UNIVERSITÄT BIELEFELD
Fakultät für Wirtschaftswissenschaften

Zentrum für Statistik

16th Young Researchers Workshop of Centre for Statistics (ZeSt)

27th November 2025

Programme

9:30 - 9:35	Welcoming
9:35 - 10:05	Jan-Ole Koslik: Automatic and efficient selection of multiple LASSO penalties for complex likelihood models
10:05 - 10:35	Maya Vienken: Inference on state occupancy in covariate-driven hidden Markov models
10:35 - 11:05	Michael Balzer: Gradient Boosting for Spatial Regression Models with Autoregressive Disturbances
11:05 - 11:20	Coffee Break
11:20 - 11:50	Jonas Bauer: Modelling football fever of Arminia Bielefeld fans during the 2025 cup final of the German football association (DFB)
11:50 - 12:20	Julian Wäsche: Experimental design for dynamic models and small sample sizes
12:20 - 12:30	Discussion and Closing Remarks

Automatic and efficient selection of multiple LASSO penalties for complex likelihood models

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

Regularisation via the LASSO is a powerful tool for variable selection and shrinkage in regression models, yielding more parsimonious representations in high-dimensional settings. However, in complex likelihood-based frameworks — such as generalised additive models for location, scale and shape (GAMLSS), hidden Markov models (HMMs), or copula-based joint regressions — its practical use is often hindered by the need to select multiple penalty parameters simultaneously. Standard heuristic approaches based on cross-validation or information criteria quickly become computationally prohibitive as the number of penalties increases.

In this talk, I present a likelihood-based alternative for automatic and stable selection of LASSO penalties, inspired by the analogy between smoothness selection for penalised splines and variance component estimation in mixed models. Viewing the LASSO as a model with Laplace-distributed random effects, we derive a restricted likelihood by marginalising over these effects using a Laplace approximation and obtain remarkably simple closed-form updating equations for the penalty parameters. The resulting algorithm is straightforward to implement and numerically efficient. Owing to the simplicity of the updating equations, the method naturally accommodates arbitrary reparameterisations of the coefficients, making it broadly applicable across a wide range of likelihood-based models. We demonstrate the method's efficacy through preliminary simulation results.

Inference on state occupancy in covariate-driven hidden Markov models

Maya Vienken
Faculty of Business Administration and Economics, Bielefeld University, Germany
maya.vienken@uni-bielefeld.de

Hidden Markov models have become immensely popular tools for analysing animal behaviour based on movement, acceleration and other sensor data. In particular, these models allow to infer how the animal decision-making process interacts with internal and external drivers, by relating the probabilities of switching between distinct behavioural states to covariates.

A key challenge arising in the statistical analysis of behavioural data using covariate-driven HMMs is the models' interpretation, especially when there are more than two states, as then several functional relationships between state-switching probabilities and covariates need to be jointly interpreted. The model-implied probabilities of occupying the different states, as a function of a covariate of interest, constitute a much simpler and hence useful summary statistic.

A pragmatic approximation of the state occupancy distribution, namely the hypothetical stationary distribution of the model's underlying Markov chain for fixed covariate values, has in fact routinely been reported in HMM-based analyses of ecological data. However, for stochastically varying covariate processes with relatively little persistence, we show that this approximation can be severely biased, hence potentially invalidating ecological inference based on the approximate version of this important summary statistic of interest.

In this contribution, we develop two alternative approaches for obtaining the state occupancy distribution as a function of a covariate of interest — one based on an additional model fitted to the covariate process, the other obtained by regression analysis of the model-implied state probabilities. The practical application of these approaches are demonstrated in simulations and a case study on giant Galápagos tortoise movement data.

Our methods enable practitioners to conduct unbiased inference on the relationship between animal behaviour and general types of covariates, thus allowing to uncover the factors influencing behavioural decisions made by animals.

Gradient Boosting for Spatial Regression Models with Autoregressive Disturbances

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

Researchers in urban and regional studies increasingly work with high-dimensional spatial data that captures spatial patterns and spatial dependencies between observations. To address the unique characteristics of spatial data, various spatial regression models have been developed. In this article, a novel model-based gradient boosting algorithm tailored for spatial regression models with autoregressive disturbances is proposed. Due to its modular nature, the approach offers an alternative estimation procedure with interpretable results that remains feasible even in high-dimensional settings where traditional quasi-maximum likelihood or generalized method of moments estimators may fail to yield unique solutions. The approach also enables data-driven variable and model selection in both low- and high-dimensional settings. Since the bias-variance trade-off is additionally controlled for within the algorithm, it imposes implicit regularization which enhances predictive accuracy on out-of-sample spatial data. Detailed simulation studies regarding the performance of estimation, prediction and variable selection in low- and high-dimensional settings support proper functionality of the proposed methodology. To illustrative the applicability of the model-based gradient boosting algorithm, a case study is presented where the life expectancy in German districts is modeled, incorporating a potential spatial dependence structure.

Modelling football fever of Arminia Bielefeld fans during the 2025 cup final of the German football association (DFB)

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

Fans are devoted supporters of a football club, whose emotional attachment to the team drives high engagement, passion, and loyalty that can seem beyond reason. We call this bond "football fever" because it manifests physiological, behavioral and psychological symptoms and can spread and evolve. Understanding the nature of football fever is of interest for many stakeholders of the global football industry as well as for researchers in sports marketing, social sciences and behavioral psychology. Previous work mainly used cross-sectional questionnaires and interviews of spectators, offering only a snapshot in time. This year, during the DFB Cup, the football fever grew gradually as the success of DSC Arminia Bielefeld rose, reaching a peak at the final in Berlin. At that moment, the tense atmosphere among fans was palpable. To study this phenomenon in detail, Bielefeld University and Wissenswerkstatt Bielefeld collected a rich wearable-technology dataset from 197 fans, recording various physiological measures (e.g. heart rate and stress levels) at high temporal resolution during the cup final. This dataset makes it possible to measure football fever and analyze how it evolves over time in ways not previously possible. In this talk, we present the data and compare four strategies to incorporate "time" in structural equation models of the latent variable football fever.

Experimental design for dynamic models and small sample sizes

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

It has become considerably easier to collect large amounts of experimental data in order to study underlying processes giving rise to scientific phenomena. However, in disciplines such as biology, medicine, or engineering, experimental data generation can be costly and labour-intensive. In these cases, careful experimental design is crucial to use limited resources efficiently and to obtain statistically informative experiments.

Focusing on biomedical applications, we compare different experimental design approaches to determine informative measurement times for reliable parameter estimation in population growth models when only few observations are available. In a simulation study inspired by preclinical trials on gene modifications in leukaemia cells, we assess how these designs affect the precision of parameter estimates. Our results indicate that profile likelihood-based approaches are more promising than a standard Fisher information matrix-based approach in terms of applicability and reducing parameter uncertainty. These findings demonstrate that appropriate design strategies can enhance the informativeness of experiments while making more efficient use of available resources.