83rd International Scientific Conference of the University of Latvia 2025

Mathematical Statistics and Data Science with Applications

Book of Abstracts

April 11, 2025

Department of Mathematics
University of Latvia

Contents

This session is organized within the framework of the project titled "Digitālie rīki uzņēmējdarbības izglītībā: ietekme uz uzņēmējdarbības nodomu un attīstības finansējuma piesaisti", grant number LU-BA-PA-2024/1-0068.

The research is funded by the Latvian Recovery and Resilience Facility Investment-supported project "Internal and External Consolidation of the University of Latvia"
(No. 5.2.1.1.i.0/2/24/I/CFLA/007).

Data Analytics-Based Digitals Tools and AI in Business Education

AKSJOŅENKO Airita, MAVĻUTOVA Inese, and LEŠINSKIS Kristaps

BA School of Business and Finance
K. Valdemara street 161, Riga, LV-1013, Latvia
E-mail: Airita.Aksjonenko@ba.lv, Inese.Mavlutova@ba.lv, Kristaps.Lesinskis@ba.lv

The integration of data analytics-based digital tools and artificial intelligence (AI) in entrepreneurial education is transforming how entrepreneurial skills and competencies are developed. As entrepreneurship becomes increasingly data-driven, educational tools that enhance decision-making, opportunity recognition, and strategic planning are essential. AI-powered business planning tools, virtual simulations, and online collaboration platforms provide structured learning environments, real-time insights, and interactive experiences that improve entrepreneurial education outcomes.

This report examines the role of digital tools in entrepreneurial learning and categorizes 14 business planning platforms into three groups: (1) Non-AI tools, which rely on manual input and predefined templates; (2) Hybrid tools, which combine AI automation with user-driven customization; and (3) AI-powered tools, which use natural language processing (NLP), machine learning, and predictive analytics to generate business plans and market insights. AI-driven solutions enhance structured business planning, improving entrepreneurial intention and decision-making efficiency. Hybrid models allow users to refine AI-generated content while maintaining control over their strategies.

Beyond business planning, AI and digital tools reshape entrepreneurial education through experiential learning. Virtual simulations, AI chatbots, and MOOCs provide immersive, personalized learning experiences, fostering engagement, adaptability, and data-driven decision-making. These technologies also expand access to entrepreneurship education, making it more inclusive and interactive.

Additionally, this report highlights the importance of digital literacy in effectively utilizing AI-driven tools. Blended learning models, integrating digital platforms with traditional education, are crucial for fostering an entrepreneurial mindset that is innovative and adaptable to technological advancements.

Application of Structural Equation Modeling to Analysis of Entrepreneurial Intention

VILUNAS Ēriks[1], VALEINIS Jānis[1], and HERMANIS Jānis[2]

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

[2] *BA School of Business and Finance*
K. Valdemara street 161, Riga, LV-1013, Latvia
E-mail: Janis.Hermanis@ba.lv

Entrepreneurial intention and factors affecting it is a popular research topic due to the benefits of entrepreneurship on economics and society. One of factors that is being researched is entrepreneurship education, however there is not a clear consensus about effect of entrepreneurship education on entrepreneurial intention. There is a discussion about whether the effects are direct or indirect and also other factors, such as demographic variables, region, must be considered.

Structural equation modeling (SEM) is a data analysis method which is widely used in social and behavioral science research in order to test whether research data support certain theoretical models. Studied models involve observed and latent variables and relations among them. In this work, structural equation modeling is applied to entrepreneurial intention research data and relation between entrepreneurship education and entrepreneurial intention is studied.

Mediation Analysis of Glycemic Variability and Systemic Inflammation in Type 1 Diabetes Patients

PAHIRKO Leonora[1], FEDULOVS Aleksejs[2], PARAMONOVA Natālija[2], and SOKOLOVSKA Jeļizaveta[2]

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

[2] *Faculty of Medicine and Life Sciences*
University of Latvia
Jelgavas street 3, Riga, LV-1004, Latvia
E-mail: aleksejs.fedulovs@lu.lv, natalia.paramonova@lu.lv, jelizaveta.sokolovska@lu.lv

This study presents a data-driven stratification framework for continuous glucose monitoring (CGM) metrics in individuals with type 1 diabetes (T1D), aiming to uncover latent structures in glycemic profiles and assess their statistical associations with systemic inflammation and liver dysfunction.

We applied unsupervised hierarchical agglomerative clustering using Ward's minimum variance criterion and Euclidean distance on a set of six CGM-derived features (mean glucose, coefficient of variation, time in/above/below range, and frequency of hypoglycemic events). Feature standardization and principal component analysis (PCA) were performed as preprocessing steps. The optimal number of clusters was determined via consensus across 30 internal validation indices implemented in the NbClust R package and corroborated by the elbow method.

To evaluate causal pathways from glycemic control to inflammation via liver dysfunction indices (Fatty Liver Index — FLI; Hepatic Steatosis Index — HSI), we conducted causal mediation analysis using two nested linear models, standardizing all continuous variables prior to analysis. The sensitivity analysis was based on nonparametric bootstrap (5,000 iterations),and implemented using the R package `mediation`.

Clustering revealed two statistically distinct glycemic patterns corresponding to poorly and moderately controlled diabetes. These latent classes were significantly associated with varying levels of insulin resistance, hepatic biomarkers, and inflammatory markers. Mediation results suggested a partially mediated effect of glycemic dysregulation on inflammation through steatotic liver indices, indicating potential mechanistic pathways.

This work demonstrates how unsupervised learning and causal inference methods can be combined to extract interpretable structures from high-dimensional biomedical data, with implications for personalized risk stratification in chronic disease.

Statistical and Machine Learning Methods for Analyzing the Progression of Diabetic Retinopathy

VALEINIS Jānis

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

Diabetic retinopathy (DR) is a leading cause of visual impairment worldwide. Duration of diabetes is a major risk factor associated with the development and progression of diabetic retinopathy. As a by-product of a recent project, a large amount of clinical and fundus imaging data has been obtained from type I and type II diabetic patients. The general aim is to evaluate risk factors and retinal changes in diabetic patients using multimodal data and to extract different features from fundus image data.

Artificial intelligence (AI) is playing an increasingly important role in medicine. There has been a lot of research into predicting the time to progression of DR from fundus images alone, even without direct knowledge of laboratory or clinical data. Deep learning with convolutional neural networks has been developed for automated detection of DR from retinal photographs. In addition, optical coherence tomography angiography (OCT-A) has become an important technique for imaging the retinal and choroidal microvasculature.

Many statistical and machine learning tools have been developed to analyse a significant range of available demographic and clinical parameters, including age, sex, systolic blood pressure, glycaemic control (HbA1c), lipid levels, smoking status and cardiovascular risk. Some risk calculators have been developed separately for clinical data and fundus image data.

Our aim is to explore and find new mathematical, statistical and machine learning methods and to develop new risk calculators based on multimodal data, combining clinical and imaging data to predict DR progression.

Time Series Clustering, Pattern Recognition and Forecast with Data-Driven Methods

VEDENNIKOVA Ekaterina

Faculty of Science and Technology
University of Latvia
Jelgavas street 3, Riga, LV-1004, Latvia
E-mail: ekaterinamarsel@gmail.com

When analyzing real-world time series data, the primary focus is often not on short-term forecasting but on gaining a qualitative understanding of the underlying processes that generate these time series. This means that the problems of clustering and pattern recognition come forth. In this project, we aim to perform a qualitative analysis of a large dataset comprising tens of thousands of real-world time series, each representing individual player activity on various iGaming platforms. Such a vast amount of data introduces significant data processing and computational challenges.

Our research has two main objectives:

1. *Clustering time series based on similarity in patterns, trends, motifs, and shifts.* To achieve this, we evaluate various clustering methods, highlight their strengths and limitations, and demonstrate how customized strategies can overcome their shortcomings.

2. *Forecasting future behavior of these time series over extended horizons.* Unlike traditional short-term prediction methods (e.g., ARIMA), our goal is to identify broader trends and their extent (e.g., anticipating upward shifts and their potential rates). To this end, we explore a number of data-driven methods, including Dynamic Time Warping (DTW) [1], Symbolic Aggregate Approximation (SAX) [2], Hidden Markov Models (HMM) [3], Singular Spectrum Analysis (SSA) [4], and Partition Around Medoids (PAM) and discuss their applicability.

Overall, we found out that a well-designed combination of multiple approaches can yield good results even for relatively short time series.

References:

[1] Alexis, S.E. Time-Series Clustering in R Using the dtwclust Package. *The R Journal*, 11(**1**), 22–43, 2019.

[2] Feasel, K. *Symbolic Aggregate Approximation (SAX).* In *Finding Ghosts in Your Data: Anomaly Detection Techniques with Examples in Python*, 293–309, Berkeley, CA:Apress, 2022.

[3] Rabiner, L.R., and Juang, B.H. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(**1**), 4–15, 1986.

[4] Golyandina, N., Nekrutkin, V., and Zhigljavsky, A.A. *Analysis of Time Series Structure: SSA and Related Techniques.* 1st edition, Chapman and Hall/CRC, 2001.

Bartlett Correction for the Empirical Likelihood Method in the Case of Dependent Observations

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

Empirical Likelihood (EL) has proved to be a powerful nonparametric inference method, combining the robustness of likelihood methods with minimal distributional assumptions [1]. A desirable feature of EL is its Bartlett correctability, which significantly improves the coverage accuracy of confidence regions by reducing coverage errors from $O(n^{-1})$ to $O(n^{-2})$ in the independent data setting [2]. While well understood in i.i.d. cases, the Bartlett correction's applicability under dependence is less studied, although some research in this direction has been done [3],[4],[5],[6].

In this work, a comprehensive review is presented of developments on Bartlett correction for EL in dependent data context. The validity of Edgeworth expansions and higher-order cumulants under short-memory and long-memory Gaussian and non-Gaussian settings is examined, based on both blockwise EL and frequency-domain EL approaches.

Despite these advances, the literature has largely focused on one-sample or parametric estimation problems. This work proposes to fill the gap by investigating Bartlett corrections for EL in two-sample inference problems under dependence. The unique challenges posed by dependence within samples is discussed.

References:

[1] Owen, A.B. *Empirical likelihood.* Chapman and Hall/CRC, 2001.

[2] DiCiccio, T., Hall, P., and Romano, J. Empirical likelihood is Bartlett-correctable. *The Annals of Statistics*, 19(**2**), 1053–1061, 1991.

[3] Chan, N.H., and Liu, L. Bartlett correctability of empirical likelihood for time series. *Journal of the Japan Statistical Society*, 40(**2**), 221–238, 2010.

[4] Chan, N.H., Chen, K., and Yau, C.Y. On the Bartlett correction of empirical likelihood for Gaussian long-memory time series. *Electronic Journal of Statistics*, 8(**1**), 1460-1490, 2014.

[5] Chen, K., Chan, N.H., and Yau, C.Y. Bartlett correction of empirical likelihood for non-Gaussian short-memory time series. *Journal of Time Series Analysis*, 37(**5**), 624–649, 2016.

[6] Chen, K., Chan, N.H., and Yau, C.Y. Bartlett correction of frequency domain empirical likelihood for time series with unknown innovation variance. *Annals of the Institute of Statistical Mathematics*, 74(**5**), 1159–1173, 2020.

U-Statistics and Their Properties

SILIŅŠ Emīls

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

The U-statistics are one of the most universal objects in modern probability theory of summation and date back to 1948, when they were introduced by Hoeffding [1]. Let $X_1, ...X_n$ be independent random variables with the same distribution P. For $n \geq m$, a U-statistic is defined as follows

$$U_n = \binom{n}{m}^{-1} \sum_{1 \leq i_1 < ... < i_m \leq n} \Phi(X_{i_1}, ..., X_{i_m}), \tag{1}$$

where $\Phi : X^m \to \mathbb{R}$ is a symmetric function of m variables and m is considered the rank of the U_n statistic. Hoeffding showed that U_n is asymptotically normal.

Depending on how we define the $\Phi(\cdot)$ function we can obtain other familiar statistics, for example if $\Phi(x) = x$ then $U_n = \overline{X}$ and if $\Phi(x_1, x_2) = 1/2(x_1^2 + x_2^2 - 2x_1x_2)$ then $U_n = s^2$.

In [3] it is shown how the jackknife pseudo-values can be linked to U-statistics and in [4] the empirical likelihood (EL) method based on jackknife pseudo-values is developed.

In this paper we recount some important results in U-statistic theory from [2] and perform a simulation study comparing the jackknife EL method on U-statistics to other similar methods.

References:

[1] Hoeffding, W. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(**3**), 293-325, 1948.

[2] Korolyuk, V.S., and Borovskich, Y.V. *Theory of U-Statistics.* Springer Science & Business Media, 2013.

[3] Arvesen, J.N. Jackknifing U-statistics. *The Annals of Mathematical Statistics*, 40(**6**), 2076-2100, 1969.

[4] Jing, B.Y., Junqing Y., and Wang Z. Jackknife empirical likelihood. *Journal of the American Statistical Association*, 104(**487**), 1224-1232, 2009.

Empirical Likelihood Residual Bootstrap Method

LUKASHEVYCH Sofiia

UL Faculty of Science and Technology
University of Latvia
Jelgavas street 3, Riga, LV-1004, Latvia
E-mail: sofiia.lukashevych@lu.lv

Empirical likelihood (EL) for linear models was presented by A. B. Owen as an extension to the original empirical likelihood method [1]. B. W. Brown and W. K. Newey later introduced an EL-based bootstrap approach, broadening the scope of empirical likelihood applications [2]. Separately, the concept of the residual bootstrapping was introduced by D. A. Freedman in the context of regression models, providing one of alternative versions of the classical bootstrap method [3].

In this work we consider the following one-factor linear fixed effects model with two treatments:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \tag{2}$$

where Y_{ij} is the response corresponding to the jth subject under the ith treatment, $i = 1, 2$ and $j = 1, \ldots, n_i$. Here, n_i is the size of the ith treatment group, and $n = n_1 + n_2$ is the total number of subjects. The parameter μ is the overall mean, and τ_i is the effect of the ith treatment. We further assume that $\{\varepsilon_{ij} : i, j\}$ are i.i.d. with $\mathbb{E}(\varepsilon_{ij}) = 0$ and $\mathrm{Var}(\varepsilon_{ij}) = \sigma_i^2$.

For identifiability and interpretability reasons, assume that $\tau_1 + \tau_2 = 0$. Consequently, τ_i can be interpreted as the difference between the ith treatment mean and the overall mean, i.e. $\tau_i = \mu_i - \mu$, where $\mu_i = \mathbb{E}(Y_{ij})$.

We propose an *empirical likelihood residual bootstrap* (ELRB) method in this setting. Then its performance is compared against the standard bootstrap, residual bootstrap, and ANOVA methods in a simulation study.

References:

[1] Owen, A.B. Empirical Likelihood for Linear Models. *The Annals of Statistics*, 19(**4**), 1725-1747, 1991.

[2] Brown, B.W., and Newey, W.K. Generalized Method of Moments, Efficient Bootstrapping, and Improved Inference. *Journal of Business & Economic Statistics*, 20(**4**), 507–517, 2002.

[3] Freedman, D.A. Bootstrapping Regression Models. *The Annals of Statistics*, 9(**6**), 1218–1228, 1981.

Plug-in Empirical Likelihood Method with Applications to Right-Censored Data

JĒKABSONS Deivids

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

In fields such as medicine and engineering an important measurement is lifespan. Examples may include the amount of time a patient has survived since the start of a study, or how long it takes for a certain machine part to fail. As such, survival data takes on the unique form of a time measurement and a censoring indicator pair (Z, δ), which is defined as

$$Z = \min(X, Y), \quad \delta = I(X \leq Y),$$

where X is the survival time and Y is the censoring time.

Since 1975 ([1]), the empirical likelihood (EL) method has been used for applications with survival data to estimate survival probabilities. With EL, the distribution of the data is not assumed, which provides a non-parametric statistic estimation procedure, and in most cases the likelihood ratio statistic converges to a chi-square distribution. However, with the introduction of a plug-in estimator (i.e., a Kaplan-Meier estimator for the survival function of the data) the statistic distribution tends to get disrupted.

In this work, we explore three different plug-in EL approaches to estimating the survival probabilities and residual mean survival times of right censored two-sample survival data. They are compared with each other and against standard approximation methods via simulated coverage accuracy, and a real data example is presented.

References:

[1] Thomas, D.R., and Grunkemeier, L.G. Confidence Interval Estimation of Survival Probabilities for Censored Data. *Journal of the American Statistical Association*, 70(**352**), 865—71, 1975.

Smoothed and Bartlett-Corrected Empirical Likelihood for Contaminated Samples

AKOPJANA Kristīne

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

Empirical likelihood is a robust method, used for different aspects of hypothesis testing, especially for asymmetrical distributions and contaminated samples. Method is appropriate for both mean and quantile cases. The coverage accuracy of this method has been proven to be improved by smoothing techniques and Bartlett correction [1]. The aim of this simulation study is to test whether empirical likelihood can show better coverage accuracy than alternative methods such as Wilcoxon's test, percentile bootstrap, Harrell-Davis quantile estimator and other methods.

In this work coverage accuracy of empirical likelihood, smoothed empirical likelihood and Bartlett-corrected empirical likelihood is analyzed and compared to other methods for small, contaminated samples. Symmetrical and asymmetrical contamination is investigated. Emphasis is put on one sample quantile case. Method's performance at different bandwidth parameters is examined. The methods are compared by simulation study using empirical coverage accuracy.

References:

[1] Chen, S.X., and Hall, P. Smoothed empirical likelihood confidence intervals for quantiles. *The Annals of Statistics*, 21(**3**), 1166-1181, 1993.