10th Young Researchers Workshop of ZeSt

11 March 2022 in lecture hall H7

Programme

9:00 - 9:05 a.m.	Welcoming Roland Langrock (Speaker of the ZeSt)
9:05 - 9:25 a.m.	Jonas Bauer: *Gliding the simplex and other tales of self-written algorithms*
9:25 - 9:45 a.m.	Rouven Michels: *Using tensor product B-splines for nonparametric inference in multivariate hidden Markov models*
9:45 - 10:05 a.m.	Sina Mews: *Continuous-time modelling of disease progression using claims data: how hard can it be?*
10:05 - 10:25 a.m.	David Winkelmann: *Capacitated Vehicle Routing with Stochastic Loading Constraints*
10:25 - 10:40 a.m.	Coffee Break
10:40 - 11:00 a.m.	Benedikt Langenberg: *Bayesian Analysis of Multi-factorial Repeated Measures Designs Using SEM*
11:00 - 11:20 a.m.	Jasper Bendler *An investigation of the reciprocal causal effect structure between individual political interest and individual political participation behaviour.*
11:20 - 11:40 a.m.	Alexander Stappert: *The role of implicit theories study-demands-resources of ability as individual preconditions in the framework. Focussing the domain of learning statistics.*
11:40 - 12:00 a.m.	Julia Dyck: *Bayesian survival analysis for signal detection of adverse drug reactions*
12:00 - 1:20 p.m.	Lunch Break
1:20 - 1:40 p.m.	Lennart Oelschläger *Bayes estimation of probit choice models with {RprobitB}*
1:40 - 2:00 p.m.	Katrin Rickmeier *Determinants of regional mobility after job loss in Germany - the importance of economic and non-economic factors in migration decisions of displaced workers.*
2:00 - 2:15 p.m.	Coffee Break
2:15 - 2:35 p.m.	Julian Wäsche: *Modelling tumor load evolution using ordinary differential equations and the challenge of parameter identifiability*
2:35 - 2:55 p.m.	Sebastian Büscher: *Weighting strategies for pairwise composite marginal likelihood estimation*
2:55 - 3:15 p.m.	Final Discussion

Gliding the simplex and other tales of self-written algorithms

Jonas Bauer

Universität Bielefeld, Fakultät für Wirtschaftswissenschaften
j.bauer@uni-bielefeld.de

Universality and efficiency are not only crucial properties of any algorithm, but also two sides of the very same coin. The wide field of MCMC variants (especially including Hamiltonian dynamics and iterative adaptions) serves perfectly to highlight the trade-off relationship between these two. In particular the curse of dimensionality and parameter restrictions, like a simplex parameter space, can become substantial burdens to an algorithms performance. We introduce several ways to deal with simplex parameter spaces using Hamiltonian Monte Carlo on high-dimensional count data.

Using tensor product B-splines for nonparametric inference in multivariate hidden Markov models

Rouven Michels

Universität Bielefeld, Fakultät für Wirtschaftswissenschaften
r.michels@uni-bielefeld.de

Hidden Markov models (HMM) comprise an observed time series that is driven by an unobserved Markov chain. The class of state-dependent distributions — e.g. normal, Poisson, or gamma — is typically chosen before modelling. An unfortunate choice of this class can lead to a poor fit, to wrong inference regarding the number of states, and to unsatisfactory state decoding. When fitting HMMs to multivariate time series, challenges already present in the univariate setting, such as skewness, heavy tails or outliers, are amplified by possibly complex dependence structures between the variables considered. To avoid the potential pitfalls associated with the use of a parametric class of distributions we estimate the state-dependent distributions nonparametrically. To this end, we discuss the use of multivariate tensor product B-splines within HMMs, thus estimating the multivariate state-dependent density in a data-driven way, i.e. without having to make a distributional assumption. In a simulation study, we demonstrate the general feasibility of the suggested nonparametric approach, and showcase a scenario in which it would be superior to a parametric approach. The practical use of the nonparametric approach is further illustrated in a case study using football data. Specifically, we model the bivariate data on length and angle of the passes of goalkeepers during the 2021 UEFA European Championship, thereby detecting match phases in which teams apply different strategies to start an attack. By incorporating covariates into the state-switching probabilities, we gain information about potential tactical adjustments by the team managers.

Continuous-time modelling of disease progression using claims data: how hard can it be?

Sina Mews

Universität Bielefeld, Fakultät für Wirtschaftswissenschaften
sina.mews@uni-bielefeld.de

Medical claims data are routinely collected for billing and reimbursement purposes, thus providing rich databases on real-life healthcare provisions. As these data sets contain information on persons' resource use, costs, and specific diagnoses, they are used, for example, to optimise health service provisions or to estimate the prevalence and incidence of diseases. However, the potential to analyse patients' disease progression over time based on claims data has hardly been explored in the literature.

The main challenge associated with modelling disease progression is that the disease stage is usually not (directly) observed. Instead, the evolution of patients' disease activity over time is inferred from available medical data. Based on claims data, a first idea is to use information on persons' drug consumption, assuming that higher drug usage indicates worsening health conditions, while lower drug usage indicates improved health conditions. A second idea is to consider the length of time intervals between patients' consultations, assuming that more frequent consultations result from an increased disease activity.

For both approaches, latent-state models offer a natural framework to analyse the evolution of patients' disease activity underlying the observed data. As claims data are collected at points irregularly sampled in time, i.e. only when a patient interacts with the healthcare system, we formulate our models in continuous time. In the first case, we apply continuous-time hidden Markov models and state-space models to patients' observed drug usage, where the observations' distribution is assumed to depend on a latent (discrete- or continuous-valued) disease process. In the second case, we use Markov-modulated Poisson processes to model patients' consultations over time, where the rate of consultations depends on a latent continuous-time Markov process with discrete (disease) states.

In my talk, I will briefly outline both modelling approaches and present some preliminary results of applying these models to claims data from patients diagnosed with chronic obstructive pulmonary disease (COPD). In particular, I will point out several challenges encountered so far while working with claims data.

Capacitated Vehicle Routing with Stochastic Loading Constraints

David Winkelmann

Universität Bielefeld, Fakultät für Wirtschaftswissenschaften
david.winkelmann@uni-bielefeld.de

The optimisation of operational processes related to the transportation of products is of crucial importance for many companies. Using homogeneous vehicles for the shipping process of packages requires the allocation of customers to tours with respect to constraints caused by the characteristics of the vehicle, e.g. volume and maximum loading weight. However, if the shape of packages is highly heterogeneous, the achievable fill rates vary and, e.g. in case of a manual packing process, are unknown when the tours are determined. This leads to an optimisation problem under uncertainty addressing the trade-off between routing costs and additional (penalty) costs if not all packages can be loaded into the trucks as planned. We propose to model the problem as a capacitated vehicle routing problem that integrates a binary regression model to estimate the probability that the packages do not fit into a truck. Using Taylor series expansion of the exponential function allows to reduce the transformation of the linear predictor to a polynomial term, which can be directly incorporated into the Gurobi optimiser.

Bayesian Analysis of Multi-factorial Repeated Measures Designs Using SEM

Benedikt Langenberg

Universität Bielefeld, Fakultät für Psychologie und Sportwissenschaft
benedikt.langenberg@uni-bielefeld.de

Latent repeated measures ANOVA (L-RM-ANOVA) has recently been proposed as an alternative to traditional repeated measures ANOVA. L-RM-ANOVA builds upon structural equation modeling and enables the researcher to investigate interindividual differences in main and interaction effects, examine custom contrasts, incorporate a measurement model, and account for missing data. However, L-RM-ANOVA uses maximum likelihood and thus cannot incorporate prior information and can have poor statistical properties in small samples. In this presentation, we show how L-RM-ANOVA can be used with Bayesian estimation to resolve the aforementioned issues. In particular, we demonstrate how to place informative priors on model parameters that constitute main and interaction effects. We further show how to place weakly informative priors on standardized parameters which can be used when no prior information is available. We compare Type 1 error, power, bias and efficiency of the Bayesian estimation to maximum likelihood estimation in two stimulation studies. We conclude that Bayesian estimation can lower Type 1 error and bias, and increase power and efficiency when priors are chosen adequately. However, statistical properties for Bayesian estimation can be worse than maximum likelihood when priors are not chosen carefully. We conclude that weakly informative priors for standardized parameters are a viable approach and, can be used when researchers have little or nil prior knowledge. Lastly, we argue that ANOVA tables and sums of squares are not sufficient information to specify informative priors, and we identify which parameter estimates should be reported in order for future research to create informative prior distributions; thereby promoting cumulative research.

An investigation of the reciprocal causal effect structure between individual political interest and individual political participation behaviour.

Jasper Bendler

Universität Bielefeld, Fakultät für Soziologie
jasper.bendler@uni-bielefeld.de

The relationship between individual political interest and individual political participation behaviour is one of the most studied relationships in political sociology. Most research about this topic focused on the (causal) effect of political interest on political participation and neglects a theoretical possible and plausible effect of political participation on political interest. Furthermore, studies that investigate both effects simultaneously are even rarer to find and are often based on a so called crossed-lagged panel model, which tends to have a problem with biased estimates because it can't control for unobserved heterogeneity. To help to close this research gap, I analysed the reciprocal effect structure between the political interest and political participation by using a so-called random intercept cross-lagged panel model (RI-CLPM) proposed by Hamaker et al. (2015) which allows analysing a longitudinal crossed effect structure and control for stable unobserved heterogeneity. The analysis is based on German Socio-Economic Panel Study (SOEP) from the years 2009 to 2019.

The role of implicit theories of ability as individual preconditions in the study-demands-resources framework. Focussing the domain of learning statistics

Alexander Stappert

Universität Bielefeld, Fakultät für Psychologie und Sportwissenschaft
alexander.stappert@uni-bielefeld.de

The Study-Demands-Resources (SDR) model - as an adapted version of the popular Job-Demands-Resources (JDR) model - poses a framework for predicting students' burnout, motivation and academic performance by means of institutional demands and resources. In this study, implicit theories of ability as individual beliefs about the malleability of individual competencies are introduced as individual preconditions into the SDR model. Structural equation models and path analysis are used to analyse (1) the relationships between implicit theories of ability, study demands and study resources (2) implicit theories of ability as predictors of students' burnout, motivation and performance (3) the moderating role of implicit theories of ability regarding the effects of study demands and study resources on students' burnout and motivation. Cross-sectional self-report data from online questionnaires of $N = 749$ students from german universities are used to test the hypotheses. To facilitate domain specificity, all constructs were measured considering the subject of learning statistics in the respective item-wordings. Data collection was conducted during the summer term of the year 2021. Hence, the data represent students' views during an online semester due to the COVID-19 pandemic.

Bayesian survival analysis for signal detection of adverse drug reactions

Julia Dyck

Universität Bielefeld, Fakultät für Wirtschaftswissenschaften
j.dyck@uni-bielefeld.de

After release of a drug on the market, pharmacovigilance monitors occurence and changes of known adverse drug reactions (ADRs) as well as detects new ADRs in the population. This is done to keep a drug's harm profile updated and can potentially result in adjustments of the prescription labelling or — in the extreme case — a recall of the product from the market.

In recent years the interest for the use of electronic health records and longitudinal data for pharmacovigilance increased. This data type provides potential for application of survival analysis tools in order to perform signal detection tests with respect to a suspected adverse drug reaction.

Cornelius et. al provided a signal detection test based on the Weibull Shape Parameter (WSP). This approach was refined by Sauzet and Cornelius leading to the Power generalized Weibull Shape Parameter (PgWSP) test.

Both approaches rely on data only. We believe that the performance of the PgWSP test can be improved by incorporating existing knowledge about the ADR profile of drugs from the same family.

Our goal is to construct a bayesian version of the PgWSP test that — besides other properties — allows for inclusion of prior knowledge about the drug family's ADR profile.

In the talk, the main idea of the PgWSP test including its limits is explained. Based on the conclusions about the existing approach, the concept for a bayesian PgWSP test for ADR signal detection given time-event data is presented.

Modelling tumor load evolution using ordinary differential equations and the challenge of parameter identifiability

Julian Wäsche

Universität Bielefeld, Fakultät für Wirtschaftswissenschaften
julian.waesche@uni-bielefeld.de

A good understanding of tumor growth is essential for the treatment of cancer patients. One crucial way of achieving it is by modelling the evolution of tumor cells and exploring how drugs are able to impede their spread. Since the amount of tumor cells within an individual are discrete, their evolution can explicitly be modelled by pure Markov jump processes. Ordinary differential equations (ODEs) haven proven to be suitable approximations of processes that facilitate parametric inference, such as in this case. Fitting a parametrized ODE to data can be achieved by maximum likelihood estimation. However, the quantities of interest, i.e. the explicit number of cells, are often only partially observed in practice. The concrete form of the observations and measurement errors need to be taken into account. Moreover, certain parametrizations as well as insufficient experimental data may lead to a situation in which parameters cannot be estimated unambiguously. An analysis of parameter identifiability must account for that. This talk presents an approach to model leukemia treatment in children based on ODEs. In order to study the issue of parameter identifiability, profile likelihoods are computed to assess the uncertainty of parameter estimates. The profile likelihoods also serve as starting points to enhance the experimental design for future experiments such that measurements can be taken more efficiently, thereby improving the reliability of statistical results.

Bayes estimation of probit choice models with {RprobitB}

Lennart Oelschläger

Universität Bielefeld, Fakultät für Wirtschaftswissenschaften
lennart.oelschlaeger@uni-bielefeld.de

Around a year ago, the initial version of RprobitB was released - an R package for fitting the multinomial probit model to discrete choice data to understand the driving factors behind decisions. Since then, we have continued to enhance and improve the package. In this talk, I present the status quo as well as future goals. The following are some of the subjects covered, both theoretically and applied directly in R: The Dirichlet process has been incorporated to flexibly address choice behavior heterogeneity. Several tools for model selection are now available, including the Bayes factor and the WAIC value. A function for choice prediction has been added to the package. These extensions are useful for empirical choice applications, where preliminary results in the fields of contraception and chess will complete the presentation.

Determinants of regional mobility after job loss in Germany - the importance of economic and non-economic factors in migration decisions of displaced workers.

Katrin Rickmeier

Universität Bielefeld, Fakultät für Wirtschaftswissenschaften
katrin.rickmeier@uni-bielefeld.de

Even though research has shown that job loss leads to a higher propensity of regional mobility, it is not yet clear how job displacement affects the decision to migrate. I am closing this research gap by studying the determinants of regional mobility after job loss in Germany. As already established in the literature, migration decisions are at least partly driven by economic reasons, such as regional disparities in job prospects. However, non-economic factors like ties to one's relatives play an important role in determining migration decisions and have been neglected in the literature so far. Therefore, I investigate spatial inequalities in local labour markets but also the relationships to one's relatives and one's own local connections as determinants for moving decisions. Using data from the German Socio-Economic Panel study, I am evaluating the importance of local labour market conditions and family ties on the regional mobility behaviour of displaced workers.

Weighting strategies for pairwise composite marginal likelihood estimation

Sebastian Büscher

Universität Bielefeld, Fakultät für Wirtschaftswissenschaften
sebastian.buescher@uni-bielefeld.de

The composite marginal likelihood (CML) estimation approach is an alternative to the maximum likelihood (ML) estimation of multinomial probit (MNP) models, usually employed to understand the effects of different variables on repeated discrete choice occasions. Whereas MNP models improve upon some of the drawbacks and restricting assumptions of the multinomial logit models, the ML estimation of MNP models suffers greatly from large computational costs. In the CML estimation approach, instead of looking at the joint probability of all decisions of one decision maker, the product of the probabilities of pairs of decisions is calculated. This drastically reduces the required computational cost. There are, however, some considerations to be made with the CML approach:

1. The number of pairs for each decision maker depends quadratically on the number of observations of the decision maker, and for different numbers of observations per decision maker, each decision is included in differing numbers of CML pairs.
2. The possibility that different pairs contain different amounts of information, depending on the distance between the involved decision instances, should be considered.

The CML framework allows us to give different weights to different CML pairs to account for these and other challenges. In this talk, we will discuss some of the reasoning behind using different weighting strategies, discuss the effects on the statistical properties of the resulting CML estimator, introduce some new weighting strategies to address these issues, and have a look at some first results obtained using synthetic data.