11th TARTU CONFERENCE ON MULTIVARIATE STATISTICS

25-28 June 2024

Programme and Abstracts

Contents

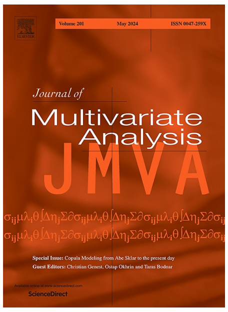
Welcome!

Welcome to the 11th Tartu Conference on Multivariate Statistics

Local organizing committee

Krista Fischer (Chair)
Raul Kangro
Anastassia Kolde
Meelis Käärik
Jüri Lember
Merli Mändul
Märt Möls
Kalev Pärna
Anne Selart
Imbi Traat
Artur Tuttar
Mare Vähi

With support


Volume 201
May 2024
ISSN 0047-259X
Journal of
Multivariate
Analysis
JMVA
Special Issue: Copula Modeling from Abe Sklar to the present day
Guest Editors: Christian Genest, Ostap Okhrin and Taras Bodnar
Available online at www.sciencedirect.com
ScienceDirect

Programme committee

Dietrich von Rosen (Chair, Sweden)
Tõnu Kollo (Co-Chair, Estonia)
Katarzyna Filipiak (Poland)
Solomon Harrar (USA)
Tapio Nummi (Finland)
Jianxin Pan (China)

Timetable

Tuesday, June 25th

KL: Keynote Lecture
IT: Invited Talk
IS: Invited Session
CS: Contributed Session

13:00	Registration is open (at info desk)		
13:45–14:00	Opening		
14:00–15:45	**Plenary Session 1,** Room 1037, Chair Dietrich von Rosen		
14:00–15:00	KL	**Simo Puntanen** *Tampere University*	C.R. Rao's Century from Tampere's perspective
15:00–15:45	IT	**Katarzyna Filipiak** *Poznań University of Technology*	Safety belt estimation under the multivariate linear model
15:45–16:15	Coffee break		
16:15–18:00	IS1	Org: **Peter Song** *University of Michigan*	New analytics for complex correlated data
18:00–19:30	Walking tour in Tartu city centre		
19:30–21:00	Welcome reception at Delta		

Invited Session 1: New analytics for complex correlated data

Tuesday 16:15–18:00, Room 1020

Organizer: Peter Song	*Chair: Éric Marchand*
Peter Song	Quantile mediation analysis
Margaret Banker	Changepoint and functional parameter estimation with an accelerometer data application
Ostap Okhrin	Addressing maximization bias in reinforcement learning with two-sample testing
Menggang Yu	Sufficient dimension reduction for populations with structured heterogeneity

Wednesday, June 26th

9:30–10:15	**Plenary Session 2,** Room 1020, Chair Christian Genest		
9:30–10:15	IT	**Adelchi Azzalini** *University of Padova*	On the use of ordered factors as explanatory variables
10:15–10:45	Coffee break		
10:45–12:25	CS1	Multivariate regression modeling	
12:25–14:00	Lunch break		
14:00–15:45	IS2	Org: **Éric Marchand** *Université de Sherbrooke*	Predictive density estimation
15:45–16:15	Coffee break		
16:15–17:55	CS2	Statistics in high dimensions	

Contributed Session 1: Multivariate regression modeling

Wednesday 10:45–12:25, Room 1020

Chair: Katarzyna Filipiak	
Jūratė Karasevičienė	Some properties of generalized subexponential distributions
Natalya Pya Arnqvist	Shape constrained additive models with smooth interactions
Stanislav Anatolyev	AIC for many-regressor heteroskedastic regressions
Evangelos Evangelou	Mixed-penalty Bayesian Gaussian graphical lasso
Ivan Žežula	Multivariate mean testing in a class of models with patterned variance matrices

Invited Session 2: Predictive density estimation

Wednesday 14:00–15:45, Room 1020

Organizer: Éric Marchand	*Chair: Adelchi Azzalini*
Éric Marchand	The search for efficient predictive density estimators in multivariate models
Takeru Matsuda	Matrix estimation and prediction via singular value shrinkage
Fumiyasu Komaki	Improving predictions based on right invariant priors for group models
Keisuke Yano	Predictive inference in linear mixed models

Contributed Session 2: Statistics in high dimensions

Wednesday 16:15–17:55, Room 1020

Chair: Kalev Pärna	
Per Arnqvist	Gaussian model-based clustering for functional data with covariates
Tõnu Kollo	Covariance structure tests for multivariate t-distribution
Taras Bodnar	Reviving pseudo-inverses: Asymptotic properties of large dimensional Moore-Penrose and Ridge-type inverses with applications
Xiaomi Hu	Multivariate order and a pseudo restricted MLE under multivariate order restrictions
Joni Virta	Spatial depth for object-valued data

Thursday, June 27th

9:00–10:45	**Plenary Session 3,** Room 1020, Chair Tõnu Kollo		
9:00–10:00	KL	**Christian Genest and Johanna Nešlehová** *McGill University*	Bayes spaces and their use in functional data analysis of multivariate probability densities
10:00–10:45	IT	**Thomas Mikosch** *University of Copenhagen*	Extreme value theory for multivariate heavy-tailed time series
10:45–11:15	Coffee break		
11:15–12:35	CS3	Applications	
12:35–14:00	Lunch break		
14:00–15:45	IS3	Org: **Wen Zhou** *Colorado State University*	Statistical modeling and inference on complex network data
15:45–16:15	Coffee break		
16:15–17:45	Panel discussion Moderator: D. von Rosen	What is multivariate statistics in 2024?	
19:00–22:00	**Conference Dinner at the restaurant Vilde & Vine**		

Contributed Session 3: Applications

Thursday 11:15–12:35, Room 1020

Chair: Krista Fischer	
Maryna Prus	Computing optimal allocation of trials to sub-regions in crop-variety testing in case of correlated genotype effects
Māra Delesa-Vēlina	NMR metabolomics data as a predictor of lifespan: challenges of modelling and interpretation in the Estonian Biobank cohort
Dace Kalsone	Socioeconomic status and reading achievements in Latvia and Lithuania: schools and individual level
Kalev Pärna	On uncertainty estimation in Estonian Forestry Inventory (NFI)

Invited Session 3: Statistical modeling and inference on complex network data

Thursday 14:00–15:45, Room 1020

Organizer: Wen Zhou	*Chair: Meelis Käärik*
Wen Zhou	Detection and statistical inference on informative core and periphery structures in weighted directed networks
Yuan Zhang	U-statistic reduction: Higher-order accurate risk control and statistical-computational trade-off, with application to network method-of-moments
Ji Zhu	A latent space model for hypergraphs with diversity and heterogeneous popularity
Yunpeng Zhao	Community detection with heterogeneous block covariance mode

Friday, June 28th

9:00–10:30	**Plenary Session 4,** Room 1020, Chair Ivan Žežula		
9:00–9:45	IT	**Solomon W. Harrar** *University of Kentucky*	Overcoming biomarker bias with finite mixtures for multivariate outcomes
9:45–10:30	IT	**Hannu Oja** *University of Turku*	Notions of dispersion, kurtosis and information: From principal components to independent components
10:30–11:00	Coffee break		
11:00–12:40	CS4	New developments in statistical estimation and modeling	
12:40–13:00	**Closing**		

Contributed Session 4: New developments in statistical estimation and modeling

Friday 11:00–12:40, Room 1020

Chair: Thomas Mikosch	
Han Bao	Maximum spacing estimation under unbounded h-functions for multivariate observations
Janis Valeinis	Smoothly trimmed mean for two and more sample statistical inference
Leonora Pahirko	Location and scale parameter estimation in two-sample problems
Mohammad Jamsher Ali	PH approximation of two-barrier ruin probability for Lévy risk having two-sided PH jumps
Artur Tuttar	On generation of tabular data using *CTGAN* and *mice* methods

List of Abstracts

C.R. Rao's Century from Tampere's perspective

Simo Puntanen

Tampere University, Tampere, Finland, simo.puntanen@tuni.fi

KL

June 25
14:00
Δ1037

Keywords: Calyampudi Radhakrishna Rao, Prasanta Chandra Mahalanobis, Indian Statistical Institute, IWMS matrix workshop series

The goal of this talk is to give a brief look at the life and career of Professor Rao (1920–2023) leading to the International Prize in Statistics Prize 2023. The Prize is being considered as the statistics' equivalent of the Nobel Prize. Moreover, some personal glimpses, illustrated with a good bunch of photographs, are given.

Bibliography

[1] Efron, B., Cox, D.R. et al. (2020). C.R. Rao's century. *Significance*, Vol. 17, No. 4, pp. 36–38.

[2] Krishnankutty, N. (1996). *Putting Chance to Work ... a life in statistics: A Biography of C.R. Rao*. Dialogue, State College, PA, USA.

[3] Puntanen, S., Styan, G.P.H. (1996). A brief biography and appreciation of C. R. Rao, with a bibliography of books and papers. *Linear Algebra and its Applications*, 237/238, 1–40.

[4] Prakasa Rao, B.L.S. (2014). C.R. Rao: A life in statistics. *Current Science*, Vol. 107, No. 5, pp. 895–901.

[5] Rao, C.R. (1992). Statistics as a last resort. *Glimpses of India's Statistical Heritage*. (J.K. Ghosh, S.K. Mitra and K.R. Parthasarathy, Eds.), Wiley. pp. 151–213.

Bayes spaces and their use in functional data analysis of multivariate probability densities

KL

June 27
9:00
Δ1020

Christian Genest[1] **and Johanna G. Nešlehová**[12]

[1] *McGill University, Montréal (Québec) Canada, christian.genest@mcgill.ca*
[2] *Wirtschaftsuniversität Wien, Wien, Austria, johanna.neslehova@mcgill.ca*

Keywords: Anomaly detection, Bayes space, compositional data, orthogonal decomposition

Bayes spaces were initially designed to provide a geometric framework for modeling and analyzing distributional data. In this talk, we will derive an orthogonal decomposition of multivariate probability densities within the structure of this Hilbert space using a distributional analog of the Hoeffding–Sobol identity. A connection between this decomposition of a multivariate density and its copula-based representation will also be detailed. To illustrate this methodology, we will perform a functional data analysis of chemical element concentration after regional post-stratification. Based on data gathered between 1990 and 2009, this approach will allow us to confirm several suspected contamination patterns in Czech agricultural soils. Open problems and research perspectives related to this approach will be highlighted. This talk is based on joint work with Karel Hron, Tomáš Matys Grygar and others; see Genest et al. (2023), Matys Grygar et al. (2024), and references therein.

Bibliography

[1] Genest, C., Hron, K., Nešlehová, J.G. (2023). Orthogonal decomposition of multivariate densities in Bayes spaces and relation with their copula-based representation. *Journal of Multivariate Analysis* 198, Article 105228, 24 pp.

[2] Matys Grygar, T., Radojičić, U., Pavlů, I., Greven, S., Nešlehová, J.G., Tůmová, Š., Hron, K. (2024). Exploratory functional data analysis of multivariate densities for the identification of agricultural soil contamination by risk elements. *Journal of Geochemical Exploration* 259, Article 107416, 17 pp.

Safety belt estimation under the multivariate linear model

Katarzyna Filipiak[1], **Dietrich von Rosen**[2,3], **Martin Singull**[3],
and Wojciech Rejchel[4]

[1] *Poznań University of Technology, Poland, Katarzyna.Filipiak@put.poznan.pl*
[2] *Swedish University of Agricultural Sciences, Uppsala, Sweden, Dietrich.von.Rosen@slu.se*
[3] *Linköping University, Sweden, Martin.Singull@liu.se*
[4] *Nicolaus Copernicus University, Toruń, Poland, iggypop@mat.umk.pl*

IT

June 25
15:00
Δ1037

Keywords: MANOVA model, maximum likelihood estimation, ridge regression, convex optimization theory

Our goal is to determine the maximum likelihood estimators of unknown parameters under the multivariate linear model with a priori information related to the maximum effects in the model. Such estimators will be called Safety Belt estimators. It is shown that in many cases the convex optimization theory can be utilized, leading to two alternative types of solutions, strongly dependent on the data and the restriction.

As a special case of multivariate model the univariate case is considered, as some similarities to the ridge regression estimation can be observed there. It is experimentally shown that the Safety belt estimates behave usually better than estimates obtained via ridge regression

On the use of ordered factors as explanatory variables

IT

June 26
10:00
Δ1020

Adelchi Azzalini[1]

[1] *Università degli Studi di Padova, Italy, azzalini@stat.unipd.it*

Keywords: ordered factors, ordered categorical variables, factor coding, transformations of the normal distributions

A methodology is proposed to deal with the presence of ordered factors used as explanatory variables, typically included in the linear predictor of a model under consideration. For an ordered factor with K levels, say, a set of K numeric values is introduced, with a given value assigned to each factor level. In the end, the original factor is effectively replaced by a numeric variable.

A simple time-honoured option to deal with ordered factors is to assign numeric scores taken from the basic sequence of integers from 1 to K, but this option is often too simplistic, since it implicitly assumes equally-spaced effects. Alternatively, the current method of choice for dealing with ordered factors employs a set of contrasts based on orthogonal polynomials. The proposed methodology is stays in the logic of the old-fashion approach, but it overcomes the limitation of equally-spaced scores by constructing scores targeted to the specific situation under consideration.

The actual construction of the numeric scores proceeds by selecting K quantiles of a distribution belonging to some suitable parametric family. The adoption of a sufficiently flexible parametric family helps to find a scoring system adequate for the data under consideration. A concomitant product of this scheme is the identification of numeric values which indicate how the K levels are “really” spaced. Combining these two features, a key feature of the proposal is interpretability of the resulting construction.

With a single numeric variable to represent a given factor, one cannot expect to achieve the same numerical fit to the data as obtained by a set of polynomial contrasts, when these contrasts involve high degrees polynomials, and correspondingly several parameters. However, a range of numerical explorations has indicated that in many cases the resulting fit is equal or similar to the one achieved via polynomial contrasts, with non-negligible simplification in the model specification, and easier interpretation.

In a nutshell, the aim of the approach is to achieve a satisfactory data fit while improving on model parsimony, with simple interpretability of the scoring scheme.

Extreme value theory for heavy-tailed time series

IT

June 27
10:00
Δ1020

Thomas Mikosch

University of Copenhagen, Denmark, mikosch@math.ku.dk

We will consider regularly varying time series. The name comes from the marginal tails which are of power-law type. Davis and Hsing (1995) and Basrak and Segers (2009) started the analysis of such sequences. They found an accompanying sequence (spectral tail process) which contains the information about the influence of extreme values on the future behavior of the time series, in particular on extremal clusters. Using the spectral tail process, it is possible to derive limit theory for maxima, sums, point processes... of regularly varying sequences, but also refined results like precise large deviation probabilities for these structures.

In this talk we will give a short introduction to regularly varying sequences and and explain how the aforementioned limit results can be derived.

Overcoming biomarker bias with finite mixtures for multivariate outcomes

IT

June 28
9:00
Δ1020

Solomon Harrar[1] **and Zi Ye**[2]

[1] *University of Kentucky, USA, solomon.harrar@uky.edu*
[2]*Lehigh University, USA, ziy421@lehigh.edu*

Keywords: diagnostic misclassification, nonparametric relative effect, ordinal data, asymptotic, bias, validated data

In personalized medicine, patients are first classified as having or not of the characteristic of interest using diagnostic biomarkers. This classification is used to decide on treatment options. However, diagnostic biomarkers are usually not perfectly accurate. This will induce bias in assessing a treatment's effect. In this talk, we consider this problem in a fully nonparametric setting. Methods for consistently estimating and testing meaningful yet nonparametric treatment differences will be presented. Along the way, we present an estimation technique for misclassification error rates and discuss their asymptotic properties. These methods are applicable for outcomes measured in ordinal, discrete, or continuous scales. Simulation results show significant advantages of the methods in bias reduction, coverage probability, and power. The applications of the methods are illustrated with transcriptome data from bronchial airway brushing in asthmatic and healthy control subjects.

Bibliography

[1] Ye,Z., Harrar, S. (2024). Nonparametric Finite Mixture: Applications in Overcoming Misclassification Bias, *Journal of the American Statistical Association* (in press).

[2] Ye, Z., Harrar, S. (2022). Estimation of Multivariate Treatment Effects in Contaminated Randomized Trials,*Pharmaceutical Statistics* 21, 535–565.

[3] Harrar, S. , Amatya, A., Kalachev, L. (2016). Assessing Treatment Efficacy in the Presence of Diagnostic Errors, *Statistics in Medicine* 35, 5338–5355.

Notions of dispersion, kurtosis and information: From principal components to independent components

IT

June 28
9:45
Δ1020

Hannu Oja

University of Turku, Finland, hannu.oja@utu.fi

Keywords: Independent component analysis, principal component analysis, projection pursuit

Partial orderings and measures of dispersion, kurtosis and information for continuous univariate random variables with a special role of the Gaussian distribution are discussed. The independent component analysis is often seen as a continuation step for the principal component analysis. Starting from standardized principal components, the information measures and measures of non-Gaussianity including third and fourth cumulants are then generally used as projection indices in the projection pursuit approach for the independent components. The connections between kurtosis, information, non-Gaussianity and statistical independence in the context of independent component analysis is discussed in detail. The presentation is based on the results in [1] and [2].

Bibliography

[1] Oja, H. (1981), On location, scale, skewness and kurtosis of univariate distributions, *Scandinavian Journal of Statistics*, 8, 154–168.

[2] Radojičić, U. , Nordhausen, K. ja Oja, H. (2020), Notion of information and independent component analysis, *Applications of Mathematics*, 65, 311–330.

Invited Session 1: New analytics for complex correlated data

Organizer: Peter Song

Quantile mediation analysis

Peter Song

School of Public Health, University of Michigan, pxsong@umich.edu

IS

June 25
16:15

Mediation analysis is widely used in practice to examine if, and if so how, an intermediate variable mediates an exposure effect on an outcome. In some biomedical studies such as biological aging and obesity, quantiles of the outcome, instead of its mean, are of central interest. We develop a new methodological framework of quantile mediation analysis to facilitate identifying, estimating, and testing quantile mediation effects under a class of directed acyclic graphs, in which resulting key estimands take some closed form expressions. Since the null hypothesis of no mediation effect is composite, traditional tests such as the Sobel test or MaxP test appear conservative and underpowered. To improve statistical power, we adopt an adaptive bootstrap (AB) method to construct a calibrated test statistic that enables to properly control type I error under a composite null hypothesis. We establish theoretical guarantees for the proposed AB test method. We examine numerically both type I error and power of our AB test method through extensive simulation experiments and illustrate it in assessing mediation effects of lipid metabolites on the association between exposure to phthalates and childhood obesity.

Changepoint and functional parameter estimation with an accelerometer data application

IS

June 25
16:15

Margaret Banker

Feinberg School of Medicine, Northwestern University margaret.banker@northwestern.edu

Accelerometry data enables scientists to extract personal digital features useful in precision health decision making. Existing analytic methods often begin with discretizing Physical Activity (PA) counts into activity categories via fixed cutoffs; however, the cutoffs are validated under restricted settings and cannot be generalized across studies. Here, we develop a data-driven approach to overcome this bottleneck in the analysis of PA data, in which we holistically summarize an individual's PA profile using Occupation-Time Curves that describe the percentage of time spent at or above a continuum of activity levels. The resulting functional curve is informative to capture time-course individual variability of PA. We investigate functional analytics under an L0 regularization approach, which handles highly correlated micro-activity windows that serve as predictors in a scalar-on-function regression model. We develop a new one-step method that simultaneously conducts fusion via change-point detection and parameter estimation through a new L0 constraint formulation, which is evaluated via simulation experiments and a data analysis assessing the influence of PA on biological aging.

Addressing maximization bias in reinforcement learning with two-sample testing

IS

June 25
16:15

Ostap Okhrin

Institute of Transport and Economics, Dresden University of Technology,
ostap.okhrin@tu-dresden.de

Value-based reinforcement-learning algorithms have shown strong results in games, robotics, and other real-world applications. Overestimation bias is a known threat to those algorithms and can lead to dramatic performance decreases or even complete algorithmic failure. We frame the bias problem statistically and consider it an instance of estimating the maximum expected value (MEV) of a set of random variables. We propose the T-Estimator (TE) based on two-sample testing for the mean, that flexibly interpolates between over- and under-estimation by adjusting the significance level of the underlying hypothesis tests. A generalization, termed K-Estimator (KE), obeys the same bias and variance bounds as the TE while relying on a nearly arbitrary kernel function. We introduce modifications of Q-Learning and the Bootstrapped Deep Q-Network (BDQN) using the TE and the KE and prove convergence in the tabular setting. Furthermore, we propose an adaptive variant of the TE-based BDQN that dynamically adjusts the significance level to minimize the absolute estimation bias. All proposed estimators and algorithms are thoroughly tested and validated on diverse tasks and environments, illustrating the bias control and performance potential of the TE and KE.

Sufficient dimension reduction for populations with structured heterogeneity

IS

June 25
16:15

Menggang Yu

School of Public Health, University of Michigan, menggang@umich.edu

A key challenge in building effective regression models for large and diverse populations is accounting for patient heterogeneity. An example of such heterogeneity is in health system risk modeling efforts where different combinations of comorbidities fundamentally alter the relationship between covariates and health outcomes. Accounting for heterogeneity arising combinations of factors can yield more accurate and interpretable regression models. Yet, in the presence of high dimensional covariates, accounting for this type of heterogeneity can exacerbate estimation difficulties even with large sample sizes. To handle these issues, we propose a flexible and interpretable risk modeling approach based on semiparametric sufficient dimension reduction. The approach accounts for patient heterogeneity, borrows strength in estimation across related subpopulations to improve both estimation efficiency and interpretability, and can serve as a useful exploratory tool or as a powerful predictive model. In simulated examples, we show that our approach often improves estimation performance in the presence of heterogeneity and is quite robust to deviations from its key underlying assumptions. We demonstrate our approach in an analysis of hospital admission risk for a large health system and demonstrate its predictive power when tested on further follow-up data.

Invited Session 2: Predictive density estimation

Organizer: Éric Marchand

The search for efficient predictive density estimators in multivariate models

IS

June 26
14:00

Éric Marchand[1]

[1] *Université de Sherbrooke, Canada, eric.marchand@ushebrooke.ca*

Keywords: $\alpha-$ divergence, Dominance, Frequentist risk, Kullack-Leibler divergence, L_1 loss, Minimax, Mixtures, Multivariate normal, Restricted parameter space, Skew-normal, Type-II censoring

This talk will address the estimation of predictive densities, Bayesian or otherwise, and their efficiency as measured by frequentist risk. For Kullback-Leibler ([5]), $\alpha-$divergence ([8]) and integrated L_1 losses ([3, 6]), we review several recent findings that bring into play improvements by scale expansion, as well as duality relationships with point estimation and point prediction problems. A range of models or situations are studied and include multivariate normal with both known and unknown covariance structure ([4]), scale mixture of normals, mean mixture of normals including skew-normal distributions ([2], future order statistics with type II censoring ([1]), as well as models with restrictions on the parameter space ([5, 7]).

Bibliography

[1] Asgharzadeh, A., Marchand, É, & Saadati Nik, A. (2024). Bayesian prediction regions and density estimation with type-2 censored data. ArXiv 2403-6718

[2] Bhagwat, P. & Marchand, É (2023). Bayesian inference and prediction for mean-mixtures of normal distributions. *Electronic Journal of Statistics*, **17**, 1893–1922

[3] Bhagwat, P. & Marchand, É (2023). Predictive density estimation with integrated L_1 loss. *Journal of Multivariate Analysis*, **197**, 105190

[4] Fourdrinier, D., Marchand, É. & Strawderman, W.E. (2019). On efficient prediction and predictive density estimation for spherically symmetric models. *Journal of Multivariate Analysis*, **173**, 18-25.

[5] Fourdrinier, D., Marchand, É., Righi, A., & Strawderman, W.E. (2011). On improved predictive density estimation with parametric constraints. *Electronic Journal of Statistics*, **5**, 172-191.

[6] Kubokawa, T., Marchand, É & Strawderman, W.E. (2017). On predictive density estimation for location families under integrated absolute value loss. *Bernoulli*, **23**, 3197-3212.

[7] Kubokawa, T., Marchand, É., Strawderman, W.E., Turcotte, J.P. (2013). Minimaxity in predictive density estimation with parametric constraints. *Journal of Multivariate Analysis*, **116**, 382-397.

[8] LMoudden, A. & Marchand, É (2019). On Predictive Density Estimation under α-divergence Loss. *Mathematical Methods of Statistics*, **28**, 127–143.

Matrix estimation and prediction via singular value shrinkage

IS

June 26
14:00

Takeru Matsuda

University of Tokyo, Japan & RIKEN Center for Brain Science, Japan,
matsuda@mist.i.u-tokyo.ac.jp.

Keywords: Bayes, nonparametric estimation, predictive density, reduced-rank regression, shrinkage, singular value, superharmonic

In the estimation of a normal mean vector under the quadratic loss, the maximum likelihood estimator (MLE) is inadmissible and dominated by shrinkage estimators (e.g. James–Stein estimator) when the dimension is greater than or equal to three (Stein's paradox). In particular, generalized Bayes estimators with respect to superharmonic priors (e.g. Stein's prior) are minimax and dominate MLE. Recently, parallel results have been obtained for the predictive density problem.

In this talk, I will introduce recent studies on generalizations of the above results to matrices and their applications. First, we develop a superharmonic prior for a normal mean matrix that shrinks singular values towards zero, which can be viewed as a natural generalization of Stein's prior. This prior is motivated from the Efron–Morris estimator, which is an extension of the James–Stein estimator to matrices. This prior dominates the Jeffreys prior in estimation under the Frobenius loss as well as prediction under the Kullback–Leibler loss. In particular, since it shrinks towards the space of low-rank matrices, it attains large risk reduction when the unknown matrix is close to low-rank (e.g. reduced-rank regression). This idea also leads to an empirical Bayes matrix completion algorithm. Next, we construct a theory of shrinkage estimation under the "matrix quadratic loss", which is a matrix-valued loss function suitable for matrix estimation. A notion of "matrix superharmonicity" for matrix-variate functions is introduced and the generalized Bayes estimator with respect to a matrix superharmonic prior is shown to be minimax under the matrix quadratic loss. The matrix-variate improper t-priors are matrix superharmonic and this class includes the above generalization of Stein's prior. Finally, we show that the blockwise Efron–Morris estimator attains adaptive minimaxity in a multivariate Gaussian sequence model, where adaptation is not only to unknown smoothness but also to arbitrary quadratic loss.

vector $(\mathbb{R}^n, n \geq 3)$	matrix $(\mathbb{R}^{n\times p}, n \geq p+2)$
James–Stein estimator $\hat{\mu}_{\mathrm{JS}} = \left(1 - \frac{n-2}{\|x\|^2}\right) x$	Efron–Morris estimator $\hat{M}_{\mathrm{EM}} = X\left(I_p - (n-p-1)(X^\top X)^{-1}\right)$
Stein's prior $\pi_{\mathrm{S}}(\mu) = \|\mu\|^{-(n-2)}$	singular value shrinkage prior $\pi_{\mathrm{SVS}}(M) = \det(M^\top M)^{-(n-p-1)/2}$

Bibliography

[1] Efron, B. and Morris, C., Empirical Bayes on vector observations: an extension of Stein's method, *Biometrika*, **59** (1972), 335–347.

[2] Matsuda, T. and Komaki, F., Singular value shrinkage priors for Bayesian prediction, *Biometrika*, **102** (2015), 843–854.

[3] Matsuda, T. and Komaki, F., Empirical Bayes matrix completion, *Computational Statistics & Data Analysis*, **137** (2019), 195–210.

[4] Matsuda, T. and Strawderman, W. E., Estimation under matrix quadratic loss and matrix superharmonicity. *Biometrika*, **109** (2022), 503–519.

[5] Matsuda, T., Adapting to general quadratic loss via singular value shrinkage. *IEEE Transactions on Information Theory*, accepted.

[6] Matsuda, T., Komaki, F. and Strawderman, W. E., Double shrinkage priors for a normal mean matrix. arXiv:2311.13137.

Improving predictions based on right invariant priors for group models

IS

June 26
14:00

Fumiyasu Komaki

University of Tokyo, Japan, komaki@g.ecc.u-tokyo.ac.jp

We demonstrate that the right invariant prior density for a group model is a harmonic function under the Fisher metric, and discuss a method to improve predictions based on the right invariant prior. In group models with multiple transformation groups, averaging predictive densities based on the right invariant priors corresponding to each transformation group presents a systematic method to improve each predictive density. The relationship between this method and Bayesian methods has not been clear. From the perspective of asymptotic theory, we clarify the meaning of this method from Bayesian theory. This talk is based on joint works with Tomonari Sei and Ryotaro Kawata.

Predictive inference in linear mixed models

IS

June 26
14:00

Gourab Mukherjee[1] **and Keisuke Yano**[2]

[1] *University of Southern California, USA, gourab@usc.edu*
[2] *The Institute of Statistical Mathematics, Japan, yano@ism.ac.jp*

Keywords: Empirical Bayes, Kullback-Leibler loss, Predictive density estimation, Oracle optimality, Shrinkage, Risk estimation, Predictive inference, Linear Mixed Models

We consider predictive inference in *Linear Mixed Models* (LMMs) and study the problem of estimating the predictive density under Kullback-Leibler (KL) loss in LMMs with large number of units. We consider flexible classes of empirical Bayes (EB) predictive densities and develop a novel risk estimation based methodology for selecting hyper-parameters of EB predictive density estimates. Our risk estimation based hyper-parameter selection methodology uses the decision-theoretic identity in Lemma 2 of [1] that connects predictive KL risk for density estimation to Stein's unbiased estimate of the quadratic risk in point estimation. Direct construction of unbiased KL risk estimates is not possible in LMMs. We leverage information in the covariates and exchangeability of the individual effects to construct asymptotically efficient estimates of the KL risks for wide classes of predictive density estimators. We derive the rates of convergences of the proposed KL risk estimates (KLRE) and show that EB predictors calibrated by minimizing KLREs are asymptotically optimal in LMMs.

Bibliography

[1] George, E., Liang, F., Xu, X. (2006). Improved minimax predictive densities under Kullback–Leibler loss. *The Annals of Statistics* 78-91.

Invited Session 3: Statistical modeling and inference on complex network data

Organizer: Wen Zhou

Detection and statistical inference on informative core and periphery structures in weighted directed networks

IS

June 27
14:00

Wen Zhou

Colorado State University, USA, riczw@stat.colostate.edu

In network analysis, noises and biases, which are often introduced by peripheral or non-essential components, can mask pivotal structures and hinder the efficacy of many network modeling and inference procedures. Recognizing this, identification of the core-periphery (CP) structure has emerged as a crucial data pre-processing step. While the identification of the CP structure has been instrumental in pinpointing core structures within networks, its application to directed weighted networks has been underexplored. Many existing efforts either fail to account for the directionality or lack the theoretical justification of the identification procedure. In this work, we seek answers to three pressing questions: (i) How to distinguish the informative and non-informative structures in weighted directed networks? (ii) What approach offers computational efficiency in discerning these components? (iii) Upon the detection of CP structure, can uncertainty be quantified to evaluate the detection? We adopt the signal-plus-noise model, categorizing uniform relational patterns as non-informative, by which we define the sender and receiver peripheries. Furthermore, instead of confining the core component to a specific structure, we consider it complementary to either the sender or receiver peripheries. Based on our definitions on the sender and receiver peripheries, we propose spectral algorithms to identify the CP structure in directed weighted networks. Our algorithm stands out with statistical guarantees, ensuring the identification of sender and receiver peripheries with overwhelming probability. Additionally, our methods scale effectively for expansive directed networks. Implementing our methodology on faculty hiring network data revealed captivating insights into the informative structures and distinctions between informative and non-informative sender/receiver nodes across various academic disciplines.

U-statistic reduction: Higher-order accurate risk control and statistical-computational trade-off, with application to network method-of-moments

IS

June 27
14:00

Yuan Zhang

The Ohio State University, USA, yzhanghf@stat.osu.edu

U-statistics play central roles in many statistical learning tools but face the haunting issue of scalability. Significant efforts have been devoted into accelerating computation by U-statistic reduction. However, existing results almost exclusively focus on power analysis, while little work addresses risk control accuracy – comparatively, the latter requires distinct and much more challenging techniques. In this paper, we establish the first statistical inference procedure with provably higher-order accurate risk control for incomplete U-statistics. The sharpness of our new result enables us to reveal how risk control accuracy also trades off with speed for the first time in literature, which complements the well-known variance-speed trade-off. Our proposed general framework converts the long-standing challenge of formulating accurate statistical inference procedures for many different designs into a surprisingly routine task. This paper covers non-degenerate and degenerate U-statistics, and network moments. We conducted comprehensive numerical studies and observed results that validate our theory's sharpness. Our method also demonstrates effectiveness on real-world data applications.

A latent space model for hypergraphs with diversity and heterogeneous popularity

IS

June 27
14:00

Ji Zhu

University of Michigan, USA, jizhu@umich.edu

While relations among individuals make an important part of data with scientific and business interests, existing statistical modeling of relational data has mainly been focusing on dyadic relations, i.e., those between two individuals. This work addresses the less studied, though commonly encountered, polyadic relations that can involve more than two individuals. In particular, we propose a new latent space model for hypergraphs using determinantal point processes, which is driven by the diversity within hyperedges and each node's popularity. This model mechanism is in contrast to existing hypergraph models, which are predominantly driven by similarity rather than diversity. Additionally, the proposed model accommodates broad types of hypergraphs, with no restriction on the cardinality and multiplicity of hyperedges. Consistency and asymptotic normality of the maximum likelihood estimates of the model parameters have been established. The proof is challenging, owing to the special configuration of the parameter space. Simulation studies and an application to the What's Cooking data show the effectiveness of the proposed model.

Community detection with heterogeneous block covariance mode

IS

June 27
14:00

Yunpeng Zhao

Colorado State University, USA, yunpeng.zhao@colostate.edu

Community detection is the task of clustering objects based on their pairwise relationships. Most of the model-based community detection methods, such as the stochastic block model and its variants, are designed for networks with binary (yes/no) edges. In many practical scenarios, edges often possess continuous weights, spanning positive and negative values, which reflect varying levels of connectivity. To address this challenge, we introduce the heterogeneous block covariance model (HBCM) that defines a community structure within the covariance matrix, where edges have signed and continuous weights. Furthermore, it takes into account the heterogeneity of objects when forming connections with other objects within a community. A novel variational expectation-maximization algorithm is proposed to estimate the group membership. The HBCM provides provable consistent estimates of memberships, and its promising performance is observed in numerical simulations with different setups and data examples.

PH approximation of two-barrier ruin probability for Lévy risk having two-sided PH jumps

CS

June 28
11:00

Mohammad Jamsher Ali[1] **and Kalev Pärna** [1]

[1] *University of Tartu, Estonia, jamsher.ali@ut.ee*

Keywords: Risk process, Lévy process, Phase type distribution, Ruin probability

In this paper, we study a Lévy risk process consisting of Brownian component together with premiums and claims that are phase-type with many phases. Our aim is to approximate the probability of ruin without touching an upper barrier a. In line with this, the study demonstrates that the described Lévy risk process can essentially be replaced with a simpler risk process in which both premiums and claims are phase-type with just few phases.

Assume that the reserve of an insurer with initial capital $u \in [b, a]$, where a and b are the upper and lower boundaries, respectively, and having both-sided jumps, can be expressed by the following equation:

$$R_t = u + \sum_{i=1}^{N_t^1} p_i - \sum_{i=1}^{N_t^2} c_i + \mu t + \sigma W_t \quad \text{with} \quad R_0 = u, \tag{1}$$

where positive jumps $\{p_n\}_{n\geq 1}$ are a family of i.i.d. random variables having distribution F_p and occur at the epochs of the Poisson(λ_p) process N_t^1, also independent of N_t^1, and are of phase-type(PH) with representation $(\boldsymbol{\alpha}_p, \mathbf{T}_p)$; and negative jumps $\{c_n\}_{n\geq 1}$ are a family of i.i.d.random variables having distribution F_c and occur at the epochs of the Poisson(λ_c) process N_t^2, also independent of N_t^2, and are of phase-type (PH) with representation $(\boldsymbol{\alpha}_c, \mathbf{T}_c)$. The parameters μ and $\sigma > 0$ are respectively the drift and the variability of the Brownian motion, and W_t is a standard Brownian motion. It is also assumed that all component processes in (1) are mutually independent.

If the distributions F_p and F_c of premiums and claims in (1) are PH with a big number of phases, then technically it is hard to handle the process of calculating the probability of up-crossing before down-crossing (ruin in case of $b = 0$) and vice-versa. To reduce the workload, we propose to replace the reserve process (1) with a modified reserve process below in such a way that the first four moments of aggregate premiums and aggregate claims of both processes (1) and (2) are correspondingly equal. The modified reserve process is defined by

$$\tilde{R}_t = u + \sum_{i=1}^{\tilde{N}_t^1} \tilde{p}_i - \sum_{i=1}^{\tilde{N}_t^2} \tilde{c}_i + \mu t + \sigma W_t, \tag{2}$$

where the notation is similar to (1) except the premiums and claims in (2) are PH with just two phases instead of many phases. For two phases PH distribution, we use some well-known PH distributions, like hyper-exponential, Coxian, and Erlang.

From our empirical example, it is evident that replacing the original risk process's premium and claim distributions, which are phase-type with multiple phases, with simpler two-phase hyper-exponential, Coxian, and Erlang distributions, results in very small relative errors for various initial capitals.

Taking into account precision and simplicity, we conclude that approximation with hyper exponential distribution is most appropriate.

Bibliography

[1] Ali M.J., and Pärna K. (2023). PH approximation of two-barrier ruin probability for Lévy risk having two-sided PH jumps. *ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA* 27, 113–129

[2] Asmussen S. , and Albrecher H. (2010). *Ruin Probabilities.* World Scientific, New Jersey.

AIC for many-regressor heteroskedastic regressions

CS

June 26
10:45

Stanislav Anatolyev[1]

[1] *CERGE-EI, Czech Republic, stanislav.anatolyev@cerge-ei.cz*

Keywords: linear regression, Akaike information criterion, heteroskedasticity, many regressors

The original Akaike information criteria (AIC) and its corrected version (AICc) have been routinely used for model selection for ages. The penalty terms in these criteria are tied to the classical normal linear regression, characterized by conditional homoskedasticity and a small number of regressors relative to the sample size, which leads to very simple and computationally attractive penalty forms.

We derive, from the same principles, a general version that takes account of conditional heteroskedasticity and regressor numerosity. The new AICm penalty takes a form of a ratio of certain weighted average error variances, and encompasses the classical ones: it is approximately equal to the AIC penalty when the regression is conditionally homoskedastic and regressors are few, and to the AICc penalty when the regression is conditionally homoskedastic but the number of regressors is not negligible. In contrast to those of AIC and AICc, the AICm penalty is stochastic and thus not immediately implementable, as it in addition depends on the pattern of conditional heteroskedasticity in the sample.

The infeasible AICm criterion, however, can be operationalized via unbiased estimation of individual variances. The feasible AICm criterion still minimizes the expected Kullback-Leibler divergence up to an asymptotically negligible term that does not relate to regressor numerosity. In simulations, the feasible AICm does select models that deliver systematically better out-of-sample predictions than the classical criteria.

Gaussian model-based clustering for functional data with covariates

Per Arnqvist[1], Sara Sjöstedt de Luna[1] and Natalya Pya Arnqvist[1]

[1] *UmeåUniversity Sweden, per.arnqvist@math.umu.se*

Keywords: Functional data, Model based Clustering, Dependence, Misalignment, Sediment Data

Climate and environmental changes are widely discussed today, particularly the impact of human activity. Historical documents, year rings from trees, ice cores from glaciers, and lake and sea sediments [4],[1], are used to understand past climate variations over extended periods; we propose a model-based method for clustering subjects for which functional data and covariates are observed. The suggested model allows the covariance structures within the clusters to be different. The model thus extends a model proposed by [3]. It also will enable us to model the dependency of the chosen basis coefficients and the covariates. We also allow for different covariance structures within each cluster and suggest how to determine how many clusters to use. For an illustration, the model is applied to annually laminated (varved) sediment covering more than 6400 years from Lake Kassjön in northern Sweden to infer past climate changes [2].

Bibliography

[1] Arnqvist, P., Bigler, C., Renberg, I., Sjöstedt de Luna, S. (2016). Functional clustering of varved lake sediment to reconstruct past seasonal climate. *Environmental and Ecological Statistics*, 23, 4, 513-529.

[2] Arnqvist, P., and Sjöstedt de Luna, S. (2019). Model based functional clustering of varved lake sediments. *arXiv preprint*, arXiv:1904.10265.

[3] James, G. and Sugar, C.A. (2003). Clustering for sparsely sampled functional data, *Journal of the American Statistical Association*, 98, 397-408.

[4] Petterson, G., Renberg, I., Sjöstedt-de Luna, S., Arnqvist, P., Anderson, N.J. (2010). Climatic influence on the inter-annual variability of late-Holocene minerogenic sediment supply in a boreal forest catchment, *Earth Surface Processes and Landforms*, 35, 4, 390-398.

Maximum spacing estimation under unbounded h-functions for multivariate observations

CS

June 28
11:00

Han Bao[1], **Kristi Kuljus**[2] **and Bo Ranneby**[3]

[1] *University of Tartu, Estonia, han.bao@ut.ee*
[2] *University of Tartu, Estonia, kristi.kuljus@ut.ee*
[3] *Swedish University of Agricultural Sciences, Sweden, bo@ranneby.nu*

Keywords: maximum spacing estimation, nearest neighbour balls, weak consistency, multivariate observations

Maximum spacing (MSP) method is a parameter estimation method for continuous distributions based on spacings. By defining the MSP method using appropriate h-functions corresponding to different divergence measures, we can consider the MSP method for different information measures such as Kullback-Leibler, Jeffrey's divergence measure, the Hellinger distance, etc. This gives parameter estimators with different properties. Strong and weak consistency of MSP estimators in the case of multivariate i.i.d. observations and for a class of strictly concave h-functions that are bounded above was proved in [1]. One example of such h-function is $h(x) = \ln x - x + 1$. To get an MSP estimator with better properties, we study a more general class of h-functions that are not necessarily bounded above, $h(x) = \ln x$ corresponding to the Kullback-Leibler information is one example in this class. Weak consistency of the MSP estimator is proved for the considered class of h-functions. The advantages of MSP estimators in the context of model validation are illustrated by examples of normal mixture data and skew-normal data.

Bibliography

[1] Kuljus, K. and Ranneby, B. (2015). Generalized maximum spacing estimation for multivariate observations. *Scandinavian Journal of Statistics*, 42(4):1092–1108.

Reviving pseudo-inverses: Asymptotic properties of large dimensional Moore-Penrose and Ridge-type inverses with applications

CS

June 26
16:15

Taras Bodnar[1], **Nestor Parolya**[2]

[1]*Stockholm University, Sweden, taras.bodnar@math.su.se*
[2]*Delft University of Technology, The Netherlands, n.parolya@tudelft.nl*

Keywords: Moore-Penrose inverse, ridge-type inverse, Bell polynomials, sample covariance matrix, random matrix theory, high-dimensional asymptotics

We derive high-dimensional asymptotic properties of the Moore-Penrose inverse and the ridge-type inverse of the sample covariance matrix (see, e.g., [1], [2]). In particular, the analytical expressions of the weighted sample trace moments are deduced for both generalized inverse matrices and are present by using the partial exponential Bell polynomials which can easily be computed in practice. The existent results are extended in several directions: (i) First, the population covariance matrix is not assumed to be a multiplier of the identity matrix; (ii) Second, the assumption of normality is not used in the derivation; (iii) Third, the asymptotic results are derived under the high-dimensional asymptotic regime. Our findings are used to construct improved shrinkage estimators of the precision matrix, which asymptotically minimize the quadratic loss with probability one. Finally, the finite sample properties of the derived theoretical results are investigated via an extensive simulation study.

Bibliography

[1] Bodnar, T., H. Dette and N. Parolya (2016). Spectral analysis of the Moore-Penrose inverse of a large dimensional sample covariance matrix. *Journal of Multivariate Analysis 148*, 160–172.

[2] Imori, S. and D. von Rosen (2020). On the mean and dispersion of the Moore-Penrose generalized inverse of a Wishart matrix. *The Electronic Journal of Linear Algebra 36*, 124–133.

NMR metabolomics data as a predictor of lifespan: challenges of modelling and interpretation in the Estonian Biobank cohort

CS

June 27
11:15

Māra Delesa-Vēlina[1] **and Krista Fischer**[2]

[1] *University of Tartu, Estonia, mara.delesa-velina@ut.ee*
[2] *University of Tartu, Estonia, krista.fischer@ut.ee*

Keywords: metabolomics, NMR, survival modelling, mortality, biological age

There has been a great interest in studying nuclear magnetic resonance (NMR) metabolomics data as predictor of overall mortality and proxy for biological (or metabolomic) age in biobanks data. As the total number of participants with the NMR data in the Estonian Biobank exceeds 200'000, with the mean follow-up time of 15 years for the first 50'000 participants and 4 years for the newest 150'000, we are interested to develop a model for all-cause mortality based on NMR data for the Estonian Biobank.

The participants have been recruited in two large waves in 2002-2010 and in 2018-2019, using somewhat different recruitment strategies. This resulted in some notable differences between the two subcohorts in the distribution of baseline sociodemographic variables as well as in the prevalence of chronic diseases and also in the overall risk for all-cause mortality and incident diseases.

These differences pose a challenge to survival modelling: the two cohorts cannot be analyzed together. However, when analyzed separately, the two cohorts produce models that lead to very different predicted risk levels for the same individual (if used for out-of-sample predictions).

Instead of either making a choice between the two cohorts or producing some average risk estimates based on the two, we propose to use the data differences for model validation. We develop the survival model based on the subset of the first wave NMR records and validate it in the second wave data.

We show that although the absolute risk estimates differ in the two cohorts, the predictions based on the biological age difference are still similar. Thus we conclude that one should find alternatives to absolute risk estimates, as the risk, interpreted as a “fraction of individuals getting the disease” would always depend on the particular subset of a population, represented by the given biobank cohort.

Mixed-penalty Bayesian Gaussian graphical lasso

CS

June 26
10:45

Evangelos Evangelou[1] **and Sandipan Roy**[1]

[1] *University of Bath, UK, ee224@bath.ac.uk*

Keywords: Bayes factors, Elastic net, Generalised lasso, Gibbs sampling, Group lasso, Markov random field

The standard graphical lasso (least absolute shrinkage and selection operator) [7, 2] is a regularisation approach for sparse estimation of Gaussian precision matrices. Its Bayesian equivalent is derived by assigning double exponential priors to the elements of the precision matrix and a block Gibbs sampling approach can be used to generate samples from the posterior distribution of the precision matrix [6]. Just like the standard graphical lasso can be generalised to include other types of regularisers, such as elastic net, group, and fused lasso [1, 3, 5, 4], so can the Bayesian graphical lasso. When multiple regularisation penalties are applied, the Gibbs sampling approach breaks down because the regularisation parameter does not have a known full-conditional distribution. I will present a novel Bayesian Gaussian graphical lasso model that allows for Gibbs sampling for some of the parameters, while the regularisation parameter is chosen via an empirical Bayes approach. This provides an efficient an general methodology for fitting such models.

Bibliography

[1] Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(2):373–397.

[2] Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

[3] Gibberd, A. J. and Nelson, J. D. (2017). Regularized estimation of piecewise constant gaussian graphical models: The group-fused graphical lasso. *Journal of Computational and Graphical Statistics*, 26(3):623–634.

[4] Kovács, S., Ruckstuhl, T., Obrist, H., and Bühlmann, P. (2021). Graphical elastic net and target matrices: Fast algorithms and software for sparse precision matrix estimation. *arXiv preprint arXiv:2101.02148*.

[5] Pircalabelu, E. and Claeskens, G. (2020). Community-based group graphical lasso. *Journal of Machine Learning Research*, 21(1):2406–2437.

[6] Wang, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, 7(4):867–886.

[7] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35.

Multivariate order and a pseudo restricted MLE under multivariate order restrictions

CS

June 26
16:15

Yujie Jia and Xiaomi Hu[1]

Wichita State University, Wichita, Kansas
[1]*Presenter: xiaomi.hu@wichita.edu*

Keywords: Closed convex cone, matrix projection, multivariate order, pseudo restricted MLE under multivariate order restrictions

In Hilbert space combining the concept of quasi order, its relation to the linear operations and its relation to the convergence, we define a binary relation $\preceq$ called order. In R^p, with $p = 1$ it is a univariate order, with $p > 1$ it is a multivariate order. There is a 1-1 mapping between the collection of all orders in R^p and the collection of all closed convex cones in R^p such that the order is generated from the cone. This cone is called the order generating cone. The diversified forms of multivariate orders allow researchers to render the prior knowledge on parameter vectors $\theta_i \in R^p$, $i = 1, ..., q$, as multivariate order restrictions. The collection of all $p \times q$ matrices whose columns satisfy the specified order restrictions is a closed convex cone in $R^{p\times q}$. This cone is called the order restricted cone. The restricted MLE of the mean matrix in MANOVA is the projection of the MLE onto the order restricted cone.

Based on the pair-wise optimization criterion, we define a closed convex subset of the order restricted cone corresponding to general grass ordering that puts simple ordering, simple tree ordering and umbrella ordering as special cases, and propose the use of the projection onto this subset as a pseudo RMLE due to its good properties and the simplicity of the algorithm. The proposed algorithm converts the problem of the matrix projection onto order restricted cone to the problem of vector projections onto order generating cone. In the simulation study the pseudo RMLE exhibits the behavior of a consistent estimator.

Socioeconomic status and reading achievements in Latvia and Lithuania: schools and individual level

CS

June 27
11:15

Dace Kalsone

Ministry of Education and Science, Latvia, dace.kalsone@izm.gov.lv

Keywords: MLM, ANCOVA, PIRLS 2021, Mplus, reading achievements, socioeconomic status

Is the school mean of socioeconomic status (SES) of families associated with student reading achievement after controlling for individual SES? Random intercepts model with Level 1 and Level 2 predictors helps to obtain compositional effect and thus explain which proportion of variance do these variables on each level. ANCOVA [1] is one of the tools to come closer to answers.

International Association for the Evaluation of Educational Achievement (IEA) provides general public with the continuous data [2] of large-scale assessments on students' readings, mathematical, science, ICT, civic and other competences assessments. Inter alia one of the latest public data sets are available about students reading skills – Progress in International Reading Literacy Study (PIRLS) on 2021 data.

The interest for data analysts from the policy makers' side that myself representing this time and my Baltic countries – here Lithuania and Latvia – is utmost importance to keep our societies competitive and happy at the same time. Therefore, finding potential success elements and also drawbacks is important task to ensure data driven decisions to tackle those elements more precisely and thus also saving resources with less targeted attempts.

Bibliography

[1] Mplus, a latent variable modeling program (1973-2024), https://www.statmodel.com/index.shtml

[2] Progress in international reading literacy study, data repository (2021), https://www.iea.nl/data-tools/repository/pirls

Some properties of generalized subexponential distributions

CS

June 26
10:45

Jūratė Karasevičienė [1], Jonas Šiaulys[1], Svetlana Danilenko[2]

[1]*Institute of Mathematics,Vilnius University, Naugarduko 24,Vilnius LT-03225,Lithuania,*
jurate.karaseviciene@mif.vu.lt,
jonas.siaulys@mif.vu.lt
[2]*Vilniaus Gedimino technikos universitetas, Saulėtekio al.11,Vilnius LT-10223,Lithuania*
svetlana.danilenko@vilniustech.lt

Keywords: subexponentiality, generalized subexponentiality, heavy tail, randomly stopped sum

We present conditions under which distributions functions of randomly stopped sum, minimum, maximum, minimum of sums and maximum of sums belong to the class of generalized subexponential distributions. The primary random variables are supposed to be independent and real-valued, but not necessarily identically distributed. The counting random variable describing the stopping moment of random structures is supposed to be nonnegative, integer-valued and not degenerate at zero. In addition, it is supposed that counting random variable and the sequence of the primary random variables are independent. We demonstrate how randomly stopped structures can be applied to the construction of new generalized subexponential distributions. The results published in the articles [1] and [2] will be discussed during the presentation.

Bibliography

[1] J. Karasevičienė, J. Šiaulys, Randomly stopped sums with generalized subexponential distribution, Axioms, 2023, 12, 641.

[2] J. Karasevičienė, J. Šiaulys, Randomly stopped minimum, maximum, minimum of sums and maximum of sums with generalized subexponential distribution, Axioms, 2024, 13, 85.

Covariance structure tests for multivariate t-distribution

CS

June 26
16:15

Katarzyna Filipiak[1] **and Tõnu Kollo**[2]

[1]*Poznań University of Technology, Poland, katarzyna.filipiak@put.poznan.pl*
University of Tartu, Estonia, tonu.kollo@ut.ee

Keywords: covariance structure, multivariate t-distribution, likelihood ratio test, Rao score test, Wald test

We derive an equation system for finding Maximum Likelihood Estimators (MLE) for the parameters of p-dimensional t-distribution with ν degrees of freedom, $t_{p,\nu}$, and use the MLEs for testing covariance structures for the $t_{p,\nu}$-distributed population. The likelihood ratio test (LRT), Rao score test (RST) and Wald test statistic (WTS) are derived under the general null-hypothesis $\mathrm{H}_0 : \Sigma = \Sigma_0$, using matrix derivative technique. Here $p \times p$-matrix Σ is a dispersion/scale parameter. Convergence to the asymptotic chi-square distribution under the null hypothesis is examined in extensive simulation experiments. Also the convergence to the chi-square distribution is studied empirically in the situation when MLEs of $t_{p,\nu}$-distribution are changed to the corresponding estimators for a normal population. In simulation study the RST behaved most adequately compared with LRT and WST in the situation when dimensionality p is growing.

Location and scale parameter estimation in two-sample problems

CS

June 28
11:00

Leonora Pahirko[1] **and Jānis Valeinis**[1]

[1] *University of Latvia, Latvia, leonora.pahirko@lu.lv, janis.valeinis@lu.lv*

Keywords: location-scale models, parameter estimation, robust estimators, empirical likelihood method, data normalization

Let us consider the two-sample problem, where X and Y are independent random variables with distribution functions F and G, respectively. Then the location-scale model is defined as

$$F(x) = G\left(\frac{x-\mu}{\sigma}\right), \quad x \in \mathbb{R},$$

where $\mu \in \mathbb{R}$ is the location and $\sigma > 0$ is the scale parameter. The model parameters can be expressed as

$$\sigma = \frac{\sigma_X}{\sigma_Y} \quad \text{and} \quad \mu = \mu_X - \sigma\mu_Y,$$

where μ_X, μ_Y, σ_X and σ_Y denote the means and standard deviations of X and Y, respectively.

In order to validate the location-scale model between two distributions F and G, we aim to test the hypothesis

$$H_0 : \text{There exist constants}\, \mu \in \mathbb{R}\, \text{and}\, \sigma > 0\, \text{such that}\, X \overset{\mathcal{D}}{=} \sigma Y + \mu,$$

where $\mathcal{D}$ denotes the equality of random variables in distribution. The test based on characteristic functions was introduced in [1] to test the above hypothesis. Their approach was based on the standardization of both samples using the sample means and standard deviations. Pahirko *et al.* [2] proposed to use the procedure based on the hypothesis $H_0 : F(x) = G^*(x)$, $x \in \mathbb{R}$, where $G^*(x)$ is the distribution function of $Y^* = \hat{\sigma}Y + \hat{\mu}$. Here $\hat{\mu}$ and $\hat{\sigma}$ denote some consistent estimators of two-sample location and scale parameters.

Several methods for two-sample location and scale parameter estimation were studied previously (see, e.g., [3], [4]). A natural choice is to directly replace the theoretical values of μ and σ with their sample estimators, which gives $\hat{\sigma} = S_X/S_Y$ and $\hat{\mu} = \bar{X} - \hat{\sigma}\bar{Y}$, where $\bar{X}$ and $\bar{Y}$ denote the corresponding sample means and S_X and S_Y denote the corresponding sample standard deviations. In [3], it was proposed to use trimmed means and standard deviations instead. The aim of this study is to compare various methods for finding $\hat{\mu}$ and $\hat{\sigma}$ in order to examine their robustness properties. The procedure of the location-scale model hypothesis test using parameter estimation is demonstrated using real data example from Latvian diabetes patient study to carry out data normalization.

Bibliography

[1] Hall, P., Lombard, F., Potgieter, C.J. (2013). A new approach to function-based hypothesis testing in location-scale families. *Technometrics*, 55(**2**), 215–223.

[2] Pahirko, L., Valeinis, J., Gredzens, J., Krumina., M. (2023). Validation of two-sample location-scale model using empirical likelihood-based statistics. In: *5th International Conference on Statistics: Theory and Applications*, paper ID 168, Avestia Publishing.

[3] Potgieter, C.J. and Lombard, F. (2012). Nonparametric estimation of location and scale parameters. *Computational Statistics & Data Analysis*, 56(**12**), 4327–4337.

[4] Potgieter, C.J. and Lombard, F. (2016). Nonparametric two-sample estimation of location and scale parameters from empirical characteristic functions. *Journal of Statistical Computation and Simulation*, 86(**16**), 3225–3242.

Computing optimal allocation of trials to sub-regions in crop-variety testing in case of correlated genotype effects

CS

June 27
11:15

Maryna Prus[1]

[1] *University of Hohenheim, Stuttgart, Germany*

Keywords: Best linear unbiased prediction, Correlated genotype effects, Experimental design, Linear mixed model, Multi-environment trials, Optimal design

New crop varieties are extensively tested in multi-environment trials in order to obtain a solid basis for recommendations to farmers. When the target population of environments is large, a division into sub-regions is often advantageous. If the same set of genotypes is tested in each of the sub-regions, a linear mixed model (LMM) may be fitted with random genotype-within-sub-region effects. The first analytical results to optimizing allocation of trials to sub-regions have been obtained in [1]. In that paper the genotype effects are assumed to be uncorrelated. However, this assumption is not always suitable for practical situations. In practice, genetic markers are often used in plant breeding for determining genetic relationships of genotypes, which helps to model their correlation. In this work a more general LMM with correlated genotype effects is considered. Analytical results and a computational approach are proposed for optimal allocation of trials.

Bibliography

[1] Prus, M. and Piepho, H.-P. (2021). Optimizing the allocation of trials to sub-regions in multi-environment crop variety testing. *Journal of Agricultural, Biological and Environmental Statistics*, 26: 267-288.

Shape constrained additive models with smooth interactions

Natalya Pya Arnqvist

Umeå University, Sweden, natalya.pya@umu.se

Keywords: regression, shape constraints, smoothing, interaction, smooth ANOVA

Any application area that analyzes the relationship between a response and multiple covariates could potentially benefit from using nonparametric and semiparametric regression models. Regression models that incorporate smooth functions of predictor variables have gained widespread usage and proved successful due to the work of [1, 2]. When analyzing the relationships between a response and predictors, it might be natural to assume that some obey certain shape constraints, such as monotonicity and convexity. Such problems are widespread in ecological and environmental studies. Shape-constrained additive models [3] offer a general framework for fitting exponential family generalized additive models with shape restrictions on smooths. The main objective of this talk is to provide extensions of the existing framework for generalized additive modelling with a mixture of unconstrained terms and various shape-restricted terms to accommodate smooth interaction of covariates, varying coefficient terms, linear functionals with or without shape constraints as model components, and data with short-term temporal or spatial autocorrelation. The practical usage of the suggested extensions will be illustrated on various examples.

Bibliography

[1] Wood, S.N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *JRSS(B)* 73(1), 3–36

[2] Wood, S.N., Pya, N., Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *JASA* 111(516), 1548–1563

[3] Pya, N., Wood, S.N. (2015). Shape constrained additive models. *Statistics and Computing* 25, 543–559

On uncertainty estimation in Estonian Forestry Inventory (NFI)

Raul Kangro[1], **Märt Möls**[1], **Kalev Pärna**[1]

[1] *University of Tartu, Estonia, raul.kangro@ut.ee, mart.mols@ut.ee, kalev.parna@ut.ee*

Keywords: National Forestry Inventory, systematic cluster sampling, uncertainty estimation, spatial correlations, simulation, post-stratification

Estonian National Forestry Inventory (NFI) is a rolling programme to provide information about the size, distribution, composition and condition of forests and woodlands. It is based on systematic cluster sampling which uses a regular grid of tracts (clusters) each consisting of 16 circular sample plots. The analysis of NFI data is focused on estimation of target variables like the area of forest and its volume, but also uncertainty (error) characteristics of the estimates must be included.

Error calculations in NFI is by no means a trivial task. Statistically correct error calculation follows both the sampling design and takes into account spatial correlations between observations. Straightforward use of plots as independent observations leads to an underestimated error. Instead, when using systematic cluster sampling, the tracts, not plots, are considered observations. Hence the data from plots should first be aggregated to the tract level and all further statistics is based on the tract level data.

In this paper we propose an alternative method for uncertainty estimation in Estonian NFI. The method differs from the existing error calculation method in two important aspects. Firstly, following the cluster sampling design, the tracts rather than the plots are considered observation units. Secondly, a special technique (with roots that go back a hundred years) is included to account for spatial correlations, thus allowing to reduce the error estimates. We describe the alternative method in details and study its usability in Estonia. The method is validated by means of two computer experiments based on local forestry data (both field and remote sensing ALS data are used). The results of the experiments show that the proposed error calculation method is more adequate than the existing simplified method and can easily be implemented.

On generation of tabular data using *CTGAN* and *mice* methods

CS

June 28
11:00

Artur Tuttar[1], **Meelis Käärik**[2] **and Yevhen Havrylenko**[3]

[1] *University of Tartu, Estonia, artur.tuttar@ut.ee*
[2] *University of Tartu, Estonia, meelis.kaarik@ut.ee*
[3] *University of Copenhagen, yh@math.ku.dk*

Keywords: synthetic data, generative adversarial networks, multivariate imputation by chained equations, actuarial science

With the rising availability of computational resources, we have seen a sharp increase in ways to apply machine learning techniques to solve predictive problems. However, to train these predictive machine learning models, a large amount of data is required. But even here, new techniques are being developed. This paper focuses on the latter problem and aims to showcase ways of utilizing machine learning techniques to augment existing data and generate synthetic data that has similar modelling efficacy as the original source data. In this paper, two methods *mice* [1] and *CTGAN* [2] are employed, combined and compared in generating synthetic insurance tabular data. The generated data is compared and assessed in terms of variable distributions, model training efficacy and stability. Additionally, an extension workflow combining the perceived strengths of the methods is considered.

Bibliography

[1] van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), 1–67. https://doi.org/10.18637/jss.v045.i03

[2] Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling Tabular data using Conditional GAN. *arXiv*. https://arxiv.org/abs/1907.00503

[3] Delong, Ł. and Kozak, A. (2023) 'The use of autoencoders for training neural networks with mixed categorical and numerical features', *ASTIN Bulletin*, 53(2), pp. 213–232. doi:10.1017/asb.2023.15.

[4] Kuo, K. (2020). Generative Synthesis of Insurance Datasets. *arXiv*. https://arxiv.org/abs/1912.02423

Smoothly trimmed mean for two and more sample statistical inference

Jānis Valeinis[1] **and Emils Silins**[1]

[1] *University of Latvia, Latvia, janis.valeinis@lu.lv*

Keywords: Smoothly trimmed mean, robust estimator, empirical likelihood

The classical trimmed mean is a robust estimator, that can be used in the presence of outliers in datasets. In 1973, Stigler [1] showed that the asymptotic normality of the trimmed mean may fail. He also proposed some smoothed versions of the trimmed mean instead.

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 the respective order statistics.

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.

In this work, we establish empirical likelihood for a more general version of the smoothly trimmed mean. We establish the asymptotic variance of our estimator and show the advantages of our approach over the established empirical likelihood method for the classical trimmed mean introduced in [2].

We also extend our approach to the two-sample and ANOVA cases. Finally, we compare all methods through a simulation study using the empirical coverage accuracy.

Bibliography

[1] Stigler, S.M. (1973). The asymptotic distribution of the trimmed mean. *The Annals of Statistics*, 1(3), 472-477.

[2] Qin, G., and Tsao, M. (2002). Empirical likelihood ratio confidence interval for the trimmed mean. *Communications in statistics-theory and methods*, 31(12), 2197-2208.

Spatial depth for object-valued data

Joni Virta[1]

[1] *University of Turku, Finland, joni.virta@utu.fi*

CS

June 26
16:15

Keywords: Distance function, object-valued data, outlier detection, statistical depth

A classical tool for quantifying the centrality/outlyingness of a point $\mu \in \mathbb{R}^p$ with respect to a distribution P is given by *statistical depth measures*. Informally, a depth $D(\mu; P)$ of μ with respect to P is a function that assigns high (low) values for points μ located near (far away from) the bulk of the distribution P. One commonly used statistical depth measure is the *spatial depth* [1, 2] which, in addition having a simple interpretation, is robust and fast to compute.

In this talk, we present a generalization of spatial depth to *object-valued data*. By object-valued data we refer to samples of data where the observations reside in an arbitrary metric space $\mathcal{X}$, instead of the space $\mathbb{R}^p$. The proposed generalization is shown to have highly interpretable geometric properties, making it appealing in object data analysis where standard descriptive statistics are difficult to compute. The proposed measure reduces to the classical spatial depth when $\mathcal{X}$ is a Euclidean space. In addition to studying its theoretical properties, to provide intuition on the concept, we explicitly compute the object spatial depths in several different metric spaces. Finally, we showcase the practical usefulness of the proposed depth measure in non-convex depth region estimation and classification.

The talk is based on the preprint [3].

Bibliography

[1] Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91(434), 862–872.

[2] Vardi, Y., Zhang, C. H. (2000). The multivariate L_1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4), 1423–1426.

[3] Virta, J. (2023). Spatial depth for data in metric spaces. *arXiv preprint*, arXiv:2306.09740.

Multivariate mean testing in a class of models with patterned variance matrices

CS

June 26
10:45

Ivan Žežula and Daniel Klein

University of P. J. Šafárik, Košice, Slovakia, ivan.zezula@upjs.sk

Keywords: multivariate linear model, special variance structure, quadratic subspace, mean testing

In recent years, multivariate models with matrix-valued observations attract more and more attention, see e.g. [1], [2] or [3]. One of the main problems of their practical use is usually the problem of small sample size, which causes numerical instability or even rank deficiency of the sample variance matrix. As a result, mean testing can be problematic or even impossible. Various special variance structures are used, when possible, to tackle the problem. However, this brings the need to derive the distribution of the test statistic for every special variance matrix structure. We will present solution for a general class of these test procedures, when the variance matrix can be decomposed into parts belonging to some commutative quadratic subspace. This covers many previous results, which can be viewed as special cases of the general model introduced here.

Acknowledgment

This work was supported by the Slovak Research and Development Agency under the Contract No. APVV-21-0369, and grant VEGA No. 1/0585/24.

Bibliography

[1] Roy, A., Zmyślony, R., Fonseca, M., Leiva, R. (2016). Optimal estimation for doubly multivariate data in blocked compound symmetric covariance structure. *Journal of Multivariate Analysis* 144, 81–90.

[2] Žežula, I., Klein, D., Roy, A. (2018). Testing of multivariate repeated measures data with block exchangeable covariance structure. *Test* 27(2), 360–378.

[3] Dai, D., Hao, C., Jin, S., Liang Y. (2023). Regularized estimation of Kronecker structured covariance matrix using modified Cholesky decomposition. *Journal of Statistical Computation and Simulation*, DOI: 10.1080/00949655.2023.2291536.

Index