvgmc.lv

[2] *Laboratory of Statistical Research and Data Analysis*
University of Latvia
Jelgavas street 3, Riga, LV-1004, Latvia
E-mail: janis.valeinis@lu.lv, reinis.alksnis@lu.lv

The task of change-point analysis is to identify abrupt shifts or anomalies in datasets, aiding in quality control, segmentation, or event detection across various fields. By locating a moment when statistical properties of a dataset change, change-point detection plays a crucial role in understanding data properties, identifying critical events, and making decisions based on this information, e.g., if a change is caused by an error, then it is possible to correct the dataset, thus improving the results of a further statistical analysis.

In this study, authors deal with detecting shifts in mean values for weakly dependent data. For the change-point detection we establish and propose to use the two-sample blockwise empirical likelihood for the difference of two-sample means. We propose not only to obtain a p-value from the testing procedure, but also to draw a p-value curve for a visual check of the statistical significance of a change-point, as well as to determine the location of the change-point. In various simulation studies, we compare the proposed algorithm with CUSUM, Hodges–Lehmann and Wilcoxon–Mann–Whitney tests, and additionally using the historical wind speed observations in Latvia, we demonstrate the applicability of the proposed method to the real datasets.

Optimizing ROI in Machine Learning: A Framework for Feature Cost Analysis

KRĒGERS Rūdolfs

Department of Mathematics, Faculty of Physics, Mathematics and Optometry
University of Latvia
Jelgavas street 3, Riga, LV-1004, Latvia
E-mail: rudolfs.kregers@gmail.com

With the expansion of big data technologies and the use of new data sources for business decision automation, it becomes crucial to evaluate how additional data can affect profits, considering the cost of data acquisition against existing information. This study presents a methodology for assessing the monetary value of data sources and individual predictors in supervised machine learning models to maximize return on investment (ROI). The methodology is founded on cross-validation and comparing returns using two datasets: one with and the other without the additional data. It provides a a guide on balancing data investments with expected profit increases, optimizing business decision-making.

Chlorophyll-a Modelling Using Generalized Additive Models in Freshwater Lakes

LŪSIS Mārtiņš

Laboratory of Statistical Research and Data Analysis
University of Latvia
Jelgavas street 3, Riga, LV-1004, Latvia
E-mail: martins.lusis@lu.lv

An important topic in water resource management is eutrophication, or an increase in the growth of algae due to an increase in phosphorus and nitrogen. Chlorophyll-a serves as a proxy for algal biomass, which is used as an indicator of overall lake health.

In this study, we model chlorophyll-a in Finnish lakes using generalized additive models (GAMs) and cluster analysis methods. Several GAM models are obtained corresponding to the grouping of lakes into clusters and are then compared to the appropriate multiple linear regression models.

A New Method for Deriving Confidence Intervals for Location and Scale Parameters

VALEINIS Jānis[1], PAHIRKO Leonora[1] and PRAVDINA Ksenia[2]

[1] *Laboratory of Statistical Research and Data Analysis*
University of Latvia
Jelgavas street 3, Riga, LV-1004, Latvia
E-mail: janis.valeinis@lu.lv, leonora.pahirko@lu.lv

[2] *Department of Mathematics, Faculty of Physics, Mathematics and Optometry*
University of Latvia
Jelgavas street 3, Riga, LV-1004, Latvia
E-mail: ksenia.pravdina15@gmail.com

Two-sample location-scale models characterize the functional relationship between two distribution functions. The simple location model describes, for example, the difference of means between control and treatment groups in medical research. Meanwhile, scale models can be found in survival analysis where the proportional hazards models are used.

To test location-scale models, one option is to conduct a formal hypothesis test. Equivalently, confidence intervals or bands for location and scale parameters can be constructed. If the constant function falls within the bands, then the null hypothesis is not rejected. In this work, a new method for constructing confidence intervals is proposed [1], which utilizes the inversion of various tests based on the empirical likelihood function.

References:

[1] Pahirko, L. and Valeinis, J. *Empirical Likelihood-Based Confidence Regions for the Structural Relationship Parameter.* In Proceedings of the 2023 6th International Conference on Mathematics and Statistics, 38-47, 2023.

Various Blocking Schemes for Blockwise Empirical Likelihood Method

ALKSNIS Reinis

Laboratory of Statistical Research and Data Analysis
University of Latvia
Jelgavas street 3, Riga, LV-1004, Latvia
E-mail: reinis.alksnis@lu.lv

In literature, there are two main approaches of applying empirical likelihood (EL) method to dependent observations. The first one is a spectral approach introduced by Monti [1], where Whittle's likelihood was utilized to obtain M-estimator of asymptotically independent periodogram ordinates. The other one is blockwise empirical likelihood (BEL) approach introduced by Kitamura [2], who, inspired from bootstrap literature, defined blockwise method which applies the usual independent data EL to averages of blocks of data.

Due to simplicity and wider applicability, the second approach has gained more attention in literature, and since its introduction, the theory of BEL method has been applied for a wide range of statistical problems. However, while some research has suggested that weighted moving averages could provide greater flexibility compared to simple averages, limited attention has been given to analyzing the impact of various blocking schemes. Hence, in this study, various schemes are investigated and through simulation analysis compared to the classical BEL method.

References:

[1] Monti, A.C. *Empirical likelihood confidence regions in time series models.* Biometrika, 84(**2**), 395-405, 1997.

[2] Kitamura, Y. *Empirical likelihood methods with weakly dependent processes.* The Annals of Statistics, 25(**5**), 2084-2102, 1997.

A New Coefficient of Correlation: Scientific Challenges

KOZIREVS Filips

Department of Mathematics, Faculty of Physics, Mathematics and Optometry
University of Latvia
Jelgavas street 3, Riga, LV-1004, Latvia
E-mail: fk17001@students.lu.lv

The most popular measures of statistical association are Pearson's, Spearman's and Kendall's correlation coefficients which can be effectively applied to identify linear or monotonic associations. Nevertheless, these coefficients are not effective when it is necessary to identify non-monotonic relationships. This problem has been addressed, for example, by defining the maximal correlation coefficient and information theoretic coefficients as well as by applying kernel-based methods. However, firstly, all of these approaches are meant for testing independence rather than measuring the strength of association and, secondly, these coefficients do not have simple asymptotic theories under the hypothesis of independence [1, 2, 3, 4].

Taking this into account, Sourav Chatterjee defines a new correlation coefficient in his paper [1] which is as simple as the classical coefficients and has a simple asymptotic theory under the hypothesis of variable independence. Furthermore, this correlation coefficient is a consistent estimator of a dependence measure which is equal to zero if and only if the variables are independent and is equal to one if and only if one of the variables is a measurable function of another one. The measure of association defined by Chatterjee has many advantages but it has also drawbacks which are discussed, for instance, in papers [5, 6].

In this work, the main properties of the new correlation coefficient are described and the current problems related to its practical application are summarized.

References:

[1] Chatterjee, S. *A new coefficient of correlation.* Journal of the American Statistical Association, 116(**536**), 2009-2022, 2021.

[2] Gebelein, H. *Das statistische Problem der Korrelation als Variations-und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung.* ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik, 21(**6**), 364-379, 1941.

[3] Linfoot, E.H. *An informational measure of correlation.* Information and control, 1(**1**), 85-89, 1957.

[4] Gretton, A., Fukumizu, K., Teo, C.H., Song, L., Scholkopf, B. and Smola, A.J. *A kernel statistical test of independence.* Advances in Neural Information Processing Systems, 20, 585-592, 2008.

[5] Shi, H., Drton, M., and Han, F. *On the power of Chatterjee's rank correlation.* Biometrika, 109(**2**), 317-333, 2022.

[6] Lin, Z., and Han, F. *On the failure of the bootstrap for Chatterjee's rank correlation.* arXiv preprint arXiv:2303.14088, 2023.

The Advantages of the Smoothly Trimmed Mean

SILIŅŠ Emīls

Laboratory of Statistical Research and Data Analysis
University of Latvia
Jelgavas street 3, Riga, LV-1004, Latvia
E-mail: es18098@students.lu.lv

The classical trimmed mean is a robust statistic, which is used primarily in cases of contaminated data, where the sample mean would fail. Let $X_1, ..., X_n$ be *iid* random variables with a common distribution F and let $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ denote ordered random variables. Then the α-trimmed mean is defined as

$$\overline{X}_\alpha = \frac{1}{n-2r} \sum_{i=r+1}^{n-r} X_{(i)},$$

where $0 \leq \alpha < 0.5$ is the trimming portion and $r = [n\alpha]$.

In 1973, Stigler [1] showed that the asymptotic normality of the trimmed mean fails, when trimming proportions are chosen wrong, and proposed a new method, where the data is trimmed and at the same time smoothed using a weight function. The smoothly trimmed mean is defined as

$$\overline{X}_{ST} = \frac{1}{n} \sum_{i=1}^{n} J\left(\frac{i}{n+1}\right) X_{(i)},$$

where $J(\cdot)$ is an appropriate weight function. Such approach ensures that the asymptotic normality does not fail in case of wrong trimming.

In this work, we analyze a more general version of a smoothly trimmed mean. We compare it with the classical trimmed mean and use also the empirical likelihood method for the inference. We compare all methods by simulation study using the empirical coverage accuracy.

References:

[1] Stigler, S.M. *The asymptotic distribution of the trimmed mean.* The Annals of Statistics, 472-477, 1973.

Choosing a Method for the Problem of Two-Sample Comparison

GREDZENS Jānis

Laboratory of Statistical Research and Data Analysis
University of Latvia
Jelgavas street 3, Riga, LV-1004, Latvia
E-mail: janis.gredzens@lu.lv

Two-sample comparison is an important problem in practice, but the choice of method for comparing different distributions and unequal variances is not straightforward. The t-test is designed for testing the hypothesis of equality of means, while the Wilcoxon test is often used in practice as a non-parametric alternative to the t-test for "comparing medians", which is not a correct interpretation of the test. Alternatively, methods such as the Kolmogorov-Smirnov test, quantile regression, non-parametric bootstrap or empirical likelihood can be used to compare distributions of two samples, medians and other quantiles.

In this study, the results of the t-test, the Wilcoxon test and the empirical likelihood method from the R package *EL* [1] are presented for comparing different distributions. The performance of each method was evaluated by running simulations for various distributions and sample sizes.

The t-test performed well, as expected, for comparing the means of normally distributed data, but was not suitable for comparing skewed distributions. The Wilcoxon test showed stable results in different scenarios for skewed distributions and unequal sample sizes, but the study considers the case where the test rejects the null hypothesis even though the sample medians are equal. The empirical likelihood method performed well when analysing asymmetric data and when inferring the equality of different quantiles. The results highlight the need to make a conscious choice of the statistic to use when comparing two samples, taking into account the distribution and sample size, as well as the hypothesis of the research. The analysis of this study highlights the importance of matching the choice of statistical method to specific data conditions in order to increase the validity of empirical research results.

References:

[1] Valeinis, J., Cers, E. and Gredzens, J. *EL: Two-Sample Empirical Likelihood*, R package version 1.2, 2022. `https://CRAN.R-project.org/package=EL`