Universiteit
Leiden

Leibniz-Institut
für Präventionsforschung
und Epidemiologie – BIPS

ZeSOB organizes the

Joint PhD Seminar in Statistics

for:

University of Bremen,
Carl von Ossietzky University, Oldenburg,
Leiden University,
University of Padova,
University of Oslo

October 4-6, 2023
Bremen, Germany

Zoom link for online participants:

`https://uni-bremen.zoom-x.de/j/61906852066?pwd=NXh5bzhPSnZ4VlV2dCtkVjBFOEx0Zz09`

Meeting-ID: 619 0685 2066
Code: 598591

Welcome

Dear attendees,

welcome to University of Bremen for the joint PhD Seminar in Statistics among the universities of Bremen, Oldenburg, Leiden, Padova and Oslo. The scope of this seminar is to bring together PhD students from the Northern universities in the mentioned fields and to foster discussions among you both on topics of your research and on general PhD related questions. Moreover, the opportunity to present your results in an extended format of 40 minutes allows the colleagues from other places to get a deeper insight into current research of the other groups and will hopefully give you valuable feedback for your work on top of your advisors'. Of course, as any pressing deadline, such a presentation will also push forward your work and get you closer to successfully finishing your thesis.

Speakers

Kristin Blesch, BIPS
Andrea Mathisen Bratsberg, University of Oslo
Justus Contzen, University of Bremen
Chiara Degan, Leiden University
Jelle Goeman, Leiden University
Vincent Jankovic, University of Bremen
Sophie Langbein, BIPS
Remi Luschei, University of Bremen
Daniel Ochieng Odipo, University of Bremen
Pascal Rink, University of Bremen
Riccardo De Santis, University of Padova
Marije Sluiskes, Leiden University
Oke Wübbenhorst, Oldenburg University

Organizing Team

Werner Brannath
Thorsten Dickhaus
Peter Ruckdeschel
Jelle Goeman
Marvin Wright

How to reach University of Bremen

By Car (via A1 or A27):

- Coming from the A1, exit "Bremer Kreuz" onto the A27 towards "Bremerhaven"
- Ausfahrt 19 "Horn-Lehe/Universität" in direction of "Universität"
- At the first traffic light turn right into "Hochschulring"
- After 200 meters turn right into "Linzer Straße".
- After 150 meters you will reach the building at Linzer Straße 4 on the left
- The entrance (Linzer Straße 4) is not easy to find.
- Cross the parking to the end (behind the company *swooop*).
- You will find us on the ground floor on the left.

By tram and bus:

- Tram line 6 from the airport or from the central station (platform E) in the direction of the university
- Get off either at the final stop "Universität-Nord" and follow the street "Hochschulring" towards the Fallturm, and continue to "Linzer Straße 4"
- Or get off at "Universität Zentralbereich" and then take bus number 22 or 31 to reach "Linzer Straße"

Programme Overview

Wednesday, October 4, 2023

15:00 - 15:10 **Welcome Session**
— *Werner Brannath*, University of Bremen

Block 1: Modelling
— Chair: *Thorsten Dickhaus*, University of Bremen

15:10 - 15:50 *Justus Contzen*, University of Bremen: An investigation of present and future climate extremes using climate model data

15:50 - 16:30 *Chiara Degan*, Leiden University: Multivariate longitudinal modeling with bounded outcomes and endogenous covariates

18:00 Dinner at Platzhirsch (Kuhgrabenweg 30, 28359 Bremen)

Thursday, October 5, 2023

Block 2: Machine Learning
— Chair: *Werner Brannath*, University of Bremen

09:30 - 10:10 *Pascal Rink*, University of Bremen: Simultaneous Bootstrap Tilting Confidence Intervals with Applications in Machine Learning
10:10 - 10:50 *Kristin Blesch*, BIPS: Statistical considerations to improve interpretable machine learning for mixed tabular data

10:50 - 11:10 Coffee Break

Block 3: Survival Data
— Chair: *Marvin Wright*, BIPS

11:10 - 11:50 *Marije Sluiskes*, Leiden University: A reduced rank proportional hazards model for multiple outcome survival data
11:50 - 12:30 *Sophie Langbein*, BIPS: Interpretable Machine Learning for Survival Analysis

12:30 - 14:10 Lunch Break at University Mensa

Block 4: Dealing with dependence
— Chair: *Jelle Goeman*, Leiden University

14:10 - 14:50 *Vincent Jankovic*, University of Bremen: Asymptotic online FWER control for dependent test statistics
14:50 - 15:30 *Andrea Mathisen Bratsberg*, University of Oslo: Conditional variable screening for ultra high dimensional longitudinal data with time interactions

15:30 - 15:50 Coffee Break

Block 5: Semiparametric Models
— Chair: *Peter Ruckdeschel*, Oldenburg University

15:50 - 16:30 *Riccardo De Santis*, University of Padova: Group invariance tests for regression models
16:30 - 17:10 *Oke Wübbenhorst*, Oldenburg University: Next Steps toward a Near to Optimally Robust Cox Regression Estimator

18:00 Dinner at Matisse (Vorstraße 79, 28359 Bremen)

Friday, October 6, 2023

Block 6: Hypothesis Testing
— Chair: *Thorsten Dickhaus*, University of Bremen

09:30 - 10:10 *Daniel Ochieng Odipo*, University of Bremen: On the usage of single and two stage randomized p-values in the Schweder Spjøtvoll estimator

10:10 - 10:50 *Remi Luschei*, University of Bremen: The effect of estimating prevalences on the population wise error rate

10:50 - 11:10 Coffee Break

ZeSOB Colloquium
— Chair: *Werner Brannath*, University of Bremen

11:10 - 12:10 *Prof. Jelle Goeman*, Leiden University: All resolutions inference: adaptively choosing the scale of your analysis

12:10 - 12:20 *Werner Brannath*, University of Bremen: Closing Remarks

12:20 Lunch at Café Unique (Enrique Schmidt Str. 7, 28259 Bremen)

Abstracts

Block 1: Modelling

An investigation of present and future climate extremes using climate model data

Justus Contzen

University of Bremen

15:10 - 15:50, Wednesday, October 4, Block 1

Climate models are used to simulate how the Earth's climate could change in the near future. We investigate results of climate models with regard to climate extremes, as a potential increase of strength and frequency of climate extremes can have severe consequences. We first analyze data from simulations of the historical climate and compare them with real-world data in order to assess the reliability of climate models with respect to extremes. Then, we investigate simulations of the future climate, and we analyze how strongly and how quickly extremes are predicted to change at different geographic locations. Finally, we discuss some methods to build statistical models for the multivariate analysis of extremes that can be used to investigate spatial dependencies of extremes.

Multivariate longitudinal modeling with bounded outcomes and endogenous covariates

Chiara Degan

Leiden University

15:50-16:30, Wednesday, October 4, Block 1

Measuring the relationship between bounded longitudinal responses and endogenous time-varying covariates is not always trivial, and the use of simple Beta Mixed Models is no longer appropriate for mainly two reasons. First, it introduces bias when it fails to properly account for the dependence between the endogenous variable and the outcome history. Second, the longitudinal response and covariate can be measured at different points in time and may contain missing values.
Multivariate models, on the other hand, could be utilized to analyze the relationship between the response and the endogenous variables. In this talk we consider two types of multivariate models, each assuming a different form of the association. One induces the association by jointly modelling the random effects, called Joint Mixed Models; the other quantifies the association using a scaling factor, called Joint Scaled Models. However, fitting these models is not straightforward, and their computational intensity, due to the high-dimensional integration over the random effects terms, limits their applicability. A flexible Bayesian estimation approach, known as INLA, will be used to fill this gap. Analytical

work on these models will be presented along with the results of a clinical study conducted at Leiden University Medical Center

Block 2: Machine Learning

Simultaneous Bootstrap Tilting Confidence Intervals with Applications in Machine Learning

Pascal Rink

University of Bremen

09:30 - 10:10, Thursday, October 5, Block 2

Bootstrap tilting is a universal technique to improve the accuracy of statistical parameter estimation and confidence interval construction. It is particularly useful when dealing with non-normally distributed data. Bootstrap tilting combines the generation of bootstrap samples from the data at hand and the application of a transformation on each of these bootstrap samples, accounting for distributional characteristics of the data. This transformation is designed such that the transformed bootstrap sample resembles a specific target distribution, for example, a specific null distribution, which is helpful when estimating a confidence interval. In this talk I will present a way how to extend bootstrap tilting confidence intervals to a multiple testing setup. In particular, we propose a maxT-type multiplicity correction in order to obtain simultaneous confidence intervals for multiple statistical parameters. The resulting statistical inference is valid, as the proposed simultaneous confidence intervals are consistent for mean-type parameters. I will also present an application of this approach to the evaluation of machine learning prediction models.

Statistical considerations to improve interpretable machine learning for mixed tabular data

Kristin Blesch

BIPS, Bremen

10:10 - 10:50, Thursday, October 5, Block 2

Interpretable machine learning techniques aim at shedding light on the behavior of opaque machine learning algorithms, but oftentimes fail to acknowledge the challenges real-world data imposes on the task. In specific, the fact that empirical tabular datasets typically incorporate dependency structures and may consist of both continuous and categorical features (mixed data) is frequently overlooked. This work takes on a statistical perspective to illuminate the far-reaching implications mixed data with a dependency structure has for interpretability. Methodology that is suitable for this kind of data is proposed and simulations as well as empirical applications reveal that favorable outcomes in interpretability and machine learning more broadly can be achieved.

Block 3: Survival Data

A reduced rank proportional hazards model for multiple outcome survival data. With applications to age-related morbidity event data

Marije Sluiskes

Leiden University

11:10 - 11:50, Thursday, October 5, Block 3

The identification of biomarkers of aging is an important biomedical research theme. Most current statistical methods that aim to capture the aging process either use chronological age or time-to-mortality as the outcome of interest. There is however a shift in the field towards the study of health span and patterns of age-related multimorbidity, as aging entails more than lifespan duration alone.
Several large epidemiological studies, such as the UK Biobank and the Leiden Longevity Study, have recently incorporated detailed age-at-disease-onset profiles, obtained from electronic health records. The availability of these data opens new analytical possibilities. Nevertheless, analyses conducted thus far oversimplify the complexity of multimorbidity patterns, for instance by ignoring information on age-at-disease-onset or by failing to acknowledge that age-related diseases are likely driven by a shared set of underlying factors.
We propose a new methodological framework for the analysis of age-related multimorbidity data, based on multiple-outcome survival modelling. Specifically, we propose to use a reduced rank proportional hazards model. This model can be fitted on the (possibly right-censored and left-truncated) age-at-disease-onset of several age-related diseases simultaneously. It assumes that there is a set of shared latent factors that drive all age-related diseases considered, thereby reducing the dimensionality of the problem and providing additional insight into different facets of the aging process. As there is a large interest in the use of high-dimensional omics data as potential biomarkers of aging, we also discuss how to include (lasso-)penalization in the reduced rank proportional hazards framework.

Interpretable Machine Learning for Survival Analysis

Sophie Langbein

BIPS, Bremen

11:50 - 12:30, Friday, February 26th, Block 1

With the spread and rapid advancement of black box machine learning models, the field of interpretable machine learning (IML) has become increasingly important over the last decade. This is particularly relevant for survival analysis, where explainability can uncover a model's potential biases and limitations and provide more mathematically sound ways to understand how and which features are influential for prediction or constitute risk factors. However, the lack of readily available IML methods may have deterred researchers from

leveraging the full potential of machine learning for analyzing time-to-event data. We formally adapted commonly used IML methods to survival outcomes and implemented methods such as PDP, ICE, ALE, LIME and SHAP in the 'survex' package in R. The software design focuses on compatibility with existing packages for machine learning survival models, thereby enhancing accessibility for researchers and practitioners in the field of survival analysis. We show how the 'survex' package can be utilized to provide human-interpretable insights into the decision-making process of black box survival machine learning models. Thus, the 'survex' can help to promote transparency and accountability in sensitive areas, such as clinical decision making processes, the development of targeted therapies, interventions or in other healthcare related contexts.

Block 4: Dealing with dependence

Asymptotic online FWER control for dependent test statistics

Vincent Jankovic

University of Bremen

14:10 - 14:50, Thursday, October 5, Block 4

In online multiple testing, an a priori unknown number of hypotheses is tested sequentially, i.e. at each time point a test decision for the current hypothesis has to be made using only the data available so far. Although many powerful test procedures have been developed for online error control in recent years, most of them are designed solely for independent or at most locally dependent test statistics. In this work, we provide a new framework for deriving online multiple test procedures which ensure asymptotical (with respect to the sample size) control of the familywise error rate (FWER), regardless of the dependence structure between test statistics. In addition, we give a few concrete examples of such test procedures and discuss their properties. Furthermore, we conduct a simulation study to investigate the behaviour regarding type I error and power of these test procedures for finite sample size.

Conditional variable screening for ultra high dimensional longitudinal data with time interactions

Andrea Mathisen Bratsberg

University of Oslo

14:50 - 15:30, Thursday, October 5, Block 4

Recently we have been able to gather large amounts of genetic data at a fast rate, creating situations where the number of variables greatly exceeds the number of observations. In the setting of longitudinal data, the correlations between responses introduce additional challenges, because most models that can deal with correlations become computationally

infeasible for truly high-dimensional variables. Hence, there is a need for a pre-screening of variables to reduce the dimension efficiently and accurately to a more moderate scale. Additionally, it is of interest to capture possible interactions between the genetic variable and time in many of these longitudinal studies. For this reason, we propose a novel screening procedure that ranks variables according to the likelihood value at the maximum likelihood estimates in a semi-marginal linear mixed model, where the gene variable and its interaction with time is included in the model. This is to our knowledge the first conditional screening approach for clustered data.

Block 5: Semiparametric Models

Group invariance tests for regression models

Riccardo De Santis

University of Padova

15:50 - 16:30, Thursday, October 5, Block 5

Regression models are a popular tool to exploit the relationship between some covariates and a given outcome of interest. The relevance of any target covariate is further exploited by means of a statistical test; the standard parametric approach is based on a full set of assumptions and, for nonlinear models, it relies on asymptotic results which may be unsatisfactory for small sample size. We propose a semiparametric method for statistical testing to relax some of those assumptions and improve the small-sample performance, based on the group-invariance approach. We will exploit the validity of our proposal through theoretical results and simulation studies.

Next Steps toward a Near-to-Optimally-Robust Cox Regression Estimator

Oke Wübbenhorst

Oldenburg University

16:30 - 17:10, Thursday, October 5, Block 5

We discuss robustness aspects for Cox regression. The overall goal is to find robust alternatives to the Cox proportional hazard estimator [1] which (a) come up with a decent breakdown point, (b) have bounded but highly efficient influence functions, and (c) retain the adaptivity from Cox's partial likelihood estimator w.r.t. an unknown fully non-parametric hazard rate.
So far, existant robust alternatives do achieve (c), but often [2] only have bounded influence functions when compactly supported weights are used, which then comes with the cost of a low efficiency at the ideal model. At the same time existant robust alternatives [2,3] have unsatisfactory breakdown points of order 1/k, k the number of parametric regressors.

We head for one-step constructions, which start from a high-breakdown, robust, but inefficient starting estimator, and then are enhanced by a reweighting step to achieve near-to-optimally robust efficiency.
To this end, a certain infinite dimensional orthogonality condition has to be fulfilled. Improving upon [3], we thin out the infinite dimensional condition to only be taken up for the most relevant, finitely many nuisance tangents through a LASSO procedure [4]. The reweighting step then takes up these most relevant nuisance tangents and gauges the estimator to a higher efficiency in the ideal model while still bounding the bias. Hence we achieve (b). At the same time we maintain the breakdown point of the starting estimator, hence achieve (a). We do not reach complete adaptivity (c) w.r.t. to the nuisance parameter, the unknown hazard rate, but the incurred bias in this case is controlled through the boundedness of the influence function. I.e., bias in our case enters through both outliers and the imcomplete adaptation.
In this talk we (i) show how to achieve a high breakdown fully adapted starting estimator, (ii) repeat how to achieve evidence-guided partial adaptation and (iii) present the form of the optimally-robust influence function.

Block 6: Hypothesis testing

On the usage of single and two stage randomized p-values in the Schweder Spjøtvoll estimator

Daniel Ochieng Odipo

University of Bremen

09:30 - 10:10, Friday, October 6, Block 6

In this article, we consider single-and two-stage randomized p-values for testing composite null hypothesis in discrete models. In particular, we investigate by use of mathematical proofs if these p-values are (i) stochastically ordered in $c \in (0, 1]$, or c^*, which is a support point for the LFC based p-value. We also prove (ii) if the p-values are strictly increasing with an increase in the sample size and if (iii) they are uniformly conservative. An analysis of the behavior of the CDF of the p-values under the alternative hypothesis for different values of θ is also considered. It is shown that it is possible for the CDF of the single-stage randomized p-value in the alternative hypothesis to exceed the CDF of the LFC-based p-value. The CDF of the two-stage randomized p-value, on the other hand, can never coincide with the one for the UMP p-value. A detailed analysis of the bias and mean square error (MSE) of the Schweder–Spjøtvoll estimator is considered. The MSE is considered for the case when the p-values are assumed to be independent and for the case when dependence is assumed to exist. Values of c that minimize the bias and MSE are also found.

The effect of estimating prevalences on the population-wise error rate

Remi Luschei

University of Bremen

10:10 - 10:50, Friday, October 6, Block 6

The population-wise error rate (PWER) is a type I error rate for clinical trials with multiple target populations. It is defined as the probability that a randomly selected, future patient will be exposed to an inefficient treatment based on the study results. By only considering type I errors that are relevant in this setting, the PWER is more liberal than the family-wise error rate. In practice, however, the prevalences of the disjoint population strata needed to compute the PWER are often not known and must be estimated from the study sample. In this talk, I will examine the impact of estimating these prevalences on the true PWER. I will present results of simulations where they are estimated by the maximum-likelihood estimator from a multinomial distribution, showing that adequate PWER control is reached for realistic sample sizes. Finally, I will also consider the maximum family-wise error rate for the disjoint population strata that results from PWER-control.

ZeSOB Colloquium

All-resolutions inference: adaptively choosing the scale of your analysis

Prof. Jelle Goeman

Leiden University

11:10 - 12:10, Friday, October 6, Block 7

Many fields of science nowadays gather data at a very fine resolution but do inference at a higher aggregated level. For example, in neuroimaging data are gathered at the level of 3 mm x 3 mm x 3 mm voxels, but the relevant biology happens at the level of cm-scale brain areas; in genetics, data are gathered at the level of single-DNA-base polymorphisms, but interesting questions happen at the level of genes or even gene groups; in spatial statistics, data may be gathered at street level but interesting questions are about neighbourhoods or provinces. Often, there is not just one natural scale or one natural way to aggregate data to prepare for inference. Multiple alternative criteria could be used to drive the grouping. Aggregation to large regions may give low specificity; more limited aggregation may give low power. Moreover, there is a huge multiplicity, since m spatial units give rise to 2^m regions of potential interest. Without correction for multiple testing, it is easy to be wrong-footed by seemingly convincing clusters of signal. This talk presents how Closed Testing can be used to tackle this huge multiple testing problem, and to analyze this type data

at all resolutions simultaneously. The method allows the choice how and how much to aggregate to be chosen freely by the researcher, in a data-dependent way, while still strictly controlling the probability of false positive findings. This allows researchers to adapt the inference to localize and quantify the signal that is present in the data: the stronger the signal, the better it will be pinpointed by the closed testing procedure.
I will review the general idea and theory of closed testing and recent progress in method development in this area. Several example contexts illustrate the wide applicability of all-resolutions inference.

Presentations