UNIVERSITÄT
BIELEFELD
Fakultät für
Wirtschaftswissenschaften

Zentrum für
Statistik

14th Young Researchers Workshop of Centre for Statistics (ZeSt)

18th December 2024

Programme

10:00 - 10:05	Dietmar Bauer: Welcoming
10:05 - 10:25	Hannah Marchi: Development of a health recommender system for targeted antibiotic therapy in sepsis
10:25 - 10:45	Sebastian Büscher: Visual Guidance for Model Specification: Introducing Score Plots for Composite Marginal Likelihood Estimated Discrete Choice Models
10:45 - 11:05	Michael Balzer: Gradient Boosting for Dirichlet Regression Models
11:05 - 11:15	Short Break
11:15 - 11:35	Ferdinand Stoye: Increasing flexibility for the meta-analysis of full ROC curves – a copula approach
11:35 - 11:55	Antoniya Dineva: A “Double Copula” Model for Semi-Competing Risks Data
11:55 - 13:15	Lunch Break
13:15 - 13:35	Kurtulus Kidik: A New Estimation Method for High-Dimensional Cointegrated Time Series Analysis
13:35 - 13:55	Ole Koslik: Extending smoothness selection for Markov-Switching models
13:55 - 14:15	Jonas Bauer: (Mis-)specification of constructs in SEM – Implications on parameter estimates and model fit
14:15 - 14:30	Coffee Break
14:30 - 14:50	Lennart Oelschläger: Modeling personal life choices
14:50 - 15:10	Julian Wäsche: Dynamic modelling and evaluation of preclinical trials in paediatric leukaemia
15:10 - 15:30	Carlina Feldmann: A multi-state capture-recapture model for estimation of weaning duration
15:30 - 15:50	Katharina Ammann: Modelling Value-At-Risk and Expected Shortfall Using Markov-Switching Generalised Additive Models for Location, Scale, and Shape
15:50 - 16:05	Coffee Break
16:05 - 16:25	Nayeli Gast Zepeda: Penalizing Infeasibility in Neural Combinatorial Optimization: An Experimental Study on Vehicle Routing Problems
16:25 - 16:45	Julia Dyck: The Bayesian Power generalized Weibull Shape Parameter test: a case study
16:45 - 17:05	Sophie Schmiegel: Multi-label and Single-label Classification for Disease Recognition With Special Consideration of Co-morbidities
17:05 - 17:10	Discussion and Closing Remarks

Development of a health recommender system for targeted antibiotic therapy in sepsis

Hannah Marchi
Faculty of Business Administration and Economics, Bielefeld University, Germany
hannah.marchi@uni-bielefeld.de

Choosing the right antibiotic therapy for sepsis patients is a complex challenge due to the need for immediate treatment and the lack of specific pathogen information. This often necessitates the use of broad-spectrum antibiotics as empirical therapy, aiming to cover a wide range of potential pathogens. However, this procedure significantly contributes to the development and spread of antibiotic resistance, a major issue for society, health policy, and the economy. Our project is addressing this challenge by identifying targeted antibiotics that are most effective in the treatment of patients newly diagnosed with sepsis. We base our considerations on data about sepsis patients who were admitted to the intensive care unit of the Evangelisches Klinikum Bethel (EvKB, Germany) between 2012 and 2023. The data includes core patient information, laboratory results and microbiological analyses, including the resistance status of the pathogens present. An essential part of our work is the comparison of different types of recommender systems applied to our data. By using microbiology data, we create a patient-therapy matrix containing resistance information for each therapy per patient. The matrix is sparse be- cause each patient has received only a limited number of antibiotics. To fill these gaps, we apply collaborative filtering techniques, both method- and model-based, as well as hybrid methods that incorporate demographic filtering. Furthermore, we extend a hybrid approach by combining both user- and item-based filtering by incorporating both patient and therapy similarity. These methods allow us to recommend the most effective treatments, guided by medical relevance and stopping criteria like allergies. Ultimately, our health recommender system aims to support clinicians in selecting effective, targeted antibiotics, thereby contributing to the reduction of antibiotic resistance, increasing patient survival and enhancing public health.

Visual Guidance for Model Specification: Introducing Score Plots for Composite Marginal Likelihood Estimated Discrete Choice Models

Sebastian Büscher
Faculty of Business Administration and Economics, Bielefeld University, Germany
sebastian.buescher@uni-bielefeld.de

Discrete Choice Models (DCMs) are widely employed to analyse decision-making processes across a range of disciplines, with a particular prevalence in transportation planning, where they provide the basis for predictions that inform high-stakes infrastructure investments and policy decisions. However, the subjective nature of the utility function specification introduces variability in model outcomes, thereby undermining the reliability of forecasts and potentially leading to discrepancies between data-driven recommendations and stakeholder decisions. We will addresses these challenges by introducing score plots, a diagnostic visualisation tool for models estimated using composite marginal likelihood (CML) methods. Drawing conceptual parallels to residual plots in linear regression, score plots employ pairwise score contributions to detect model misspecifications, thereby providing a guided approach to enhance model fit and consequently reduce subjectivity in the model selection process. The potential of score plots to identify issues such as omitted non-linearities, missing variables, structural breaks, and dynamic error processes is demonstrated through their application to both synthetic datasets and a real-world transportation case study. The integration of score plots into DCM workflows enables analysts to more effectively navigate the intricacies of model specification, thereby enhancing the reliability of predictions and improving the transparency of the model specification process, thus reinforcing the credibility of data-driven recommendations.

Gradient Boosting for Dirichlet Regression Models

Michael Balzer
Center for Mathematical Economics, Bielefeld University
michael.balzer@uni-bielefeld.de

In many real-world settings, applied researchers have to deal with data that is compositional. For instance, in ecology, economics and medicine compositional data occurs as proportions, amounts or rates. To deal with compositional data, the framework of Dirichlet regression models has been proposed. Specifically, Dirichlet regression models can be described in the framework of multivariate generalized additive models with the Dirichlet distribution providing the unknown parameters to be estimated through the additive predictors. We propose a novel model-based boosting approach for Dirichlet regression models embedded in the framework of generalized additive models for location, scale and shape. It directly provides an alternative estimation procedure besides the well-established approach based on the maximum likelihood principle with inherent data-driven variable selection for low- as well as high- dimensional data settings. Additionally, we present a real-world application setting concerning the changes in election results in the Great Recession using a large-scale European data set. Using our proposed approach, we investigate the effect of protests on voting proportions of party families while identifying important socio-economic variables and their effect on those voting proportions via variable selection.

Increasing flexibility for the meta-analysis of full ROC curves – a copula approach

Ferdinand Stoye
Biostatistics and Medical Biometry, Medical School OWL, Bielefeld University, Germany
ferdinand.stoye@uni-bielefeld.de

The development of new statistical methods for the meta-analysis of diagnostic test accuracy (DTA) studies is a vivid field of research, especially with respect to summarizing full receiver operating characteristic (ROC) curves. Most current approaches to this task utilize Gaussian random effects to account for between-study heterogeneity. To increase flexibility in the meta-analysis of ROC curves, we substitute Gaussian random effects with copulas, leading to the ability to directly model the dependence between sensitivity and specificity and an increased control over the estimation procedure. While the resulting models are numerically challenging, they lead to much more flexible and modular model structures when compared to Gaussian random effects. Combined with re-arranging the results reported by DTA studies as being bivariate interval-censored time-to-event data and clinically plausible parametric assumptions for the resulting mixtures in the marginal distributions, this leads to a powerful model to estimate summary ROC curves. An additional advantage of using copulas is the ability to provide a closed-form likelihood, enabling the possibility to use general purpose likelihood optimization strategies. In a simulation study, our copula models are able to create very flexible model fits with high convergence probabilities and perform similarly to competing models. However, they are also numerically unstable, leading to larger variations in bias as well as low empirical coverages in the simulation. This behavior gives rise to the need for a more robust estimation procedure. We also show the practical applicability of our copula models to data from a meta-analysis for the screening of type 2 diabetes, leading to plausible estimates for summary ROC curves.

A “Double Copula” Model for Semi-Competing Risks Data

Antoniya Dineva, Oliver Kuss, Annika Hoyer
Biostatistics and Medical Biometry, Medical School OWL, Bielefeld University, Germany
antoniya.dineva@uni-bielefeld.de

Semi-competing risks describe the general setting in which the primary focus is modeling the age at a non-terminal event (e.g., disease occurrence) when the investigated subjects are also at risk for experiencing a terminal event (e.g., death). That is, the terminal event might censor the non-terminal event, but not vice versa. As a result, the observed ages at the two events within the same individual are correlated. The statistical approaches for this setting can be divided in at least two general groups: 1) copula-based models and 2) illness-death models. A common approach in the first group involves modeling the dependency between the two events by one bivariate copula with two marginal distributions for age at disease onset and age at death. However, such approaches are limited in their ability to distinguish between the two different modes of mortality, the one with and the one without the disease. We propose a “double copula” model, that estimates the three marginal distributions in the semi-competing risks framework: 1) age at disease onset, 2) age at death for individuals with the disease and 3) age at death for individuals without the disease. The model is defined based on two bivariate copulas modeling the joint distribution of age at disease onset with each of the two ages at death separately. Model parameters are estimated by maximum likelihood, accounting for the complex censoring and truncation mechanisms in a cohort study. We performed a simulation study, that demonstrated promising results in terms of accuracy and numerical robustness.

A New Estimation Method for High-Dimensional Cointegrated Time Series Analysis

Kurtulus Kidik
Faculty of Business Administration and Economics, Bielefeld University, Germany
kurtulus.kidik@uni-bielefeld.de

In many economic studies, data sets often include cointegrated time series, which must be accounted for in statistical analysis. The Vector Error Correction Model (VECM) is a classic method for examining cointegration among multiple non-stationary time series. Traditional estimation techniques, such as maximum likelihood estimation (Johansen, 1995) and Johansen's cointegration test, are commonly used to estimate and identify the cointegrating rank. However, these econometric methods encounter difficulties, particularly due to the "curse of dimensionality," as the number of parameters to estimate grows significantly with the dimensionality of the data. This talk will focus on cointegration in high-dimensional settings, where there are many variables of interest that may potentially be cointegrated, along with a substantial number of time-based observations. The talk is divided into two sections. First, we will review existing econometric methods for VECM, highlighting their limitations in high-dimensional contexts and demonstrating through simulations, that they may fail to produce accurate and reliable estimates in practical scenarios. In the second part, a new estimation approach and model are introduced, with discussions on their effectiveness based on simulation results across various data-generating processes.

Extending smoothness selection for Markov-Switching models

Ole Koslik
Faculty of Business Administration and Economics, Bielefeld University, Germany
jan-ole.koslik@uni-bielefeld.de

Markov-switching models are powerful tools for capturing complex patterns in time series data influenced by latent states. Recently, we developed a framework for estimating components of these models nonparametrically using penalised splines within a quasi restricted maximum likelihood (qREML) algorithm. However, the current approach only accommodates smooths with penalties representable as quadratic forms that are linear in the penalty strength. This restriction includes one-dimensional smooths, simple i.i.d. random effects, and tensor-product smooths with isotropic smoothing.

In this talk, I will explore two potential extensions of this methodology to address these limitations. The first extension considers tensor-product smooths with anisotropic smoothing, enabling different levels of smoothness across dimensions to better capture directional variations in the data. The second adapts the qREML framework to accommodate LASSO-type penalties, potentially opening the door to efficient covariate selection in complex models. While both extensions require modifications of the current algorithm and still demand practical evaluation, they hold considerable promise for advancing the practical utility of Markov-switching models in complex time series analysis.

(Mis-)specification of constructs in SEM – Implications on parameter estimates and model fit

Jonas Bauer
Faculty of Business Administration and Economics, Bielefeld University, Germany
j.bauer@uni-bielefeld.de

Many disciplines face research questions involving constructs that cannot be observed directly, such as (buying) intentions, (leadership) behavior and (personality) traits. Statistical models such as structural equation modelling (SEM) enable researchers to nevertheless validate their theories under these circumstances. By now SEM has become a statistical cornerstone with sound methodology and user-friendly software implementations (e.g., `R` packages `lavaan` and `csem`).

As constructs cannot be observed directly, they have to interact with the observed indicators somehow. Three variants to specify such indicator-construct models (ICM) in SEM can be distinguished. If an ICM is misspecified the parameter estimates can be biased and the model fit poorly, as highlighted by plenty Many Monte Carlo studies conducted over the past 30 years. However, previous research is unconclusive because the two issues of parameter estimation and ICM (mis-)specification have often been lumped together or conflated.

Against this background, we present a Monte Carlo study that separates ICM (mis-)specification from parameter estimation. Our study considers all ICMs for the data generation as well as in the data analysis. By doing so, we can answer the questions 'which particular construct misspecification leads to biased estimates?' And 'would such misspecification be detected by the goodness-of-fit measures?'.

Modeling personal life choices

Lennart Oelschläger
Faculty of Business Administration and Economics, Bielefeld University, Germany
oelschlaeger.lennart@gmail.com

The German family panel pairfam provides information on parenting, partnership, and social life among individuals in Germany. Based on this data, we explore the factors shaping personal choices, such as life goals and marriage, and examine how these factors vary across individuals. For the analysis, we use various types of probit models within a Bayesian framework, including normally mixed, latent class, ordered, and ranked multinomial probit models. Alongside the findings, we discuss practical hurdles that come with analyzing choice data and present an R-based solution to manage these challenges effectively.

Dynamic modelling and evaluation of preclinical trials in paediatric leukaemia

Julian Wäsche
Faculty of Business Administration and Economics, Bielefeld University, Germany
julian.waesche@uni-bielefeld.de

Dynamic models play a crucial role in mathematically representing biological processes over time, particularly in leukaemia research. In preclinical studies, genetically modified leukaemia cells are examined in mice to identify new therapeutic targets aimed at curbing cell population growth. These experiments yield time-resolved data that can uncover growth-inhibiting effects. However, standard analyses often rely on statistical tests comparing only two time points, which limits the data's temporal richness and overlooks biological mechanisms. By incorporating biological insights into mathematical models, we enhance our ability to analyze the impact of modifications on these mechanisms. Our population growth model captures cell dynamics throughout the entire experimental duration, enabling a comprehensive evaluation of all measurement times. We show that this model detects simulated scenarios more effectively than conventional statistical tests and proves to be a valuable tool for assessing patient-derived CRISPR-Cas9 screening experiments in leukaemia research.

A multi-state capture-recapture model for estimation of weaning duration

Carlina Feldmann
Faculty of Business Administration and Economics, Bielefeld University, Germany
carlina.feldmann@uni-bielefeld.de

Investigating so-called "super suckler" Galápagos sea lions - individuals who extend their weaning period until older ages - poses multiple challenges: The partially latent nature of the dependence status of a pup or juvenile, sparse and irregular observations of suckling behaviour and animal emigration allow neither pre-classifying super sucklers nor their control groups. Instead, modelling the probability of being weaned in a latent model framework is necessary. To address this, we propose a multistate capture-recapture model with irreversible transitions between three states: unweaned, weaned, and absent/dead. By framing the model as a hidden Markov model, we leverage inferential tools like the forward algorithm and the Viterbi algorithm to estimate the latent weaning process, offering a versatile method to explore weaning dynamics and their influencing factors in wild populations. We analyse data from 1,891 Galápagos sea lions tagged at birth, with observation periods of up to 20 years (minimum one year). Biannually aggregated field observations categorize individuals as either suckling, not suckling, or not observed during each season. The model incorporates age as a covariate to estimate weaning age and further examines how weather conditions and maternal fitness influence weaning duration.

Modelling Value-At-Risk and Expected Shortfall Using Markov-Switching Generalised Additive Models for Location, Scale, and Shape

Katharina Ammann
Faculty of Business Administration and Economics, Bielefeld University, Germany
katharina.ammann@uni-bielefeld.de

Quantifying the risks associated with financial assets is essential for effective risk management. Common risk metrics include value-at-risk (VaR) and expected shortfall (ES). VaR represents the threshold below which a pre-determined percentage of return observations fall, typically focusing on extreme losses, while ES measures the expected loss in cases where returns fall below the VaR threshold, providing a more sensitive measure of tail risk.

In this talk, we propose Markov-switching generalised additive models for location, scale, and shape (GAMLSS) to model these metrics, using NASDAQ returns as an example. In contrast to conventional time series models typically used for VaR and ES forecasting—such as generalised autoregressive conditional heteroscedasticity (GARCH) models—the proposed methodology accounts for covariate effects on distribution parameters beyond mean and variance, such as skewness and kurtosis. This approach improves risk prediction and tail-event accuracy under varying market conditions.

To evaluate model performance and ensure compliance with Basel regulatory standards, VaR and ES forecasts are compared with historical values using backtesting.

Penalizing Infeasibility in Neural Combinatorial Optimization: An Experimental Study on Vehicle Routing Problems

Nayeli Gast Zepeda
Faculty of Business Administration and Economics, Bielefeld University, Germany
nayeli.gast@uni-bielefeld.de

Neural Combinatorial Optimization (NCO) methods have shown promise for vehicle routing problems (VRPs), with advances in problem scale and architectural improvements. However, these methods have primarily been tested on problems where feasible solutions are easily found. While previous work assumed neural networks could learn to respect constraints through penalty terms, we demonstrate experimentally that such penalty schemes fail to ensure solution feasibility, limiting the applicability of current NCO approaches to many real-world routing problems.

The Bayesian Power generalized Weibull Shape Parameter test: a case study

Julia Dyck
Faculty of Business Administration and Economics, Bielefeld University, Germany
j.dyck@uni-bielefeld.de

After release of a drug on the market, pharmacovigilance monitors the occurrence and changes of potential ADRs in the population. Thereby, signal detection methods investigate possible associations between selected drugs and adverse events (AE) by raising a signal if the result of a statistical test is positive on a pair (drug-AE).

Sauzet and Cornelius (2022) provided a test based on the power generalized Weibull (PgW) distribution shape parameters (PgWSP). If both shape parameters of the PgW distribution are equal to one, the corresponding hazard function is constant over time. This is interpreted as no temporal association between drug and AE. Such an approach however does not make use of any clinical expertise or prior knowledge about a drug's harm profile. A Bayesian signal detection method would allow such an incorporation of prior knowledge. To do this, we propose a Bayesian PgWSP test. The test compares a region of practical equivalence (ROPE) reflecting the null hypothesis with the estimated credibility intervals (CI) (Kruschke, 2015). The single shape parameters' ROPE test outcomes and a chosen combination rule define the Bayesian Pg-WSP.

First, we present the method and a summary of the simulation study conducted to develop and tune the Bayesian PgWSP. The main focus of the talk lies on the case study demonstrating the application of the tuned Bayesian PgWSP test. We illustrate the test and R package implemented for application using data from the THIN database containing primary care data from patients in the UK. The cohort under investigation are women prescribed with bisphosphonates typically prescribed for the treatment of osteoporosis. Four AE (headache, musculoskeletal pain, alopecia, carpal tunnel syndrome)- bisphosphonate pairs are considered.

Multi-label and Single-label Classification for Disease Recognition With Special Consideration of Co-morbidities

Sophie Schmiegel
Faculty of Business Administration and Economics, Bielefeld University, Germany
sophie.schmiegel@uni-bielefeld.de

A patient suffering from an illness is often examined for several possible dis- eases, as the symptoms cannot always be attributed to one single disease. The necessary examinations may take some time; to enhance diagnostic speed and accuracy, data-driven approaches such as single-label classification (SLC) and multi-label classification (MLC) are employed. These methods differ primarily in label assignment: SLC assigns each sample to one exclusive class while MLC allows samples to belong to multiple classes or none, acknowledging the poten- tial for co-morbidities. By comparing SLC and MLC, we investigate whether disease recognition improves when co-morbidities are considered. We aim to clarify differences in model formulation, decision spaces and class imbalance handling between these approaches. To empirically assess their performance, the methods are applied to data of chronic pain patients using various clas- sifiers: decision trees (DTs), random forests (RFs), logistic regression models (LRMs), k-nearest neighbors (k-NN) and multi-layer perceptrons (MLPs). For MLC, classifier chains (CCs) are utilised to account for disease correlations. Our findings suggest that incorporating co-morbidities does not consistently enhance the recognition of the main disease. The suitability of SLC or MLC depends on factors such as the correlation strength between diseases and the severity of class imbalance. In some cases, SLC may outperform MLC in recognising the main disease. This highlights the importance of considering the specific char- acteristics of the data when choosing between SLC and MLC approaches for disease recognition.