SAM 2024 Program

Naturwissenschaftliche Fakultät Building

Hellbrunner Strasse 34, 5020 Salzburg, Austria

Day 1: July 9, 2024

Opening Remarks: 8:50-9:00, July 9, 2024

Room: HS 401

Lei Liu and Arne Bathke

IBS-ROeS Keynote Speech Keynote Speech I: 9:00-10:00, July 9, 2024

Room: HS 401

Chair: Prof. Yichuan Zhao

Presenter: Ian McKeague, Columbia University

Title: Nonparametric methods for wearable device data

Abstract: This talk discusses a nonparametric inference framework for occupation time curves derived from wearable device data. Such curves provide the total time a subject maintains activity above a given level as a function of that level. Taking advantage of the monotonicity property of these curves, we develop a likelihood ratio approach to construct confidence bands for mean occupation time curves. An extension to fitting concurrent functional linear regression models is also developed. Application to wearable device data from an ongoing study of an experimental gene therapy for mitochondrial DNA depletion syndrome will be discussed. Based on joint work with Hsin-Wen Chang (Academia Sinica).

Break 10:00-10:15, July 9, 2024

Parallel Invited Sessions 1-4: 10:15-12:00, July 9, 2024

Invited Session 1: Joint Modeling of Complex Survival Data

Organizer: Xinyuan Song. Chair: Lan Luo. Room: HS 402

Presenter: Xingqiu Zhao, The Hong Kong Polytechnic University

Title: Deep Nonparametric Inference for Conditional Hazard Function

Abstract: We propose a novel deep learning approach to nonparametric statistical inference for the conditional hazard function of survival time with right-censored data. We use a deep neural network (DNN) to approximate the logarithm of a conditional hazard function given covariates and obtain a DNN likelihood-based estimator of the conditional hazard function. Such an estimation approach grants model flexibility and hence relaxes structural and functional assumptions on conditional hazard or survival functions. We establish the consistency, convergence rate, and functional asymptotic normality of the proposed estimator. Subsequently, we develop new one-sample tests for goodness-of-fit evaluation and two-sample tests for treatment comparison. Both simulation studies and real application analysis show superior performances of the proposed estimators and tests in comparison with existing methods.

Presenter: Liming Xiang, Nanyang Tech University

Title: Multiple Imputation for Flexible Modelling of Interval-censored Data with Covariates Subject to Missingness and Detection Limits

Abstract: Interval-censored failure time data is popular in biomedical studies when a failure time is not observed exactly but only known to lie in an interval obtained from a sequence of examination times. The presence of covariates subject to missingness and detection limits poses challenges for regression analysis of interval-censored data and necessitates an effective statistical method. We propose a novel multiple imputation approach via rejection sampling for analysis of such data under semiparametric transformation models. Our proposal alleviates strong dependence of the usual imputation methods on the choice of imputation models and yields consistently and asymptotically normal estimators of the regression parameters. Simulation studies demonstrate that the proposed approach is flexible and leads to more efficient estimation than the complete analysis and augmented inverse probability weighting analysis in various practical situations. Finally, we apply the proposed approach to an Alzheimer's disease data set that motivates this study.

Presenter: Jun Ma, Macquarie University

Title: Joint modelling of longitudinal covariates and partly-interval censored survival data - a penalized likelihood approach

Abstract: This talk will focus on a joint modelling of longitudinal covariates and partly interval censored time-to-event data. Longitudinal time-varying covariates play a crucial role in achieving accurate dynamic predictions using a survival regression model. However, these covariates are often measured at limited time points and may contain measurement errors. Moreover, they are usually specific to each individual. On the other hand, the event times of interest are often interval-censored. Accounting for all these factors is essential when constructing a survival model. We will present a new approach for joint modelling of the longitudinal time-varying covariates and the time-to-event Cox model, where the latter is subject to interval censoring. We will develop a novel maximum penalized likelihood approach for estimation of all the model parameters including the random effects. A profile likelihood is used to obtain the covariance matrix of the estimated parameters.

Presenter: Xinyuan Song, Chinese University of Hong Kong

Title: Bayesian tree-based heterogeneous mediation analysis with a time-to-event outcome

Abstract: Mediation analysis aims at quantifying and explaining the underlying causal mechanism between an exposure and an outcome of interest. In the context of survival analysis, mediation models have been widely used to achieve causal interpretation for the direct and indirect effects on the survival of interest. Although heterogeneity in treatment effect is drawing increasing attention in biomedical studies, none of the existing methods have accommodated the presence of heterogeneous causal pathways pointing to a time-to-event outcome. In this study, we consider a heterogeneous mediation analysis for survival data based on a Bayesian tree-based Cox proportional hazards model with shared topologies. Under the potential outcomes framework, individual-specific conditional direct and indirect effects are derived on the scale of the logarithm of hazards, survival probability, and restricted mean survival time. A Bayesian approach with efficient sampling strategies is developed to estimate the conditional causal effects through the Monte Carlo implementation of the mediation formula. Simulation studies show the satisfactory performance of the proposed method. The proposed model is then applied to an HIV dataset extracted from the ACTG175 study to demonstrate its usage in detecting heterogeneous causal pathways.

Invited Session 2: Analysis of multi-dimensional correlated data

Organizer: Peter Song. Chair: Peter Song. Room: HS 401

Presenter: Annie Qu, University of California, Irvine

Title: Optimal Individualized Treatment Rule for Combination Treatments under Budget Constraints

Abstract: The individualized treatment rule (ITR), which recommends an optimal treatment based on individual characteristics, has drawn considerable interest from many areas such as precision medicine, personalized education, and personalized marketing. Existing ITR estimation methods mainly adopt one of two or more treatments. However, a combination of multiple treatments could be more powerful in various areas. In this paper, we propose a novel Double Encoder Model (DEM) to estimate the individualized treatment rule for combination treatments. The proposed double encoder model is a nonparametric model which not only flexibly incorporates complex treatment effects and interaction effects among treatments, but also improves estimation efficiency via the parameter-sharing feature. In addition, we tailor the estimated ITR to budget constraints through a multi-choice knapsack formulation, which enhances our proposed method under restricted-resource scenarios. In theory, we provide the value reduction bound with or without budget constraints, and an improved convergence rate with respect to the number of treatments under the DEM. Our simulation studies show that the proposed method outperforms the existing ITR estimation in various settings. We also demonstrate the superior performance of the proposed method in PDX data that recommends optimal combination treatments to shrink the tumor size of the colorectal cancer.

Presenter: Michael Elliott, University of Michigan

Title: Using Variability in Longitudinally-Measured Variables as a Predictor of Health Outcomes

Abstract: Longitudinal data has become a major part of the landscape for clinical and epidemiological research. While variance is typically understood as nuisance – the "noise" in "signal-to-noise" – there is increasing evidence that underlying variability in subject-level measures over time may also be important in predicting future health outcomes of interest. However, most statistical methods development has been focused on the use of mean trends obtained from longitudinal data; approaches that incorporate subject-level variability are far rarer and consequently use of such information is rare. I will provide a review of several methods developed to incorporate variability of predictors in a range of statistical modeling settings, with a deeper dive on a specific application where we consider how a woman's mean and variability trends of multivariate hormonal measures during menopausal transition can impact measures of post-menopausal health outcomes.

Presenter: Ji Zhu, University of Michigan

Title: A Latent Space Model for Hypergraphs with Diversity and Heterogeneous Popularity

Abstract: 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.

Presenter: Jian Kang, University of Michigan

Title: Bayesian Methods for Brain-Computer Interfaces

Abstract: A brain-computer interface (BCI) is a system that translates brain activity into commands to operate technology. BCIs help people with disabilities use technology for communication. A common design for an electroencephalogram (EEG) BCI relies on the classification of the P300 event-related potential (ERP), which is a response elicited by the rare occurrence of target stimuli among common non-target stimuli. Existing studies have focused on constructing the ERP classifiers, but few provide insights into the underlying mechanism of the neural activity. In this talk, I will discuss several new Bayesian methods for analyzing brain signals from BCI systems based on Gaussian Processes (GP). Our proposed methods can make statistical inferences about the spatial-temporal differences and dependence of the neural activity in response to external stimuli, which provides statistical evidence of P300 ERP responses and helps design user-specific profiles for efficient BCIs. Our inference results demonstrate the importance of ERPs from several brain regions for P300 speller performance. The robustness of our analysis is justified by cross-participant comparisons and extensive simulations.

Invited Session 3: Statistical Methods for Metagenomic Data

Organizer: Gen Li. Chair: Xiaonan Xue. Room: HS 414

Presenter: Hongzhe Li, University of Pennsylvannia

Title: Transfer Learning with Random Coefficient Ridge Regression for Microbiome Applications

Abstract: Ridge regression with random coefficients provides an important alternative to fixed coefficients regression in high dimensional setting when the effects are expected to be small but not zeros. Such models are particularly appropriate for microbiome-based prediction. This paper considers estimation and prediction of random coefficient ridge regression in the setting of transfer learning, where in addition to observations from the target model, source samples from different but possibly related regression models are available. The informativeness of the source model to the target model can be quantified by the correlation between the regression coefficients. This paper proposes two estimators of regression coefficients of the target model as

the weighted sum of the ridge estimates of both target and source models, where the weights can be determined by minimizing the limiting estimation risk or prediction risk. Using random matrix theory, the limiting values of the optimal weights are derived under the setting when $\frac{p}{n} \to \gamma$, where p is the number of the predictors and n is the sample size, which leads to an explicit expression of the estimation or prediction risks. We present results for several microbiome-based disease prediction, including IBD and colon cancer.

Presenter: Jiyuan Hu, New York Univesity

Title: Joint Modeling of Longitudinal Microbiome Data and Survival Outcome

Abstract: Recently more and more longitudinal microbiome studies are conducted to identify candidate microbes as biomarkers for the disease prognosis. We propose a novel joint modeling framework JointMM for longitudinal microbiome and time-to-event data to investigate the effect of dynamic changes of microbiome abundance profile on disease onset. JointMM comprises of two sub-models, i.e., the zero-inflated scaled-Beta mixed-effects regression sub-model aimed at depicting the temporal structure of microbial abundances among subjects; and the survival sub-model to characterize the occurrence of disease and its relationship with microbiome abundances changes. JointMM is specifically designed to handle the zero-inflated and highly skewed longitudinal microbiome abundance data and exhibits better interpretability that JointMM can examine whether the temporal microbial presence/absence pattern and/or the abundance dynamics would alter the time to disease onset. Comprehensive simulations and real data analyses demonstrated the statistical efficiency of JointMM compared with competing methods.

Presenter: Zhigang Li, University of Florida

Title: Estimating Equations with Inverse Probability Weighting for Microbiome Analysis

Abstract: Human microbiome data is collected in many research studies to investigate the role of microbiome in association with diseases or conditions. Sequencing technologies including 16S rRNA sequencing and metagenome shotgun sequencing are commonly used for quantifying microbiome data. However, it remains challenging to appropriately analyze microbiome data due to its unique features such as zero-inflated structure and compositional structure. We develop a novel approach to analyze microbiome data for differential abundance analyses or regression analyses. This approach employes inverse probability weighting techniques to account for the mixture of true and false zeros. GEE is used to account for the complicated inter-taxa correlation structure. The method does not require imputing zeros with a positive value for the data analysis. It has a good performance in comparison with existing methods in the simulation study. Application of the new approach in a real data set is also presented.

Presenter: Gen Li, University of Michigan

Title: Analysis of Microbiome Differential Abundance by Pooling Tobit Models

Abstract: Differential abundance analysis identifies microbiome taxa whose abundances differ between two or more conditions. Compositionality and sparsity of metagenomics sequencing data pose statistical challenges. We propose ADAPT (Analysis of Microbiome Differential Abundance by Pooling Tobit Models) as a solution. Count ratios between taxa satisfy subcompositional coherence. Zero counts can be regarded as left-censored at one. The tobit model is suitable for modeling left-censored count ratios. ADAPT first fits tobit models to relative abundances of individual taxa. It then selects a subset of non-differentially abundant taxa as the reference taxa set based on the estimated effect sizes and the distribution of p-values. Finally, tobit models for the count ratios between individual taxa and the reference taxa set reveal differentially abundant taxa. Simulation studies show that ADAPT has higher power than alternative methods while controlling false discovery rates. Application of ADAPT to early childhood dental caries data reveals differentially abundant oral bacteria species and functional genes between the saliva samples of children with and without dental caries.

Invited Session 4: Flexible statistical learning for complex data

Organizer: Yufeng Liu. Chair: Emily Hector. Room: HS 403

Presenter: Eric Chi, Rice University

Title: Proximal MCMC for Bayesian Inference of Constrained and Regularized Estimation

Abstract: Proximal Markov Chain Monte Carlo (MCMC) is a flexible and general Bayesian inference framework for constrained or regularized parametric estimation. The basic idea of proximal MCMC is to approximate non-smooth regularization terms via the Moreau-Yosida envelope. Initial proximal MCMC strategies, however, fixed nuisance and regularization parameters as constants and relied on the Langevin algorithm for the posterior sampling. Proximal MCMC is extended to a fully Bayesian framework with modeling and data-adaptive estimation of all parameters including regularization parameters. More efficient sampling algorithms such as the Hamiltonian Monte Carlo are employed to scale proximal MCMC to high-dimensional problems. The proposed proximal MCMC offers a versatile and modularized procedure for the inference of constrained and non-smooth problems that are mostly tuning parameter-free. Its utility is illustrated in various statistical estimation and machine-learning tasks.

Presenter: Ali Shojaie, University of Washington

Title: Learning Causal Effects of Multiple Covariates on Multiple Outcomes in High Dimensions

Abstract: We consider the problem of learning causal effects of multiple covariates on multiple outcomes. The problem is cast as a special instance of learning directed acyclic graphs (DAGs)

from partial or set orderings. We show that, unlike the simpler problem of learning DAGs from full causal orderings, DAG learning from partial orderings is computationally NP-hard. Building on recent developments for learning DAGs in high dimensions, we propose an efficient algorithm that learns the (direct) causal effects of covariates on outcome by leveraging the partial ordering and illustrate the advantages of the proposed algorithm over general-purpose DAG learning algorithms.

Presenter: Zhengyuan Zhu, Iowa State University

Title: Maximizing Benefits under Harm Constraints: A Generalized Linear Contextual Bandit Approach

Abstract: In many contextual sequential decision-making scenarios, such as dose-finding clinical trials for new drugs or personalized news article recommendation systems in social media, each action can simultaneously carry both benefits and potential harm. This could manifest as efficacy versus side effects in clinical trials or increased user engagement versus the risk of radicalization and psychological distress in news recommendation. These multifaceted situations can be modeled using the multi-armed bandit (MAB) framework. Given the intricate balance of positive and negative outcomes in these contexts, there is a compelling need to develop methods which can maximize benefits while limiting harm within the MAB framework. This paper addresses this gap by proposing a novel generalized linear contextual MAB model which balances the objectives of optimizing reward potential while limiting the harm and developing an ε-greedy-based policy which achieves a sublinear regret. Extensive experimental results are presented to support our theoretical analyses and validate the effectiveness of our proposed model and policy.

Presenter: Ping Ma, University of Georgia

Title: Analyzing CITE-seq Data via a Quantum Algorithm

Abstract: With the rapid development of quantum computers, researchers have shown quantum advantages in physics-oriented problems. Quantum algorithms tackling computational biology problems are still lacking. In this talk, I will demonstrate the quantum advantage of analyzing CITE-seq data. CITE-seq, a single-cell technology, enables researchers to simultaneously measure expressions of RNA and surface protein detected by antibody-derived tags (ADTs) in the same cells. CITE-seq data hold tremendous potential for elucidating RNA-ADT co-expression networks and identifying cell types effectively. However, both tasks are challenging since the best subset of ADTs needs to be identified from enormous candidate subsets. To surmount the challenge, I will present a quantum algorithm for analyzing CITE-seq data.

Lunch 12:00-1:00, July 9, 2024 (FREE for registrants)

Parallel Invited Sessions 5-8: 1:00-2:45, July 9, 2024

Invited Session 5: Analysis of data with multiple treatments and mixed outcomes

Organizer: Lan Luo. Chair: Menggang Yu. Room: HS 414

Presenter: Emily Hector, North Carolina State University

Title: Turning the data-integration dial: efficient inference from different data sources

Abstract: A fundamental aspect of statistics is the integration of data from different sources. Classically, Fisher and others were focused on *how* to integrate homogeneous sets of data. More recently, the question of *if* data sets from different sources should be integrated is becoming more relevant. The current literature treats this as a yes/no question: integrate or don't. Here we take a different approach, motivated by information-sharing principles coming from the shrinkage estimation literature. In particular, we deviate from the binary, yes/no perspective and propose a *dial parameter* that controls the extent to which two data sources are integrated. How far this dial parameter should be turned is shown to depend on the informativeness of the different data sources as measured by Fisher information. This more-nuanced data integration framework leads to relatively simple parameter estimates and valid tests/confidence intervals. We demonstrate both theoretically and empirically that setting the dial parameter according to our recommendation leads to more efficient estimation compared to other binary data integration schemes.

Presenter: Lan Luo, Rutgers University

Title: Efficient quantile covariate adjusted response adaptive experiments

Abstract: In program evaluation studies, understanding the heterogeneous distributional impacts of a program beyond the average effect is crucial. Quantile treatment effect (QTE) provides a natural measure to capture such heterogeneity. While much of the existing work for estimating QTE has focused on analyzing observational data based on untestable causal assumptions, little work has gone into designing randomized experiments specifically for estimating QTE. In this talk, we propose two covariate adjusted response adaptive design strategies–fully adaptive designs and multi-stage designs–to efficiently estimate the QTE. We demonstrate that the QTE estimator obtained from our designs attains the optimal variance lower bound from a semiparametric theory perspective, which does not impose any parametric assumptions on underlying data distributions. Moreover, we show that using continuous covariates in multi-stage designs can improve the precision of the estimated QTE compared to the classical fully adaptive setting. We illustrate the finite-sample performance of our designs through Monte Carlo experiments and one synthetic case study on charitable giving. Our proposed designs offer a new

approach to conducting randomized experiments to estimate QTE, which can have important implications for policy and program evaluation.

Presenter: Ling Zhou, Southwestern University of Finance and Economics

Title: High-dimensional subgroup learning for multiple mixed outcome

Abstract: In survey research, it is interesting to infer the grouped association patterns between risk factors and questionnaire responses, where the grouping is shared across multiple response variables that jointly capture one's underlying status. In particular, based on a survey study named the China Health and Retirement Survey (CHRS), our aim is to identify the important risk factors that are simultaneously associated with the health and well-being of senior adults. While earlier studies have pointed to several known risk factors, heterogeneity in the outcome-risk factor association exists, motivating us to analyze this data through the lens of subgroup analysis. We devise a subgroup analysis procedure that models a multiple mixed outcome which describe one's general health and well-being while tackling additional challenges that have arisen in our data analysis, including high-dimensionality, collinearity, and weak signals in covariates. Computationally, we propose an efficient algorithm that alternately updates a set of estimating equations and likelihood functions. Theoretically, we establish the asymptotic consistency and normality of the proposed estimators. The validity of our proposal is corroborated by simulation experiments. An application of the proposed method to the CHRS data identifies caring for grandchildren as a new risk factor for poor physical and mental health.

Presenter: Liangyuan Hu, Rutgers University

Title: Estimating the causal effect of multiple intermittent treatments on censored survival outcomes

Abstract: To draw real-world evidence about the comparative effectiveness of multiple time-varying treatments on patient survival, we develop a joint marginal structural survival model and a novel weighting strategy to account for time-varying confounding and censoring. Our methods formulate complex longitudinal treatments with multiple start/stop switches as the recurrent events with discontinuous intervals of treatment eligibility. We derive the weights in continuous time to handle a complex longitudinal dataset without the need to discretize or artificially align the measurement times. We further use machine learning models designed for censored survival data with time-varying covariates and the kernel function estimator of the baseline intensity to efficiently estimate the continuous-time weights. Our simulations demonstrate that the proposed methods provide better bias reduction and nominal coverage probability when analyzing observational longitudinal survival data with irregularly spaced time intervals compared to conventional methods that require aligned measurement time points. We apply the proposed methods to a large-scale COVID-19 dataset to estimate the causal effects of several COVID-19 treatments on the composite of in-hospital mortality and ICU admission.

Invited Session 6: Precision Medicine and Survival Analysis

Organizer: Simon Hirländer. Chair: Simon Hirländer. Room: HS 401

Presenter: Danyu Lin, University of North Carolina

Title: Evaluating Treatment Efficacy in Hospitalized Covid-19 Patients

Abstract: The clinical status of a COVID-19 patient is typically rated on a 7- or 8-point ordinal scale, ranging from resumption of normal activities to death, and the clinical status of a hospitalized COVID-19 patient may improve or deteriorate by different levels over the course of a clinical trial. The efficacy endpoints that have been used in clinical trials of hospitalized COVID-19 patients are the time to a specific change in clinical status or the clinical status on a particular day. For example, in the Adaptive COVID-19 Treatment Trials (ACTTs) and the Accelerating COVID-19 Therapeutic Interventions and Vaccines (ACTIV)-1 trial, the primary endpoints were time to recovery, and the secondary endpoints included 28-day mortality and clinical status at day 15 or day 28. However, these endpoints do not fully represent important clinical outcomes or make efficient use of available data. In this talk, I will present several methods that comprehensively characterize the treatment effects on the entire clinical course of a hospitalized COVID-19 patient and illustrate the advantages of these methods with the ACTT-1, ACTT-2, ACTT-3, and ACTIV-1 data.

Presenter: Limin Peng, Emory University

Title: Dynamic Regression of Longitudinal Trajectory Features

Abstract: Chronic disease studies often collect data on biological and clinical markers at follow-up visits to monitor disease progression. Viewing such longitudinal measurements governed by latent continuous trajectories, we develop a new dynamic regression framework to investigate the heterogeneity pattern of certain features of the latent individual trajectory that may carry substantive information on disease risk or status. Employing the strategy of multi-level modeling, we formulate the latent individual trajectory feature of interest through a flexible pseudo B-spline model with subject-specific random parameters, and then link it with the observed covariates through quantile regression, avoiding restrictive parametric distributional assumptions that are typically required by standard multi-level longitudinal models. We propose an estimation procedure from adapting the principle of conditional score and develop an efficient algorithm for implementation. Our proposals yield estimators with desirable asymptotic properties as well as good finite-sample performance as confirmed by extensive simulation studies. An application of the proposed method to a cohort of participants with mild cognitive impairment (MCI) in the Uniform Data Set (UDS) provides useful insights about the complex heterogeneous presentations of cognitive decline in MCI patients.

Presenter: Lei Liu, Washington University in Saint Louis

Title: Deep Learning Models to Predict Primary Open-Angle Glaucoma

Abstract: Glaucoma is a major cause of blindness and vision impairment worldwide, and visual field (VF) tests are essential for monitoring the conversion of glaucoma. While previous studies have primarily focused on using VF data at a single time point for glaucoma prediction, there has been limited exploration of longitudinal trajectories. Additionally, many deep learning techniques treat the time-to-glaucoma prediction as a binary classification problem (glaucoma Yes/No), resulting in the misclassification of some censored subjects into the non-glaucoma category and decreased power. To tackle these challenges, we propose and implement several deep-learning approaches that naturally incorporate temporal and spatial information from longitudinal visual field data to predict time-to-glaucoma. When evaluated on the Ocular Hypertension Treatment Study (OHTS) dataset, our proposed CNN-LSTM emerged as the top-performing model among all those examined.

Invited Session 7: Advanced methods for complicated data analysis

Organizer: Ying Wei. Chair: Cheng Zheng. Room: HS 402

Presenter: Nanhua Zhang, University of Cincinnati

Title: Bayesian Dirichlet Regression for Correlated Compositional Outcomes with Application to an Experimental Sleep Study

Abstract: It is common to observe compositional data in many fields and we consider the compositional data as the outcome in a regression setting. The motivation for this paper is a study of the effect of sleep restriction on physical activity outcomes where the compositional outcomes were measured under both short sleep and healthy sleep for the same participants. To address the dependence of the compositional outcomes we develop a Mixed-Effects Dirichlet Regression (MEDR) model for the compositional outcomes when the outcomes are correlated due to repeated measurements on the same subject or clustering within a group. We use an alternative parameterization of the Dirichlet distribution which allows modeling both the mean and dispersion components. Our method promises Markov chain Monte Carlo (MCMC) tools that are readily implementable in the programming language Stan and R. We apply the proposed MEDR model to the experimental sleep study and we illustrate properties of the methods through simulation studies.

Presenter: Tian Gu, Columbia University

Title: A Robust Angle-based Transfer Learning

Abstract: Transfer learning aims to improve the performance of a target model by leveraging data from related source populations which is especially helpful in cases with insufficient target data. In this paper, we study the problem of how to train a high-dimensional ridge regression

model using limited target data and existing regression models trained in heterogeneous source populations. We consider a practical setting where only the parameter estimates of the fitted source models are accessible instead of the individual-level source data. Under the setting with only one source model, we propose a novel flexible angle-based transfer learning (angleTL) method which leverages the concordance between the source and the target model parameters. We show that angleTL unifies several benchmark methods by construction including the target-only model trained using target data alone, the source model fitted on source data, and the distance-based transfer learning method that incorporates the source parameter estimates and the target data under a distance-based similarity constraint. We also provide algorithms to effectively incorporate multiple source models, accounting for the fact that some source models may be more helpful than others. Our high-dimensional asymptotic analysis provides interpretations and insights regarding when a source model can be helpful to the target model and demonstrates the superiority of angleTL over other benchmark methods. We perform extensive simulation studies to validate our theoretical conclusions and show the feasibility of applying angleTL to transfer existing genetic risk prediction models across multiple biobanks.

Presenter: Yanyuan Ma, Pennsylvania State University

Title: Doubly Flexible Estimation under Label Shift

Abstract: In studies ranging from clinical medicine to policy research, complete data are usually available from a population P but the quantity of interest is often sought for a related but different population Q which only has partial data. In this paper, we consider the setting that both outcome Y and covariate X are available from P whereas only X is available from Q under the so-called label shift assumption i.e., the conditional distribution of X given Y remains the same across the two populations. To estimate the parameter of interest in population Q via leveraging the information from population P, the following three ingredients are essential: (a) the common conditional distribution of X given Y, (b) the regression model of Y given X in population P, and (c) the density ratio of the outcome Y between the two populations. We propose an estimation procedure that only needs some standard nonparametric regression technique to approximate the conditional expectations with respect to (a) while by no means needs an estimate or model for (b) or (c); i.e., doubly flexible to the possible model misspecifications of both (b) and (c). This is conceptually different from the well-known doubly robust estimation in that double robustness allows at most one model to be misspecified whereas our proposal here can allow both (b) and (c) to be misspecified. This is of particular interest in our setting because estimating (c) is difficult if not impossible by virtue of the absence of the Y-data in population Q. Furthermore, even though the estimation of (b) is sometimes off-the-shelf, it can face curse of dimensionality or computational challenges. We develop the large sample theory for the proposed estimator and examine its finite-sample performance through simulation studies as well as an application to the MIMIC-III database.

Presenter: Ying Wei, Columbia University

Title: A Double Projection Approach for Safe and Efficient Semi-Supervised Data-Fusion

Abstract: Advances in data collection and transmission technologies have made larger amounts of data readily available. However, there are differences in the data collection capabilities of different data centers, or there are inevitable data missing. Many previous approaches to handling missing information have solely focused on either missing predictors or missing responses. In this paper, we will consider both types of missing and incorporate more information by projecting score functions into subsets, thus proposing algorithms that have ensured efficiency relative to the complete-case analysis. By generalizing the algorithm of this paper, it is promising to be able to handle more complex missing data structures in the future. This is joint work with Molei Liu, Yiming Li and Sean Yang.

Invited Session 8: Recent developments in Survival Analysis and Statistical Machine Learning

Organizer: Tony Sun. Chair: Xingqiu Zhao. Room: HS 403

Presenter: Jianwen Cai, University of North Carolina, Chapel Hill

Title: Feature Screening for Case-Cohort Studies with Failure Time Outcome

Abstract: Case-cohort design has been demonstrated to be an economical and effective approach in large cohort studies when the measurement of some covariates on all individuals is expensive. Various methods have been proposed for case-cohort data when the dimension of covariates is smaller than sample size. However, limited work has been done for high-dimensional case-cohort data which are frequently collected in large epidemiological studies. We propose a variable screening method for ultrahigh-dimensional case-cohort data under the framework of proportional hazards model which allows the covariate dimension increases with sample size at exponential rate. Our procedure enjoys the sure screening property and the ranking consistency under some mild regularity conditions. We further extend this method to an iterative version to handle the scenarios where some covariates are jointly important but are marginally unrelated or weakly correlated to the response. The finite sample performance of the proposed procedure is evaluated via both simulation studies and an application to a real data from the breast cancer study.

Presenter: Wen Su, City University of Hong Kong

Title: Deep Generative Estimation of Conditional Survival Function

Abstract: Current status data are commonly encountered in modern medicine, econometrics, and social science. Its unique characteristics pose significant challenges to the analysis of such data, and the existing methods often suffer grave consequences when the underlying model is misspecified. To address these difficulties, we propose a model-free two-stage generative

approach for estimating the conditional cumulative distribution function given predictors. We first learn a conditional generator nonparametrically for the joint conditional distribution of observation times and event status and then construct the nonparametric maximum likelihood estimators of conditional distribution functions based on samples from the conditional generator. Subsequently, we study the convergence properties of the proposed estimator and establish its consistency. Simulation studies under various settings show the superior performance of the deep conditional generative approach over the classical modeling approaches and an application to Parvovirus B19 seroprevalence data yields reasonable predictions.

Presenter: Yifan Cui, Zhejiang University

Title: Fiducial Inference in Survival Analysis

Abstract: Censored data where the event time is partially observed are challenging for survival probability estimation. In this paper, we introduce a novel nonparametric fiducial approach to interval-censored data, including right-censored, current status, case II censored, and mixed case censored data. The proposed approach, leveraging a simple Gibbs sampler, has a useful property of being "one size fits all," i.e., the proposed approach automatically adapts to all types of non-informative censoring mechanisms. As shown in the extensive simulations, the proposed fiducial confidence intervals significantly outperform existing methods in terms of both coverage and length. In addition, the proposed fiducial point estimator has much smaller estimation errors than the nonparametric maximum likelihood estimator.

Presenter: Qixian Zhong, Xiamen University

Title: Hypothesis Testing for the Deep Cox Model

Abstract: Deep learning has become enormously popular in the analysis of complex data, including event time measurements with censoring. To date, deep survival methods have mainly focused on prediction. Such methods are scarcely used in matters of statistical inference, such as hypothesis testing. Due to their black-box nature, deep-learned outcomes lack interpretability, which limits their use for decision-making in biomedical applications. This paper provides estimation and inference methods for the nonparametric Cox model—a flexible family of models with a nonparametric link function to avoid model misspecification. Here, we assume the nonparametric link function is modeled via a deep neural network. To perform statistical inference, we split the data into an estimation set and a set for statistical inference. This inference procedure enables us to propose a new significance test to examine the association of certain covariates with event times. We establish convergence rates of the neural network estimator and show that deep learning can overcome the curse of dimensionality in nonparametric regression by learning to exploit low-dimensional structures underlying the data. In addition, we show that our test statistic converges to a normal distribution under the null hypothesis and establish its consistency in terms of the Type II error under the alternative hypothesis. Numerical studies demonstrate the usefulness of the proposed test.

Break 2:45-3:00, July 9, 2024

Parallel Invited Sessions 9-12: 3:00-4:45, July 9, 2024

Invited Session 9: Advance Development and Application of Joint modeling

Organizer: Cheng Zheng. Chair: Zhigang Li. Room: HS 403

Presenter: Ying Zhang, University of Nebraska Medical Center

Title: Semiparametric Inference for Misclassified Semi-Competing Risks Data under Gamma-Frailty Conditional Markov Model

Abstract: There has been increasing interest in semi-competing risks data modeling to jointly study disease progression and death for the illness-death problem. Identification of risk factors for the benchmark events will provide insight to detect the high-risk group according to personal-level characteristics which is critical to develop a personalized prevention strategy to delay the progression from illness to death. However in many applications event ascertainment is incomplete resulting in event misclassification that complicates the statistical inference with semi-competing risks data. In this work we consider a Gamma frailty conditional Markov model to study the misclassified semi-competing risk data and propose a two-stage semiparametric maximum pseudo-likelihood estimation approach equipped with a pseudo-EM algorithm to make unbiased statistical inference. Extensive simulation studies show the proposed method is numerically stable and performs well even with a large amount of event misclassification. The method is applied to a multi-center HIV cohort study in East Africa to measure the impact of interruption of lifelong antiretroviral therapy (ART) on HIV mortality at the personal level.

Presenter: Cheng Zheng, University of Nebraska Medical Center

Title: Investigating Multiple Causal Mechanisms with Multiple Mediators and Estimating Direct and Indirect Effects: A Joint Modeling Approach for Recurrent and Terminal Events

Abstract: Understanding the diverse causal mechanisms between primary exposure and outcomes has garnered significant interest in the social and medical fields. In the context of HIV patients over 20 distinct opportunistic infections (OIs) present complex effects on the health trajectory and associated mortality. It is crucial to differentiate among these OIs to devise tailored strategies to enhance patients' survival and quality of life. However existing statistical frameworks for studying causal mechanisms have limitations either focusing on single mediators or lacking the ability to handle unmeasured confounding especially for the survival outcomes. In this work we propose a novel joint modeling approach that considers multiple recurrent events as

mediators and survival endpoints as outcomes relaxing the assumption of "sequential ignorability" by utilizing the shared random effect to handle unmeasured confounders. We assume the multiple mediators are not causally related to each other given observed covariates and the shared frailty. Simulation studies demonstrate good finite sample performance of our methods in estimating both model parameters and multiple mediation effects. We apply our approach to an AIDS study and evaluate the mediation effects of different types of OIs. We find that distinct pathways through the two treatments and CD4 counts impact overall survival via different types of recurrent opportunistic infections.

Presenter: Danping Liu, National Cancer Institute

Title: Dynamic Risk Prediction for Cervical Precancer Screening with Continuous and Binary Longitudinal Biomarkers

Abstract: Dynamic risk prediction that incorporates longitudinal measurements of biomarkers is useful in identifying high-risk patients for better clinical management. Our work is motivated by the prediction of cervical precancers. Currently, Pap cytology is used to identify HPV+ women at high-risk of cervical precancer but cytology lacks accuracy and reproducibility. HPV DNA methylation is closely linked to the carcinogenic process and shows promise of improved risk stratification. We are interested in developing a dynamic risk model that uses all longitudinal biomarker information to improve precancer risk estimation. We propose a joint model to link both the continuous methylation biomarker and binary cytology biomarker to the time to precancer outcome using shared random effects. The model uses a discretization of the time scale to allow for closed-form likelihood expressions thereby avoiding high-dimensional integration of the random effects. The method handles an interval-censored time-to-event outcome due to intermittent clinical visits, incorporates sampling weights to deal with stratified sampling data, and can provide immediate and 5-year risk estimates that may inform clinical decision-making.

Invited Session 10: Statistical and machine learning methods in biomedical research

Organizer: Qi Long. Chair: Nanhua Zhang. Room: HS 414

Presenter: Ying Guo, Emory University

Title: A Regularized Blind Source Separation Framework for Unveiling Hidden Sources of Brain Functional Connectome

Abstract: Brain connectomics has become increasingly important in neuroimaging studies to advance the understanding of neural circuits and their association with neurodevelopment, mental illnesses, and aging. These analyses often face pinmajor challenges including the high dimensionality of brain networks, unknown latent sources underlying the observed connectivity,

and the large number of brain connections leading to spurious findings. In this talk, we will introduce a novel regularized blind source separation (BSS) framework for reliable mapping of neural circuits underlying static and dynamic brain functional connectome. The proposed LOCUS methods achieve more efficient and reliable source separation for connectivity matrices using low-rank factorization, a novel angle-based sparsity regularization, and a temporal smoothness regularization. We develop a highly efficient iterative Node-Rotation algorithm that solves the non-convex optimization problem for learning LOCUS models. Simulation studies demonstrate that the proposed methods have consistently improved accuracy in retrieving latent connectivity traits. Application of LOCUS methods to the Philadelphia Neurodevelopmental Cohort (PNC) neuroimaging study generates considerably more reproducible findings in revealing underlying neural circuits and their association with demographic and clinical phenotypes, uncovers dynamic expression profiles of the circuits and the synchronization between them, and generates insights on gender differences in the neurodevelopment of brain circuits.

Presenter: Suprateek Kundu, The University of Texas at MD Anderson Cancer Center

Title: Flexible Bayesian Product Mixture Models for Vector Autoregressions

Abstract: Bayesian non-parametric methods based on Dirichlet process mixtures have seen tremendous success in various domains and are appealing in being able to borrow information by clustering samples that share identical parameters. However, such methods can face hurdles in heterogeneous settings where objects are expected to cluster only along a subset of axes or where clusters of samples share only a subset of identical parameters. We overcome such limitations by developing a novel class of product of Dirichlet process location-scale mixtures that enable independent clustering at multiple scales, which result in varying levels of information sharing across samples. First, we develop the approach for independent multivariate data. Subsequently, we generalize it to multivariate time-series data under the framework of multi-subject Vector Autoregressive (VAR) models that is our primary focus which go beyond parametric single-subject VAR models. We establish posterior consistency and develop efficient posterior computation for implementation. Extensive numerical studies involving VAR models show distinct advantages over competing methods in terms of estimation, clustering, and feature selection accuracy. Our resting state fMRI analysis from the Human Connectome Project reveals biologically interpretable connectivity differences between distinct intelligence groups while another air pollution application illustrates the superior forecasting accuracy compared to alternate methods.

Presenter: Youngjoo Cho, Konkuk University

Title: Estimation of Heterogeneous Treatment Effects for Competing Risks Data Using Random Forests

Abstract: Estimation of heterogeneous treatment effect for uncensored data has been studied extensively. However, efforts to develop methods of estimating heterogeneous treatment effect using machine learning have begun comparatively recently. In this paper, we propose a novel approach to estimating heterogeneous treatment effect with respect to cumulative incidence

curves in the competing risks data using random forests. The proposed methods employ orthogonal estimating equations and augmented functions based on semiparametric efficiency theory. Simulation studies show the utility of our approach.

Invited Session 11: Generalizability consideration in heterogeneous data

Organizer: Menggang Yu. Chair: Michael Elliot. Room: HS 401

Presenter: Jeremy Taylor, University of Michigan

Title: James-Stein approach for improving prediction of linear regression models by integrating external information from heterogeneous populations

Abstract: We consider the setting where (i) an internal study builds a linear regression model for prediction based on individual-level data, (ii) some external studies have fitted similar linear regression models that use only subsets of the covariates and provide coefficient estimates for the reduced models without individual-level data, and (iii) there is heterogeneity across these study populations. The goal is to integrate the external model summary information into fitting the internal model to improve prediction accuracy. We adapt the James-Stein shrinkage method to propose estimators that have guaranteed improvement in the prediction mean squared error after information integration regardless of the degree of study population heterogeneity. We conduct comprehensive simulation studies to investigate the numerical performance of the proposed estimators. We also apply the method to enhance a prediction model for patella bone lead level in terms of blood lead level and other covariates by integrating summary information from published literature.

Presenter: Lu Tian, Stanford University

Title: Adaptive Prediction Strategy with Individualized Variable Selection

Abstract: Today, physicians have access to a wide array of tests for diagnosing and prognosticating medical conditions. Ideally, they would apply a high-quality prediction model utilizing all relevant features as input to facilitate appropriate decision-making regarding treatment selection or risk assessment. However, not all features used in these prediction models are readily available to patients and physicians without incurring some costs. In practice, predictors are typically gathered as needed in a sequential manner while the physician continually evaluates information dynamically. This process continues until sufficient information is acquired and the physician gains reasonable confidence in making a decision. Importantly, the prospective information to collect may differ for each patient and depend on the predictor values already known. In this paper, we present a novel dynamic prediction rule designed to determine the optimal order of acquiring prediction features in predicting a clinical outcome of interest. The objective is to maximize prediction accuracy while minimizing the cost

associated with measuring prediction features for individual subjects. To achieve this, we employ reinforcement learning where the agent must decide on the best action at each step: either making a clinical decision with available information or continuing to collect new predictors based on the current state of knowledge. To evaluate the efficacy of the proposed dynamic prediction strategy, extensive simulation studies have been conducted. Additionally, we provide two real data examples to illustrate the practical application of our method.

Presenter: Menggang Yu, University of Michigan

Title: Entropy Balancing for Causal Generalization with Target Sample Summary Information

Abstract: In this talk, we focus on estimating the average treatment effect (ATE) of a target population when individual-level data from a source population and summary-level data (e.g., first or second moments of certain covariates) from the target population are available. In the presence of heterogeneous treatment effects, the ATE of the target population can be different from that of the source population when distributions of treatment effect modifiers are dissimilar in these two populations, a phenomenon also known as covariate shift. Many methods have been developed to adjust for covariate shift, but most require individual covariates from a representative target sample. We develop a weighting approach based on summary-level information from the target sample to adjust for possible covariate shift in effect modifiers. In particular, weights of the treated and control groups within a source sample are calibrated by the summary-level information of the target sample. Our approach also seeks additional covariate balance between the treated and control groups in the source sample.

Presenter: Ruth Pfeiffer, National Cancer Institute, NIH, HHS

Title: Accommodating population differences in model validation

Abstract: Validation of risk prediction models in independent data provides a rigorous assessment of model performance. However, several differences between the populations that gave rise to the training and the validation data can lead to seemingly poor performance of a risk model. We formalize the notions of “similarity” of the training and validation data and define reproducibility and transportability. We address the impact of different predictor distributions and differences in verifying the outcome on model calibration, accuracy, and discrimination. When individual-level data from both the training and validation datasets are available, we propose and study weighted versions of the validation metrics that adjust for differences in the predictor distributions and in outcome verification to provide a more comprehensive assessment of model performance. We give conditions on the model and the training and validation populations that ensure a model’s reproducibility or transportability and show how to check them. We discuss approaches to recalibrate a model. As an illustration, we develop and validate a prostate cancer risk model using data from two large North American prostate cancer prevention trials, the SELECT and PLCO trials. Joint work with Yiyao Chen, Mitchell H. Gail, and Donna P. Ankerst.

Invited Session 12: Survival analysis and its applications

Organizer: Yichuan Zhao. Chair: Jonas Beck. Room: HS 402

Presenter: Gang Li, University of California, Los Angeles

Title: A Semiparametric Bayesian Instrumental Variable Analysis Method for Partly Interval-Censored Time-to-Event Outcome

Abstract: This paper develops a semiparametric Bayesian instrumental variable (IV) analysis method for estimating the causal effect of an endogenous variable when dealing with unobserved confounders and measurement errors with partly interval-censored time-to-event data where event times are observed exactly for some subjects but left-censored, right-censored, or interval-censored for others. Our method is based on a two-stage Dirichlet process mixture instrumental variable (DPMIV) model which simultaneously models the first-stage random error term for the exposure variable and the second-stage random error term for the time-to-event outcome using a Gaussian mixture of the Dirichlet process (DPM) model. The DPM model can be broadly understood as a mixture model with an unspecified number of Gaussian components which relaxes the normal error assumptions and allows the number of mixture components to be determined by the data. We develop an MCMC algorithm for the DPMIV model tailored for partly interval-censored data and conduct extensive simulations to assess the performance of our DPMIV method in comparison with some existing methods. Our simulations revealed that our proposed method is robust under different error distributions and can have far superior performance over some competing methods under a variety of scenarios. We further demonstrate the effectiveness of our approach on the UK Biobank study to investigate the causal effect of systolic blood pressure (SBP) on time-to-development of cardiovascular disease (CVD) from the diagnosis of diabetes mellitus (DM).

Presenter: Amita Manatunga, Emory University

Title: Noninvasive Monitoring for Anemia in Very Low Birth Weight Infants Using Smartphone Images

Abstract: Monitoring for anemia is an important part of clinical care for very low birth weight (VLBW) infants in the neonatal intensive care unit. Collecting fingernail imaging using smartphone device has been experimented as a noninvasive alternative to standard repeated invasive blood draws that may paradoxically contribute to anemia through accumulated blood loss. In this talk, we will present novel preliminary analysis of a longitudinal study of VLBW infants from three Atlanta-area hospitals to evaluate the prognostic value of utilizing the fingernail imaging data which takes the form of three 51 X 51 matrices of RGB values for predicting an infant's need for red blood cell transfusion. Our analyses effectively utilize the repeated measurements of the fingernail imaging data and moreover properly account for the potential dependency between the timing of taking fingernail images and the underlying risk of

requiring blood transfusion. Survival analysis techniques are also employed to handle censoring to the observation of transfusions. We will conclude the talk with sensible interpretation of our analysis results.

Presenter: Yi Li, University of Michigan

Title: Multi-task Learning for Gaussian Graphical Regressions with High Dimensional Covariates

Abstract: Gaussian graphical regression is a powerful approach for regressing the precision matrix of a Gaussian graphical model on covariates which permits the response variables and covariates to outnumber the sample size. However, traditional approaches of fitting the model via separate node-wise lasso regressions overlook the network-induced structure among these regressions, leading to high error rates, particularly when the number of nodes is large. To address this issue, we propose a multi-task learning estimator for fitting Gaussian graphical regression models which incorporates a cross-task group sparsity penalty and a within-task element-wise sparsity penalty to govern the sparsity of active covariates and their effects on the graph, respectively. We also develop an efficient augmented Lagrangian algorithm for computation which solves subproblems with a semi-smooth Newton method. We further prove that our multi-task learning estimator has considerably lower error rates than the separate node-wise regression estimates as the cross-task penalty enables borrowing information across tasks. To address the main challenge of entangled tasks in a complicated correlation structure, we establish a new tail probability bound for dependent heavy-tailed, for example, sub-exponential variables with an arbitrary dependence structure which is a useful theoretical result in its own right. We examine the utility of our method through simulations and an application to a gene co-expression network study with brain cancer patients.

Presenter: Yichuan Zhao, Georgia State University

Title: Weighted Empirical Likelihood Inference for the Difference Between the Areas Under Two Correlated ROC Curves with Right-Censored Data

Abstract: In this article, we propose the two-sample weighted empirical likelihood building upon Chrzanowski's (2014) method to compare the area under the ROC curve of two correlated ROC curves. A normal approximation method is derived. We define a weighted empirical likelihood ratio and demonstrate that the resulting statistic follows a scaled chi-square distribution. Additionally, to improve the accuracy of confidence intervals for small sample sizes, we employ a calibration method known as the adjusted empirical likelihood. Extensive simulations are conducted to assess the excellent finite-sample performance of our proposed weighted empirical likelihood method. To further illustrate the practical applicability of our methodology, we provide a real-world example showcasing its effectiveness.

Break 4:45-5:00, July 9, 2024

Parallel Contributed Sessions 1-4: 5:00-6:00, July 9, 2024

Contributed Session 1: Applications of Machine Learning and Artificial Intelligence.

Chair: Arne Bathke. Room: HS 414

Presenter: Nicolò Biasetton, University of Padova

Title: Advancing Consumer Understanding in Sustainable Product Development: exploring Novel Methodologies Integrating Aspect-Based Sentiment Analysis and NLP

Abstract: The integration of aspect-based sentiment analysis (ABSA) and advanced natural language processing (NLP) methodologies encompassing semantic analysis and the generation of semantic spaces using embeddings signifies a notable advancement in understanding consumer behavior, especially within the domain of Sustainable Product Development (SPD). ABSA refines conventional sentiment analysis by parsing consumer feedback into specific aspects such as product features or attributes, providing a more nuanced comprehension of sentiments related to distinct elements of environmentally friendly products. This approach enables a detailed understanding of consumer perceptions, facilitating targeted enhancements in the design and marketing strategies of eco-friendly products. Additionally, the application of NLP techniques, particularly semantic analysis, assists in extracting meaning from consumer reviews by discerning contextual nuances and subtle intricacies of language. The development of semantic spaces through embeddings further elevates this process by translating words or phrases into multi-dimensional vectors, capturing semantic relationships and contextual associations. These advanced techniques empower researchers and businesses to delve deeper into the complex interplay of emotions, perceptions, and linguistic nuances that shape consumer attitudes toward sustainable products. As SPD aims for more tailored and culturally sensitive approaches, the amalgamation of aspect-based sentiment analysis and NLP methodologies explored and applied in the present work opens up novel avenues for the formulation of personalized and effective strategies in the development of eco-friendly products.

Presenter: Elena Barzizza, University of Padova

Title: A Framework for Automated Web Scraping and Sentiment Analysis of Product Reviews using ChatGPT

Abstract: The impact of online product reviews on consumer decisions nowadays cannot be overstated. This work introduces a cutting-edge tool that merges advanced Natural Language Processing and sentiment analysis techniques with ChatGPT technology. This tool automates web scraping and conducts thorough sentiment analysis on product reviews, emphasizing aspect-based sentiment analysis and aggregating sentiments at the document level. Utilizing the powerful ChatGPT language model, our tool performs aspect-based sentiment analysis,

extracting nuanced sentiments associated with specific product attributes. This approach provides a detailed comprehension of consumer opinions. Additionally, the tool incorporates sophisticated models for document-level sentiment analysis, allowing the aggregation of overall sentiment scores from various reviews. The framework consists of three main components: the Web Scraping Module automates data collection from two major e-commerce platforms, ensuring a systematic extraction of reviews and metadata. The Aspect-Based Sentiment Analysis Module, leveraging ChatGPT, categorizes opinions as positive, negative, or neutral and identifies specific product attributes. The Document-Level Sentiment Analysis Module provides an overarching sentiment rating for the product. This versatile tool finds applications in business research and consumer decision-making, enabling real-time monitoring of product performance, conducting market studies, and facilitating informed purchase decisions. Our integrated framework offers a robust solution for comprehending and responding to consumer sentiments in online product reviews. By combining aspect-based and document-level sentiment analysis with the capabilities of ChatGPT, it provides comprehensive insights into the digital marketplace.

Presenter: Alberto Molena, University of Padova

Title: Predicting football match attendance and optimizing revenues using Machine Learning

Abstract: Football is one of the most widely followed sports worldwide, with teams having evolved into comprehensive entities akin to full-fledged businesses. As in any industry, the forecasting phase and its optimization are, therefore, of fundamental importance. In this context, it translates into predicting stadium attendance for home games and optimizing the revenues derived from this activity. The objective of this contribution is twofold: firstly, thanks to the latest advancements in Machine Learning, we aim to determine a framework that allows obtaining reliable predictions regarding match attendance. Subsequently, building upon this framework, we seek to develop an optimizer that maximizes revenues from ticket sales operating on match prices. The validity of this approach will be tested through a real case study involving an Italian football team participating in the top national league.

Presenter: Riccardo Ceccato, University of Padova

Title: Choosing the optimal number of fuzzy clusters: a nonparametric approach

Abstract: The determination of the number of clusters emerges as a crucial decision in partitioning-based clustering algorithms. In the present work, we especially focus on fuzzy clustering. Hence, we introduce a robust nonparametric procedure designed for identifying the optimal number of clusters, leveraging the Nonparametric Combination (NPC) methodology. Specifically, we generate C different partitions and assess their quality by computing relevant measures derived from the membership of each unit. Subsequently, we advocate for the combined use of NPC-based tests and a ranking procedure to pinpoint the best partition based on the aforementioned measures. To illustrate the efficacy of this approach, we apply it to a case study involving fuzzy clustering on customer satisfaction data.

Contributed Session 2: Estimation and Inference with Complex Data Structures

Chair: Patrick Langthaler. Room: HS 401

Presenter: Li-Hsiang Lin, Georgia State University

Title: High-Dimensional Multivariate Linear Regression with Weighted Nuclear Norm Regularization

Abstract: We consider a low-rank matrix estimation problem when the data is assumed to be generated from the multivariate linear regression model. To induce the low-rank coefficient matrix, we employ the weighted nuclear norm (WNN) penalty defined as the weighted sum of the singular values of the matrix. The weights are set in a non-decreasing order, which yields the non-convexity of the WNN objective function in the parameter space. Such objective function has been applied in many applications but studies on the estimation properties of the estimator from the objective function are limited. We propose an efficient algorithm under the framework of alternative directional method of multipliers (ADMM) to estimate the coefficient matrix. The estimator from the suggested algorithm converges to a stationary point of an augmented Lagrangian function. Under the orthogonal design setting, effects of the weights for estimating the singular values of the ground-truth coefficient matrix are derived. Under the Gaussian design setting, a minimax convergence rate on the estimation error is derived. We also propose a generalized cross-validation (GCV) criterion for selecting the tuning parameter and an iterative algorithm for updating the weights. Simulations and a real data analysis demonstrate the competitive performance of our new method.

Presenter: Soham Bakshi, University of Michigan

Title: Selective Inference for Time-Varying Effect Moderation

Abstract: The scientific community is increasingly interested in developing data analysis techniques that can improve mobile health interventions. A key aspect of this effort involves assessing the impact of time-varying causal effect moderators. Effect modification, a scenario where the impact of treatment on outcomes varies based on other covariates, plays a significant role in decision-making processes. When there are hundreds or thousands of covariates, it becomes necessary to use observed data to select a simpler model for effect modification by extracting a smaller set of true moderators. This is generally achieved by applying Lasso type penalties with weighted centered least squares (WCLS). The selected model is much more interpretable compared to a full model consisting of all covariates. We apply selective inference methods to make valid post-selection inferences with the selected model; specifically, we take the conditional approach. We construct an exact selective pivot that is asymptotically distributed as a uniform random variable. We provide both simulations and real data study to demonstrate our method's performance and compare it with naive approaches, data splitting, and other baseline methods.

Presenter: Mali Abdollahian, RMIT University

Title: Application of Multivariate Statistical Quality Control and Profile Monitoring in the medical area

Abstract: Statistical process control has emerged in medical literature after wide expansion in the industry. In clinical monitoring, there are always more than one quality characteristic of interest which are usually correlated. In such cases, multivariate control charts and profile monitoring must be deployed to monitor the medical process. This talk will review the application of the Multivariate Exponentially Weighted Moving Average control chart (MEWMA) to monitor the patient's progress in the Intensive Care Unit, characterized by nine quality characteristics. We will discuss utilizing profile monitoring procedures to monitor and develop upper and lower profile limits for maternal and newborn mortality and childhood type one diabetes (T1D) and their associated significant factors.

Contributed Session 3: Innovative Analysis of Clinical Trial Data

Chair: Yi Li. Room: HS 402

Presenter: Zhichen Xu, Washington University in Saint Louis

Title: Subgroup Identification Based on Mixed Model for Repeated Measures for Alzheimer's Disease Trial

Abstract: In precision medicine, the identification of subgroups is pivotal for designing personalized treatments. While current methods for subgroup identification primarily concentrate on either the total treatment effect in cross-sectional studies or the mean effect in longitudinal studies, the FDA typically evaluates treatment effects at the end of longitudinal studies through the Mixed Model for Repeated Measures (MMRM). Therefore, we introduce the Interaction Tree with Mixed Model for Repeated Measures (IT-MMRM) as a novel approach for identifying subgroups. The mixed effect model with a nonlinear time trend allows us to model the dependence of longitudinal measures flexibly and single out a time-dependent treatment interaction to build up the interaction tree. Our IT-MMRM demonstrates superior performance over existing subgroup identification techniques in simulations, especially when the time-dependent treatment effect is the focal point. We also explore different tuning parameter options and use bootstrap methods to prune the trees, thereby mitigating the risk of overoptimism. The IT-MMRM model is applied to an Alzheimer's disease study aiming to uncover subgroups with varying long-term treatment responses.

Presenter: Ashwini Joshi, Helsinki University Hospital

Title: Statistical analysis of irregularly observed multiple outcomes of treatments in patients with age-related chronic diseases: Application to wet age-related macular degeneration (AMD)

Abstract: Treatment of age-related chronic diseases differs drastically between clinical trials and real-life. In real-life, the 'treat-and-extend' protocol is preferred with a patient-centric treatment regime instead of the 'treatment when needed (PRN)' protocol in a clinical trial with randomized treatment allocation. For the diseases that need treatment till up to the end of life, it is not possible to assess all possible effects of long-term treatment in the limited time period of clinical trials. Generally, the main response is the indicator of the underlying latent state of the disease. There are many other markers which mediate the effect of the treatment on the main response. Due to reasons such as cost, challenges in measurements, etc., not all markers are measured at each clinical visit. During a long follow-up, the situation of the patient may change in terms of available physical help, financial situation, and treatment response pattern. All these factors lead to missing clinical visits and may result in under-treatment or termination of treatment (either by the patient or by the clinician). To help the decision-making in treatment continuation, recognizing the treatment response pattern at that time is important. Various clustering methods using mixed models are available for fixed cluster membership. Patients with varying treatment responses over time are rarely studied. The aim of this research is to suggest methods for understanding the dynamic treatment response and for deriving inference in such long-term longitudinal multi-outcomes analysis. The effect of both the treatment since the diagnosis and the latest treatment frequency are studied using the duration-response analysis. A case-study of wet-AMD patients will be discussed.

Presenter: S. Schoenen, Sigmund Freud Private University

Title: The Impact of Allocation Bias on the Test Decisions in Clinical Trials with Multiple Endpoints

Abstract: In a clinical trial, prior knowledge or at least an idea of the next treatment allocations may induce allocation bias even within randomized trials, impacting the trial's validity. Quantifying the impact of allocation bias involves considering the type-I-error probability. For clinical trials with normally distributed single endpoints, there exists a model to assess allocation bias and a formula to calculate the corresponding type-I-error probability. However, in specific clinical contexts, the incorporation of multiple endpoints into the trial yields benefits. Multiple endpoints lead to a more comprehensive understanding of treatment effects and may increase power or reduce sample sizes, respectively. The current investigation aims to quantify the impact of allocation bias in clinical trials with multiple endpoints. Therefore, we introduce a new biasing policy tailored for multivariate normally distributed data and derive formulas for computing the family-wise error rate of the Bonferroni-Šidák procedure and the type-I-error probability of the all-or-none decision rule under bias. The developed biasing policy for multiple endpoints quantifies endpoint-specific allocation bias effects on test decisions. Therefore, we assume that the investigator is aware of previous allocations and that better-responding patients are associated with higher positive scores in each component of the multiple endpoint. This assumption enables us to distinguish between good, neutral, and worse responders. According to the convergence strategy of Blackwell and Hodges, the next patient is assigned to the group that has been assigned less frequently. If this is the experimental group, a good responder is recruited; otherwise, a worse responder. A neutral responder is allocated only if the group sizes are balanced. Using this biasing policy, we derive exact formulas for calculating the family-wise error probability of the Bonferroni-Šidák procedure and the type-I-error probability of the all-or-

none decision rule regarding different randomization procedures. The analyzed randomization procedures are complete randomization, big stick design, random allocation rule, permuted block randomization, Chen's urn design, maximal procedure, and Efron's bias coin design. Simulations demonstrate that bias inflates the error rates of the Bonferroni-Šidák procedure and the all-or-none decision rule depending on the randomization procedure. Analyzing bias effects during the trial's planning phase and opting for a bias-mitigating randomization procedure improves trial validity. Thus, the developed method can substantiate the selection of a randomization procedure and empower the design of more robust and valid trials. In addition, the method can also be used in the trial's analysis phase to perform bias-corrected statistical tests. Its applicability extends to clinical trials, even those with small sample sizes.

Contributed Session 4: Bayesian Statistics

Chair: Lu Tian. Room: HS 403

Presenter: David M. Hughes, University of Liverpool

Title: Scalable Bayesian modelling of multivariate longitudinal data: A comparison of computational options

Abstract: Multiple longitudinal responses are now routinely collected in many areas of health research. Analysing these outcomes within a single model allows the opportunity to capture correlation between each outcome. A useful way to model longitudinal data on multiple outcomes is through the use of multivariate generalised linear mixed models (MGLMMs) where correlation between repeated measurements on the same patient is modelled through the use of patient-specific random effects. MGLMMs can be estimated within a Bayesian framework using MCMC methods. However, with the increasing availability of data on many repeatedly measured outcomes on many thousands of patients, fitting MGLMMs by MCMC can be very computationally intensive. In recent years, various approximation methods have been proposed to make Bayesian inference feasible in large datasets. These include integrated nested Laplace approximations (INLA) and various forms of variational inference such as the automatic differentiation variational inference (ADVI) option available within Stan and a recently proposed semiparametric variational Bayes algorithm. These approximation methods offer much faster options to fit complex multivariate longitudinal models. However, there are few theoretical guarantees about approximation accuracy and little work has been done to assess the reliability of these approaches. In this talk, we compare computational approaches for modelling multivariate longitudinal data. We assess the speed of computation relative to MCMC and the accuracy of approximation methods both relative to MCMC and to the "truth" in simulation studies. We also present examples from clinical practice related to diabetes and liver cancer.

Presenter: Javier Enrique Aguilar, TU Dortmund University

Title: Generalized Decomposition Priors on R2

Abstract: In recent years, the adoption of shrinkage priors in high-dimensional linear models has gained momentum, driven by their theoretical and practical advantages. A novel approach to alleviate the complexity of assigning individual priors to each regression term involves employing a shrinkage prior on the model fit R2 that spans the entire set of coefficients. The core idea is to specify a prior on R2, conduct a Dirichlet decomposition to unfold the total variance, and subsequently transfer the uncertainty to the regression terms by specifying the respective variances. However, these priors disregard the intricate dependence structures among variance components during the decomposition step. In the absence of sufficient control, the competition among variance components tends to gravitate towards negative dependence structures. Yet, in reality, specific coefficients or groups may compete differently for the total variability than the Dirichlet would assume. In this work, we address this limitation by exploring the capacity to account for diverse dependence structures in the decomposition step. Through simulations and real-world case studies, we demonstrate that alternative prior choices can yield more favorable results. To this end, we introduce a novel family of prior distributions termed *Generalized Decomposition R2* (GDR2) priors that provides enhanced flexibility in capturing intricate relationships among the proportions of explained variance.

Presenter: Victor De Oliveira, The University of Texas at San Antonio

Title: On Inference about the Smoothness Parameter in Gaussian Mat\'ern Random Fields

Abstract: The Mat\'ern family of covariance functions is currently the most commonly used for the analysis of geostatistical data due to its ability to describe different smoothness behaviors. Yet, in many applications the smoothness parameter is set at an arbitrary value. This practice is due partly to computational challenges faced when attempting to estimate all covariance parameters and partly to unqualified claims in the literature stating that geostatistical data have little or no information about the smoothness parameter. In this talk I describe new class of easy-to-compute default priors for the smoothness parameter. These priors approximate reference priors, but their analysis and computation is considerably simpler. It is shown that the posterior distribution of the parameters based on these priors is proper, and Bayesian inferences about the covariance parameters based on these priors have satisfactory frequentist properties, much better than those based on maximum likelihood. The methodology is illustrated with a data set of rainfall totals in Switzerland.

Banquet/InflaRx Junior Research Award: 6:45-9:00, July 9, 2024
Peter Song & Yichuan Zhao

Day 2: July 10, 2024

Parallel Invited Sessions 13-16: 9:00-10:45, July 10, 2024

Invited Session 13: Analysis of multi-outcomes and multi-source data

Organizer: Zhezhen Jin. Chair: Liang Li. Room: HS 414

Presenter: Xiaonan Xue, Albert Einstein University

Title: Jointly modeling of sleep variables that are objectively measured by wrist actigraphy

Abstract: Recently developed actigraphy devices have made it possible for continuous and objective monitoring of sleep over multiple nights. Sleep variables captured by wrist actigraphy devices include sleep onset, sleep end, total sleep time, wake time after sleep onset, number of awakenings, etc. Currently available statistical methods to analyze such actigraphy data have limitations. First, averages over multiple nights are used to summarize sleep activities, ignoring variability over multiple nights from the same subject. Second, sleep variables are often analyzed independently. However, sleep variables tend to be correlated with each other. For example, how long a subject sleeps at night can be correlated with how long and how frequently he/she wakes up during that night. It is important to understand these inter-relationships. We propose a joint mixed effect model on total sleep time, number of awakenings, and wake time. We develop an estimating procedure based on a sequence of generalized linear mixed effects models which can be implemented using existing software. The use of these models not only avoids computational intensity and instability that may occur by directly applying a numerical algorithm on a complicated joint likelihood function but also provides additional insights on sleep activities. We demonstrated in simulation studies that the proposed estimating procedure performed well in estimating both fixed and random effects' parameters. We applied the proposed model to data from the Women's Interagency HIV Sleep Study to examine the association of employment status and age with overall sleep quality assessed by several actigraphy measured sleep variables.

Presenter: Yongzhao Shao, New York University

Title: Assessing heterogeneous effects of biomarkers on multi-outcomes in a competing-risk survival analysis

Abstract: There is often a need to evaluate heterogeneous effects on competing survival events due to different causes, increasingly so in competing-risk survival analysis of late-onset Alzheimer’s disease and cancers. In such problems, it is of interest to identify factors that have different effects among the multi-outcomes of competing survival events. We propose a semi-competing risk regression for multi-center multi-outcome data and develop a computing algorithm. Simulation studies are used to demonstrate the effectiveness of such a method under various practical scenarios. We will discuss applications to mortality analysis of multi-outcomes in a multi-center late-onset Alzheimer's disease research to understand the heterogeneous effects

of the allele 4 of the ApoE gene in the context of competing risk survival analysis of multi-outcomes.

Presenter: Shanshan Ding, University of Delaware

Title: Nonconvex-regularized integrative sufficient dimension reduction for multi-source data

Abstract: As advances in high-throughput technology significantly expand data availability, integrative analysis of multiple data sources has become an increasingly important tool for biomedical studies. An integrative and nonconvex-regularized sufficient dimension reduction method is proposed to achieve simultaneous dimension reduction and variable selection for multi-source data analysis in high dimensions. The proposed method aims to extract sufficient information in a supervised fashion and the asymptotic results establish a new theory for integrative sufficient dimension reduction and allow the number of predictors in each data source to increase exponentially fast with sample size. The promising performance of the integrative estimator and efficient numerical algorithms is demonstrated through simulation and real data examples.

Presenter: Yun Li, University of Pennsylvania

Title: Impact of Unmeasured Confounding on Clustered Data Analyses

Abstract: As the health care system becomes more digitized, large administrative databases become increasingly available. This provides valuable opportunities to conduct observational studies to evaluate the effectiveness and quality of care in actual practice and on a large scale. However, the validity of these studies is often threatened by unmeasured confounding, a source of bias which frequently occurs but is particularly difficult to remedy. Instrumental variable methods are popular choices for obtaining robust estimates in the presence of unmeasured confounding. In this talk, I will discuss the impact of unmeasured confounding on the bias of effect estimators from instrumental variable analyses and several common alternative methods for clustered data. I will show that the alternative methods may be more robust depending on the nature of unmeasured confounding. These findings provide evidence in selecting the best methods to combat the dominant types of unmeasured confounders and help interpret statistical results in the context of unmeasured confounding.

Invited Session 14: Recent Development of Statistical Methods for Complex Survival Data in Medical Studies

Organizer: Yanqing Sun. Chair: Sebastian Fuchs. Room: HS 403

Presenter: Wenqing He, University of Western Ontario

Title: Parametric and semiparametric estimation methods for survival data under a flexible class of models

Abstract: In survival analysis, accelerated failure time models are useful in modeling the relationship between failure times and the associated covariates where covariate effects are assumed to appear in a linear form in the model. Such an assumption of covariate effects is, however, quite restrictive for many practical problems. To incorporate the flexible nonlinear relationships between covariates and transformed failure times, we propose partially linear single-index models to facilitate the complex relationship between transformed failure times and covariates. We develop two inference methods that handle the unknown nonlinear function in the model from different perspectives. The first approach is weakly parametric, which approximates the nonlinear function globally, whereas the second method is a semiparametric quasi-likelihood approach which focuses on picking up local features. We establish the asymptotic properties of the proposed methods. A real example is used to illustrate the usage of the proposed methods, and simulation studies are conducted to assess the performance of the proposed methods for a broad variety of situations.

Presenter: Yanqing Sun, University of North Carolina at Charlotte

Title: Regression analysis of semiparametric Cox-Aalen transformation models with partly interval-censored data

Abstract: Partly interval-censored data comprising exact and interval-censored observations are prevalent in biomedical, clinical, and epidemiological studies. This paper studies a flexible class of semiparametric Cox-Aalen transformation models for partly interval-censored data. The model offers greater flexibility and has the potential to enhance statistical power. It extends the semiparametric transformation models by allowing potentially time-dependent covariates to work additively on the baseline hazard and extends the Cox-Aalen model through a transformation function. We construct a set of estimating equations and propose an Expectation-Solving (ES) algorithm to facilitate efficient computation. The variance estimators are computed using the weighted bootstrap via the ES algorithm. The proposed ES algorithm is an extension of the Expectation-Maximization (EM) algorithm that can handle general estimating equations beyond those derived from the loglikelihood. The proposed estimators are shown to be consistent and asymptotically normal based on theories of the empirical processes. Simulation studies show that the proposed methods work well. The proposed method is applied to analyze data from a randomized HIV/AIDS trial.

Presenter: Dung-Tsa Chen, Moffitt Cancer Center

Title: How adverse events data in cancer clinical trial could be utilized by statistical methods as a potential biomarker in predicting clinical outcomes

Abstract: Adverse event (AE) is a critical component of clinical trials served as an evaluation tool for treatment toxicity in cancer clinical trials. However, in the era of precision medicine, there is a pressing need for a transformative approach where AE not only fulfills the role of safety assessment but also functions as biomarkers for evaluating treatment efficacy. Unfortunately, AE data has been underutilized to date, mainly due to suboptimal AE reporting methods such as relying solely on the worst grade, which limits the comprehensive assessment of patient safety and efficacy profiles. In this study, given the complexity of AE, we integrate various key AE parameters to derive innovative AE biomarkers. Since AE is likely associated with clinical outcomes (e.g., improved survival outcomes associated with longer treatment duration and potentially more AE experiences), direct use of AE data could lead to biased results. To address this issue, we utilize landmark analysis at an early time point to define early AE biomarkers to increase the predictive value. Results from multiple immunotherapy cancer trials demonstrated the potential clinical utility of early AE-derived biomarkers in predicting clinical outcomes.

Presenter: Grace Yi, University of Western Ontario

Title: Graphical proportional hazards measurement error models

Abstract: In survival data analysis, the Cox proportional hazards (PH) model is perhaps the most widely used model to feature the dependence of survival times on covariates. While many inference methods have been developed under such a model or its variants, those models are not adequate for handling data with complex structured covariates. High-dimensional survival data often entail several features: (1) many covariates are inactive in explaining the survival information, (2) active covariates are associated with a network structure, and (3) some covariates are error-contaminated. To handle such kinds of survival data, we propose graphical PH measurement error models and develop inferential procedures for the parameters of interest. Our proposed models significantly enlarge the scope of the usual Cox PH model and have great flexibility in characterizing survival data. Theoretical results are established to justify the proposed methods. Numerical studies are conducted to assess the performance of the proposed methods.

Invited Session 15: Recent Advances in Neuroimaging Analysis

Organizer: Lexin Li. Chair: Shuangge Ma. Room: HS 401

Presenter: Hernando Ombao, King Abdullah University of Science and Technology

Title: Overview of Functional Dependence in Brain Networks

Abstract: Brain activity is complex. A full understanding of brain activity requires careful study of its multi-scale spatial-temporal organization (from neurons to regions of interest; and from

transient events to long-term temporal dynamics). Motivated by these challenges, we will explore some characterizations of dependence between components of a brain network. This is potentially interesting because alterations in functional brain connectivity are associated with mental and neurological diseases. In this talk, we provide an overview of functional dependence measures. We present a general framework for exploring dependence through the oscillatory activities derived from each component of the tine series. The talk will draw connections of this framework to some of the classical notions of spectral dependence such as coherence, partial coherence, and dual-frequency coherence. Moreover, this framework provides a starting point for exploring potential non-linear cross-frequency interactions. These interactions include the impact of the phase of one oscillatory activity in one component on the amplitude of another oscillation. The proposed approach captures lead-lag relationships and hence can be used as a general framework for spectral causality. Under this framework, we will also present some recent work on inference using spectral mutual information and entropy measures. This is joint work with Marco Pinto (UC Irvine), Paolo Redondo (KAUST), and Raphael Huser (KAUST).

Presenter: Jaroslaw Harezlak, Indiana University

Title: Novel penalized regression method applied to study the association of brain functional connectivity and alcohol drinking

Abstract: The intricate associations between brain functional connectivity and clinical outcomes are difficult to estimate. Common approaches used do not account for the interrelated connectivity patterns in the functional connectivity (FC) matrix which can jointly and/or synergistically affect the outcomes. In our application of a novel penalized regression approach called SpINNEr (Sparsity Inducing Nuclear Norm Estimator) we identify brain FC patterns that predict drinking outcomes. Results dynamically summarized in the R shiny app indicate that this scalar-on-matrix regression framework via the SpINNEr approach uncovers numerous reproducible FC associations with alcohol consumption.

Presenter: Haoda Fu, Eli Lilly

Title: LLM Is Not All You Need. Generative AI on Smooth Manifolds

Abstract: Generative AI is a rapidly evolving technology that has garnered significant interest lately. In this presentation, we'll discuss the latest approaches organizing them within a cohesive framework using stochastic differential equations to understand complex high-dimensional data distributions. We'll highlight the necessity of studying generative models beyond Euclidean spaces, considering smooth manifolds essential in areas like robotics and medical imagery, and for leveraging symmetries in the de novo design of molecular structures. Our team's recent advancements in this blossoming field, ripe with opportunities for academic and industrial collaborations, will also be showcased.

Presenter: Lexin Li, University of California, Berkeley

Title: Kernel Ordinary Differential Equations

Abstract: Ordinary differential equation (ODE) is widely used in modeling biological and physical processes in science. In this talk, we propose a new reproducing kernel-based approach for estimation and inference of ODE given noisy observations. We do not assume the functional forms in ODE to be known, or restrict them to be linear or additive, and we allow pairwise interactions. We perform sparse estimation to select individual functionals, and construct confidence intervals for the estimated signal trajectories. We establish the estimation optimality and selection consistency of kernel ODE under both the low-dimensional and high-dimensional settings, where the number of unknown functionals can be smaller or larger than the sample size. Our proposal builds upon the smoothing spline analysis of variance (SS-ANOVA) framework, but tackles several important problems that are not yet fully addressed, and thus extends the scope of existing SS-ANOVA as well. We demonstrate the efficacy of our method through numerous ODE examples.

Invited Session 16: The integration of high dimensional data in joint models

Organizer: Virginie Rondeau. Chair: Dimitris Rizopoulos. Room: HS 402

Presenter: Pedro Miranda Afonso, Erasmus MC

Title: A fast approach to analyzing large datasets with joint models for longitudinal and time-to-event outcomes

Abstract: The joint modeling of longitudinal and time-to-event outcomes has become a popular tool in follow-up studies. However fitting Bayesian joint models to large datasets such as patient registries can require extended computing times. To speed up sampling we divided a patient registry dataset into subsamples analyzed them in parallel and combined the resulting Markov chain Monte Carlo draws into a consensus distribution. We used a simulation study to investigate how different consensus strategies perform with joint models. In particular we compared grouping all draws together with using equal- and precision-weighted averages. We considered scenarios reflecting different sample sizes numbers of data splits and processor characteristics. Parallelization of the sampling process substantially decreased the time required to run the model. We found that the weighted-average consensus distributions for large sample sizes were nearly identical to the target posterior distribution. The proposed algorithm has been made available in an R package for joint models JMbayes2. This work was motivated by the clinical interest in investigating the association between ppFEV1 a commonly measured marker of lung function and the risk of lung transplant or death using data from the US Cystic Fibrosis Foundation Patient Registry (35153 individuals with 372366 years of cumulative follow-up). Splitting the registry into five subsamples resulted in an 85% decrease in computing time from 9.22 to 1.39 hours. Splitting the data and finding a consensus distribution by precision-weighted averaging proved to be a computationally efficient and robust approach to handling large datasets under the joint modeling framework.

Presenter: Cécile Proust Lima, Bordeaux Population Health Research Center

Title: Analysis of multivariate longitudinal and survival data: what about random forests?

Abstract: Health studies usually involve the collection of variables repeatedly measured over time. This includes exposures (e.g. treatment blood pressure nutrition) markers of progression (e.g. brain volumes blood tests cognitive functioning tumor size) and times to clinical endpoints (e.g. death diagnosis). Joint models for longitudinal and survival data are now widely used in biostatistics to analyze such longitudinal data and address a variety of etiological and predictive questions. However they quickly encounter numerical limitations as the number of repeated variables substantially increases making it challenging for them to address the in-depth medical research questions raised by the complex longitudinal information collected nowadays. In this talk we tackle the challenge of the prediction of clinical endpoints from repeated measures of a large number of markers using another paradigm: the random survival forests. Random survival forests constitute a flexible method of prediction that can handle high-dimensional predictors and capture complex relationships with the outcome to predict. However they are limited to time-invariant predictors. We show how random survival forests (possibly with competing causes of events) can be extended to time-varying noisy predictors by incorporating a modeling step into the recursive tree building procedure. The performances of the methodology implemented in the DynForest R package are assessed in simulations both in a small dimensional context (in comparison with joint models) and in a large dimensional context (in comparison with a regression calibration method that ignores the missing data informative mechanism in the time-varying covariates). The methodology is also illustrated in dementia research to (i) predict the individual probability of dementia using multi-modal repeated information (e.g. cognition brain structure) and (ii) quantify the relative importance of each type of markers.

Presenter: Denis Rustand, King Abdullah University of Science and Technology

Title: Efficient Inference for Joint Models of Multivariate Longitudinal and Survival Data Using INLAjoint

Abstract: Clinical research often requires the simultaneous study of longitudinal and survival data. Joint models which can combine these two types of data are essential tools in this context. A joint model involves multiple regression submodels (one for each longitudinal/survival outcome) usually linked together through correlated or shared random effects. This makes their estimation process rather complex time-consuming and sometimes even unfeasible especially when dealing with many outcomes. In this context we introduce INLAjoint a user-friendly and flexible R package designed to leverage the Integrated Nested Laplace Approximation (INLA) method from the INLA R package renowned for its computational efficiency and speed (Rue et al. 2009). INLAjoint can handle various model formulations and simplifies the application of INLA to fit joint models ensuring fast and accurate parameter estimation. Our simulation studies show that INLA reduces the computation time substantially compared to alternative strategies such as Bayesian inference via Markov Chain Monte Carlo without compromising on accuracy. (Rustand et al. 2023). A key application of joint models is the dynamic prediction of the risk of an event such as death or disease progression based on changes in the longitudinal outcome(s)

over time. INLAjoint allows for the estimation of dynamic risk predictions and can incorporate changes in the longitudinal outcome(s) to update future risk predictions. This makes INLAjoint a valuable tool for analyzing complex health data.

Presenter: Manel Rakez, Bordeaux Population Health Center

Title: Evolution of breast density over time and its impact on breast cancer diagnosis during screening

Abstract: Breast cancer (BC) is the leading cause of cancer death in women worldwide. Mammography-based screening programs reduce breast cancer (BC) mortality by promoting earlier detection. The mammography's sensitivity depends on breast density (BD). The latter is subject to changes over time, affecting the risk of a BC diagnosis. Women with high BD are likelier to develop BC, and their mammography's sensitivity is lessened. Thus, to better understand the impact of temporal BD changes on BC diagnosis risk during screening, we propose a new methodology to predict BC risk accounting for the deep learning assessment of the sequential BD. From the sequential and complete mammography exams of 131,209 women participating in the BC screening program, the percent density (PD), a quantitative estimation of a woman's BD at each visit, is estimated using MammoDL. This segmentation model comprises two successive modified U-Nets allowing for breast identification from the entire mammogram first and dense tissue region delineation from the breast region second. A ResNet-34 replaces the U-Net encoder to alleviate training challenges. In addition, this model is fine-tuned to extend its use to processed GE and Hologic vendors' images. Then, a joint model for a linear biomarker and a time-to-event outcome is implemented using the consensus Monte Carlo algorithm. First, the temporal trajectory of the PD is described using a linear mixed-effect model, adjusted on factors impacting the BD, such as age. This sub-model is flexible in dealing with irregular intervals between screening visits and outcome-dependent drop-out. Second, the individual and dynamic prediction of BC diagnosis is estimated conditionally on the biomarker's intermediate longitudinal measurements and is defined over the screening period. This probability is derived for each woman and is dynamically updated as PD measurements accumulate. We propose a reproducible method to estimate BD's temporal evolution and its impact on BC diagnosis. The segmentation model gives a quantitative estimation of the BD at each screening visit. The joint model uses the biomarker's repeated measurements to dynamically update the BC diagnosis prediction throughout the screening period.

Break 10:45-11:00, July 10, 2024

Parallel Invited Sessions 17-20: 11:00-12:45, July 10, 2024

Invited Session 17: Recent developments in Survival Analysis and Statistical Machine Learning

Organizer: Gang Li. Chair: Gang Li. Room: HS 402

Presenter: Xiaowu Dai, University of California, Los Angeles

Title: Kernel ordinary differential equations

Abstract: The ordinary differential equation (ODE) is widely used in modelling biological and physical processes in science. A new reproducing kernel based approach is proposed for the estimation and inference of ODE given noisy observations. The functional forms in ODE are assumed to be known or restricted to be linear or additive, and pairwise interactions are allowed. Sparse estimation is performed to select individual functionals and construct confidence intervals for the estimated signal trajectories. The estimation optimality and selection consistency of kernel ODE are established under both the low-dimensional and high-dimensional settings, where the number of unknown functionals can be smaller or larger than the sample size. The proposal builds upon the smoothing spline analysis of variance (SS-ANOVA) framework, but tackles several important problems that are not yet fully addressed, and thus extends the scope of existing SS-ANOVA too.

Presenter: Douglas Schaubel, University of Pennsylvania

Title: Dynamic Risk Assessment by Landmark Modeling of the Restricted Mean Survival Time

Abstract: Dynamic risk assessment is an important tool for healthcare providers to inform treatment selection for the purposes of optimizing patient outcomes and/or avoiding over-treatment. The risk of adverse events is regularly assessed based on changes in particular biomarkers or vital signs in order to evaluate a patient's medical status and perhaps urgency of treatment receipt. Landmark analysis is a useful dynamic prediction approach which obviates the need to jointly model the time-dependent biomarkers and time-to-event outcome. The majority of landmark methods for survival analysis utilize a hazard regression model (typically, a Cox model) to quantify the association between the longitudinal predictors and outcomes. In addition, most methods assume independent censoring. In order to broaden the scope of landmark analysis, we propose landmark methods which directly model the restricted mean survival time (RMST) and allow for dependent censoring. Advantages of RMST models include removing the assumption that the hazard model is correct at every time point, since RMST models consider a single restriction time as opposed to a process. Moreover, many investigators prefer the area under the survival curve over hazard rate as a clinical endpoint due to interpretability. Asymptotic properties of the proposed estimators are derived, and a comprehensive simulation study demonstrates their decent performance in finite samples. The proposed methods are illustrated using national registry data on a cohort of end-stage liver disease patients. This is joint work with Yuan Zhang.

Presenter: Ying Lu, Stanford University

Title: Using the Desirability of Outcome Ranking (DOOR) Approach to Construct Multicomponent Endpoints

Abstract: Complex disorders affect multiple symptom domains measured by multiple outcomes. Successful treatments may affect one or several domains that may vary among patients. Multiple component (MC) endpoints that integrate outcomes across multiple domains help the evaluation of totality of treatment benefits. In this talk, we present a general approach to construct a MC endpoint from multiple domains according to their relative ranking in an evaluation system. The ranking of outcome variables can be defined in a protocol (a shared decision making (SDM) trial) or vary by treatment approaches (such as for traditional Chinese medicine trials), or by patient preferences (such as the Patient-ranked Order of Function (PROOF) score for Amyotrophic Lateral Sclerosis (ALS) trials). Using the desirability of outcome ranking (DOOR) approach, we can construct the Mann-Whitney U-statistics to estimate the probability of a treated participant having more desirable outcomes than a control participant. This approach has the advantage of flexibility in how many domains to be integrated, independent of measurement units, and improvement in relevance of efficacy and statistical power. We demonstrated this approach using the results from the ENHANCE-AF trial (NCT04096781), which evaluated a novel SDM pathway for patients considering anticoagulation for stroke prevention, and the follow-up data from the development of the PROOF in prediction of ALS patient survival. We will discuss challenges in using this approach and strategies to address them. The presentation is based on collaborations with Professor Lu Tian, Paul Wang, and Randal l Stafford at Stanford University and Professors Ruben van Eijk and Leonard vd Berg at the University Medical Center, Utrecht, the Netherlands.

Presenter: Yuhua Zhu, University of California, Los Angeles

Title: Continuous-in-time Reinforcement Learning

Abstract: When the data is discrete-in-time, how can we solve the continuous-in-time reinforcement learning problems? Given the prevalence of continuous-time dynamics in various real-world applications, our objective is to solve the optimal control problem in the presence of unknown dynamics. First, we show that the Bellman equation serves as a first-order approximation to continuous-time problems. Then, we derive higher-order equations based on partial differential equations (PDEs). To efficiently solve continuous-time reinforcement learning problems with discrete-in-time data, we further propose algorithms for solving the PDE-based Bellman equations.

Invited Session 18: New developments for analyzing EHR and other observational data

Organizer: Shuangge Ma. Chair: Cecile Proust-Lima. Room: HS 401

Presenter: Hao Mei, Renmin University of China

Title: Clinical Human Disease Networks with Healthcare Administrative Claims Data

Abstract: Clinical treatment outcomes are the quality and cost targets that healthcare providers aim to improve. Most existing outcome analysis focuses on a single disease or all diseases combined, which ignores the complex interconnections among diseases. Motivated by the success of molecular and phenotypic human disease networks (HDNs), we develop clinical HDNs that describe the interconnections among diseases in terms of multiple clinical treatment outcomes. In this framework, one node represents one disease, and two nodes are linked with an edge if their outcomes are conditionally dependent. Along this direction, we also develop a time-dependent clinical HDN to investigate the temporal variation of disease interconnection from a clinical point of view. Our data experiments validate the performance of the proposed models in identifying correct edges. Analyzing key network properties such as connectivity, module/hub, and temporal variation using healthcare administrative claims data, the findings are not only biomedically sensible but also uncover information that is less/not investigated in the literature. Overall, clinical HDNs can provide additional insight into diseases' properties and their interconnections and assist more efficient disease management and healthcare resource allocation.

Presenter: Shuangge Ma, Yale School of Public Health

Title: Heterogeneous Network Analysis of Disease Clinical Treatment Measures via Mining Electronic Medical Record Data

Abstract: The analysis of clinical treatment measures has been extensively conducted and can facilitate more effective resource management and planning and also assist in better understanding diseases. Most existing analyses have focused on a single disease or a large number of diseases combined. Partly motivated by the successes of gene-centric and phenotypic human disease network (HDN) research, there has been growing interest in the network analysis of clinical treatment measures. However, existing studies have been limited by a lack of attention to heterogeneity and relevant covariates, ineffectiveness of methods, and low data quality. In this study, our goal is to mine the Taiwan National Health Insurance Research Database (NHIRD), a large population-level electronic medical record (EMR) database, and construct HDNs for the number of outpatient visits and medical costs. Significantly advancing from the existing literature, the proposed analysis accommodates heterogeneity and the effects of covariates (for example, demographics). Additionally, the proposed method effectively accommodates the zero-inflation nature of data, Poisson distribution, high-dimensionality, and network sparsity. Computational and theoretical properties are carefully examined. Simulation demonstrates the competitive performance of the proposed approach. In the analysis of NHIRD data, two and five subject groups are identified for outpatient visits and medical costs, respectively. The identified interconnections, hubs, and network modules are found to have sound implications.

Presenter: Kai-Yuan Hsiao, Fu Jen Catholic University

Title: Few-shot Learning of Tabular Medical Records with Large Language Models

Abstract: The medical field is replete with abundant tabular data encompassing everything from patient records to the results of clinical trials. Traditional approaches to the analysis of such data often involve complex statistical techniques, feature engineering, and conventional machine learning methods. However, the integration of large language models (LLMs) offers a transformative method for this analysis. This study examines the application of LLMs for zero-shot and few-shot classification of medical tabular data, employing techniques for generating additional feature information and natural language serialization of tabular data. In this study, the methodology encompasses the serialization of medical tabular data into natural language strings, complete with descriptions of the specific analytical tasks. For instance, tables detailing patient symptoms and diagnoses are transformed into formats readily interpretable by Large Language Models (LLMs). Where only a few examples are present, the generative capacity of LLMs is harnessed to bolster their comprehension of medical contexts. Iterative generation of additional semantically meaningful features is based on the medical background of the dataset, with an investigation into the efficacy of the generated features. In the experimental process, besides testing large language models for downstream tasks, comparisons with traditional machine learning methods are made, with efforts to interpret the results ensuring that medical professionals can understand and trust the findings.

Invited Session 19: Dependence Modeling

Organizer: Wolfgang Trutschnig & Sebastian Fuchs. Chair: Wolfgang Trutschnig. Room: HS 414

Presenter: Damjana Kokol Bukovsek, University of Ljubljana

Title: Exact upper bound for bivariate copulas with a given diagonal section

Abstract: For any bivariate copula or quasi-copula $C: \mathbb{I}^2 \to \mathbb{I}$ its diagonal section $\delta_C: \mathbb{I} \to \mathbb{I}$, defined by $\delta_C(x) = C(x,x)$, is increasing, 2-Lipschitz, satisfies $\delta_C(1) = 1$ and $\delta_C(x) \leq x$ for all $x \in \mathbb{I}$. On the other hand, given any function $\delta: \mathbb{I} \to \mathbb{I}$, which is increasing, 2-Lipschitz, satisfies $\delta_C(1) = 1$ and $\delta_C(x) \leq x$ for all $x \in \mathbb{I}$, there exists a copula, such that δ is its diagonal section. Given a function δ with these properties, it is known what is the exact (pointwise) lower bound of all bivariate copulas with diagonal section δ; it is the Bertino copula. The same function is also the lower bound for quasi-copulas. Furthermore, the exact upper bound of all bivariate quasi-copulas with diagonal section δ is also known. The main goal of this talk is to answer a question of the exact upper bound for bivariate copulas with a given diagonal section δ by giving an explicit formula for this bound. We achieve this by constructing a new copula with prescribed diagonal section, which attains the bound on the entire upper-left triangle of the unit square. We also answer the question for which diagonal sections this exact bound is a copula. As an application of our main result, we determine the maximal asymmetry of bivariate

copulas with a given diagonal section and construct a copula that attains it. This is joint work with Blaž Mojškerc (University of Ljubljana) and Nik Stopar (University of Ljubljana).

Presenter: Nik Stopar, University of Ljubljana

Title: Infima and Suprema of Multivariate Cumulative Distribution Functions

Abstract: A multivariate probability box is a set of cumulative distribution functions bounded point-wise by two standardized functions. It can be used to model imprecision in the knowledge about the true joint distribution function of a random vector. A probability box is coherent if its bounding functions are equal to the point-wise infimum and supremum of the distribution functions contained in the box. This is the main motivation for investigating infima and suprema of sets of multivariate distribution functions. A coherent probability box can be constructed by composing several univariate (marginal) probability boxes with a coherent imprecise copula (i.e., a coherent box of copulas bounded by two quasi-copulas). In this talk, we discuss the question of whether any coherent probability box can be obtained in such a way. If we are only interested in multivariate distributions with fixed marginals (i.e., each marginal probability box contains a single distribution function), then the answer is positive as a consequence of Sklar's theorem. On the other hand, the answer is negative for more general probability boxes if we insist on the standard way of representing a multivariate cumulative distribution function with a copula. Nevertheless, we show that with a slightly modified representation, a positive answer can be achieved under a mild condition on the probability boxes. In particular, we demonstrate how the point-wise infimum and supremum of a family of multivariate distribution functions can be represented with copulas that correspond to the members of the family .

Presenter: Jonathan Ansari, Paris Lodron University of Salzburg

Title: A model-free multi-output variable selection

Abstract: As a direct extension of Azadkia & Chatterjee's rank correlation T to a set of q outcome variables, the novel measure T^q, introduced and investigated in Ansari \& Fuchs, quantifies the scale-invariant extent of functional dependence of a multi-output vector $Y = (Y1,..., Yq)$ on a number of p input variables $X = (X1,..., Xp)$ and fulfils all the desired characteristics of a measure of predictability, namely $0 <= Tq(Y|X) <= 1$, $Tq(Y|X) = 0$ if and only if Y and X are independent, and $Tq(Y|X) = 1$ if and only if Y is perfectly dependent on X. Based on various useful properties of $Tq(Y|X)$, a model-free and dependence-based feature ranking and forward feature selection of data with multiple output variables is presented, thus facilitating the selection of the most relevant explanatory variables.

Presenter: Patrick Langthaler, Paris Lodron University of Salzburg

Title: Quantifying and estimating dependence via sensitivity of conditional distributions

Abstract: Recently established, directed dependence measures for pairs (X, Y) of random variables build upon the natural idea of comparing the conditional distributions of Y given X = x with the marginal distribution of Y . They assign pairs (X, Y) values in [0, 1], the value is 0 if and only if X, Y are independent, and it is 1 exclusively for Y being a function of X. We show that comparing randomly drawn conditional distributions with each other instead or, equivalently, analyzing how sensitive the conditional distribution of Y given X = x is on x, opens the door to constructing novel families of dependence measures $\Lambda\varphi$ induced by general convex functions $\varphi : R \to R$, containing, e.g., Chatterjee's coefficient of correlation as special case. After establishing additional useful properties of $\Lambda\varphi$ we focus on continuous (X, Y), translate $\Lambda\varphi$ to the copula setting, consider the Lp-version and establish an estimator which is strongly consistent in full generality. A real data example and a simulation study illustrate the chosen approach and the performance of the estimator. Complementing the afore-mentioned results, we show how a slight modification of the construction underlying $\Lambda\varphi$ can be used to define new measures of explainability generalizing the fraction of explained variance.

Invited Session 20: Extreme value analysis.

Organizer: Judy Wang. Chair: Elizabeth Juarez-Colunga. Room: HS 403

Presenter: Olivier Wintenberger, University of Vienna

Title: On the asymptotics of extremal blocks cluster inference

Abstract: Extremes occur in stationary regularly varying time series as short periods with several large observations. To infer cluster statistics such as the extremal index, cluster size probabilities, and other cluster indices, we focus on extremal blocks that have a large norm. The choice of the norm is crucial and optimally depends on the tail index of the marginal distribution. We state the asymptotic normality of block estimators for cluster inference based on consecutive observations with large ℓ^α-norms, approximating the tail index $\alpha>0$ using the Hill estimator. Regarding linear models, we prove that the asymptotic variance is null as first conjectured by Hsing (1993). We illustrate our findings on simulations. Based on a joint work with G. Buriticá (Geneva University).

Presenter: Sebastian Engelke, University of Geneva

Title: Machine learning beyond the data range: an extreme value perspective

Abstract: Machine learning methods perform well in prediction tasks within the range of the training data. These methods typically break down when interest is in (1) prediction in areas of the predictor space with few or no training observations; or (2) prediction of quantiles of the response that go beyond the observed records. Extreme value theory provides the mathematical foundation for extrapolation beyond the range of the training data both in the dimension of the predictor space and the response variable. In this talk, we present recent methodology that

combines this extrapolation theory with flexible machine learning methods to tackle the out-of-distribution generalization problem (1) and the extreme quantile regression problem (2). We show the practical importance of prediction beyond the training observations in environmental and climate applications where domain shifts in the predictor space occur naturally due to climate change and risk assessment for extreme quantiles is required.

Presenter: Tiandong Wang, Fudan University

Title: Testing for Strong VS Full Dependence

Abstract: Preferential attachment models of network growth are bivariate heavy-tailed models for in- and out-degree with limit measures which either concentrate on a ray of positive slope from the origin or on all of the positive quadrant depending on whether the model includes reciprocity or not. Concentration on the ray is called full dependence. If there were a reliable way to distinguish full dependence from not-full, we would have guidance about which model to choose. This motivates investigating tests that distinguish between (i) full dependence; (ii) strong dependence (limit measure concentrates on a proper subcone of the positive quadrant); (iii) concentration on the positive quadrant. We give two test statistics and discuss their asymptotically normal behavior under full and not-full dependence. This is a joint work with Prof. Sidney Resnick at Cornell University.

Presenter: Chen Zhou, Erasmus University Rotterdam

Title: Tail copula estimation for heteroscedastic extremes

Abstract: Consider independent multivariate random vectors which follow the same copula, but where each marginal distribution is allowed to be non-stationary. This non-stationarity is for each marginal governed by a scedasis function (see Einmahl et al. (2016)) that is the same for all marginals. We establish the asymptotic normality of the usual rank-based estimator of the stable tail dependence function, or, when specialized to bivariate random vectors, the corresponding estimator of the tail copula. Remarkably, the heteroscedastic marginals do not affect the limiting process. Next, under a bivariate setup, we develop nonparametric tests for testing whether the scedasis functions are the same for both marginals. Detailed simulations show the good performance of the estimator for the tail dependence coefficient as well as that of the new tests. In particular, novel asymptotic confidence intervals for the tail dependence coefficient are presented and their good finite-sample behavior is shown. Finally an application to the S&P500 and Dow Jones indices reveals that their scedasis functions are about equal and that they exhibit strong tail dependence.

Lunch 12:45-1:45, July 10, 2024 2024 (FREE for registrants)

Parallel Invited Sessions 21-24: 1:45-3:30, July 10, 2024

Invited Session 21: Dynamic Prediction

Organizer: Jessica Barrett. Chair: Jessica Barrett. Room: HS 402

Presenter: Hein Putter, Leiden University

Title: Dynamic prediction with many biomarkers: combining landmarking 2.0 with multivariate Functional Principal Component Analysis

Abstract: Predicting patient survival based on longitudinal biomarker measurements poses a common statistical challenge. In many cases, the volume of longitudinal data exceeds what can be practically managed with a joint model. For this reason, numerous methods employ a multi-step landmarking approach where the longitudinal data up to a specific landmark time is summarized and used in a subsequent landmark model. In our previous research, we utilized multivariate Functional Principal Component Analysis (mFPCA) to summarize the available longitudinal data and used a proportional hazards landmark model for prediction. We demonstrated the effectiveness of a "strict" landmarking approach where only the information preceding the landmark is utilized. In this presentation, we explore the advancements in landmarking 2.0 to further improve on this approach. Our approach involves using the mFPCA results up to the landmark to forecast the progression of the longitudinal biomarkers from the landmark time onward until the prediction horizon. We then fit a time-dependent Cox model incorporating these predictable time-dependent covariates as the foundation for a landmark model. We demonstrate the utility of this method for dynamic prediction, assess its performance through simulation studies, and illustrate it in real data.

Presenter: Dimitris Rizopoulos, Erasmus University Rotterdam

Title: Optimizing Dynamic Predictions from Joint Models Using Super Learning

Abstract: Joint models for longitudinal and time-to-event data are often employed to calculate dynamic individualized predictions used in numerous applications of precision medicine. Two components of joint models that influence the accuracy of these predictions are the shape of the longitudinal trajectories and the functional form linking the longitudinal outcome history to the hazard of the event. Finding a single well-specified model that produces accurate predictions for all subjects and follow-up times can be challenging, especially when considering multiple longitudinal outcomes. In this work, we use the concept of super learning and avoid selecting a single model. In particular, we specify a weighted combination of the dynamic predictions calculated from a library of joint models with different specifications. The weights are selected to optimize a predictive accuracy metric using V-fold cross-validation. We use as predictive accuracy measures the expected quadratic prediction error and the expected predictive cross-entropy. In a simulation study, we found that the super learning approach produces results

similar to the Oracle model which performed best in the test datasets. All proposed methodology is implemented in the freely available package JMbayes2.

Presenter: Danilo Alvares, University of Cambridge

Title: A two-stage approach for Bayesian joint modelling of competing risks and multiple longitudinal outcomes

Abstract: Recent trends in personalized healthcare have motivated great interest in the dynamic prediction of survival and other clinically important events by using baseline characteristics and the evolving history of disease progression. Some of the methodological developments were motivated by case studies in multiple myeloma (a type of bone marrow cancer) where progression is assessed by several biomarker trajectories and patients may experience multiple regimen changes over time. To understand the dynamic interplay between biomarkers and their connections to the survival process, a two-stage Bayesian joint model is developed for competing risks and multiple longitudinal outcomes. The proposal is applied to an observational study from the US nationwide Flatiron health electronic health record (EHR)-derived de-identified database where patients diagnosed with multiple myeloma from January 2015 to February 2022 were selected. The data is split into training and test sets in order to assess the performance of the proposal in making dynamic predictions of times to events of interest (time to next line of therapy or time to death) using baseline variables and longitudinally measured biomarkers available up to the time of prediction. Residuals validated the robustness of the model and the calibration supported its good predictive accuracy.

Presenter: Liang Li, MD Anderson Cancer Center

Title: Backward Joint Model for the Dynamic Prediction of Multivariate Longitudinal and Survival Outcomes

Abstract: Joint modeling is an important approach to dynamic prediction of clinical outcomes using longitudinally measured predictors, such as biomarkers. We consider the situation where the predictors include baseline covariates and the longitudinal trajectories of many correlated biomarkers, measured asynchronously at irregularly spaced time points. The outcomes of predictive interest include both the terminal clinical event with or without competing risks, and the future longitudinal biomarker trajectories if the terminal or competing risk events do not occur. We propose a novel backward joint model (BJM) to solve this problem. The BJM can be flexibly specified to optimize the prediction accuracy. Its likelihood-based estimation algorithm is robust, fast, and stable regardless of the dimension of longitudinal biomarkers. We illustrate the BJM methodology with simulations and a real dataset from the African American Study of Kidney Disease and Hypertension.

Invited Session 22: Multivariate and Complex Longitudinal Data

Organizer: Georg Zimmermann. Chair: Georg Zimmermann. Room: HS 403

Presenter: Geert Molenberghs, Hasselt University

Title: A Broad Framework for Likelihood Alternatives, in View of Small, Very Large, and Variable-Size Studies with Multivariate and/or Repeated Measures

Abstract: We consider a number of data settings where the use of standard maximum likelihood or other estimation method is complicated for a number of reasons: data structures are complex, there are very large data streams, in reverse, there are very small trials (like in orphan diseases), or there are non-standard design features (sequential trials, missing data, clustered data with variable size, etc.). Specific challenges arise when data are multivariate and/or longitudinal. The use of alternatives to maximum likelihood are explored with particular emphasis on pseudo-likelihood, split-sample methods, and even closed-form estimators in settings where one would not expect them. Specific attention is devoted to the computational feasibility of the proposed methods. We pay particular attention to the existence of closed forms in our modified procedures. All settings are illustrated using real-life examples.

Presenter: Frank Konietschke, Charité medical university of Berlin

Title: Statistical Planning and Evaluation of Translational Trials

Abstract: Any trial should start with its careful planning and especially with sample size calculations in particular with regard to sample size and power considerations. The planning phase of an experiment is key since errors in the statistical planning can have severe consequences on both the results and conclusions drawn from the data. In translational research (preclinical and early clinical) false conclusions highly affect subsequent trials and thus mistakes proliferate a rather unethical outcome. In statistical practice, most studies are planned based on t-tests and Wald-type statistics (including ANOVA) and make some strict distributional assumptions. Sample sizes are typically small and if planning assumptions are not met, the trials are either underpowered or too large, result in wrong conclusions and waste resources and might even be misleading. On the other hand, nonparametric ranking methods (such as Wilcoxon-Mann-Whitney test, Brunner-Munzel test, multiple contrast tests, and their generalizations) are excellent alternatives to such parametric approaches. However, sample size formulas as well as detailed power analyses are yet to be implemented for broad classes of such tests. In this talk, we discuss statistical planning and evaluation methods for translational trials. Real data sets illustrate the methods.

Presenter: Somnath Datta, University of Florida

Title: Specialized Statistical Analyses of Iowa Fluoride Study Data

Abstract: We present both Bayesian and frequentist analysis for longitudinal data that are clustered and non-continuous (more specifically count and ordinal) and exhibit zero inflation patterns. The ultimate goal is to undertake a comprehensive and unified statistical examination of the total accumulation of dental caries and fluorosis data obtained from Iowa Fluoride Study participants. More specifically, we fit longitudinal statistical models to caries and fluorosis scores data obtained at ages five, nine, thirteen, seventeen, and twenty-three for the participants in this cohort study of Iowa children. The ultimate goal is to study the time-varying (in particular long-term) and joint effects of various risk and protective factors for dental caries and fluorosis outcomes.

Presenter: Kelly van Lancker, University of Gent

Title: Ensuring valid inference for Cox hazard ratios after variable selection

Abstract: The problem of how to best select variables for confounding adjustment forms one of the key challenges in the evaluation of exposure effects in observational studies, and has been the subject of vigorous recent activity in causal inference. A major drawback of routine procedures is that there is no finite sample size at which they are guaranteed to deliver exposure effect estimators and associated confidence intervals with adequate performance. In this work, we will consider this problem when inferring conditional causal hazard ratios from observational studies under the assumption of no unmeasured confounding. The major complication that we face with survival data is that the key confounding variables may not be those that explain the censoring mechanism. In this presentation, we overcome this problem using a novel and simple procedure that can be implemented using off-the-shelf software for penalized Cox regression. In particular, we will propose tests of the null hypothesis that the exposure has no effect on the considered survival endpoint which are uniformly valid under standard sparsity conditions. Simulation results show that the proposed methods yield valid inferences even when covariates are high-dimensional.

Invited Session 23: Topics in Advanced Multi-outcome Data Analysis

Organizer: Lei Liu. Chair: Lei Liu. Room: HS 401

Presenter: Chong He and Dongchu Sun, University of Missouri

Title: Bayesian Analysis of Multivariate One-Way ANOVA Model

Abstract: Multivariate one-way ANOVA model is of substantial importance in contemporary statistical theory and application. It could be used to data fusion for analyzing data from different resources, allowing us to assess differences and correlations among different groups or resources. One primary interest is to estimate unknown overall mean and two covariance components (matrices). The usual frequentist estimates for covariance components, such as MLE or moment

estimators, may be non-positive definite. On account of the inapplicability of the usual frequentist method, we deal with this problem using Bayesian hierarchical modeling. Bayesian analysis under various subjective and objective priors is studied. For two covariance components, a new class of commutative priors is proposed. Interestingly, the commutative prior is also a conjugate prior. Propriety and moment existence are derived for both the priors and their posteriors. Moreover, a new and computationally effective MCMC algorithm is developed for the proposed commutative priors. Simulation and real data analysis show the advantages of the commutative priors.

Presenter: Donglin Zeng, University of Michigan

Title: Fusing Individualized Treatment Rules Using Auxiliary Outcomes

Abstract: An individualized treatment rule (ITR) is a decision rule that recommends treatments for patients based on their covariates. In practice, the optimal ITR that maximizes its associated value function is also expected to cause little harm to other non-primary outcomes. Hence, one goal is to learn the ITR that not only maximizes the value function for the primary outcome but also approximates the optimal rule for the other auxiliary outcomes as closely as possible. In this work, we propose a fusion penalty to encourage ITRs based on the primary outcome and auxiliary outcomes to yield similar recommendations. We then optimize a surrogate loss function using empirical data for estimation. We derive the non-asymptotic properties for the proposed method and show that the agreement rate between the estimated ITRs for primary and auxiliary outcomes converges faster to the true agreement rate as compared to methods without using auxiliary outcomes. Finally, simulation studies and a real data example are used to demonstrate the finite-sample performance of the proposed method.

Presenter: Yuanjia Wang, Columbia University

Title: A Hierarchical Random Effects State-space Model for Multi-channel EEG Data

Abstract: Mental disorders present challenges in diagnosis and treatment due to their complex and heterogeneous nature. Electroencephalogram (EEG) has shown promise as a source of potential biomarkers for these disorders. However, existing methods for analyzing EEG signals have limitations in addressing heterogeneity and capturing complex brain activity patterns between regions. This paper proposes a novel random effects state-space model (RESSM) for analyzing large-scale multi-channel resting-state EEG signals accounting for the heterogeneity of brain connectivities between groups and individual subjects. We incorporate multi-level random effects for temporal dynamical and spatial mapping matrices and address nonstationarity so that the brain connectivity patterns can vary over time. The model is fitted under a Bayesian hierarchical model framework coupled with a Gibbs sampler. Compared to previous mixed-effects state-space models, we directly model high-dimensional random effects matrices without structural constraints and tackle the challenge of identifiability. Through extensive simulation studies, we demonstrate that our approach yields valid estimation and inference. We apply RESSM to a multi-site clinical trial of Major Depressive Disorder (MDD). Our analysis uncovers

significant differences in resting-state brain temporal dynamics among MDD patients compared to healthy individuals. In addition, we show the subject-level EEG features derived from RESSM exhibit a superior predictive value for the heterogeneous treatment effect compared to the EEG frequency band power, suggesting the potential of EEG as a valuable biomarker for MDD.

Presenter: Haibo Zhou, University of North Carolina at Chapel Hill

Title: Semiparametric regression analysis of case-cohort studies with multiple interval-censored disease outcomes

Abstract: Interval-censored failure time data commonly arise in epidemiological and biomedical studies where the occurrence of an event or a disease is determined via periodic examinations. In this work, we formulate the case-cohort design with multiple interval-censored disease outcomes and also generalize it to non-rare diseases where only a portion of diseased subjects are sampled. We develop a marginal sieve weighted likelihood approach which assumes that the failure times marginally follow the proportional hazards model. We consider two types of weights to account for the sampling bias and adopt a sieve method with Bernstein polynomials to handle the unknown baseline functions. We employ a weighted bootstrap procedure to obtain a variance estimate that is robust to the dependence structure between failure times. The proposed method is examined via simulation studies and illustrated with a dataset on incident diabetes and hypertension from the Atherosclerosis Risk in Communities (ARIC) study.

Invited Session 24: General topics in multi-outcome data

Organizer: Zhezhen Jin. Chair: Yichuan Zhao. Room: HS 414

Presenter: Xuewen Lu, University of Calgary

Title: Variable Selection in Joint Frailty Model of Recurrent and Terminal Events with Diverging Number of Covariates

Abstract: In many biomedical applications, the recurrent event data are subject to an informative terminal event, for example, death. Joint modeling of recurrent and terminal events has attracted many research interests, however, very few works have been done for simultaneous estimation and variable selection for joint frailty proportional hazards models, moreover, it is lacking a theoretical justification and a validity when the dimension of covariates is diverging with the sample size. To fill this gap, we propose a broken adaptive ridge (BAR) regression procedure that combines the strengths of the quadratic regularization and the adaptive weighted bridge shrinkage. We establish the oracle property of the BAR regression. In the simulation study, the results indicate that the BAR regression outperforms the existing variable selection methods. Finally, the proposed method is applied to a real dataset for illustration.

Presenter: Yingwei Paul Peng, Queen's University

Title: Joint Analysis of Longitudinal Ordinal Categorical Item Response Data and Survival Times with Cure Fraction

Abstract: For longitudinal ordinal categorical item response data that may not be observable after a subject develops a terminal event, some statistical models were proposed for the joint analysis of the longitudinal item responses and times to the development of a terminal event. All of these models used an accelerated failure time or Cox proportional hazards model for the survival times, which may not be suitable when some of the subjects are considered cured and will therefore never develop an event. In this talk, I will present a new joint model that uses a promotion time cure model for survival times. Statistical estimation procedures are developed for the inference of the parameters in the model. The proposed model and inference procedures are assessed through a simulation study and application to data from a randomized clinical trial for patients with early breast cancer. This is joint work with Ming Chi, Xiaogang Wang, Hui Song, and Dongsheng Tu.

Presenter: Yuping Wang, Paris-Lodron-University Salzburg

Title: Hierarchical variable clustering based on the predictive strength

Abstract: A rank-invariant clustering of variables is introduced that is based on the predictive strength between groups of variables, i.e., two groups are assigned a high similarity if the variables in the first group contain high predictive information about the behavior of the variables in the other group and/or vice versa. The method presented here is model-free, dependence-based, and does not require any distributional assumptions. Various general invariance and continuity properties are investigated with special attention to those that are beneficial for the agglomerative hierarchical clustering procedure. A fully non-parametric estimator is considered whose excellent performance is demonstrated in several simulation studies and by means of real-data examples.

Presenter: Nicolas Dietrich, Paris-Lodron-University Salzburg

Title: Revisiting the Williamson transform in the context of multivariate Archimedean copulas

Abstract: Motivated by a recently established result saying that within the family of bivariate Archimedean copulas, standard pointwise convergence implies the generally stronger weak conditional convergence (i.e., convergence of almost all conditional distributions), this result is extended to the class of multivariate Archimedean copulas. Working with the fact that generators of Archimedean copulas are d-monotone functions, pointwise convergence within the family of multivariate Archimedean copulas is characterized in terms of convergence of the corresponding generators, derivatives of the generators, marginal copulas, as well as marginal densities. Furthermore, weak conditional convergence is a consequence of any of the aforementioned

properties. Utilizing that generators of Archimedean copulas can be represented via Williamson transforms of one-dimensional probability measures, it is established that weak convergence of the probability measures is equivalent to uniform convergence of the Archimedean copulas. Using Markov kernels, Archimedean copulas inherit absolute continuity, singularity, and discreteness from the aforementioned probability measures, leading to the surprising result that absolutely continuous, singular, as well as discrete copulas are dense in the class of Archimedean copulas with respect to the uniform metric.

Break 3:30-3:45, July 10, 2024

Parallel Contributed Sessions 5-8: 3:45-4:45, July 10, 2024

Contributed Session 5: Survival Analysis in Practice.

Chair: Hao Mei. Room: HS 402

Presenter: Peyton Smith, Harvard University

Title: Improving Polygenic Risk Scores for Alzheimer's Disease

Abstract: A polygenic risk score (PRS) is a metric which captures an individual's genetic susceptibility to a particular trait, condition, or disease based on multiple genetic variants or single nucleotide polymorphisms (SNPs) across their genome. PRS is a concept primarily used in the field of genomics and genetics, and it has gained significant attention in recent years due to advances in genetic research and the availability of large-scale genome-wide association studies (GWAS). To date, it is being used to predict an individual's risk of developing Alzheimer's disease or the individual Alzheimer-free survival. Here, we will evaluate several machine learning approaches to incorporate ancestry information and several genomic platforms, developing ancestry-specific predictions for both overall disease risk and disease-free survival. We will compare the approaches in simulation studies and by application to the NIAGEDS data set. We developed an improved PRS with respect to predicting the development of Alzheimer's disease and an individual's Alzheimer-free survival. These improvements have applications in calculating PRSs to predict the onset of diseases other than Alzheimer's, including cardiovascular disease and type 1 diabetes.

Presenter: Merle Munko, Otto-von-Guericke University Magdeburg

Title: Surviving the multiple testing problem: RMST-based tests in general factorial designs

Abstract: Several methods in survival analysis are based on the proportional hazards assumption. However, this assumption is very restrictive and often not justifiable in practice.

Therefore, effect estimands that do not rely on the proportional hazards assumption such as the restricted mean survival time (RMST) are highly desirable in practical applications. The RMST is defined as the area under the survival curve up to a prespecified time point and thus summarizes the survival curve into a meaningful estimand. For two-sample comparisons based on the RMST, there is an inflation of the type-I error of the asymptotic test for small samples and therefore a two-sample permutation test has already been developed. The first goal is to further extend the permutation test for general factorial designs and general contrast hypotheses by considering a Wald-type test statistic and its asymptotic behavior. Additionally, a groupwise bootstrap approach is considered. In a second step, multiple tests for the RMST are developed to infer several null hypotheses simultaneously. Hereby, the asymptotically exact dependence structure between the local test statistics is incorporated to gain more power. The small sample performance of the proposed global and multiple testing procedures is analyzed in simulations and finally illustrated by analyzing a real data example.

Presenter: Wieske Katharina de Swart, Radboud University

Title: A Comparative Study of Methods for Survival Analysis with Longitudinal Data

Abstract: Survival analysis is vital in understanding the progression of neurological and neurodegenerative diseases, particularly in predicting dementia risk. Recently, the availability of more longitudinal data and better computing resources has prompted the development of new techniques for survival analysis. Among these are machine learning methods that offer new ways to extract information from longitudinal data as well as flexible models for survival prediction. However, it is not yet clear in which situations these complex methods lead to improved performance over conventional alternatives. To study this, we consider different combinations of longitudinal and survival models. As the longitudinal model, we consider Multivariate Functional Principal Component Analysis, a Recurrent Neural Network, and two standard approaches (using only the baseline data or the data from the last visit). For the survival prediction, we use the Cox Proportional Hazards model, a Random Survival Forest, and a neural network approach. We focus on dynamic risk prediction where model predictions are made at various time intervals as new longitudinal data becomes available. Therefore, we choose a set of landmark times and evaluate all models by using all longitudinal data up until this time as input to a longitudinal model and using the output of this model in a survival model to make risk predictions for all future time points. For training, several options exist ranging from training one model on all data to creating a separate model for each landmark time which is trained only on data up until that point. We explore various options and illustrate how these choices influence the final performance. We compare these methods on different simulated datasets with specific characteristics and on real data from the Alzheimer's Disease Neuroimaging Initiative. With this study, we aim to provide researchers and practitioners with a clearer understanding of the strengths and weaknesses of different methods for survival analysis and the nuances of their implementation. By systematically evaluating these models on a diverse range of datasets and using consistent evaluation metrics, we hope to shed light on their relative performance and highlight the key factors contributing to their success or limitations.

Contributed Session 6: Nonparametric Methods

Chair: Paul Peng. Room: HS 403

Presenter: Jonas Beck, Paris Lodron University of Salzburg

Title: Combining Stochastic Tendency and Distribution Overlap Towards Improved Nonparametric Effect Measures and Inference

Abstract: A fundamental functional in nonparametric statistics is the Mann-Whitney functional $\theta = P(X < Y)$ which constitutes the basis for the most popular nonparametric procedures. The functional θ measures a location or stochastic tendency effect between two distributions. A limitation of θ is its inability to capture scale differences. If differences of this nature are to be detected, specific tests for scale or omnibus tests need to be employed. However, the latter often suffer from low power, and they do not yield interpretable effect measures. In this manuscript, we extend θ by additionally incorporating the recently introduced distribution overlap index (nonparametric dispersion measure) I2 that can be expressed in terms of the quantile process. We derive the joint asymptotic distribution of the respective estimators of θ and I2 and construct confidence regions. Extending the Wilcoxon-Mann-Whitney test, we introduce a new test based on the joint use of these functionals. It results in much larger consistency regions while maintaining competitive power to the rank sum test for situations in which θ alone would suffice. Compared with classical omnibus tests, the simulated power is much improved. Additionally, the newly proposed inference method yields effect measures whose interpretation is surprisingly straightforward.

Presenter: Lukas Mödl, Charité medical university of Berlin

Title: Wild Bootstrapping the (asymptotic) Joint Distribution of Wilcoxon-Mann-Whitney Test Statistics

Abstract: The Wilcoxon-Mann-Whitney (WMW) test is a widely used non-parametric test for comparing the distribution of a single outcome between two independent groups. However, when analyzing multivariate outcomes, it is more appropriate to consider the joint distribution of WMW test statistics to accurately control the Type-I error rate and to account for potential interdependencies between the outcomes. Common approaches to address both aspects are non-parametric Multivariate Analysis of Variance (MANOVA)-type tests or Multiple Comparison Testing Procedure (MCTP)-type tests. These tests are essentially Lp norm variants applied to a vector of WMW test statistics with advantages and disadvantages associated with the chosen norm. However, any norm has the limitation that statistical inference relying on the (asymptotic) joint distribution of WMW test statistics exhibits diminished reliability in high dimensions (i.e., when the number of comparisons exceeds the sample size) with regard to Type-I and Type-II error rate control. This issue has significant practical implications for studies involving multiple comparisons and limited sample sizes, such as animal trials or rare disease studies. Researchers may face challenges in achieving appropriate Type-I error rate control, appropriate Type-II rate

error control, or both. In such cases, it may be necessary to reduce the number of comparisons or increase the sample size to maintain statistical power while controlling the Type I error rate. However, restricting the number of comparisons may compromise the study's objectives and increasing the sample size may be impractical or impossible in some cases. To address this issue, we propose a wild bootstrap resampling procedure that effectively mimics the (asymptotic) joint distribution of WMW test statistics. Simulations demonstrate that our resampling procedure maintains accurate Type-I error control across a wide range of scenarios, including scenarios characterized by a large number of comparisons and small sample sizes. A real data set illustrates the application.

Presenter: Elizabeth Juarez-Colunga, University of Colorado Anschutz Medical Campus

Title: Application of Bayesian non-parametric modeling to longitudinal seizure data

Abstract: Understanding the long-term trajectory of seizures over time among patients with focal epilepsy remains a significant challenge for physicians and researchers alike due to the lack of large-scale studies and difficulty in tracking and modeling longitudinal seizure outcomes. Seizure occurrences are often difficult to predict given their irregular pattern. It is hypothesized that subgroups of patients exist, such as some that have very few seizures, some that have regular constant seizures, and some that have sporadic seizures. Traditional statistical modeling frameworks in which a single distribution is chosen to model the outcome of interest do not accurately capture the intricacies of seizure event data. We propose a Bayesian non-parametric (BNP) model to longitudinal daily seizure data from over 400 patients who tracked daily seizure occurrences in the Human Epilepsy Project, a 5-year prospective study. Specifically, we use a Dirichlet process mixture of binary distributions to model the daily seizure occurrence. The stick-breaking algorithm is used to jointly cluster the many correlated observations per individual, inducing dependence within individuals. We demonstrate how the BNP model can flexibly model a longitudinal binary outcome and we provide compelling data to suggest subgroups in long-term trajectories of seizures in patients with focal epilepsy.

Contributed Session 7: Multiple Outcome Analysis
Chair: Donglin Zeng. Room: HS 401

Presenter: Taban Baghfalaki, University of Bordeaux

Title: Dynamic Prediction through Joint Modeling of Longitudinal Markers and Time-to-Event Data: A new Bayesian Two-Stage Approach

Abstract: In the realm of clinical and epidemiological research, it is essential to collect longitudinal markers and time-to-event data simultaneously to predict disease progression and outcomes. However, as the number of longitudinal markers increases, traditional joint modeling methods often fail due to computational complexity and convergence issues. To address these

challenges, we propose an innovative two-stage joint modeling approach. In the first stage, we estimate individual marker trajectories using one-marker Bayesian joint models to avoid biases induced by informative dropouts. Next, we will estimate a proportional hazard model that incorporates the current estimated values and/or slopes of all markers as dynamic covariates over time. This two-step approach can accommodate numerous longitudinal markers without bias. Simulation studies and real-world applications demonstrate the effectiveness of the proposed method for both parameter estimation and dynamic risk prediction.

Presenter: Tiphaine Saulnier, University of Bordeaux

Title: Joint analysis of disease progression markers and death using individual temporal recalibration

Abstract: Establishing the natural history of a disease permits to better understand its progression over time. However, when the disease is difficult to diagnose, uncertainty remains around the onset time and patients are potentially recruited in cohorts at different disease stages. Occurrence of clinical events, such as death, also interrupts follow-ups, inducing missing data potentially not at random. The present work introduces a joint model combining a disease progression model based on an individual temporal recalibration to describe markers progression according to the latent disease time and a survival model to assess the association with death. The methodology is motivated by the study of Multiple system atrophy (MSA), a rare neurodegenerative disease. Annual data of 663 patients from the French MSA cohort were analyzed over 10.8 years. MSA progression was described by the Unified MSA Rating Scale sumscores I (functional sphere) and II (motor sphere). Once time recalibrated, their progressions spanned over 12 years. Compared to non-dependent patients at inclusion, mean time gaps between moderately-dependent and helpless patients at inclusion were 2.56 (95\%CI=2.36,2.76) and 5.84 (95\%CI=4.92,6.77) years, respectively. Risk of death highly depended on markers' dynamics and individual shift (with higher risk for more advanced patients). This latent disease time approach has potential to describe complex disease progression while accounting for heterogeneity of patients' profiles and informative dropout.

Presenter: L. Courcoul, University of Bordeaux

Title: A flexible location-scale joint model to study the effect of blood pressure variability on competing events

Abstract: A high level of blood pressure is a well-known risk factor for several major health issues (cardiovascular, dementia, etc.), but an increasing number of studies suggest that blood pressure variability may also be an independent risk factor for these events. However, these studies suffer from significant methodological weaknesses and most often assume that the blood pressure variability does not change with time. Motivated by assessing the association between blood pressure variability and cardiovascular events, we developed a location-scale joint model with a flexible modeling of within-subject variance for repeated measures of a longitudinal marker and time-to-events. The proposed joint model combines a location-scale mixed model and a cause-specific model with proportional intensities for the competing events. In the mixed model, the residual variance may depend on subject-specific random effects, covariates, and

time, or may be the sum of two components: within-visit and between-visits variances. The risk of events may depend simultaneously on the subject-specific residual variance(s), the current value, and the current slope of the marker. The model is estimated by maximizing the likelihood function using the Marquardt-Levenberg algorithm. The estimation procedure is implemented in the R-package "FlexVarJM" and is validated through a simulation study. The model was applied on the PROGRESS clinical trial data for the prevention of the recurrence of stroke that includes 6105 subjects followed over 5 years with 12 measurement times for blood pressure. Finally, some goodness-of-fit and dynamic prediction tools have been used to illustrate the effect of blood pressure variability.

Presenter: H. Vermeulen, Universiteit Hasselt

Title: Analysis of proliferation assay data for immunomonitoring: A bivariate modelling approach

Abstract:

Introduction

Immunomonitoring of antigen-specific responses can be done using assays in which a radioactive nucleoside is incorporated into proliferating T-cells. The amount of radioactivity, and thus T-cell proliferation, can then be measured by a scintillation counter in counts per minute (cpm). Typically, these assays are performed for multiple replicates of antigen stimulated cells, as well as on multiple replicates of unstimulated cells, where the number of replicates may differ between both conditions, resulting in two sets of data with a different sample size for each patient. To evaluate the effect of immunomodulatory therapies in a patient group, it is important to model T-cell proliferation over time in the stimulated and unstimulated condition simultaneously, accounting for the variability between patients as well as within patients, while allowing for different sample sizes under both conditions. We therefore propose a mixed model for bivariate outcomes for the analysis of proliferation assay data.

Methods

The ln transformed cpms in the stimulated and unstimulated condition ($ln(Scpm)$ and $ln(Ucpm)$, respectively) were modelled by linear mixed models, with correlated random effects, to flexibly couple the subject specific deviations from the average T-cell response in the stimulated and unstimulated condition. We defined the stimulation index (SI), a measure for T-cell responses in antigen stimulated cells relative to unstimulated cells, as: $ln(SI) = ln\left(\frac{Scpm}{Ucpm}\right) = ln(Scpm) - ln(Ucpm)$ and tested the null hypothesis of no T-cell response, i.e. $ln(SI) = 0$, for three different proliferation assays. In a first proliferation assay, T-cells were stimulated with antigens that were expected to provoke a strong immune response, while in a second proliferation assay T-cells were stimulated with antigens that were expected to provoke a weak immune response. In a third assay, all T-cells remained unstimulated. The proliferation assays were performed in 12 multiple sclerosis patients treated with an experimental immunomodulatory therapy. Antigen provoked T-cell responses were measured at the start of the immunomodulatory therapy and at three consecutive follow-up visits after the start of the therapy, i.e. at 6, 14 and 26 weeks.

Results

The mixed model for bivariate outcomes described the observed data well for all three proliferation assays. A significant T-cell immune response was found at the start of the experimental immunomodulatory therapy, and at all follow-up visits for the antigen expected to provoke a strong immune response. For the antigen expected to provoke a weak immune response, a significant immune response was found at the start of the treatment, and at the first and third follow-up visit, while no significant immune response was found for the proliferation assay performed on unstimulated T-cells.

Conclusion

Our results show that the proposed mixed model for bivariate outcomes, with correlated random effects, performs well and offers an adequate and flexible modelling framework for the analysis of T-cell immune responses in proliferation assays.

Contributed Session 8: Applied Bayesian Analysis

Chair: Chong He. Room: HS 414

Presenter: Sangita Kulathinal, University of Helsinki

Title: Bayesian Hidden Markov Model for Natural History of Colorectal Cancer: Handling Misclassified Observations, Varying Observation Schemes and Unobserved Data

Abstract: Simulation models for colorectal cancer (CRC) have provided important tools for medical and public health decision making. These models typically contain information and uncertainty from multiple sources. Choice of parameters are crucial in simulation models. We use the observed individual-level event histories to estimate the parameters of a 5-state progressive model for the natural history of CRC. Individual-level event histories are obtained by combining data from a randomised control trial on CRC screening and population-level cancer register. These two data sources have different observation schemes, and hence, the combined data have unobserved as well as misclassified states. We employ Bayesian continuous-time Hidden Markov Model (HMM) because of these features of the combined data. We use simulation-based calibration method to ensure that the posterior distributions can be reliably estimated. We carry out Bayesian computation to obtain posterior distributions of the quantities of interest using two sampling algorithms; Automatic Differentiation Variational Inference (ADVI) and Hamiltonian Monte Carlo (HMC). We also evaluate the ability of these algorithms to recover parameters of the data generating process using simulated data sets. We then apply the algorithms to a real data set. Our approach will help in policy decisions for the "real" population because it is based on individual-level data from that population.

Presenter: Daniel J. Phillips, University of Oxford

Title: Multiple imputation for joint modelling of COVID-19 antibody decay and risk of infection

Abstract: After receiving a COVID-19 vaccine, the body produces antibodies which protect against future infection. We aim to improve understanding of how antibody levels affect risk of infection and how risk of infection changes over time. We are not aware of any previous work applying joint modelling to understand how vaccine-induced immune responses correlate with risk of an infectious disease. We develop a novel two-stage joint model for post-vaccination COVID-19 antibody levels and risk of infection. We fit a Bayesian hierarchical model for antibody levels over time, then impute the antibody levels as a time-changing covariate in a proportional hazards model. We observe non-Gaussian parameter estimates, so Rubin’s rules for multiple imputation do not apply. We instead show that the asymptotic distribution of the MLEs from the Cox model can be interpreted as an approximate Bayesian posterior with non-

informative priors. This allows us to calculate multiply imputed credible intervals via sampling. We fit the model to data from the COVID-19 Oxford-AstraZeneca vaccine study (COV002, NCT04400838), where blood samples were retrospectively tested for antibodies based on a case-cohort design. Initial results indicate COVID-19 antibodies significantly decrease the risk of infection in the trial, and protection wanes over time.

Presenter: Weining Shen, University of California, Irvine

Title: Bayesian biclustering and its application in education data analysis

Abstract: Motivated by an English proficiency assessment study, we propose a novel nonparametric Bayesian item response theory model that estimates clusters at the question level while simultaneously allowing for heterogeneity at the examinee level under each question cluster characterized by a mixture of binomial distributions. We present a tractable sampling algorithm to obtain valid posterior samples from our proposed model. We also show that our model is identifiable under a set of conditions and establish its asymptotic properties. Compared to the existing methods, our model manages to reveal the multi-dimensionality of the examinees' proficiency level in handling different types of questions parsimoniously by imposing a nested clustering structure. The data analysis example nicely illustrates how our model can be used by test makers to distinguish different types of students and aid in the design of future tests.

Break 4:45-5:00, July 10, 2024

IBS-ROeS Keynote Speech II: 5:00-6:00, July 10, 2024

Room: HS 401

Chair: Arne Bathke

Presenter: Markus Pauly, TU Dortmund University

Title: Can’t see the forest for the trees? On theoretical and numerical results for Random-Forest-type methods

Abstract: Today, tree-based methods are an integral part of the statistical toolbox and serve as essential analysis and modeling techniques. This is especially true for Random Forests, which are often used as a reliable benchmark method. In fact, there is a lot of empirical evidence for the good performance of these methods. In this talk, we will provide different approaches to analyze their applicability for specific problems. To this end, we will discuss both, (i) numerical results for specific applications such as imputation, and (ii) theoretical properties such as their consistency for regression tasks. In particular, we will provide exciting mathematical insights on Random Forests and related approaches before discussing what the proven statements actually mean in real applications.

Closing Remarks: 6:00-6:10. July 10, 2024

Room: HS 401

Lei Liu and Virginie Rondeau