PROGRAM
6th International Workshop on Statistical Analysis on Multi-Outcome data
June 29July 1st, 2026, Bordeaux, FR

KEYNOTE SPEAKER 1:
Dimitris Rizopoulos d.rizopoulos@erasmusmc.nl

Title: Harnessing Variational Auto-Encoders for Fitting Deep Mixed-Effects Models
Authors: Dimitris Rizopoulos, Nina van Gerwen, and Stem Willemsen
Institution: Department of Biostatistics, Erasmus MC Rotterdam, the Netherlands
Abstract: Nowadays, there is a growing interest in using electronic health record data to monitor, advise, and treat patients. This data often entails a plethora of longitudinal outcomes collected at different resolutions for the same subject. The size and complexity of this data often make the use of traditional statistical frameworks, such as mixed-effects and multivariate regression models, computationally challenging or even infeasible. On the contrary, machine learning (ML) procedures are mainly designed to accommodate the complexity of such data designs. Here, we focus on deep neural networks to analyze complex longitudinal data. The standard deep learning architectures to model temporal data are recurrent neural networks (RNNs) and their variants, such as long short-term memory RNNs and temporal convolutional networks. However, these ML algorithms do not share the advantageous properties of statistical models to accommodate the challenging features of longitudinal data. In particular, a common issue in longitudinal data analysis is that subjects provide measurements at irregular time intervals, or they may completely drop out of the study. Current approaches to handling missing data in RNN variants include last-observation-carried-forward methods or the inclusion of missing-data indicators as covariates, which are not guaranteed to yield valid inferences under missing-at-random mechanisms. In this work, we propose using the random-effects models paradigm and assume that the correlations in the longitudinal responses stem from a set of shared latent variables. In particular, we define a deep latent variable model with two feed-forward neural network architectures corresponding to the classic fixed- and random-effects parts of mixed models. We estimate the parameters of our proposed deep latent variable model using an extension of variational autoencoders that handles unbalanced missing-at-random longitudinal outcome data. In our approach, we use a third neural network architecture to specify a variational density for the conditional distribution of the random effects given a set of imputed longitudinal outcome data. The imputation of the missing longitudinal outcomes is performed via a data augmentation algorithm. Our estimation procedure alternates between data-augmentation and estimation steps and shares similarities with the stochastic approximation expectation-maximization algorithm.

KEYNOTE SPEAKER 2:
Xihong Lin xlin@hsph.harvard.edu

Title: Navigating the Crossroads of Statistics, Generative AI, and Genomic Health
Authors: Xihong Lin**,** Department of Biostatistics and Department of Statistics
Harvard University
Abstract: Rapid advancements in generative AI have revolutionized data utilization and enabled machines to learn from data more effectively. Statistics, as the science of learning from data while accounting for uncertainty, plays a pivotal role in addressing complex real-world problems and facilitating trustworthy decision-making. Massive multi-modal data have been generated in genomic health. In this talk, I will discuss the challenges and opportunities in building a ecosystem that integrates statistics, generative AI, and genomics and health to accelerate trustworthy discovery. I will illustrate key points using the analysis of large scale biobanks, whole genome sequencing data, and electronic health records (EHRs), and demonstrate how to empower integrated statistical and AI analysis of genomic data by leveraging synthetic data generated from generative AI models. A key feature of the synthetic data powered approach is that it ensures valid statistical analysis while allowing blackbox generative ML models to be misspecified, and improves the power of statistical inference when generative AI models

are useful. Building an end-to-end data science ecosystem is pivotal for accelerating biological and health science discoveries and translational science. I will also discuss how to build such an ecosystem that allows for affordable, scalable and interpretable analysis of the whole genome sequencing and EHR data of the UK biobank of 500,000 subjects in the cloud platform RAP and the All of Us data of 400,000 subjects in the NIH cloud platform AnVIL.

Session 1: Deep learning in survival analysis

Organizer: Lei Liu lei.liu@wustl.edu

- Jon Steingrimsson jon_steingrimsson@brown.edu

Title: Deep Learning with Time-to-event Outcomes
Authors Jon Steingrimsson, Samantha Morrison, Constantine Gatsonis, Ani Eloyan.
Institution: Brown University
Abstract: Deep learning is a class of algorithms that uses multiple layers to build risk prediction models. Each layer involves an unknown weight vector, which is usually estimated by minimizing a loss function. We extend deep learning methods to account for censoring by replacing the standard loss function with censoring-unbiased loss functions. We then examine the theoretical properties of these loss functions as well as practical challenges in implementing the algorithms. The performance of the proposed methods is evaluated through simulation studies, and we further explore their applications in medical imaging analysis using convolutional neural networks.

- Peter X Song pxsong@umich.edu

Title: Neural Network Machine Regression (NNMR): A Deep Learning Framework for Uncovering High-order Synergistic Effects
Authors Jiuchen Zhang, Peter X Song
Institution: University of Michigan
Abstract: We propose a new neural network framework, termed ***Neural Network Machine Regression (NNMR),*** which integrates trainable input gating and adaptive depth regularization to jointly perform feature selection and function estimation in an end-to-end manner. By penalizing both gating parameters and redundant layers, NNMR yields sparse and interpretable architectures while capturing complex nonlinear relationships driven by high-order synergistic effects. We further develop a post-selection inference procedure based on split-sample, permutation-based hypothesis testing, enabling valid inference without restrictive parametric assumptions. Compared with existing methods, including Bayesian kernel machine regression and widely used post hoc attribution techniques, NNMR scales efficiently to high-dimensional feature spaces while rigorously controlling type I error. Simulation studies demonstrate its superior selection accuracy and inference reliability. Finally, an empirical application reveals sparse, biologically meaningful food group predictors associated with somatic growth among adolescents living in Mexico City.

- Kevin He kevinhe@umich.edu

Title: Flexible Deep Learning Techniques for Survival Analysis with Data Integration
Authors*:* Kevin He, Feiyang Deng and Lingfeng Luo
Institution: Department of Biostatistics, School of Public Health, University of Michigan
Abstract:
Prognostic prediction using survival analysis faces challenges due to complex relationships between risk factors and time-to-event outcomes. Deep learning methods have shown promise in overcoming these challenges, but their effectiveness often relies on large datasets. In contrast, when implemented on moderate or small data sets, these methods often suffer from severe problems, such as insufficient training data, overfitting, and difficulty in tuning hyperparameters. To address these issues and improve prognosis predictions, this talk introduces a flexible deep learning framework for integrating external risk models with internal time-to-event data using a generalized Kullback-Leibler divergence penalty. Our approach accommodates both homogeneous and heterogeneous settings, allows for the integration of risk scores derived from various external

models, and leverages the power of deep neural networks to capture complex non-linear relationships.

Session 2: Multiple/intercurrent events in clinical trials

Organizer: Catherine Legrand catherine.legrand@uclouvain.be

- Philippe Lambert p.lambert@uliege.be

Title: Joint Modeling of Longitudinal Health-Related Quality of Life Data in the Presence of Competing Dropout Risks

Authors *Hortense DOMS[1], Philippe LAMBERT [1,2], Catherine LEGRAND[1]*
Institution: UCLouvain[1] and ULiège[2], Belgium

Abstract:
In cancer clinical trials, health-related quality of life (HRQoL) is an important endpoint, providing information about patients' well-being and daily functioning. However, missing data due to premature dropout can lead to biased estimates, especially when dropouts are informative. This paper introduces the extJMIRT approach, a novel tool that efficiently analyzes multiple longitudinal ordinal categorical data while addressing informative dropout. Within a joint modeling framework, this approach connects a latent variable, derived from HRQoL data, to cause- specific hazards of dropout. Unlike traditional joint models, which treat longitudinal data as a covariate in the survival submodel, our approach prioritizes the longitudinal data and incorporates the log dropout risks as covariates in the latent process. This leads to a more accurate analysis of longitudinal data, accounting for potential effects of dropout risks. Through extensive simulation studies, we demonstrate that extJMIRT provides robust and unbiased parameter estimates and highlight the importance of accounting for informative dropout. We also apply this methodology to HRQoL data from patients with progressive glioblastoma, showcasing its practical utility. If time permits, preliminary results using two-step estimation approaches—designed to improve computational efficiency—will also be presented.

- Georgios Kazantzidis georgios.kazantzidis@roche.com

Title: A Joint Multi-State Model for Causal Mediation Analysis adjusting for treatment switching in Oncology

Authors: Georgios Kazantzidis, Virginie Rondeau, Francois Mercier

Institution: University of Bordeaux, Inserm; F. Hoffmann-La Roche AG

Abstract: Causal mediation analysis using joint models of longitudinal tumor size and survival is a powerful tool for understanding how anti-cancer treatments work and for validating surrogate endpoints. However, a critical limitation of existing methods is the failure to adequately account for treatment switching. These intercurrent events can bias the estimation of natural direct and indirect effects, complicating the interpretation of the mediator's role.

This work proposes a Bayesian semi-competing risk joint modeling framework to address this challenge by integrating a multi-state model. Our model jointly analyzes the non-linear tumor growth inhibition (TGI) dynamics for the longitudinal mediator (tumor size) and its association with overall survival (OS),addressing the potential bias introduced by a treatment switch. The transition intensities in the multi-state model are linked to the underlying tumor dynamics, allowing the effect of the tumor growth inhibition on both treatment switch and overall survival to be estimated simultaneously.

This framework allows us to define a hypothetical estimand to quantify treatment effects in the absence or in presence of subsequent therapies. We can then propose a novel causal decomposition of the total treatment effect on survival into three distinct pathways: a natural direct effect (NDE), a natural indirect effect (NIE) mediated by tumor growth dynamics, and an NIE mediated through the treatment switch. By applying this model to clinical trial data where treatment switching is common, we anticipate to provide less biased estimates of the proportion of the treatment effect mediated by tumor dynamics.

- Luc Boone luc.boone@eortc.org

Title: Estimation of inverse probability of censoring weights using shared parameter joint models for longitudinal and survival outcomes.
Authors *Luc Boone, Catherine Legrand*
Institution: ISBA/LIDAM, Université Catholique de Louvain (UCLouvain)
Abstract:
Randomized clinical trials with time to event primary endpoints give rise to censored observations. For instance, progression-free survival and recurrence-free survival in cancer clinical trials. The censoring can be a direct consequence of a pre-specified 'censoring rule' following the definition of the targeted (hypothetical) estimand. In this case, an intercurrent event such as a post-protocol treatment or toxicity might occur prior to the occurrence of the outcome of interest, whilst the estimand of interest is to evaluate the protocol treatment in the absence of such intercurrent events. Alternatively, these intercurrent events might lead to deviations or in some cases prevent further disease assessments, crucial for the evaluation of the outcome of interest, and thus indirectly result in censored observations.
The assumption of independent censoring is in both cases generally violated such that common approaches to estimate the survival function or compare treatment arms using the Cox proportional hazards model result in a biased estimation of the survival function and the treatment effect estimate, respectively. Inverse probability of censoring weighting (IPCW) was proposed by Robins (1993) to adjust for dependent censoring using auxiliary baseline and time-varying covariate data collected during the clinical trial.
We propose and study an extension to the application of IPCW to adjust for dependent censoring in the context of (open label) cancer clinical trials in light of the estimands framework. In Robins' (1993) initial proposal of IPCW, the censoring weights were estimated using the Cox proportional hazards model. We propose the use of joint models for longitudinal and time-to-event data to estimate the censoring weights in the presence of endogenous time-varying covariates. This presentation provides a rationale for the use IPCW and its extension using joint models.

- Tomasz Burzykowski tomasz.burzykowski@iddi.com

Title: Early-outcome-based interim analyses in randomized clinical trials with long-term clinical endpoints
Authors: Tomasz Burzykowski
Institution: Hasselt University (Belgium) & IDDI (Belgium)
Abstract:
In randomized clinical trials that use a long-term efficacy endpoint *T* (e.g., overall survival), the follow-up time necessary to observe *T* may be substantial. In such trials, an attractive option is to consider an interim analysis based solely on an early outcome *S* (e.g., tumor response) that could be used to expedite the evaluation of treatment's efficacy.
In this paper, we present a methodology that allows introducing such an early interim analysis for any combination of *S* and *T* types. It appears that such a design may offer substantial gains in terms of both the expected trial duration and the expected sample size. A prerequisite, though, is that the treatment effect on *S* has to be strongly correlated with the treatment effect on *T*, i.e., the early outcome has to be a good trial-level surrogate for the long-term endpoint. Importantly, in practice, the strength of the correlation has to be evaluated based on historical clinical-trial data. The developed methodology does take the uncertainty resulting from the real-life evaluation into account. We illustrate the application of the methodology by using real-life examples of trials in different disease areas.

Session 3: Trustworthy AI and Statistical Innovation for Complex Health Data

Organizer: Ying Wei yw2148@cumc.columbia.edu
and TBD

- Ying Wei yw2148@cumc.columbia.edu

Title: Time-varying Quantile Regression with Multi-outcome Latent Groups
Authors Tianying Wang, Yanyuan Ma and Ying Wei
Institution: Columbia University

Abstract: Aging is a complex, multidimensional process, with biological, behavioral, and clinical factors evolving at different rates across individuals. Understanding the heterogeneity in these dynamic changes is critical for elucidating the mechanisms of aging and for developing targeted interventions. We introduce a novel time-varying quantile regression framework for jointly modeling multiple aging-related outcomes, such as biomarkers, physiological measurements, and clinical traits. Our approach simultaneously identifies latent groups of outcomes that share temporal patterns while estimating heterogeneous, time-dependent covariate effects—without requiring full specification of the joint multivariate distribution. Theoretical guarantees are provided for both coefficient estimation and latent group recovery, and our algorithm scales to large-scale cohort data. Applied to UK Biobank data, our method reveals distinct temporal genetic influence patterns on lipid traits and body composition measures, highlighting differential aging trajectories, particularly among postmenopausal women. This distribution-free framework offers a powerful tool for uncovering heterogeneous patterns in aging processes across multiple domains. This is joint work with Ying Wei and Yanyuan Ma.

- Kaizheng Wang kaizheng.wang@columbia.ca

Title: Quantifying Fidelity in AI Persona Simulations
Authors: Chengpiao Huang, Yuhang Wu, Kaizheng Wang (kw2934@columbia.edu)
Institution: Columbia University
Abstract: Large language models (LLMs) are increasingly used to simulate human behaviors such as public opinions and consumer choices. Yet, the unknown distribution shift between AI personas and the real populations they aim to emulate undermines the validity of conclusions. To tackle this challenge, we introduce an uncertainty quantification approach that converts imperfect simulated responses into confidence sets for population parameters of human responses. A key innovation lies in determining the optimal number of simulated responses: too many produce overly narrow confidence sets with poor coverage, while too few yield excessively loose estimates. Our method adaptively selects the simulation sample size and ensures valid average-case coverage guarantees. It is broadly applicable to any LLM, irrespective of its fidelity, and any procedure for constructing confidence sets. Crucially, the selected sample size also serves as an interpretable fidelity metric that directly quantifies the simulator-population misalignment. We illustrate this method on real-world datasets and LLMs.

- Nan Lin nlin@wustl.edu

Title: Trajectory Clustering via Spatial-Graph and LLM-Semantic Embeddings
Authors: Sizhe Zhang, Xinyu Zhang, Yunhan Jiang, Patrick Fowler, Michael Podgursky, Guangli Zhang, Kenan Li, and Nan Lin (nlin@wustl.edu)
Institution: Washington University at St. Louis
Abstract: Many modern datasets record multiple outcomes that evolve jointly over space and time. I present an unsupervised framework that learns low-dimensional, interpretable representations of such multivariate trajectories and enables cluster analysis. For each spatial unit, two complementary feature channels are constructed: (i) a spatial-graph embedding capturing spatial connectivity, and (ii) a semantic embedding that maps domain metadata (e.g., points-of-interest categories) into a continuous space using a large language model. Trajectories are represented as multivariate time series over these channels and encoded with an attention-based transformer autoencoder. In empirical studies, these factors exhibit clear cluster structure; clustering in the latent space produces groups with distinct temporal deviation profiles across functional categories. Our framework provides a general recipe for (a) constructing multi-outcome features by combining spatial graphs with semantic embeddings, (b) representation learning for multivariate trajectories, and (c) interpretable clustering via category-wise temporal profiles.

- Jinbao Chen jinboche@upenn.edu

Title: A Semiparametric Method for Addressing Under-Diagnosis Using Electronic Health Record Data
Authors: Weidong Ma, Jordana Cohen, Jinbo Chen(jinboche@pennmedicine.upenn.edu)
Institution: University of Pennsylvania

Abstract: Effective treatment of medical conditions begins with an accurate diagnosis. However, many conditions are often underdiagnosed, either being overlooked or diagnosed after significant delays. Electronic Health Records (EHRs) contain extensive patient health information, offering an opportunity

to probabilistically identify underdiagnosed individuals. The rationale is that both diagnosed and underdiagnosed patients may display similar health profiles in EHR data, distinguishing them from condition-free patients. Thus, EHR data can be leveraged to develop models that assess an individual's risk of having a condition. To date, this opportunity has largely remained unexploited, partly due to the lack of suitable statistical methods. The key challenge is the positive-unlabeled EHR data structure, which consists of data for diagnosed (``positive'') patients and the remaining (``unlabeled'') that include underdiagnosed patients and many condition-free patients. Therefore, data for patients who are unambiguously condition-free, essential for developing risk assessment models, is unavailable. To overcome this challenge, we propose ascertaining condition statuses for a small subset of unlabeled patients. We develop a novel statistical method for building accurate models using this supplemented EHR data to estimate the probability that a patient has the condition of interest. We study the asymptotic properties of our method and assess its finite-sample performance through simulation studies. Finally, we apply our method to develop a preliminary model for identifying potentially underdiagnosed non-alcoholic steatohepatitis patients using data from Penn Medicine EHRs.

Session 4: Joint models with Bayesian inference

Organizer: Sangita Kulathinal (Finland) sangita.kulathinal@helsinki.fi

- Etienne Sebag etienne.sebag@helsinki.fi

Ttile: Bayesian estimating Transition Rates in Two-State Non-Homogeneous Markov Jump Processes with Intermittent Observations: An Honest Time Data Augmentation Approach
Authors: Dario Gasbarra, Sangita Kulathinal, Etienne Sebag
Institution: Department Mathematics and Statistics, University of Helsinki, Finland
Abstract :
A possibly time-dependent transition intensity matrix or generator (Q(t)) characterizes the law of a Markov jump process (MP). For a time homogeneous MP, the transition probability matrix (TPM) can be expressed as a matrix exponential of Q. However, when dealing with a time non-homogeneous MP, there is often no simple analytical form of the TPM in terms of Q(t), unless all the Q(t) commute. This poses a challenge because when a continuous MP is observed intermittently, a TPM is required to build a likelihood. In this paper, we show that the Bayesian estimation of the transition intensities of a two-state non-homogeneous Markov model can be carried out by augmenting the intermittent observations with honest random times associated with two independent driving Poisson point processes, and that sampling the full path is not required. We propose a Bayesian data augmentation algorithm using the Gibbs sampler to estimate the transition rates. Finally, we illustrate our approach by simulating a continuous MP and by using observed (intermittent) time grids extracted from real clinical visits data.

- Marion Kerioui marion.kerioui@inserm.fr

Title: A Bayesian partition model for multivariate functional data: discovering biological pathways by modelling the time course of gene expression level
Authors*:* Marion Kerioui[1,2], Daniel Temko[1], Shahin Tavakoli[3], Hélène Ruffieux[1]
Institution: [1]MRC Biostatistics Unit, Cambridge University, United Kingdom
[2]Université Paris Cité and Université Sorbonne Paris Nord, Inserm, IAME F-75018 Paris, France
[3]RISIS, GSEM, University of Geneva, Geneva, Switzerland
Abstract:
Longitudinal data is increasingly collected in multiple variables (e.g., genomics data). One approach to characterize the coordinated evolution of several variables is the Multivariate Functional Principal Component Analysis (MFPCA)[1]. It assumes that each variable trajectory is a combination of eigen functions, weighted by a set of individual scores capturing the individual deviation from the mean trajectory. This set of scores is shared across all variables to account for a common latent dynamic i.e., assuming all variables are driven by the same biological pathway. This assumption becomes unreasonable when the number of variables is large, as several biological pathways may be involved. Here, we introduce a Bayesian Partition FPCA (PFPCA) framework by combining MFPCA with a mixture model, allowing to simultaneously group variables driven by the same pathway and disentangle different sources of variability in the dynamics. We consider an overfitted mixture with a data-driven selection of the optimal number of groups. We implemented a mean-field variational algorithm[2] for joint inference of all model parameters on large datasets within reasonable timeframes, coupled with a simulated annealing procedure to increase chances of converging to the global mode. The approach showed good

empirical performances in simulations, with minimum loss of estimation accuracy when comparing to MFPCA applied separately to each group i.e., assuming groups are known. Thanks to information borrowing across variables within a group, simulations also demonstrated improved estimation accuracy when using PFPCA compared to applying univariate FPCA to each variable separately. We applied the model to gene expressions levels data collected in 17 healthy individuals inoculated with H3N2 influenza. In most groups, the first component scores were able to discriminate between symptomatic and asymptomatic individuals. Enrichment analysis showed consistency between PCA groups and known pathways, including the immune response to influenza, confirming the biological relevance of the approach.

- Trinh Dong Huu trinh.dong@ncl.ac.uk

Title: Double inverse probability weighting for modeling sparsely measured longitudinal markers in presence of left and right truncation by death

Authors: Trinh Dong, Helene Jacqmin-Gadda, Denis Rustand, Alexandra Samier Foubert, Pierrick Coupé, Cécile Proust-Lima

Institution: Univ. Bordeaux, Inserm Bordeaux Population Health Research Center, Bordeaux Hospital and CNRS Labri Research Center

Abstract: In epidemiological research, it is common for a cohort participant to be recruited into a sub-cohort after their initial enrolment. For instance, in the French Multiple System Atrophy (MSA) cohort where more than 750 patients are followed annually from their MSA diagnosis, sub-cohorts have been recruited to explore neurodegeneration biomarkers. For disease with high mortality rate such as MSA, this delayed enrolment can induce a left-truncation bias, in which recruited participants are healthier than those left out due to death before recruitment. Even after enrolment, participants may drop out due to premature death, leading to right truncation. In the context of sparsely measured biomarkers, the ability of joint modelling to extrapolate the trajectory is limited. In this talk, we propose a double inverse probability weighting method to be applied on mixed models in order to correct for both left truncation and informative dropout. The weights are estimated using a joint model that leverages a densely measured longitudinal mediator. This method is applicable to both low and high dimensional repeated biomarkers. We illustrate the performances of the method in simulations, and in the French MSA cohort to assess sub-type differences in the progression of brain subcortical volumes, using an annually measured clinical progression marker as the instrumental mediator for correcting the double selection by death at sub-cohort entry and over time during follow-up.

- Satrajit Roychoudhury satrajit.roychoudhury@pfizer.com

Title: Robust dynamic borrowing designs for randomized basket trials: a case study from the ultra-rare invasive mold infections

Authors: *Satrajit Roychoudhury*

Institution: Pfizer Inc., New York, NY

Abstract : Invasive mold infections (IMI) are rare but can be severe, especially for people with weakened immune systems. Though rare, the all-cause mortality rate is high for IMI. Timely diagnosis and appropriate treatment of IMIs can mitigate the impact on morbidity and mortality. There are a limited number of antifungal drugs available to treat IMIs, and there is a need for new drugs with different mechanisms of action. The regulatory requirement for a new antifungal needs a traditional randomized and well-powered noninferiority clinical trial, which is difficult due to small participant pools. Moreover, the mortality rate for IMI can vary considerably depending on the type of mold and the patient's characteristics. Such heterogeneity across tumor types poses an additional challenge to plan for appropriate sample size and treatment effect interpretation after pooling different mold types. Recent FDA and EMA guidelines encourage using high-quality external control in an ultra-rare disease setting where patient recruitment is challenging. A basket trial is an innovative trial design that enables drug developers to investigate the effect of a novel treatment in multiple diseases. Despite its popularity in oncology, statistical methods for randomized basket trials in non-oncologic areas still need to be improved. This talk will discuss a robust borrowing strategy based on the Bayesian model averaging method for randomized basket trials. The proposed design increases the efficiency and precision of

treatment effect estimates for various molds. It also increases the ethical appeal by reducing the number of subjects required for the control group. Using simulation and real-life examples, we have demonstrated that the proposed approach can significantly increase statistical power and precision while maintaining the family-wise type I error rate at acceptable levels. The proposed approach greatly benefited from current practice, where different models are pooled together for inference and decision-making.

Session 5: Some Advanced Statistical Learning Methods for Modern Scientific Data

Organizer: Lexin Li lexinli@berkeley.edu

- Ji Zhu jizhu@umich.edu

Title: Modeling Non-Uniform Hypergraphs Using Determinantal Point Processes
Authors: Yichao Chen, Emma Zhang, Ji Zhu
Institution: University of Michigan, Emory University, University of Michigan
Abstract: Most statistical models for networks focus on pairwise interactions between nodes. However, many real-world networks involve higher-order interactions among multiple nodes, such as co-authors collaborating on a paper. Hypergraphs provide a natural representation for these networks, with each hyperedge representing a set of nodes. The majority of existing hypergraph models assume uniform hyperedges (i.e., edges of the same size) or rely on diversity among nodes. In this work, we propose a new hypergraph model based on non-symmetric determinantal point processes. The proposed model naturally accommodates non-uniform hyperedges, has tractable probability mass functions, and accounts for both node similarity and diversity in hyperedges. For model estimation, we maximize the likelihood function under constraints using a computationally efficient projected adaptive gradient descent algorithm. We establish the consistency and asymptotic normality of the estimator. Simulation studies confirm the efficacy of the proposed model, and its utility is further demonstrated through edge predictions on several real-world datasets.

- Yuhua Zhu yuhua.zhu@stat.ucla.edu

Title: A trade-off between discretization error and sample complexity in continuous-time reinforcement learning
Authors Yuhua Zhu
Institution: University of California, Los Angeles
Abstract: We consider continuous-time reinforcement learning (CTRL) where the system dynamics are governed by an unknown stochastic differential equation, and only discrete-time observations are available. We analyze the fundamental trade-off between discretization error and sample complexity for classical reinforcement learning algorithms and for the recently proposed Optimal-PhiBE framework. Our results reveal a surprising phenomenon in the linear–quadratic regulator problem: while classical methods suffer from the bias–variance trade-off induced by discretization, the Optimal-PhiBE framework can eliminate this barrier, yielding a “free lunch” in sample efficiency.

- Moo K. Chung mkchung@wisc.edu

Title: Covariate-Adjusted Topological Inference and Learning
Authors: Moo K. Chung (mkchung@wisc.edu)
Institution: University of Wisconsin-Madison
Abstract: Many structured datasets—from neuroimaging to genomics—are influenced by nuisance covariates such as age, sex, or acquisition site. Conventional machine learning methods for embedding and clustering often overlook these effects, leading to confounded comparisons where biological signals are obscured by covariate-driven variation. We address this limitation by formulating covariate adjustment, including data harmonization, as an equivalence-class problem in the sense of group theory: two data objects are equivalent if they differ only by the action of a covariate operator. This defines a latent quotient space that partitions observations into intrinsic structures after removing nuisance effects, providing a statistically principled framework for non-Euclidean and structured data. We demonstrate this framework in topological data analysis (TDA), where persistent features are embedded in Hilbert space and adjusted via orthogonal projection to construct the quotient space. This enables downstream

tasks such as embedding and clustering to be performed free of confounding influences. As a primary application, we analyze resting-state fMRI brain networks, adjusting for covariates including sex, age, and acquisition site. Covariate-adjusted TDA markedly reduces variability and enhances interpretability, offering a generalizable approach for reproducible, robust neuroscience and a template for structured data analysis in other scientific domains.

- Junwei Lu junweilu@hsph.harvard.edu

Title: Preference Inference for Language Model Evaluation
Authors: Yichi Zhang, Alexander Belloni, Ethan X. sesssse, Junwei Lu, Xiaoan Xu
Institution: Harvard T.H. Chan School of Public Health
Abstract: Human feedback tuning has been shown to effective to train the large language models (LLM). It allows the LLM understand the human feedback and preference. Despite the extensive literature dealing with algorithms aligning the rank of human preference, uncertainty quantification for the ranking estimation still needs to be explored and is of great practical significance. For example, it is important to overcome the problem of hallucination for LLM in the medical domain, and an inferential method for the ranking of LLM answers becomes necessary. In this talk, we will present a novel inferential framework for testing the ranking of human feedback and illustrate its application in the language models for medical knowledges.

Session 6: Joint models for risk prediction according to marker variability

Organizer: Helene Jacqmin-Gadda Helene.Jacqmin-Gadda@u-bordeaux.fr

- Jessica Barrett jessica.barrett@mrc-bsu.cam.ac.uk

Title: Flexible Bayesian semi-parametric approaches for modelling within-individual variability in joint models
Authors: Sida Chen[1], Jessica Barrett[1], Marco Palma[1,2], Jianxin Pan[3], Brian Tom[1]
Institution: [1]MRC Biostatistics Unit, University of Cambridge, [2] Great Ormond Street Institute of Child Health, University College London, [3]Beijing Normal–Hong Kong Baptist University
Abstract:
Joint models of longitudinal biomarker trajectories and time-to-event data allow us to estimate the association between the biomarker value and the risk of the event. But in many applications not only the mean value of the biomarker, but also the variability of the biomarker trajectory is predictive of the event risk. While many clinical applications have used naïve measures of within-individual variability (WIV), such as the standard deviation of observed measurements, there is a lack of agreement on how to define WIV statistically. We explore the definition of Wang et al (Biostatistics, 2023) which is based on a functional of the second-order derivative of the mean biomarker trajectory function with respect to time. We propose two semiparametric Bayesian approaches to flexibly model the biomarker trajectory and its WIV: the first approach uses Bayesian penalized splines and the second uses Bayesian function principal component analysis. We also consider two association structures with the time-to-event sub-model: the current curvature and the cumulative curvature. The approaches are evaluated and compared using simulation studies and applied to data from the UK Cystic Fibrosis Registry to estimate the association between lung function variability and mortality for adult women with cystic fibrosis.

- Leonie Courcoul leonie.courcoul@u-bordeaux.fr

Title: Joint models with heteroscedastic residual variance: an application to the impact of blood pressure variability on competing health events.
Authors: Léonie Courcoul, Antoine Barbieri, Hélène Jacqmin-Gadda
Institution: Univ. Bordeaux, INSERM U1219, Bordeaux Population Health
Abstract:
We propose a flexible joint modelling framework that rigorously accounts for individual variability in a longitudinal biomarker as a risk factor for health events. The longitudinal component is described using a linear mixed-effects model with heteroscedastic residual variance, allowing the residual variability to depend on individual random effects, covariates, or time. This formulation enables two approaches: modelling a time-dependent residual variability or explicitly distinguishing inter-visit and intra-visit residual variabilities when multiple measurements are collected at each visit.
The survival component can accommodate various event structures. It may take the form of a standard survival model with one or two competing events, or an illness-death model that accounts for an interval-

censored intermediate event in competition with a terminal event, such as death. It also allows for taking into account delayed entry. The event risks can be adjusted for the current value, slope, and variability of the biomarker.
The proposed joint model framework is illustrated using data from the Three-City cohort, assessing the impact of inter- and intra-visit blood pressure variability on the risks of dementia and death, while accounting for interval censoring and the semi-competing events.
Finally, the R package LSJM has been developed to implement these models, offering flexible specifications for residual variability and survival components, supporting one or two (semi-) competing events, interval censoring, and delayed entry. The package also provides tools for graphical model assessment and dynamic prediction of health event risks.

- Jianxin Pan,China jianxinpan@uic.edu.cn

Title: Measuring Biomarker Variability for Survival Data
Authors: Jianxin Pan
Institution: Beijing Normal-Hong Kong Baptist University, China
Abstract: It is well known that biomarker variability is prognostically significant in predicting associated disease risk such as cardiovascular disease. However, existing measures of biological variability are criticized for being entangled with random variability resulted from measurement error and/or being unreliable due to a likely limited number of measurements for each individual. We propose a new measure, defined as an integrated squared second derivative of the biomarker function over a period of time, so as to quantify the fluctuation of each individual biomarker trajectory. A Cox-type model is then proposed for survival data by incorporating the new variability measure as well as the underlying longitudinal trajectory level as predictors, forming a joint modelling framework for longitudinal data and survival data. A weighted version of the biomarker variability measure is also considered. Asymptotic properties of maximum likelihood estimators are studied and the parameter estimation is implemented via an EM algorithm. Simulation studies are conducted to assess the modelling strategy. The proposed model is applied to investigating the effects of systolic blood pressure variability on cardiovascular events by analyzing an elderly trial data of the UK, which motivates actually this research project. Extensions to other model scenarios are also discussed.

- Michael Elliott mrelliot@umich.edu

Title: Joint Modeling of Multiple Longitudinal Biomarkers and Survival Outcomes via Threshold Regression: Variability as a Predictor
Authors: Mingyan Yu, Zhenke Wu, Michelle M. Hood, Carrie Karvonen-Gutierrez, Sioban D. Harlow, Michael R. Elliott
Institution: University of Michigan
Abstract: Longitudinal biomarker data and health outcomes are routinely collected in many studies to assess how biomarker trajectories predict health outcomes. Existing methods primarily focus on mean biomarker profiles, treating variability as a nuisance. However, excess variability may indicate system dysregulations that may be associated with poor outcomes. In this paper, we address the long-standing problem of using variability information of multiple longitudinal biomarkers in time-to-event analyses by formulating and studying a Bayesian joint model. We first model multiple longitudinal biomarkers, some of which are subject to limit-of-detection censoring. We then model the survival times by incorporating random effects and variances from the longitudinal component as predictors through threshold regression that admits non-proportional hazards. We demonstrate the operating characteristics of the proposed joint model through simulations and apply it to data from the Study of Women's Health Across the Nation (SWAN) to investigate the impact of the mean and variability of follicle-stimulating hormone (FSH) and anti-Mullerian hormone (AMH) on age at the final menstrual period (FMP).

Session 7: Recent Developments in Survival Analysis

Organizer: Yifan Cui yifanc095@gmail.com

- Ingrid Van Keilegom ingrid.vankeilegom@kuleuven.be

Title: On an extension of the Cox model for time-dependent covariates under dependent censoring with unknown association
Authors: Negera Wakgari Deresa, Ingrid Van Keilegom

Institution: KU Leuven
Abstract:
This paper proposes a Cox model to address dependent censoring in the presence of time-dependent covariates, in which the dependence between the survival and the censoring time is modeled by a copula function. Unlike existing approaches, the parameter defining the copula function is considered unknown and is estimated from the data. While a partial likelihood approach can be used to estimate the regression coefficients in standard Cox models with time-dependent covariates, this approach is no longer suitable in the presence of dependent censoring. Instead, the likelihood depends on the nonparametric hazard function, and we approximate it using an expansion of M-spline basis functions. The model is then estimated using both maximum likelihood and penalized likelihood methods. The latter method employs a penalty function to regularize the baseline hazard estimator, giving a numerically stable estimator of the variance–covariance matrix. The finite-sample performance of the proposed estimator is evaluated via simulation studies, and the methodology is further illustrated using a real data example.

- Danyu Lin lin@bios.unc.edu

Title:Semiparametric Regression Analysis of Interval-Censored Data
Authors: Yangjianchen Xu, Donglin Zeng and Danyu Lin
Institution: University of North Carolina at Chapel Hill, USA
Abstract : Interval censoring arises frequently in clinical, epidemiological, financial, and sociological studies, where the event or failure of interest is not observed at an exact time point but is rather known to occur within a time interval induced by periodic monitoring. We formulate the effects of potentially time-dependent covariates on the interval-censored failure time through semiparametric regression models, such as the Cox proportional hazards model. We study nonparametric maximum likelihood estimation with an arbitrary number of monitoring times for each study subject. We develop an EM algorithm that involves very simple calculations and converges stably for any dataset, even in the presence of time-dependent covariates. We show that the estimators for the regression parameters are consistent, asymptotically normal, and asymptotically efficient with an easily estimated covariance matrix. In addition, we extend the EM algorithm and asymptotic theory to competing risks, multivariate failure times, and other types of data. Finally, we demonstrate the desirable performance of the proposed numerical and inferential procedures through simulation studies and applications to real medical studies.

- Donglin Zeng dzeng@umich.edu

Title: Semiparametric Regression Analysis of Interval-Censored Data
Authors*:*Mingya Long, Danyu Lin, Donglin Zeng
Institution: University of Michigan
Abstract : In many longitudinal studies, the occurrence of the event of interest is not directly observed but is rather known to lie within a random time interval. Such interval-censored data arise frequently in clinical and epidemiological studies because the occurrence of an asymptomatic disease can only be assessed periodically. Another complication in longitudinal studies is that the effects of covariates on the event time may vary over time. To address these challenges, we propose an approach to construct simultaneous confidence bands for time-varying hazard ratios with interval-censored data. With an initial estimator based on splines, we solve weighted estimating equations based on a subsample of the entire data set and then adopt a simple resampling technique to construct confidence bands. We show that the estimators for the time-varying coefficients converge weakly to Gaussian processes and that the confidence bands have correct coverage probabilities asymptotically. In addition, we evaluate the performance of the proposed methods through extensive simulation studies. Finally, we estimate the time-varying effect of baseline glucose level on the risk of diabetes in a major epidemiological cohort study.

- Zhe Fei zhef@ucr.edu

Title: Adaptive Mixture-of-Experts Transformers for Joint Modeling of Irregular Longitudinal and Survival Data
Authors*: Shijie Mao, Zhe Fei, Esra Kurum*
Institution: UC Riverside

Abstract: We present a general end-to-end framework for joint modeling of irregularly sampled longitudinal measures and time-to-event outcomes. Prior deep joint models are often multi-stage (e.g., MFPCA) or, when trained end-to-end, rely on restrictive design choices—such as fixed-interval observation grids (TransformerJM, MATCHNET) or pre-specified dependencies among longitudinal outcomes (TransformerLSR). Our approach embeds a mixture-of-experts (MoE) to a Transformer-based joint modeling architecture, enabling adaptation to diverse sampling patterns and outcome dynamics with minimal structural assumptions. The proposed method is compared with existing ones across a range of simulations. On two real-world datasets, the Mayo Clinic PBC2 randomized trial and the multicenter ADNI cohort, the model delivers robust predictive performance for both longitudinal reconstruction and survival risk prediction. The proposed architecture offers a novel scalable alternative to existing pipelines for joint modeling under irregular sampling.

Session 8: Trustworthy analysis of healthcare and medical data

Organizer: Fang Liu Fang.Liu.131@nd.edu
and Boris Hejblum boris.hejblum@u-bordeaux.fr

- Chuan Hong chuan.hong@duke.edu

Title: **An Agentic AI System for Multi-Framework Communication Coding**

Authors: Bohao Yang, Rui Yang, Joshua M. Biro, Haoyuan Wang, Jessica L. Handley, Brianna Richardson, Sophia Bessias, Nicoleta Economou-Zavlanos, Armando D. Bedoya, Monica Agrawal, Michael M. Zavlanos, Anand Chowdhury, Raj M. Ratwani, Kai Sun, Kathryn I. Pollak, Michael J. Pencina, Chuan Hong

Abstract:
Clinical communication is central to patient outcomes, yet large-scale human annotation of patient–provider conversations remains labor-intensive, inconsistent, and difficult to scale. Existing large language model (LLM)–based approaches often rely on static prompts or single-task pipelines that lack adaptability, interpretability, and reliability, particularly when applied across diverse communication frameworks and clinical domains.
We developed MOSAIC (Multi-framewOrk Structured Agentic AI system for Clinical Communication), a modular LangGraph-based system that orchestrates multiple specialized agents to automate the annotation of clinical conversations. Core components include a Plan Agent for codebook selection and workflow planning, an Update Agent for dynamic document retrieval and index maintenance, an Annotation Agent using retrieval-augmented generation (RAG) with codebook-guided prompts, and a Verification Agent that enforces consistency and integrates feedback.
To assess performance, we evaluated MOSAIC on 50 transcripts from rheumatology and OB/GYN encounters using gold-standard human annotations across multiple communication frameworks, including Patient Behavior, Bias, Relational Quality, and Shared Decision Making. MOSAIC achieved an overall F1 score of 92.8%, with strongest performance in Rheumatology (F1 = 96.2%) and in Patient Behavior and Bias domains. In ablation analyses, MOSAIC outperformed prompting-only and static-retrieval baselines by 6.8% in F1. The system also offers full provenance tracking and configurable RAG policies, supporting transparency and auditability.
As ambient AI scribes continue to generate large volumes of clinical text, MOSAIC provides essential infrastructure for transforming raw transcripts into structured, auditable, and multi-framework insights to support quality improvement, clinician education, and patient engagement at scale.

- Xinghua Mindy Shi mindyshi@temple.edu

Title: Mathematical Foundations of Generative AI for Human Genetics and Epigenetics

Authors: Xinghua Shi
Institution: Temple University

Abstract: Generative artificial intelligence (GenAI) is revolutionizing human genetics, epigenetics, and disease research by tackling long-standing analytical challenges and unlocking new opportunities for discovery. In this talk, I will highlight how key GenAI frameworks, such as variational autoencoders (VAEs), generative adversarial networks (GANs), Transformers, and diffusion models, are being leveraged to model complex genetics and epigenetics data. Drawing from examples in our own research, I will discuss both the mathematical underpinnings and practical applications of these models in addressing classical and emerging questions in human genetics and epigenetics.

- Yajuan Si yajuan@umich.edu

Title: Towards Differentially Private Finite Population Estimation: An Approach Based on Survey Weight Regularization
Authors: Jeremy Seeman, Yajuan Si, Jerome P. Reiter
Institution: University of Michigan
Abstract:
In general, it is challenging to release differentially private versions of survey-weighted statistics with low error and acceptable privacy loss. In part, this is because weighted statistics from complex sample survey data can be highly sensitive to the values of the survey weights, resulting in differentially private mechanisms that can add substantial noise to unbiased estimates of finite population quantities. On the other hand, simply disregarding the survey weights can introduce bias, which also can result in inaccurate estimates. Thus, the problem of releasing an accurate survey-weighted estimate essentially involves a trade-off between accuracy and privacy. We propose a strategy for managing this trade off to produce differentially private estimates of finite population quantities. The key step is to privately estimate a hyperparameter that determines how much to regularize or shrink unequal survey weights to equality as a function of privacy loss. We present an instantiation of the strategy under certain conditions, most notably for scenarios where the survey weights can be considered fixed features of the units in a finite population of interest. We illustrate the strategy using data from the Panel Study of Income Dynamics. We show that optimal strategies for releasing DP survey-weighted mean income estimates can require orders-of-magnitude less noise than naively using the original survey weights without modification. We conclude with discussions of the conditions underpinning the instantiation, highlighting future work needed to extend the estimation strategy to contexts often seen in practice.

- Fang Liu Fang.Liu.131@nd.edu

Title: A Differentially Private Weighted Empirical Risk Minimization Procedure and its Application to Outcome Weighted Learning
Authors: Fang Liu
Institution: University of Notre Dame
Abstract: Data used to build predictive models in the empirical risk minimization (ERM) framework often contain personal information. While these models can be highly accurate in prediction, shared model results may be susceptible to privacy attacks. Differential privacy (DP) offers a rigorous framework for privacy protection with mathematically provable bounds on the privacy loss incurred from releasing information based on sensitive data. Previous work has focused on applying DP to unweighted ERM. We consider weighted ERM (wERM), an important generalization, where each individual's contribution to the objective function may not be weighted equally. We propose the first DP algorithm for general wERM, with proven privacy guarantees. The general wERM framework creates a pathway for deriving privacy-preserving learning methods for individualized treatment rules, including the popular outcome weighted learning (OWL), a special case of wERM. We evaluate the performance of DP-wERM applied to OWL in experiments with simulated and real data. All empirical results demonstrate the feasibility of training OWL models via wERM with DP guarantees while maintaining sufficiently robust model performance, providing strong evidence for the practicality of implementing the proposed procedure in real-world scenarios involving sensitive data.

Session 9: Recent developments in reinforcement learning methods for precision medicine

Organizer: Donglin Zeng. dzeng@umich.edu

- Alex Luedtke aluedtke@uw.edu

Title: DoubleGen: Debiased Generative Modeling of Counterfactuals
Authors: Alex Luedtke, Kenji Fukumizu
Institution: Harvard University and Institute of Statistical Mathematics
Abstract: Generative models for counterfactual outcomes face two key sources of bias. Confounding bias arises when approaches fail to account for systematic differences between those who receive the intervention and those who do not. Misspecification bias arises when methods attempt to address confounding through estimation of an auxiliary model, but specify it incorrectly. We introduce DoubleGen, a doubly robust framework that modifies generative modeling training objectives to mitigate these biases. The new objectives rely on two auxiliaries -- a propensity and outcome model -- and successfully address confounding bias even if only one of them is correct. We provide finite-sample

guarantees for this robustness property. We further establish conditions under which DoubleGen achieves oracle optimality -- matching the convergence rates standard approaches would enjoy if interventional data were available -- and minimax rate optimality. We illustrate DoubleGen with three examples: diffusion models, flow matching, and autoregressive language models.

- Nilanjana Laha. nilanjanaaa.laha@gmail.com

Title: Model-free dynamic treatment regimes with arbitrary number of treatments and stages
Authors: Nilanjana Laha, Nilson Chapagain, Victoria Cicherski, Aaron Sonabend-W
Institution: University of Texas A&M
Abstract: Patients suffering from chronic diseases may receive treatments at multiple time-points or stages. Our aim is to learn the best individualized treatment policy, or dynamic treatment regimes (DTR), in this longitudinal setting using existing patient data. It is known that the above optimization problem reduces to a sequential weighted classification problem. This paper considers simultaneously solving the above-mentioned classification problem through Fisher consistent surrogate losses, when the number of treatment-levels per stage can be arbitrary. Although computationally feasible Fisher consistent losses are available for special settings, e.g., the binary treatment case, a general theory of surrogate losses for the sequential DTR classification remained undeveloped. To address this, we establish necessary and sufficient conditions for DTR Fisher consistency within the class of non-negative, stagewise separable surrogate losses. Furthermore, we show that many convex surrogate losses fail to be Fisher consistent for the DTR classification problem, and we formally establish this inconsistency for smooth, permutation equivariant, and relative-margin-based convex losses. Next, we use the sufficient conditions to construct smooth Fisher consistent surrogates. Since they are non-convex, the resulting simultaneous direct search method (SDSS) reduces to a smooth but non-convex optimization, for which we design a computationally fast, gradient-based algorithm. When the optimization error is small, we establish a sharp upper bound on the regret decay of SDSS for rich policy classes. We assess the numerical performance of SDSS through simulations. Finally, we demonstrate its real-world performance by applying it to estimate optimal fluid resuscitation strategies for severe septic patients using electronic health record data.

- Yuanjia Wang yw2016@cumc.columbia.edu

Title: Learning Optimal Early Decision Treatment Rules with Multi-domain Intermediate Outcomes
Authors: Wenbo Fei, Yuan Chen, Zexi Cai, Donglin Zeng and Yuanjia Wang
Institution: Columbia University
Abstract: Implementing precision medicine for mental disorders presents challenges due to disease complexity and heterogeneity in patient responses. Empirical studies suggest that early indicators, such as interim measures (e.g., interim patient self-reports) of disease improvement or relapse, can predict longer-term outcomes, serving as proxies when final outcomes (e.g., in-clinic assessments) are less accessible. However, existing approaches for deriving individualized treatment rules (ITRs) often ignore these early signals, instead focusing only on a final outcome as the reward. In this work, we propose a new method incorporating intermediate outcomes from various domains into a personalized composite outcome, serving as the reward for learning ITRs. This composite is a weighted sum of inferred latent states from observed measures, with weights personalized for each patient, ensuring consistency with the long-term final response. Our simulations show that this approach not only provides early detection of non-responders but also improves long-term treatment outcomes. Applying our framework to a randomized clinical trial on major depressive disorder (MDD) demonstrates its effectiveness and advantages in ITR learning.

Session 10: Distributional and Quantile Regressions for complex response data

Organizer: Antoine Barbieri antoine.barbieri@u-bordeaux.fr

- Angelo Alcaraz angelo.alcaraz@ens-paris-saclay.fr
  Title: On asymmetric Laplace regression models: application to trophic diversity
  Authors: Alcaraz Angelo, Durrieu Gilles, Poterie Audrey
  Institution: Univ. Bretagne Sud, CNRS UMR 6205, LMBA, France

Abstract: Measures of diet diversity in certain animal species can be based on determining an area within a space of carbon and nitrogen isotopic signature data. By comparing this area with the dispersion of these isotopic signatures, ecological research questions lead us to develop a model that focuses not on predictive quality, but on accurately representing data variability. The requirement for similarity between actual observations and modelling leads us to consider distributional regression models based on the asymmetric Laplace distribution, permissive models that allow for heteroscedasticity or asymmetry and demonstrate advantageous penalisation properties. We can see how an ecological problem that undermines traditional statistical methods can lead to the development of new statistical estimators, which in turn fuel
ecological questions.

- Lei Liu lei.liu@wustl.edu

Title: Efficient Estimation in Quantile Mixed Models via Smooth Check-Loss Approximation

Authors: Lei Liu
Institution: Washington University in St. Louis

Abstract: Quantile regression is a valuable framework for examining how predictors influence different quantiles of an outcome distribution. Unlike mean regression, it provides a flexible approach that is less sensitive to skewness and heteroscedasticity. However, the traditional check-loss function used in quantile regression is non-differentiable at the origin, complicating the computation of standard errors. While linear programming methods are commonly applied, incorporating random effects in mixed effects quantile model makes estimation less straightforward. To address this challenge, we propose a smooth approximation to the check-loss function, enabling estimation via maximum likelihood under the Asymmetric Laplace Distribution. The proposed method is implemented in SAS PROC NLMIXED for random effects quantile models. Simulation studies at different quantiles (τ = 0.25, 0.5, and 0.75) demonstrate that our method yields consistent parameter estimates and standard errors without requiring bootstrap procedures. This substantially reduces the computational burden compared to existing approaches, such as the *lqmm* package in R, which relies on bootstrapping for standard error estimation. Overall, our approach provides a smooth and computationally efficient alternative to traditional quantile regression with random effects, delivering valid inference without the need for resampling. This is a joint work with Tina Shih, Lexy Xu, and Lili Liu.

- Mouna Abed mouna.abed@u-bordeaux.fr

Title: A joint model based on distributional regression to study the link between blood pressure and the risk of cerebral vasospasm.

Authors: Abed M, De Courson H, Jacqmin-Gadda H, Barbieri A.
Institution: Univ. Bordeaux, Bordeaux Population Health center, Inserm U1219, Bordeaux, France

Abstract: Cerebral vasospasm is a serious complication of subarachnoid hemorrhage (SAH) associated with high morbidity and mortality. Although the mechanisms underlying vasospasm remain unclear, a growing number of studies suggest that blood pressure characteristics, such as variability, may play a key role in cerebrovascular complications. Blood pressure distribution shows large heterogeneity between subjects, both at the central level and in terms of variability and asymmetry of measurements. To account for this heterogeneity, we propose a joint model for heterogeneous longitudinal data and time-to-event data, assessing the impact of several features of the entire distribution of the longitudinal outcome on event risk. It combines a proportional hazard model and a distributional regression model. The last one is based on the asymmetric Laplace distribution where its location, dispersion and skewness parameters are defined by linear predictors depending on covariates and random effects. This model offers two main advantages. First, the interpretation of the parameters is intuitive. Second, it provides access to a wide range of time-varying individual distribution features to include as covariates in the survival model. This is made possible by the explicit form of the distribution quantiles and the cumulative distribution function which, depend on the parameters and random effects. Based on data from 201 patients hospitalized for SAH and monitored hourly for up to 14 days, the joint model will be applied to investigate the link between blood pressure and the risk of vasospasm.

- Gauthier Thurin Gauthier.thurin@ens.fr

Title: Optimal-transport based conformal prediction

Authors*:* Gauthier Thurin, Kimia Nadjahi, Claire Boyer

Institution: CNRS, École Normale Supérieure, Paris
Abstract: Conformal Prediction is a principled framework for quantifying uncertainty in blackbox learning models, by constructing prediction sets with finite-sample coverage guarantees. We introduce a novel procedure for multi-output settings, through the lens of optimal transport. Specifically, we leverage Monge-Kantorovich vector ranks and quantiles to construct prediction regions with flexible shapes better suited to the complex uncertainty patterns encountered in multivariate learning tasks. We prove that our approach ensures finite-sample, distribution-free coverage properties, similar to typical methods. We then apply our method for multi-output regression and multiclass classification. In the regression case, we use multivariate quantile regression to generate prediction regions adaptive to the inputs with asymptotic conditional coverage guarantees. Finally, we evaluate our method on practical regression and classification problems, illustrating its advantages in terms of (conditional) coverage and efficiency.

Session 11: Innovative Statistical Methods for Complex Health and Biomedical Data

Organizer: Mei Hao hao.mei@ruc.edu.cn

- Hao Zhang haozhang@arizona.edu

Title: A Performance-Based Framework for Transfer Learning Measurement and Guidance
Authors: Shudong Sun, Hao Helen Zhang
Institution: University of Arizona
Abstract: Transfer learning has become a cornerstone of modern machine learning, enabling the transfer of knowledge from data-rich source domains to data-scarce target domains. However, determining whether transfer learning will be beneficial prior to implementation remains a critical challenge. This work proposes a novel performance-based framework to measure similarity between source and target datasets and quantify transferability. Theoretical justifications are provided by connecting the new measure to the cosine similarity between decision boundaries in supervised learning. Key advantages of the proposed framework include its nonparametric and flexible nature, easy implementation, and computational scalability without requiring estimation of the underlying data distribution. We further suggest practical guidance that categorizes source datasets into positive, ambiguous, or negative zones based on their transferability. Finally, we extend this approach to encoder-head architectures in deep learning. Numerical results and real-world applications are presented to demonstrate the empirical effectiveness of the framework.

- Kai Zhang zhangk@email.unc.edu

Title: BELIEF in Dependence
Authors: Benjamin Brown, Kai Zhang, Xiao-Li Meng
Institution: University of North Carolina, Chapel Hill
Abstract: Two linearly uncorrelated binary variables must be also independent because non-linear dependence cannot manifest with only two possible states. This inherent linearity is the atom of dependency constituting any complex form of relationship. Inspired by this observation, we develop a framework called binary expansion linear effect (BELIEF) for understanding arbitrary relationships with a binary outcome. Models from the BELIEF framework are easily interpretable because they describe the association of binary variables in the language of linear models, yielding convenient theoretical insight and striking Gaussian parallels. With BELIEF, one may study generalized linear models (GLM) through transparent linear models, providing insight into how the choice of link affects modeling. For example, setting a GLM interaction coefficient to zero does not necessarily lead to the kind of no-interaction model assumption as understood under their linear model counterparts. Furthermore, for a binary response, maximum likelihood estimation for GLMs paradoxically fails under complete separation, when the data are most discriminative, whereas BELIEF estimation automatically reveals the perfect predictor in the data that is responsible for complete separation. We explore these phenomena and provide related theoretical results. We also provide preliminary empirical demonstration of some theoretical results.

- Lexin Li lexinli@berkeley.edu

Title: Brain Encoding and Decoding: Some Examples
Author: Lexin Li
Institution: University of California, Berkeley

Abstract: Brain encoding and decoding aims to understand the relationship between external stimuli and brain activity: encoding models how the brain transforms external information into neural signals, while decoding infers stimuli or cognitive states from recorded neural activities. It is a fundamental problem in cognitive and computational neuroscience, as it reveals how the brain represents, processes, and interprets information, providing crucial insights into the neural mechanisms underlying perception, cognition, and behavior. In this talk, we discuss a few examples of applying statistical methods to facilitate brain encoding and decoding tasks.

- Hao Mei hao.mei@ruc.edu.cn

Title: Functional Latent Space Model for Time-Varying Networks: Application to Taiwan Administrative Claims Data

Authors Guojun Zhu, Sanguo Zhang, Shuangge Ma, and Hao Mei
Institution: Renmin University of China

Abstract: Human disease network (HDN) analysis, which jointly considers a large number of diseases and focuses on their interconnections, is getting increasingly popular and can shed important insight not possessed by individual-disease-based analysis. Multiple network analysis techniques have been developed for HDNs, although new developments are still strongly needed. In this article, we adopt latent space modeling, which has proven powerful in other network analysis contexts and offers unique, insightful interpretations, but has been limitedly applied in HDN analysis. Different from some other types of network analysis and some other HDN analyses (such as gene-centric ones), in this article, we pay unique attention to modeling temporal variations in a functional way. For this purpose, a penalization approach is developed, which can identify time regions with constant network structures (that correspond to ignorable changes) as well as those with smooth variations. The statistical and computational properties are rigorously established. With Taiwan administrative claims data, we analyze the hospital records of 400 thousand patients from January 2008 to December 2013. Sensible findings are made on disease interconnections and clustering structures. Additionally, the temporal variations, which have not been revealed in the literature, are found to be interpretable. The analysis can provide a new way for connecting and grouping diseases and assist in understanding and planning medical resources.

Session 12: High-dimensional surrogate markers in multi-outcome settings

Organizer: Boris Hejblum boris.hejblum@u-bordeaux.fr

- Layla Parast parast@austin.utexas.edu

Title: Resilience Measures for the Surrogate Paradox

Authors:: Emily Hsiao, Lu Tian, Layla Parast
Institution: University of Texas at Austin, United States (Parast)

Abstract: Surrogate markers are often used in clinical trials to evaluate treatment effects when primary outcomes are costly, invasive, or take a long time to observe. However, reliance on surrogates can lead to the “surrogate paradox,” where a treatment appears beneficial based on the surrogate but is actually harmful with respect to the primary outcome. We propose formal measures to assess resilience against the surrogate paradox. Our setting assumes an existing study in which the surrogate markers and outcomes have been measured (Study A) and a new study (Study B) in which only the surrogates are measured. Rather than assuming transportability of the conditional mean functions across studies, we consider a class of functions for Study B that deviate from those in Study A. Using these, we estimate the distribution of potential treatment effects on the unmeasured multi-outcome and define resilience measures including a resilience probability, resilience bound, and resilience set. Our approach complements traditional surrogate validation methods by quantifying the plausibility of the surrogate paradox under controlled deviations from what is known from Study A. We investigate the performance of our proposed measures via a simulation study and application to a diabetes clinical trial.

- Florian Stijven florian.stijven@kuleuven.be

Title: Meta-Analytic Evaluation of Complex Surrogate Endpoints based on the Surrogate Index

Authors: Florian Stijven, Peter B. Gilbert

Institution: Florian Stijven: I-BioStat, KU Leuven, Leuven, Belgium
Peter B. Gilbert: Fred Hutchinson Cancer Center and University of Washington, Seattle, WA, USA
Abstract: The meta-analytic (MA) framework is widely regarded as the gold standard for evaluating trial-level surrogate endpoints. However, the canonical MA framework has two important limitations: it cannot handle complex surrogates, and it cannot flexibly adjust for differences in the distribution of baseline covariates across trials. By contrast, in the context of data fusion, the surrogate-index (SI) framework can handle complex surrogates and allows for complex relationships between baseline covariates, surrogates, and clinical endpoints. Yet, the SI framework is not designed for evaluating surrogates and relies on strong identifying assumptions.
To address the limitations of the MA framework, we propose an extension that incorporates ideas from the SI framework. Specifically, we estimate a transformation of the baseline covariates and the surrogate, the so-called surrogate index. This is a new univariate surrogate. Under certain conditions, the surrogate index is a perfect trial-level surrogate and the estimated version—the estimated surrogate index—is expected to be an almost perfect trial-level surrogate.
Our estimand is the correlation between treatment effects on the estimated surrogate index and on the clinical endpoint, a so-called trial-level correlation. This data-adaptive estimand depends on the observed data through the estimated transformation. We show that, under weak regularity conditions, asymptotically valid inference about this estimand is possible. Importantly, these conditions allow for using machine learning to estimate the surrogate index, enabling the evaluation of complex (e.g., high-dimensional) surrogates by reducing them to the univariate estimated surrogate index. Until now, it was not possible to evaluate high-dimensional surrogates in the MA framework.
We illustrate this approach with a set of COVID-19 vaccine trials where antibody markers are evaluated as trial-level surrogates.

- Tianxi Cai tcai@hsph.harvard.edu

Title: Stable Multi-Surrogate Transformation for Robust and Generalizable Surrogacy
Authors: Keyao Zhan, Yihong Gu, and Tianxi Cai
Institution: Harvard University
Abstract: In randomized clinical trials, the primary outcome Y often requires long-term follow-up or is costly to measure. For such settings, it is desirable to use some surrogate markers S to infer the treatment effect on the primary outcome. Programs often collect many candidate markers across heterogeneous studies; naively combining them yields surrogacy measures that fail to generalize to new target population. In this paper, we propose a single, model-free optimal transformation of stable surrogates S, which optimizes the proportion of treatment effect explained (PTE) and generalizes well under distribution shift. To ensure distributional robustness across trials, we use a maximin objective and identify its solution as a convex combination of the optimal transformation from single source. We then propose a stability-aware selection pipeline that screens surrogate subsets via an invariance gap enforcing cross-source stability, and selects via a distributionally robust objective evaluated on the adversarial setting. We develop nonparametric estimators accommodating flexible machine learning and high-dimensional extensions, and establish selection consistency and near-optimal worst-case risk; we also give conditions under which the approach avoids the surrogate paradox within our DRO uncertainty set. Simulations demonstrate stable selection, improved worst-case predictiveness and generalizable PTE on diverse targets.

- Peter Gilbert peterg@uw.edu

Title: Surrogate endpoint evaluation for outcomes with high-dimensional features, with application to vaccines
Authors : Peter B. Gilbert, Florian Stijven
Institution: Fred Hutchinson Cancer Center, Seattle, WA USA
Abstract: sequence of the outer surface protein
A basic principle for developing a surrogate endpoint involves pairing a candidate surrogate with the specific outcome under investigation. Many infectious pathogens circulate as multiple immunologically distinct types—typically serotypes. In vaccine research, standard practice uses an antibody response measured against a specific type as a surrogate for the disease outcome caused by that same type (“strain-matched surrogate”).
However, for certain infectious diseases, there exist relevant pathogen types not defined by a low-dimensional categorization of strains. Instead, they are high-dimensional, characterized by the amino acid sequence of a protein of the pathogen causing the disease outcome or a continuous

immunophenotype such as experimentally measured sensitivity to neutralization by sera collected from vaccine recipients.
This work explores how surrogate endpoint evaluation frameworks can be adapted to accommodate outcomes with high-dimensional features. We illustrate this using data from vaccine (or monoclonal antibody) efficacy trials for dengue, HIV, and SARS-CoV-2.
One challenge is identifying the most useful pathogen features for integration into the antibody-based surrogate endpoint evaluation. Features meriting consideration should both (a) modify vaccine efficacy and (b) modify the association between the antibody marker and the disease outcome.
Analyzing data from six SARS-CoV-2 vaccine efficacy trials, we present an example of trial-level surrogate endpoint evaluation. Candidate surrogates are defined by measuring or predicting neutralizing antibody titers against a panel of SARS-CoV-2 strains that represent circulating lineages, as well as predicting each individual's relative level of potential exposure to each circulating lineage.
A tradeoff in identifying an optimal surrogate balances surrogate quality against ease of measurement—specifically, deciding which experimental neutralization measurements to include versus relying on predictions from external studies. The talk also considers the use of predicted viral envelope structures, derived from envelope amino acid sequences using deep neural networks such as AlphaFold3, to potentially improve surrogate endpoints.

Session 13: Recent advances in transfer learning for biomedical applications

Organizer: Lan Luo ll1118@sph.rutgers.edu

- Rui Duan

Title: Unsupervised Aggregation of Multiple Learning Algorithms
Authors: Rui Duan
Institution: Department of Biostatistics, Harvard T.H. Chan School of Public Health
Abstract: Across various domains, the growing advocacy for open science and open-source machine learning has made an increasing number of models publicly available. These models allow practitioners to integrate them into their own contexts, reducing the need for extensive data labeling, training, and calibration. However, selecting the best model for a specific target population remains challenging due to issues like limited transferability, data heterogeneity, and the difficulty of obtaining true labels or outcomes in real-world settings. In this paper, we propose an unsupervised model aggregation method, U-aggregation, designed to integrate multiple pre-trained models for enhanced and robust performance in new populations. Unlike existing supervised model aggregation or super learner approaches, U-aggregation assumes no observed labels or outcomes in the target population. Our method addresses limitations in existing unsupervised model aggregation techniques by accommodating more realistic settings, including heteroskedasticity at both the model and individual levels, and the presence of adversarial models. Drawing on insights from random matrix theory, U-aggregation incorporates a variance stabilization step and an iterative sparse signal recovery process. These steps improve the estimation of individuals' true underlying risks in the target population and evaluate the relative performance of candidate models. We provide a theoretical investigation and systematic numerical experiments to elucidate the properties of U-aggregation. We demonstrate its potential real-world application by using U-aggregation to enhance genetic risk prediction of complex traits, leveraging publicly available models from the PGS Catalog.

- Emily C Hector

Title: Heterogeneity-adaptive meta-analysis
Authors: Elizabeth M Davis, Emily C Hector
Institution: University of Michigan
Abstract: Meta-analytic methods tend to take all-or-nothing approaches to study-level heterogeneity, assuming all studies are heterogeneous or homogeneous, leading to inefficiency and/or bias in estimation and inference. In this paper, we develop a heterogeneity-adaptive meta-analysis in linear models that adapts to the amount of information shared between datasets. The primary mechanism for the information-sharing is a shrinkage of dataset-specific distributions towards a new ``centroid'' distribution through a Kullback-Leibler divergence penalty. The Kullback-Leibler divergence is uniquely geometrically suited for measuring relative information between datasets, and leads to relatively simple closed form estimators with intuitive interpretations. We establish our estimator's desirable inferential properties without assuming homogeneity of dataset parameters. Among other results, we show that our

estimator has a provably smaller mean squared error than the dataset-specific maximum likelihood estimators, and establish asymptotically valid inference procedures. A comprehensive set of simulations highlights our estimator's versatility, and an analysis of data from the eICU Collaborative Research Database illustrates its performance in a real-world setting.

- Tian Gu

Title: Hierarchical Projection for Adaptive Knowledge Transfer
Author: Samhita Pal, Amir Asiaee, Tian Gu
Institution: Department of Biostatistics, Columbia University Mailman School of Public Health
Abstract: Modern studies often face the challenge of limited sample sizes in a primary dataset while related information exists across multiple external sources. Simply combining these sources can be harmful if they are only partially relevant or contain spurious signals. We propose Projection Transfer Learning (Projection-TL), a new framework that selectively integrates information from multiple auxiliary models while protecting against negative transfer. Projection-TL adaptively identifies which sources and variables align with the target dataset and projects them into a sparse, interpretable model that remains robust even when many sources are misaligned. Through simulations and real data applications, our method demonstrates improved accuracy, stability, and interpretability compared to existing transfer learning approaches. This work provides a practical and generalizable strategy for harnessing heterogeneous external knowledge in high-dimensional settings.

- Ling Zhou

Title: On a synergistic learning phenomenon in nonparametric domain adaptation
Authors: Ling Zhou and Yuhong Yang
Institution: Southwestern University of Finance and Economics
Abstract: Consider nonparametric domain adaptation for regression, which assumes the same conditional distribution of the response given the covariates but different marginal distributions of the covariates. An important goal is to understand how the source data may improve the minimax convergence rate of learning the regression function when the likelihood ratio of the covariate marginal distributions of the target data and the source data are unbounded. A previous work of Pathak, Ma and Wainwright (2022) show that the minimax transfer learning rate is simply determined by the faster rate of using either the source or the target data alone. In this paper, we present a new synergistic learning phenomenon (SLP) that the minimax convergence rate based on both data may sometimes be faster (even much faster) than the better rate of convergence based on the source or target data only. The SLP occurs when and only when the target sample size is smaller (in order) than but not too much smaller than the source sample size in relation to the smoothness of the regression function and the nature of the covariate densities of the source and target distributions. Interestingly, the SLP happens in two different ways according to the relationship between the two sample sizes. One is that the target data help alleviate the difficulty in estimating the regression function at points where the density of the source data is close to zero and the other is that the source data (with its larger sample size than that of the target data) help the estimation at points where the density of the source data is not small. The SLP is illustrated via simulations.

Session 14: Clustering from multivariate data

Organizer: Anais Rouanet anais.rouanet@u-bordeaux.fr

- Julien Jacques julien.jacques@univ-lyon2.fr

Title: Clustering of Multivariate Longitudinal Data of mixed type
Authors: Francesco Amato, Julien Jacques
Institution: Université de Lyon
Abstract: Multivariate longitudinal data of mixed types are increasingly collected across scientific domains. However, clustering algorithms for such data remain scarce due to the challenge of simultaneously modeling within-time and between-time dependence structures for multivariate data of mixed types.We introduce the Mixture of Mixed-Matrices (MMM) model, which reorganizes the data into a three-way structure. The model assumes that non-continuous variables (ordinal, binary,

nominal, or count data) are observations of underlying latent continuous variables. By relying on a mixture of matrix-variate normal distributions, the MMM model performs clustering in the latent dimension. It can thus handle continuous, ordinal, binary, nominal, and count data, while concurrently modeling heterogeneity, associations among responses, and temporal dependencies —without assuming conditional independence—in a parsimonious way.
Inference is conducted via an MCMC-EM algorithm, which we detail. The model's performance is evaluated using synthetic data, demonstrating its inference capabilities. Finally, a real-world application to financial data is presented.

- Julie Fendler julie.fendler@mrc-bsu.cam.ac.uk

Title: Consensus Monte Carlo for mixtures of categorical distributions

Authors: Julie Fendler, Paul Kirk
Institution: MRC Biostatistics Unit, University of Cambridge

Abstract: Motivated by privacy-sensitive electronic health record (EHR) data, we present a novel Consensus Monte Carlo (CMC) algorithm for Bayesian categorical mixture models in a federated learning setting, i.e. when data cannot be fully shared/pooled across compute nodes.
CMC assumes that the dataset is split into S shards, and an MCMC algorithm is run independently within each shard. For a given shard, the target posterior of the MCMC algorithm, called the local posterior, is proportional to the likelihood computed within the shard multiplied by a fractionated prior.
To draw samples from the global posterior (i.e. the posterior over the entire dataset), different methods exist for combining the local posteriors. A naïve approach is to average the draws from the local posteriors at each iteration. Alternatively, since the data may be heterogeneously distributed across shards, a weighted mean can be considered. Rabinovich, Angelino and Jordan (2015) propose to infer these weights, called shard weights, within a variational inference framework. However, in the context of finite mixture models, their approach assumes that the mixture weights are known, and inferring the shard weights requires access to the entire dataset.
We extend their work to more general settings and show that, in the context of mixtures of categorical distributions, only a summary of the dataset is required to infer the shard weights, and consider an application to large-scale EHR data.

- Anais Rouanet anais.rouanet@u-bordeaux.fr

Title: Non-parametric clustering of multivariate longitudinal data with variable selection: Identifying sub-phenotypes of Alzheimer's Disease

Authors*:* Anaïs Rouanet, Carole Dufouil, Cécile Proust-Lima
Institution:U1219 Bordeaux Population Health research Center, Bordeaux, France

Abstract: Many diseases are characterized by highly heterogeneous progression patterns across patients. This is the case in Alzheimer's disease and related dementias (ADRD). Despite the abundance of biomarkers now available in ageing cohorts to describe pathological changes involved in AD (such as neurodegeneration, cognitive impairment, brain atrophy), this heterogeneity is statistically difficult to apprehend. It requires clustering methods that could handle a large number of biomarker data measured repeatedly at irregular visits over time. In this work, we developed a non-parametric Bayesian clustering model to identify latent clusters of subjects from multivariate longitudinal outcomes and additional cross-sectional variables. The objective was to uncover latent ADRD sub-phenotypes from repeated biomarker data and to characterize their specific physio pathological pathways.
We extended the profile regression approach developed by Liverani et al. (2015), that links non-parametrically a response vector and cross-sectional variables through cluster membership, to handle multiple longitudinal outcomes measured irregularly over time. The cluster-specific trajectories were described by mixed models, and the profiles of cross-sectional variables were modeled using appropriate Gaussian or discrete models. A Dirichlet Process prior was adopted for the mixture distribution to deal with an unconstrained number of clusters, and a variable selection procedure based on importance weighting was used to identify both repeated and cross-sectional markers that best discriminate between clusters. Parameter estimation was achieved using Monte Carlo Markov Chains.
We illustrate this method using data from the French MEMENTO study (Dufouil et al., 2017) which provides longitudinal cognitive evaluations together with cross-sectional measures of brain imaging

volumes. The objective is to identify sub-phenotypes of ADRD that differ according to the sequence and speed of neuropathological degradations.

- Qendresa Selimi qendresa.selimi@manchester.ac.uk

Title: A Dynamic Joint Latent Class Model With Time-Dependent Transition Probabilities
Authors: Qendresa Selimi, Christiana Charalambous, Taban Baghfalaki.
Institution: University of Manchester
Abstract: Joint models are studied for analysing the association between longitudinal and survival data, which often arise in clinical studies. To account for heterogeneity within the population, joint latent class models have been developed. We propose a new Dynamic Joint Latent Class Model (DJLCM) by extending the traditional JLCM to incorporate a time-varying class membership model. We use a Hidden Markov Model to model the probability of transitioning between classes over time. The longitudinal outcome, given class membership at each time point, is modelled using a linear mixed-effects model. Additionally, we include a survival submodel conditional on the class membership, capturing the association between longitudinal trajectories and time-to-event outcomes. For parameter estimation, we adopt a Bayesian framework using Hamiltonian Monte Carlo (HMC). Instead of the traditional likelihood function, we employ the Forward Algorithm to account for the dynamic nature of the latent classes. We evaluated the proposed model through a simulation study and applied it to a study on children suffering from inflammatory diseases at Great Ormond Street Hospital (GOSH) in London. We analysed the association between repeated creatinine measurements and time to chronic kidney disease or acute kidney injury, while also providing individualised risk predictions.

Session 15: The use of functional data analysis in multivariate settings

Organizer: Helene Jacqmin-Gadda helene.jacqmin-gadda@u-bordeaux.fr

- Luo Xiao lxiao5@ncsu.edu

Title: Functional joint model for longitudinal and survival data with a functional covariate with application to a longitudinal study of objective physical activity
Authors: Luo Xiao[1], Wenyi Wang[1], Ilsuk Kang[3], Michael LaMonte[2], Chongzhi Di[3]
Institution: [1]North Carolina State University, [2]University at Buffalo, and [3]Fred Hutchinson Cancer Center
Abstract: The Objective Physical Activity and Cardiovascular Health (OPACH) study collects objectively measured physical activity (PA) by accelerometers in elderly women and aims to determine the association of PA with risks of incidents such cardiovascular diseases and mortality. We convert the accelerometry data into daily profiles of sitting bouts of different lengths and propose a joint model for longitudinal (physical function via survey) and survival data (all-cause mortality) with a baseline functional covariate. The longitudinal data is modeled by functional data methods and links to the survival data via functional principal component scores. In both sub-models, linear scalar-on-function regression forms are used to account for the association of the baseline functional covariate. For model estimation, an expectation-maximization (EM) algorithm coupled with penalized splines is proposed. When applied to the OPACH study, the proposed model gives scientifically interpretable associations of sitting-bout-length profiles with physical function and mortality. A simulation study is also conducted to confirm the good performance of the proposed model.

- Andrew Simpkin andrew.simpkin@universityofgalway.ie

Title: Estimation and Application of Derivative Multivariate Functional Principal Component Analysis
Authors: Yueyun Zhu, Andrew Simpkin
Institution: School of Mathematical and Statistical Sciences, University of Galway, Ireland

Abstract: Dai and Müller (2018) proposed a Derivative Functional Principal Component Analysis (DFPCA) approach to recover the derivatives of univariate functional data, by estimating the derivative functional principal components (DFPCs) and the corresponding scores. Happ and Greven (2018) proposed a Multivariate Functional Principal Component Analysis (MFPCA) approach for multivariate functional data. Following their studies, we aim to develop a Derivative Multivariate Functional Principal Component Analysis (DMFPCA) approach which estimates the derivatives of multivariate functional data.
The DMFPCA approach starts from applying Mercer's theorem to the covariance of derivatives of multivariate functional data. The eigenvalues of the covariance represent the amount of variability in the derivatives explained by the derivative multivariate functional principal components (DMFPCs), while the derivative multivariate functional principal component scores (DMFPC-scores) serve as weights of DMFPCs in the multivariate Karhunen-Loève expansion of the derivatives. We establish a relationship between DFPCA and DMFPCA, which offers an estimation strategy to calculate DMFPCs and DMFPC-scores based on their univariate counterparts. The asymptotic properties for the proposed estimates are derived. It emerges that DMFPCA provides more accurate recovered derivatives and a more reasonable explanation for understanding data dynamics.
Analysing the dynamics of functional data through their derivatives is of interest in biostatistics. For example in cardiology, angiograms provide two functional variables – the quantitative flow ratio (QFR) and diameter – measured across the length of a vessel. The dynamics of these functional data are of key interest to assess coronary artery disease. However, there are currently no methods to deal with the derivatives of multivariate functional data. The DMFPCA we propose extends the existing methods for estimating the derivatives of univariate functional data to multivariate cases. By applying

DMFPCA to the angiogram data, it allows us to explore the joint dynamics between QFR and diameter. It turns out that DMFPCA effectively captures important modes of variation in the change of diameters and the velocity of QFR, which helps to assess coronary artery disease.

- Robin Genuer robin.genuer@u-bordeaux.fr

Title: Random Survival Forests for the prediction of survival outcomes from time-varying markers using functional principal component analysis

Authors: Corentin Segalas, Robin Genuer, Cécile Proust-Lima
Institution: Université de Bordeaux, Bordeaux Population Health

Abstract: Individual dynamic prediction of survival events from medical history is statistically challenging, notably because the history includes time-dependent predictors measured irregularly and with error. A variety of models have been proposed to this end. However, the landmark approach does not make use of all available data, the joint models estimation becomes intractable when the number of longitudinal predictors increases, and two-stage regression calibration methods ignore the informative truncation of the longitudinal process due to the event. In this work, we propose to combine functional principal component analysis (FPCA) and random survival forests to predict a survival outcome from longitudinal predictors.
Random forests can capture complex and nonlinear relations between predictors and the outcome. It has been extended to handle survival outcome and competitive risks under the framework of random survival forests (RSF). However, they cannot include time-dependent predictors. We propose an extension of RSF where, for each tree and at each split, individual trajectories of longitudinal predictors are summarized by their FPCA scores, which are then used for splitting. This method has two main advantages: informative truncation of predictors by the event is taken into account as it considers survival and longitudinal data jointly and it is fully nonparametric.

A simulation study illustrates the performance of the methodology implemented in the R package DynForest. The method is also applied to predict the risk of severe complication in patients hourly monitored in neuro-intensive care unit after a subarachnoid hemorrhage.
This fully non-parametric approach allows to predict clinical events from many irregularly measured predictors. Moreover, thanks to the random forests structure, one can retrieve variable importance assessment and perform variable selection, making this method more explainable.

- Sophie Dabo sophie.dabo@univ-lille.fr

Title: Preliminary Test Estimation in ULAN models for Functional Data

Authors: Sophie Dabo-Niang
Institution: University of Lille, Université Libre de Bruxelles
Abstract: Functional data, in which each observation takes the form of a curve, appear in a wide range of scientific contexts. In these settings, estimation procedures can be improved by using preliminary tests to adaptively choose estimators according to specified model structures. We introduce a general approach to preliminary test estimation for ULAN (Uniform Local Asymptotic Normality) models, extending classical asymptotic methods to encompass infinite-dimensional functional parameters. We establish the estimators' asymptotic behavior, investigate their risk behavior, and document performance improvements via numerical studies.

Session 16: New advances in survival analysis and machine learning

Organizer: Yichuan Zhao yichuan@gsu.edu
And Liang Li lli15@mdanderson.org

- Edsel Pena pena@stat.sc.edu

Title: Prediction Machines Based on Concordance Correlation
Authors: Edsel A. Peña and Md. Ariful Islam Sanim
Institution: Department of Statistics University of South Carolina Columbia, SC 29208 USA
Abstract: Let **V** and **W** be $p \times 1$ random vectors with the same measurement units. The (sample) concordance correlation coefficient (CCC) between these vectors is defined via $\mathrm{CCC}(\mathbf{V}, \mathbf{W}) = \frac{2rS_V S_W}{S_V^2 + S_W^2 + (\bar{V} - \bar{W})^2}$

where r is the sample Pearson's correlation between **V** and **W**, S_V^2 and S_W^2 are their respective sample variances, and $\bar{V}$ and $\bar{W}$ are their respective sample means. This correlation measure was introduced in Lin (1989) and it measures the degree of agreement between **V** and **W**. Let $(\mathbf{X}, Y)$ where $X \in \Re_p$ and $Y \in \Re$ be a random vector, and let $(\mathbf{X},\mathbf{Y}) = \{(\mathbf{X}_i, Y_i) : i=1,2,...,n\}$ be IID copies of $(\mathbf{X}, Y)$. Consider the problem of predicting the value of Y, given $\mathbf{X} = x$. Let $\mathrm{C} = \{\delta(\cdot : \theta) : \Re_p \to \Re, \theta \in \Theta\}$ be a postulated class of predictors, which is parametrized by a parameter θ. Define. $\theta^\circ(\mathbf{X},\mathbf{Y}) = \mathrm{argmax}\,\mathrm{CCC}[\mathbf{Y}, \{\delta(\mathbf{X}_i;\theta), i = 1,2,...,n\}]$. Consider then the predictor $\delta^\circ(x;(\mathbf{X},\mathbf{Y})) = \delta(x : \theta^\circ(\mathbf{X},\mathbf{Y}))$.
In this talk we examine the properties of this predictor and its viability for predicting the value of Y, given an observed value of X. We also examine the impact of the choice of the postulated class C in terms of predictive ability, and potentially in the context of variable selection

- Yifan Cui yifanc095@gmail.com

Title : Double Machine Learning of Continuous Treatment Effects with General Instrumental Variables
Authors : Yifan Cui
Institution : Zhejiang University
Abstract: Estimating causal effects of continuous treatments is a common problem in practice, for example, in studying dose-response functions. Classical analyses typically assume that all confounders are fully observed, whereas in real-world applications, unmeasured confounding often persists. In this article, we propose a novel framework for local identification of dose-response functions using instrumental variables, thereby mitigating bias induced by unobserved confounders. We introduce the concept of a uniform regular weighting function and consider covering the treatment space with a finite collection of open sets. On each of these sets, such a weighting function exists, allowing us to identify the dose-response function locally within the corresponding region. For estimation, we develop an augmented inverse probability weighting score for continuous treatments under a debiased machine learning framework with instrumental variables. We further establish the asymptotic properties when the dose-response function is estimated via kernel regression or empirical risk minimization. Finally, we

conduct both simulation and empirical studies to assess the finite-sample performance of the proposed methods.

- Liang Li lli15@mdanderson.org

Title: Modeling the Association Between Multivariate Nonlinear Longitudinal Data and Survival Outcome: A Measurement Error Perspective
Authors: Liang Li
Institution: MD Anderson Cancer Center
Abstract: In longitudinal cohort studies, the repeated measures data and the time-to-event data are correlated outcomes that quantify different aspects of the study subject's health status. It is of research interest to explore their associations. From a measurement error modeling perspective, this can be viewed as a Cox model with intermittently measured error-prone time-dependent covariates. Of the two mainstream approaches to measurement error models, the structural approach, which uses distributional assumptions for the longitudinal trajectories, resembles the shared random effect joint models. In this paper, we study the functional approach, which does not use such distributional assumptions. We develop a non-likelihood-based approach to estimating individual-level longitudinal trajectories that are possibly nonlinear and multivariate, avoiding the potential statistical challenge from informative dropout. These estimated trajectories are used in a Cox model with time-dependent covariates with a measurement error correction step to produce unbiased estimation. The proposed method is computationally less intensive than the structural approach and more robust against mis-specifying the shapes of the longitudinal trajectories. We demonstrate that the proposed error correction method works with a wide range of association structures between the longitudinal trajectories and the survival outcome.

- Yichuan Zhao yichuan@gsu.edu

Title: Novel empirical likelihood method for the cumulative hazard ratio under stratified Cox models
Authors: Dazhi Zhao, Yichuan Zhao
Institution: Georgia State University
Abstract: Evaluating the treatment effect is a crucial topic in clinical studies. Nowadays, the ratio of cumulative hazards is often applied to accomplish this task, especially when those hazards may be nonproportional. The stratified Cox proportional hazards model, as an important extension of the classical Cox model, has the ability to flexibly handle nonproportional hazards. In this article, we propose a novel empirical likelihood method to construct the confidence interval for cumulative hazard ratio under the stratified Cox model. The large sample properties of the proposed profile empirical likelihood ratio statistic are investigated, and the finite sample properties of the empirical likelihood-based estimators under some different situations are explored in simulation studies. The proposed method was finally applied to perform statistical analysis on a real world dataset on the survival experience of patients with heart failure.

Session 17: Statistical Learning for Complex Data

Organizer: Xingqiu Zhao xingqiu.zhao@polyu.edu.hk

- Qixian Zhong qxzhong@xmu.edu.cn

Title: Semiparametric Inference for Functional Survival
Authors: Hongyi Zhou, Wenqing Su, Qixian Zhong, Ying Yang
Institution: Xiamen University
Abstract: In recent years, there has been growing interest among researchers in modeling right-censored survival data with functional covariates. While existing functional methods primarily focus on

the Cox model, its proportional hazards assumption can be challenging to verify and may be violated in practice. To address this issue, we extend the ordinary differential equation (ODE) framework for survival data to incorporate functional covariates and develop an inference procedure for both scalar and functional parameters. Specifically, we establish asymptotic normality and semiparametric efficiency for the scalar coefficient estimators, enabling a valid inference procedure. Additionally, we derive an asymptotic simultaneous confidence band for the functional parameter. Simulations are conducted to evaluate the finite sample performance of the proposed method.

- Qiang Wu qiangwu@mail.bnu.edu.cn

Title: Deep Nonparametric Inference for Interval-Censored Data

Authors*:* Qiang Wu, Wen Su, Shuangge Ma and Xingqiu Zhao
Institution: The Hong Kong Polytechnic University

Abstract: Interval-censored failure time data pose significant challenges in survival analysis, particularly when capturing complex covariate interactions. Traditional statistical approaches, such as the Cox proportional hazards model, are often limited by strict structural assumptions. Conversely, fully nonparametric methods, like interval-censored recursive forests, relax these assumptions but lack the capacity for rigorous statistical inference. To overcome these limitations, we propose a novel deep neural network-based approach to directly estimate the conditional cumulative hazard function. A key contribution of this work is establishing a rigorous theoretical foundation for deep survival models. Specifically, we derive nonasymptotic error bounds and establish nonparametric asymptotic normality, thereby facilitating valid statistical inference, a critical component often absent in deep learning applications. Furthermore, we develop novel procedures for two-sample testing and goodness-of-fit evaluation. Extensive simulation studies demonstrate that our proposed estimator achieves superior predictive accuracy and maintains high performance across diverse data-generating mechanisms compared to state-of-the-art methods. Finally, an analysis of the Atherosclerosis Risk in Communities study confirms its strong empirical performance.

- Xiangbin Hu xiang-bin.hu@connect.polyu.hk

Title: Deep Conditional Generative Learning for Optimal Individualized Treatment Rules

Authors: Xiangbin Hu, Wen Su, Zhisheng Ye, Xingqiu Zhao.
Institution: Beijing Institute of Technology

Abstract: Personalized treatment regimes that account for individual characteristics offer substantial potential to minimize treatment risks and enhance patient outcomes. Existing methods for estimating individualized treatment rules in multi-arm settings often face limitations due to model misspecification. To overcome the challenges, we propose a novel conditional generative learning framework, CG-Learning, which employs a Wasserstein generative adversarial network to estimate optimal decision rules that minimize a specified risk measure in multi-armed treatment scenarios. By modeling the conditional distribution of rewards through conditional generative adversarial networks, our approach avoids restrictive structural assumptions. We establish key theoretical properties of the resulting decision rules, including nonasymptotic bounds on regret and mis-assignment probabilities. We show that these bounds can be tightened when the covariate space has a low Minkowski dimension. Through comprehensive simulations and an application to data from the AIDS Clinical Trials Group, we demonstrate that CG-Learning outperforms existing benchmarks.

Session 18: Software for the Analysis of Multi-Outcome Models (INLAjoint, JMbayes2, frailtypack)

Organizer: Denis Rustand denis.rustand@u-bordeaux.fr

- Elias Krainski elias.krainski@kaust.edu.sa

Title: graphpcor: Models for correlation matrices based on graphs.

Authors: Elias Teixeira Krainski, Haavard Rue, Denis Rustand, Janet van Niekerk, Anna Freni Sterrantino.
Institution: King Abdullah University of Science and Technology, Saudi Arabia.

Abstract: Correlation matrices are fundamental in multivariate statistical analysis; modeling their structure is a way to understand complex dependencies between random variables, but assigning priors for them presents challenges. Standard choices, such as the Wishart distribution, suffer from restrictive features, including an undesirable dependency between variances and correlations. Similarly, the LKJ prior is limited by its tendency to shrink estimations toward an uncorrelated identity matrix. In the graphpcor R package, we implement a novel framework that addresses these limitations by constructing correlation matrix whose structure is deduced from a user-defined graph. This graphical approach enhances interpretability, making the dependency structure explicit and intuitive, and promotes parsimony by reducing the number of parameters to be estimated, which eases the computational burden. We propose the use of the multivariate penalized complexity prior in order to control for model complexity and define informative priors. These priors provide a principled approach to preventing overfitting by penalizing the model's divergence from a simpler or base model. A key strength of this framework is the flexibility in defining this base model and we take this advantage to actually be able to take prior information in the prior. Rather than being restricted to the identity matrix, our approach facilitates shrinkage towards a sequence of simpler, structured models derived by progressively removing parent nodes from the graph, or just to a base correlation matrix. This capability is particularly powerful for creating informative priors that robustly incorporate expert knowledge by penalizing deviations from a specified reference correlation structure. As a practical tool for multi-outcome analysis, the graphpcor package allows researchers to model complex and interpretable correlation structures with greater flexibility.

- Virginie Rondeau Virginie.Rondeau@inserm.fr

Title: Assessing surrogacy from joint modeling and mediation analysis when surrogates are either censored event times or longitudinal biomarker: cancer application

Authors: Quentin Lecoent, Catherine Legrand, Virginie Rondeau

Institution: Department of Biostatistics, Bordeaux Population Health Research Center, INSERM U1219, Université de Bordeaux, France

Abstract: The use of surrogate endpoints instead of final endpoints may be attractive for clinical trials. However, before a candidate surrogate can be used in a clinical trial, it must be statistically validated. Joint modeling is a powerful approach to better understand the treatment effect when time-to-event and longitudinal endpoints commonly co-occur in clinical trials.

We focus on the cases where the final endpoint is a time-to-event endpoint (such as time-to-death) and the surrogate is either a time-to-event or a longitudinal biomarker. Two new joint models and different approaches were proposed depending on the nature of the surrogate to validate surrogate endpoints. A mediation analysis can be proposed to decompose the total effect of the treatment on the final endpoint as a direct effect and an indirect effect through the surrogate. The ratio of the indirect effect over the total effect of the treatment on the final endpoint can be computed from the parameters of the model and used as a measure of surrogacy.

We present applications of survival mediation analysis using real oncology datasets, both with and without a meta-analytic structure. The proposed analyses investigate time-to-relapse as a surrogate for overall survival in gastric cancer, and tumor size as a surrogate biomarker for overall survival in colorectal cancer.

The R package **frailtypack**, available at https://CRAN.R-project.org/package=frailtypack, provides a user-friendly implementation of several methods for validating surrogate biomarkers or surrogate time-to-event endpoints.

We proposed a valuable tool to inform decision-making and advance our understanding of different treatment effects in mediation analysis with either a longitudinal mediator or a time-to-event mediator for a final survival endpoint.

- Janet van Niekerk Janet.vanNiekerk@up.ac.za

Title: INLAjoint and beyond: INLA for multi-outcome modeling.

Authors: Janet van Niekerk, Elias Teixeira Krainski, Haavard Rue, Denis Rustand.

Institution: University of Pretoria, South Africa.

Abstract: The Integrated Nested Laplace Approximation (INLA) methodology has been shown to produce accurate and efficient approximations for Bayesian inference of Generalized Additive MIxed type models, when compared to sampling-based approaches like MCMC. In this talk we show how INLA can also be used for full Bayesian inference of multi-outcome models through the INLAjoint user interface. INLAjoint facilitates the setup of the model and performs Bayesian inference using INLA.

Multiple examples are used to illustrate the impact of INLA and INLAjoint for Bayesian inference of multi-outcome models. In complex models or large data, INLA can perform Bayesian inference with speedups of 100X+ when compared to HMC and MCMC, with the same accuracy. This framework also enables the inclusion of random effects such as frailties, temporal and spatial effects, splines and more.

- Pedro Miranda-Afonso p.mirandaafonso@erasmusmc.nl

Title: The JMbayes2 R package for joint models of longitudinal and time-to-event data: From flexible fitting to dynamic prediction and validation
Authors: Dimitris Rizopoulos[1], Pedro Miranda-Afonso[1], Grigorios Papageorgiou[2]
Institution: [1]Department of Epidemiology and Biostatistics, Erasmus University Medical Center, Rotterdam, the Netherlands; [2]Statistics and Data Science Innovation Hub, GSK, the Netherlands
Abstract: Joint models for longitudinal and survival data are powerful frameworks for analyzing repeatedly measured biomarkers and clinical events that may have complex relationships. *JMbayes2* is an open-source R package that implements a broad class of extended joint models, covering multiple longitudinal outcomes (continuous, binary/categorical, count), competing risks, multi-state processes, recurrent events, interval censoring, and flexible association structures. It includes tools for generating dynamic predictions and evaluating predictive performance.
This talk presents an end-to-end workflow in JMbayes2 for both inference and prediction. On the inference side, we first demonstrate how to estimate and interpret marker–event associations using flexible links (e.g., current level, short-term change, cumulative history) and time-varying relationships. We then cover the process of conducting sensitivity analyses across alternative specifications. On the prediction side, we show how to compute personalized dynamic predictions at a chosen landmark time, using each subject's observed history to estimate their risk over user-defined horizons. Finally, we address how to measure the quality of these predictions, including their discrimination, calibration, and overall accuracy, and provide practical guidance for internal and external validation. Throughout, we focus on interpreting and validating models in clinically meaningful terms.
Concise, reproducible R code will be provided so that researchers can adapt the workflow to their own studies to obtain clinically interpretable marker–event associations and well-calibrated, discriminative dynamic predictions.

Session 19: Advanced Analytical Methods in Multi-outcome Data

Organizer: Lei Liu lei.liu@wustl.edu

- Grace Y. Yi gyi5@uwo.ca

Title: Function-on-Scalar Regression with Ultra high-Dimensional Error-Prone Covariates
Authors: Yifan Sun, Grace Y. Yi
Institution: University of Western Ontario
Abstract: In real-world applications, data are often contaminated by measurement error; naively applying conventional methods without accounting for these effects typically yields inconsistent estimates. Bias can be further exacerbated when covariates are ultrahigh-dimensional. Focusing on the widely used **function-on-scalar linear regression** model, I will discuss methods for **simultaneous parameter estimation and variable selection** when covariates are **error-prone** and ultrahigh-dimensional, with the framework accommodating **multiple measurement-error models**. I will present **asymptotic properties under identified regularity conditions.** In particular, I will highlight key differences between ultra high- and finite-dimensional settings and the distinct impacts of different error processes. **Numerical studies will be presented to** illustrate finite-sample performance of the proposed methods.

- Shalini Ramachandra shalini.ramachandra@pennmedicine.upenn.edu

Title: Time-varying frailty model for analyzing recurrent/terminal event data

Authors: Douglas Schaubel, Shalini Ramachandra and Eric Tchetgen Tchetgen
Institution: Division of Biostatistics, University of Pennsylvania Perelman School of Medicine
Abstract: We consider the data structure wherein the event of chief interest is recurrent (can occur multiple times for the same subject) and subject to a terminating event (e.g., death) which stops the recurrent event process. Shared frailty models have been employed several times in the literature as a means of analyzing recurrent/terminal event data. We propose a model which, unlike most predecessors, allows the frailty variable to differ in distribution during the follow-up period. Likelihood-based methods are proposed and evaluated through simulation. The methods are applied to a study hospitalization (recurrent) and graft failure (terminating) among liver transplant recipients from the University of Pennsylvania. This is joint wok with Shalini Ramachandra and Eric Tchetgen Tchetgen.

- Yi Li yili@umich.edu

Title: Inference for Deep Neural Network Estimators
Authors: Yi Li
Institution: University of Michigan
Abstract: While deep neural networks (DNNs) are used for prediction, inference on DNN-estimated subject-specific means for categorical or exponential family outcomes remains underexplored. We address this by proposing a DNN estimator under generalized nonparametric regression models (GNRMs) and developing a rigorous inference framework. Unlike existing approaches that assume independence between estimation errors and inputs to establish the error bound, a condition often violated in GNRMs, we allow for dependence and our theoretical analysis demonstrates the feasibility of drawing inference under GNRMs. To implement inference, we consider an Ensemble Subsampling Method (ESM) that leverages U-statistics and the Hoeffding decomposition to construct reliable confidence intervals for DNN estimates. We show that, under GNRM settings, ESM enables model-free variance estimation and accounts for heterogeneity among individuals in the population. Through simulations under nonparametric logistic, Poisson, and binomial regression models, we demonstrate the effectiveness and efficiency of our method. We further apply the method to the electronic Intensive Care Unit (eICU) dataset, a large-scale collection of anonymized health records from ICU patients, to predict ICU readmission risk and offer patient-centric insights for clinical decision-making.

Session 20: Causal Learning for Optimal Health Decision-Making with Complex Outcomes

Organizer: Yuanjia Wang yw2016@cumc.columbia.edu

- Erica Moodie erica.moodie@mcgill.ca

Title: Estimating the optimal allocation of kidneys in the presence of competing risks from clustered data
Authors: Misha Dolmatov[1], Erica E. M. Moodie[1], David A. Stephens[1], and Dipankar Bandyopadhyay[2]
Institution: [1]McGill University, [2]Virginia Commonwealth University
Abstract: The precision medicine framework has been used to discover tailored treatment strategies in a variety of settings and has largely been derived within a causal framework due to the need for large, typically observational, datasets. Existing methods are extended to address the question of how to optimally allocate kidneys for transplantation when the pool of donors includes individuals who were living with hepatitis C virus (HCV). The proposed approach accounts for the non-random allocation of kidneys from people with or without HCV, multiple (competing risks) endpoints, and clustering of data due to side effects. The proposed method is applied to data from the US National Organ Procurement and Transplant Network registry from 1994 to 2014.

- Ashkan Ertefaie Ashkan.Ertefaie@PennMedicine.upenn.edu

Title: A structural nested rate model for estimating the effects of time-varying exposure on recurrent event outcomes
Authors: Daniel Mork, Robert L. Strawderman, Michelle Audirac, Francesca Dominici, Ashkan Ertefaie

Institution: Harvard T.H. Chan School of Public Health, University of Rochester, University of Pennsylvania
Abstract: Assessing the causal effect of time-varying exposures on recurrent event processes is challenging in the presence of a terminating event. Our objective is to estimate both the short-term and delayed marginal causal effects of exposures on recurrent events while addressing the bias of a potentially correlated terminal event. Existing estimators based on marginal structural models and proportional rate models are unsuitable for estimating delayed marginal causal effects for many reasons, and furthermore, they do not account for competing risks associated with a terminating event. To address these limitations, we propose a class of semiparametric structural nested recurrent event models and two estimators of short-term and delayed marginal causal effects of exposures. We establish the asymptotic linearity of these two estimators under regularity conditions through the novel use of modern empirical process and semiparametric efficiency theory. We present the utility of our methods in the context of a large epidemiological study of 299,661 Medicare beneficiaries, where we estimate the effects of fine particulate matter air pollution on recurrent hospitalizations for cardiovascular disease.

- Yuan Chen cheny19@mskcc.org

Title: Learning Locally Interpretable Individualized Treatment Rules

Authors*:* Yasin Khadem Charvadeh, Katherine Panageas, Yuan Chen
Institution: Memorial Sloan Kettering Cancer Center

Abstract: Individualized treatment rules (ITRs) aim to optimize healthcare by tailoring therapy decisions to patient-specific characteristics to achieve precision medicine. Existing methods for deriving ITRs often rely on either regression-based models, which are interpretable but limited in flexibility, or black-box machine learning approaches, which capture complex treatment-response mechanisms but sacrifice interpretability. Furthermore, most current approaches impose a single, universal treatment rule, which may fail to capture nuanced, patient-specific variations in treatment effects. To address these challenges, we propose the Locally Interpretable Individualized Treatment Rule (LI-ITR) framework. LI-ITR constructs patient-specific treatment rules by combining flexible predictive models, such as neural networks, with locally interpretable surrogate models. Unlike existing post hoc interpretability methods, LI-ITR generates synthetic patient-specific samples using variational autoencoders to preserve correlations among features, ensuring realistic local neighborhoods. A mixture-of-linear-experts model with a gating network is then employed to select the most relevant local expert for each patient, improving robustness to sample variability and accommodating heterogeneous treatment-response structures. By integrating flexibility with transparency, LI-ITR provides treatment recommendations that are more accurate and interpretable.

- Cheng Zheng cheng.zheng@unmc.edu

Title: Joint Model for Mediation Analysis with Causally Related Longitudinal and Recurrent Event Mediators for Survival Outcome.

Abstract: Recurrent events and repeated measures are commonly encountered in clinical longitudinal studies, often holding strong associations with patient outcomes. Although joint models for repeated measures, recurrent events, and a terminal event have been developed to account for their correlation, limited methodologies exist to rigorously examine causal mediation mechanisms involving multiple types of mediators, especially when mediators are causally related. This study addresses this gap by proposing a novel causal mediation analysis framework to quantify natural direct and indirect effects when both recurrent events and repeated measures act as mediators with causal dependencies. We extend joint modeling approaches by incorporating shared random effects (frailties) structures, relaxing the commonly used ``sequential ignorability'' assumption, and accounting for unmeasured time-independent confounders through shared random effects. We apply our method to the Terry Beirn Community Programs for Clinical Research on AIDS (CPCRA) study and demonstrate that both recurrent opportunistic infections (OIs) and repeated CD4 count measurements mediate the effects of prior AIDS-defining conditions on survival outcomes. Additionally, the shared random effects between repeated CD4 and survival models highlight the presence of unmeasured confounding between CD4 counts and mortality. Simulation studies demonstrate the robustness and finite sample performance of our estimators for natural direct and indirect effects. The proposed methodology enables a more comprehensive investigation of causal

pathways in longitudinal studies with multiple mediators, providing insights into treatment mechanisms and informing clinical decision-making.

Session 21: Statistical and machine learning methods for complex biomedical data

Organizers: QI Long, qlong@pennmedicine.upenn.edu
and Kevin He, kevinhe@umich.edu

- Hongzhe Li hongzhe@upenn.edu

Title: A Robust Local Frechet Regression Using Unbalanced Neural Optimal Transport with Applications to Dynamic Single-cell Genomics Data
Authors: Binghao Yan and Hongzhe Li
Institution: University of Pennsylvania
Abstract: Single-cell RNA sequencing (scRNA-seq) technologies have enabled the profiling of gene expression for a collection of cells across time during a dynamic biological process. Given that each time point provides only a static snapshot, modeling and understanding the underlying cellular dynamics remains a central yet challenging task in modern genomics. To associate biological time with single cell distributions, we develop UOTReg, a robust local Fr\'echet regression for estimating the high-dimensional cellular distribution at any given time point using data observed over a finite time points. To allow for robustness in cell distributions, we propose to apply the unbalanced optimal transport-based Wasserstein distance in our local Fr\'echet regression analysis. We develop a computationally efficient algorithm to generate the cell distribution for a given time point using generative neural networks. The resulting single cell regression model and the corresponding transport plans facilitate the estimation of the cell distributions at any unobserved time point and the estimation of the cell trajectory during the cell differentiation process. We apply UOTReg to two single cell differentiation data sets, including differentiation of human embryonic stem cells into embryoids and mouse hematopoietic stem and progenitor cell differentiation. We show that UOTReg leads to better single cell distribution estimates, reveals different cell differential trajectories, and identifies early genes that regulate these cell trajectories.

- Sandra Safo ssafo@umn.edu

Title: iDeepViewLearn+: Interpretable AI for Robust and Fair Multimodal Data Integration
Authors: Sinian Zhang, Hengkang Wang, Ju Sun, Sandra E. Safo
Institution: University of Minnesota
Abstract: Complex diseases such as heart failure (HF), coronary artery disease (CAD), and hypertension (HTN) often exhibit significant class imbalance, making accurate prediction difficult. Deep learning models, while powerful, can lack fairness across sensitive attributes and struggle to generalize across multimodal data with differing statistical distributions. Their limited interpretability further restricts their clinical utility. Addressing these challenges is essential for building robust, equitable, and interpretable models for biomedical data integration. We propose iDeepViewLearn+, an interpretable deep learning framework for multimodal data integration that learns nonlinear relationships across heterogeneous data types while enabling biologically meaningful feature selection. The model incorporates tailored loss function to improve predictive performance across imbalanced classes and promote equitable representation. It accommodates diverse statistical distributions across views and leverages regularization penalties on reconstructed data to guide variable selection. We applied iDeepViewLearn+ to simulated and molecular data from the COPDGene study to identify molecular signatures predictive of HF, CAD, and HTN. The framework demonstrated competitive performance in classification, fairness, robustness, and variable selection, even under small-sample-size conditions. The selected features uncovered biologically relevant patterns associated with cardiovascular diseases, supporting the model's interpretability. These results highlight the potential of iDeepViewLearn+ for robust, fair, and interpretable integration of multimodal biomedical data in complex disease settings.

- Nicholas Henderson nchender@umich.edu

Title: Ranking-based Data Integration for Robust Outcomes Prediction and Biomarker Detection

Authors : Nicholas C. Henderson
Institution: University of Michigan, Department of Biostatistics
Abstract : Utilizing established risk factors and prognostic models can often improve the construction of a newer risk model that uses novel biomarkers in a smaller, internal study. However, directly borrowing information from an established prognostic model is often unsuitable due to differences in study populations, patient outcomes measured, and other specific features of the internal study design. To better enable the use of established prognostic information when constructing a novel risk model, we propose an estimation approach centered around the idea that the risk rankings rather than the risk scores from an established prognostic model are often more transportable to the internal study context. To leverage external ranking information, our approach introduces the ranking parameters associated with the regression coefficients of an internal risk model and estimates the internal risk model parameters by penalizing a ranking-based discrepancy measure between the ranking parameters and the rankings implied by the established prognostic model. We also describe how this framework can be used to harness patients with large discrepancies in risk rankings to better identify predictive biomarkers. We demonstrate the use of our approach through the development of a prognostic model for advanced prostate cancer patients treated with an immune checkpoint inhibitor.

Session 22: Causal inference in presence of truncation by death

Organizer: Cecile Proust-Lima cecile.proust-lima@u-bordeaux.fr

- Daniel Nevo danielnevo@gmail.com

Title: Causal effects on non-terminal event time with application to antibiotic usage and future resistance
Authors: Tamir Zehavi, Uri Obolski, Michal Chowers and Daniel Nevo
Institution: Tel Aviv University
Abstract: Comparing future antibiotic resistance levels resulting from different antibiotic treatments is challenging because some patients may survive only under one of the antibiotic treatments. We embed this problem within a semi-competing risks approach to study the causal effect on resistant infection, treated as a non-terminal event time, while also being aware of the other outcome – death. We argue that existing principal stratification estimands for such problems exclude patients for whom a causal effect is well-defined and is of clinical interest. Therefore, we present a new principal stratum, the infected-or-survivors (ios). The ios is the subpopulation of patients who would have survived or been infected under both antibiotic treatments. This subpopulation is more inclusive than previously defined subpopulations. We target the causal effect among these patients, which we term the feasible-infection causal effect (FICE). We develop large-sample bounds under novel assumptions, and discuss the plausibility of these assumptions in our application. As an alternative, we derive FICE identification using two illness-death models with a bivariate frailty random variable. These two models are connected by a cross-world correlation parameter. Estimation is performed EM algorithm followed by a Monte Carlo procedure. We apply our methods to detailed clinical data obtained from a hospital setting.

- Catherine Legrand catherine.legrand@uclouvain.be

Title: Bias-Corrected Two-Stage Modeling for Joint Analysis of Multidimensional HRQoL and Survival
Authors*:* Catherine Legrand, Hortense Doms, Philippe Lambert
Institution: LIDAM/ISBA, UCLouvain, Louvain-la-Neuve, Belgium
Abstract: Health-related quality-of-life (HRQoL) outcomes are increasingly incorporated into oncology research to capture patients' well-being beyond traditional survival endpoints. However, the joint analysis of multivariate ordinal HRQoL data and time-to-event outcomes remains challenging, particularly when multiple longitudinal outcomes and complex random-effects structures are involved. We propose a slope-corrected two-stage approach for multidimensional latent trait joint models, in which information from the longitudinal model is propagated through informative priors for random effects during the survival stage. This strategy improves estimation accuracy while retaining the computational advantages of two-stage procedures. Simulation studies show that the method substantially reduces bias in both longitudinal and survival parameters and closely approaches the performance of fully joint Bayesian estimation, with markedly reduced computation time. We further illustrate the method using HRQoL data from patients with progressive glioblastoma, demonstrating its practical utility. The

proposed approach provides a robust and practical solution for analysing complex multidimensional ordinal patient-reported outcomes jointly with a terminal event.

- Arce Domingo-Relloso ad3531@cumc.columbia.edu

Title: A path-specific effect approach to mediation analysis with time-varying mediators and time-to-event outcomes accounting for competing risks

Authors: Arce Domingo-Relloso, Yuchen Zhang, Ziqing Wang, Astrid M Suchy-Dicey, Dedra S Buchwald, Ana Navas-Acien, Joel Schwartz, Kiros Berhane, Brent A. Coull, Linda Valeri
Institution: IE University School of Science and Technology

Abstract: Not accounting for competing events in survival analysis can lead to biased estimates, as individuals who die from other causes do not have the opportunity to develop the event of interest. Formal definitions and considerations for causal effects in the presence of competing risks have been published, but not for the mediation analysis setting. We propose, for the first time, an approach based on the path-specific effects framework to account for competing risks in longitudinal mediation analysis with time-to-event outcomes. We do so by considering the pathway through the competing event as another mediator, which is nested within our longitudinal mediator of interest. We provide a theoretical formulation and related definitions of the effects of interest based on the mediational g-formula, as well as a detailed description of the algorithm. We also present a simulation study and an application of our algorithm to data from the Multi-Ethnic Study of Atherosclerosis. In this application, we evaluated the mediating role of coronary artery calcification (measured in five visits) on the association between arsenic and cadmium, with time to cardiovascular disease, accounting for competing risks by death. Identifying the effects through different paths enables us to evaluate the impact of metals on the outcome of interest, as well as through competing risks, more transparently.

- Stijn Vansteelandt stijn.vansteelandt@ugent.be

Title: Evaluating Treatment Effects in Longitudinal Trials with Truncation by Death: An Assumption-Lean Approach

Authors: Georgi Baklicharov, Kelly Van Lancker & Stijn Vansteelandt
Institution: Ghent University, Belgium

Abstract: Intercurrent events such as treatment switching, rescue medication, or death often complicate the interpretation of intention-to-treat (ITT) analyses in randomized clinical trials. Although causal inference frameworks propose alternative estimands—such as hypothetical or principal-stratum estimands—to address these issues, their identification typically relies on strong, unverifiable assumptions about counterfactual outcomes and time-varying confounding. Moreover, the strict nature of clinical trial protocols can induce near-violations of the positivity assumption, further limiting identification and interpretability.

We introduce a novel approach that enables testing of treatment effects in longitudinal randomized trials without requiring such strong assumptions. The key idea is to compare treated and untreated participants matched on baseline covariates at their **pairwise last observation time**—the most recent time point before either member of a matched pair experiences an intercurrent event. We term the resulting contrasts *Pairwise Last Observation Time (PLOT) estimands*.

PLOT estimands can be identified directly from randomized data without modeling post-baseline treatment or survival processes, thereby sidestepping untestable structural and positivity assumptions. While they can exhibit a mild form of residual selection bias, we show analytically that this bias vanishes under the assumptions typically required by existing causal estimands, and simulation studies demonstrate that it is generally negligible in realistic settings.

Building on this identification strategy, we construct asymptotically efficient, model-free test statistics using data-adaptive estimators of nuisance parameters. Through extensive simulations, we show that the proposed tests maintain nominal type-I error and achieve competitive power even when traditional causal estimands are non-identifiable. We illustrate the method by re-analyzing data from a recent diabetes trial affected by truncation by death, demonstrating how PLOT-based inference provides robust evidence of treatment effects under minimal assumptions.

Session 23: Innovative Statistical Methods for Longitudinal and Survival Analysis in Medical Studies

Organizer: Yanqing Sun yasun@charlotte.edu

- Lang Wu lang@stat.ubc.ca

Title: Joint models in longitudinal studies for efficient and robust inferences
Authors: Lang Wu
Institution: University of British Columbia
Abstract: In many longitudinal studies, the within individual repeated measurements often exhibit large variations and these variations appear to change over time. A good understanding of the nature of the within-individual systematic and random variations allows us to conduct more efficient statistical inference. We consider a nonlinear mixed effects (NLME) model for modelling the longitudinal means, together with a model for the within-individual variances which also allows us to address outliers in the repeated measurements. In addition, we consider joint models for longitudinal data collected at different phases of the studies. Statistical inference is based on the joint likelihood for all models. The proposed methods are applied to a recent AIDS dataset, with interesting new findings.

- Jimin Ding jmding@wustl.edu

Title: Enhanced survival trees with applications to Alzheimer's disease study
Authors: Jimin Ding
Institution: Washington University at St. Louis
Abstract: Motivated by the challenges of risk stratification in Alzheimer's disease (AD) research, we introduce SurvTreeFuL, a new survival tree framework for censored failure time data that integrates methodological innovation with clinical relevance. The proposed method incorporates three key advancements over traditional survival tree approaches. First, we develop a computationally efficient splitting strategy that effectively mitigates the end-cut preference problem and propose an intersected validation scheme to reduce the variable selection bias inherent in greedy searches. Second, we present a novel framework for determining tree structures through fused regularization, which, in combination with conventional pruning, enables the merging of non-adjacent terminal nodes to produce more parsimonious and interpretable models. Third, we address the inference challenge by constructing valid confidence intervals for hazard ratios that compare the subgroups identified by the final tree, using bootstrap-based bias correction to improve standard error estimation. Extensive simulation studies demonstrate the robustness and reliability of the proposed framework, and its application to the Alzheimer's Disease Neuroimaging Initiative (ADNI) data illustrates how SurvTreeFuL facilitates clinically meaningful risk stratification among individuals with mild cognitive impairment (MCI) as they progress toward AD.

- Wenqing He whe@stats.uwo.ca

Title: Boosting methods for interval censored data with regression and classification
Authors: Yuan Bian, Grace Y. Yi and Wenqing He
Institution: University of Western Ontario
Abstract: Boosting has gained significant interest across both machine learning and statistical communities. Traditional boosting algorithms, designed for fully observed random samples, often struggle with real-world problems, particularly with interval censored data. This type of data is common in survival analysis where exact event times are unobserved but fall within known intervals. Effective handling of such data is crucial in fields like medical research, reliability engineering, and social sciences. In this work, we introduce novel nonparametric boosting methods for regression and classification tasks with interval censored data. Our approaches leverage censoring unbiased transformations to adjust loss functions and impute transformed responses while maintaining model accuracy. Implemented via functional gradient descent, these methods ensure scalability and adaptability. We rigorously establish their theoretical properties. Our proposed methods not only offer a robust framework for enhancing predictive accuracy in domains where interval censored data are common but also complement existing work, expanding the applicability of existing boosting techniques. Empirical studies demonstrate robust performance across various finite-sample scenarios, highlighting the practical utility of our approaches.

- Yanqing Sun yasun@charlotte.edu

Title: Estimation and Inference of Semiparametric Temporal Intensity Models for Recurrent Events with Application to a Malaria Vaccine Trial
Authors: Jing Xu, Yanqing Sun, Fei Heng, Peter B. Gilbert
Abstract: Motivated by a malaria vaccine efficacy trial, we propose a class of generalized semiparametric temporal intensity models that simultaneously capture both constant and time-varying effects across two distinct time scales. These models provide flexibility in incorporating time-dependent effects and event history while allowing for various choices of link functions and parametric forms. Specifically, our framework includes both semiparametric multiplicative and additive temporal intensity models. To estimate the effects across the two-time scales, we develop a profile maximum likelihood estimation procedure using local linear kernel smoothing. We introduce a cross-validation bandwidth selection by using the logarithm of likelihood as a measure of prediction accuracy. The asymptotic properties of the estimators are derived using the martingale theory and empirical process techniques. Additionally, we develop hypothesis testing procedures to assess model specifications, thereby providing practical tools for constructing more efficient models. Through simulation studies, we demonstrate the effectiveness of our estimation and testing procedures for both multiplicative and additive intensity models, showing strong finite-sample performance. Finally, we apply the proposed methods to a data set from the MAL-094/MAL-095 malaria vaccine efficacy trial to examine how malaria infection risk evolves over time, how recency of malaria infection impacts future infection risk, as well as protective effects of the RTS,S/$AS01_E$ malaria vaccine at different doses.

Session 24: New Estimation and Inference Methods for Analyzing Composite Survival Outcomes

Organizer: TBD

- Chen Hu huc@jhu.edu

Title: On statistical inference of multiple competing risks in oncology clinical trials
Authors : Chen Hu
Institution: Division of Quantitative Sciences, Department of Oncology, Johns Hopkins University School of Medicine; Baltimore, MD
Abstract: Competing risks data are commonly encountered in comparative clinical trials. Ignoring competing risks in survival analysis leads to biased risk estimates and improper conclusions. Often, one of the competing events is of primary interest and the rest competing events are handled as nuisances. These approaches can be inadequate when multiple competing events have important clinical interpretations and thus of equal interest. For example, in hospitalized critical care treatment trials, the outcomes are either death or discharge from hospital, which have completely different clinical implications and are of equal interest. In oncology trials, while composite endpoints, such as disease-free survival, are used frequently, it is often concerned that novel interventions do not necessarily impact all components of a composite endpoint equally. We develop nonparamteric estimation and simultaneous inferential methods for multiple cumulative incidence functions (CIFs) and corresponding restricted mean times. Based on Monte Carlo simulations and a data analysis of a completed clinical trial, we demonstrate that the proposed method provides global insights of the treatment effects across multiple endpoints. This is a joint work with Dr. Mei-Cheng Wang and Dr. Jiyang Wen.

- Yu Cheng yucheng@pitt.edu

Title: Win odds for Sequential Multiple Assignment Randomized Trials
Authors : Zi Wang, Jiazhi He, Abdus Wahed, and Yu Cheng
Institution: University of Pittsburgh, Departments of Statistics and Biostatistics
Abstract: Sequential multiple-assignment randomized trials (SMARTs) are designed to construct and evaluate dynamic treatment regimes (DTRs)—sequential treatment decisions tailored to a patient's evolving status across multiple stages of care. In a two-stage SMART, patients are randomized twice: first among initial treatments, and subsequently among second-stage options conditional on response to the initial treatment, resulting in multiple embedded DTRs. In biomedical settings, patients often experience multiple possibly competing events (e.g., an adverse event occurring before a therapeutic response), complicating personalized decision-making. We adopt the win odds (WO) to assess the

comparative effectiveness of one DTR versus another. WO and related win statistics have gained traction as intuitive metrics that prioritize clinically important outcomes and can aggregate multiple endpoints; they are typically estimated from win/loss proportions in matched or unmatched pairs. However, existing methods do not directly extend to SMARTs because embedded DTRs can share patient paths, inducing dependence. To address this, we use inverse probability weighting to handle systematic missingness arising from subsequent randomizations and inverse probability of censoring weighting to address indeterminacy due to censoring. We consider two inference strategies: (i) deriving the WO variance via U-statistic theory and the delta method and (ii) a hull-based confidence approach. Both methods are shown to perform well in numerical studies.

- Ariane Bercu ariane.bercu@u-bordeaux.fr

Title : Variable Selection for Time-Varying Covariates in Multistate Survival Models with Interval Censoring.
Authors :
Institution:
Abstract:

- Yingye Zheng yzheng@fredhutch.org

Title: Joint Modeling of True and Misclassified Event Times for Biomarker Evaluation

Authors: Yingye Zheng, Xuan Wang, Tianxi Cai
Institution: Department of Biostatistics, Fred Hutchinson Cancer Center, Seattle WA 98109
Department of Population Health Sciences, University of Utah
Biomedical Informatics, Harvard Medical School

Abstract: Outcome misclassification in time-to-event analyses can substantially attenuate biomarker effect estimates and distort predictive accuracy. Existing approaches for handling misclassified outcomes in survival analysis typically assume that the sensitivity and specificity of the imperfect outcome are known or specified in advance, an assumption that is rarely satisfied in practice. In many clinical settings, procedures treated as gold standards may themselves be error-prone. For example, in prostate cancer active surveillance, disease progression is commonly defined by surveillance biopsy, which is subject to sampling error and inter-reader variability.
Motivated by this setting, we propose a likelihood-based framework for evaluating biomarkers using panel current status data when event times are subject to misclassification. The approach jointly models the distribution of the true event time and the observed misclassified time through a discrete proportional hazards model coupled with a probabilistic misclassification mechanism, allowing simultaneous estimation of biomarker effects, baseline hazards, and misclassification parameters without requiring known sensitivity or specificity. Using the estimated survival function, we further construct estimators for time-dependent predictive accuracy measures, including ROC curves and the area under the ROC curve. Simulation studies demonstrate good finite-sample performance of the proposed estimators. The effectiveness of the method is demonstrated in a biomarker study in prostate cancer surveillance.

Session 25: Machine Learning and Survival Analysis

Organizer: Jiaoguo Sun sunj@missouri.edu

- Yu Gu yugu@hku.hk

Title: Semiparametric Functional Multi-State Models: Estimation and Testing with Application to Alzheimer's Disease
Authors: Chenrui Qi[1], Yu Gu[1], Kai Kang[2]
Institution: [1]The University of Hong Kong, [2]Sun Yat-sen University
Abstract: Recent advances in brain imaging have greatly enhanced early and accurate prediction of Alzheimer's disease (AD) risk. However, existing imaging-based methods focus solely on the progression from mild cognitive impairment (MCI) to AD, and improperly assume that the exact time of AD conversion is known or right-censored. In this work, we study, for the first time, the entire trajectory of AD progression, which spans transitions from cognitively normal to MCI to AD. We model this disease trajectory as a multi-state process under intermittent observation, and formulate the effects of imaging and clinical covariates via functional mixed proportional intensity models, with imaging data as a functional covariate. We integrate functional principal component analysis (FPCA) with nonparametric maximum likelihood estimation and develop a stable EM algorithm for efficient computation. Moreover, we propose a profile score test to assess the association between the functional covariate and the multi-state process. We establish the asymptotic properties of the proposed estimators and test statistic through novel use of theories of FPCA, empirical processes, and semiparametric efficiency. Simulation studies and application to the Alzheimer's Disease Neuroimaging Initiative demonstrate the satisfactory performance of our proposed methods.

- Jiaqi Zhao jiaqi2025.zhao@connect.polyu.hk

Title: Spline Learnable Neural Networks with Applications to Generative Diffusion Models
Authors: Chenghao Li and Yuanyuan Lin:
Institution: The Chinese University of Hong Kong
Abstract: In this paper, we introduce a data-augmented nonparametric noise contrastive estimation method to density estimation using deep neural networks. By leveraging the idea of contrastive learning, our density estimator exhibits efficiency with a one-step and simulation-free evaluation process, imposes no constraints on the neural network, and is shown to be consistent and asymptotically automatically normalized. A novel data augmentation procedure allows us to mitigate the influence of the choice of reference distribution on our method, as well as to avoid the strong density assumption on the target distribution. Non-asymptotic upper bounds for the expected L2-risk and the expected total variation distance have been established, which achieve minimax optimal rates. Moreover, our new method exhibits inherent adaptivity to low-dimensional structures of data with a faster convergence rate under a compositional structure assumption. Numerical experiments show the competitiveness of our new method compared with the state-of-the-art nonparametric density estimation methods.

- Li Liu lliu.math@whu.edu.cn

Title: Automatic structure identification and variable selection for additive accelerated failure time model with ultra high dimensional covariates
Authors: Li Liu, Jiaxiang Chen, Wen Su, Xingqiu Zhao
Institution: Wuhan University
Abstract: We propose a novel approach called doubly penalized regularization to address the challenges of variable selection, model structure identification, and parameter estimation in ultra-high dimensional additive accelerated failure time (AFT) models. In this method, we introduce penalties on a weighted sieve least squares loss function. Our method allows for an exponential increase in the dimension of covariates as the sample size grows. We establish the asymptotic properties of the estimators, including the identifiability of the model structure and the oracle property for variable selection. To mitigate computational complexity, we utilize the blockwise majorization descent algorithm. This algorithm is implemented in an accessible and user-friendly R package called AFTBMD, which we have made publicly available. Simulation studies are conducted to evaluate the performance of our proposed method. Additionally, we apply our method to analyze the breast cancer data, further demonstrating its effectiveness and practicality.

- Rujia Cao rujia.cao@connect.polyu.hk

Title: Deep Semiparametric and Nonparametric Inference for Longitudinal Data
Authors*:* Rujia Cao, Shirong Deng and Xingqiu Zhao
Institution: The Hong Kong Polytechnic University

Abstract: While deep learning has achieved remarkable success across diverse fields, its use for longitudinal data analysis remains underexplored. To address this gap, we propose a flexible semiparametric modeling framework in longitudinal settings, and then develop a deep neural network (DNN) estimation procedure, which can simultaneously estimate linear coefficients and a nonlinear multivariate function. We establish nonasymptotic and asymptotic properties of the resulting estimators, including nonasymptotic error bound, convergence rate, and asymptotic normality, via empirical process techniques combined with recent theory for DNN. We further develop a general goodness-of-fit testing procedure that integrates the DNN-based estimator with data splitting and provide theoretical guarantees. The performance of the estimation and testing procedures is demonstrated through simulations, and the methodology is illustrated with a longitudinal yeast cell cycle gene expression dataset.

Session 26: Area of causal inference

Organizer: Cheng Zheng cheng.zheng@unmc.edu

- Ying Zhang ying.zhang@unmc.edu

Title: Joint Analysis of Multivariate Longitudinal Data and Interval-Censored Event Time Data with Application to Huntington's Disease Progression
Authors: Yue Zhan, Cheng Zheng, Ying Zhang
Institution: University of Nebraska Medical Center
Abstract: Huntington's disease is an autosomal dominant neurodegenerative disorder, characterized by motor dysfunction, psychiatric disturbances, and cognitive decline. The onset of Huntington's disease is diagnosed by severe motor impairment, which can be predicted by cognitive decline and may also worsen cognitive impairment. However, clinical data are often collected at discrete times, and therefore, disease onset is subject to interval censoring. We develop a joint model of multivariate longitudinal biomarkers with a change point anchored at an interval-censored event time. Our model allows us to simultaneously study the effect of longitudinal biomarkers on the event time and the changes in trajectories of the longitudinal biomarkers post the event. We conduct a comprehensive simulation study to demonstrate the satisfactory finite-sample performance of the proposed method for making inferences. We applied the method to the PREDICT-HD data from a multisite observational cohort study of prodromal Huntington's disease individuals to ascertain how cognitive impairment and motor dysfunction interact during disease progression.

- Ran Dai ran.dai@unmc.edu

Title: Model-free High Dimensional Mediator Selection with False Discovery Rate Control
Authors: Runqiu Wang, Ran Dai, Jieqiong Wang, Kah Meng Soh, Ziyang Xu, Mohamed Azzam, Hongying Dai, and Cheng Zheng
Institution: University of Nebraska Medical Center
Abstract: There is a challenge in selecting high-dimensional mediators when the mediators have complex correlation structures and interactions. In this work, we frame the high-dimensional mediator selection problem into a series of hypothesis tests with composite nulls, and develop a method to control the false discovery rate (FDR) which has mild assumptions on the mediation model. We show the theoretical guarantee that the proposed method and algorithm achieve FDR control. We present extensive simulation results to demonstrate the power and finite sample performance compared with existing methods. Lastly, we demonstrate the method for analyzing the Alzheimer's Disease Neuroimaging Initiative (ADNI) data, in which the proposed method selects the volume of the hippocampus and amygdala, as well as some other important MRI-derived measures as mediators for the relationship between gender and dementia progression.

- Mladen Kolar mkolar@marshall.usc.edu

Title: Confidence Sets for Causal Orderings
Authors: Y. Samuel Wang, Mladen Kolar, Mathias Drton
Institution: USC & MBZUAI
Abstract: Understanding causal relationships among variables is critical for effective decision-making in diverse fields such as economics, healthcare, social sciences, and business analytics. While numerous

methods for causal discovery have been proposed, existing approaches typically yield single-point estimates or equivalence classes, often neglecting the quantification of uncertainty. In practice, however, it is crucial to understand the reliability of inferred causal structures, especially in high-stakes scenarios where incorrect causal inferences can lead to misguided interventions.
In this talk, we introduce a novel approach to quantify uncertainty in causal discovery by constructing confidence sets for causal orderings. Recognizing that establishing a causal ordering among variables is fundamental yet challenging, our methodology focuses on generating sets of causal orderings that remain plausible given observed data. Specifically, we consider identifiable structural equation models with additive errors and employ a residual bootstrap procedure designed to assess the goodness-of-fit of candidate causal orderings. Our proposed method provides asymptotically valid confidence sets, enabling practitioners to explicitly incorporate model uncertainty into their causal inference process.
We demonstrate how these confidence sets not only allow researchers to identify ancestral relationships with specified confidence but also facilitate constructing confidence intervals for causal effects that inherently reflect underlying model uncertainty.
Joint work with Y. Samuel Wang and Mathias Drton. https://arxiv.org/abs/2305.14506

- Youngjoo Cho yvc5154@konkuk.ac.kr

Title: Doubly robust inference for the heterogeneous treatment effect of multiple treatments on time-to-event outcomes

Authors: Heejun Shin[1], Youngjoo Cho[2]
Institution: [1]Harvard T.H. Chan School of Public Health, [2]Konkuk University

Abstract: Doubly robust (DR) estimators are widely used in causal inference for estimating treatment effects, providing consistent estimates when either the outcome model or the propensity score model is correctly specified. While the estimational performance of DR estimators has been extensively studied, their inferential performance---especially in survival analysis---remains underexplored. In particular, their performance can be limited when one of the models is misspecified or when working with finite samples. In this paper, we propose a novel extension of DR estimators that achieves doubly robust inference, ensuring valid confidence intervals with nominal coverage even in finite samples and under model misspecification for survival analysis with multiple treatment levels. We provide theoretical justification for the proposed method and demonstrate its performance through simulation studies, showing that it offers reliable inference in challenging settings.

Session 27: Joint analysis of time-to-event or longitudinal markers in the presence of mortality as a competing risk

Organizer: Sebastien HANEUSE shaneuse@hsph.harvard.edu

- Eleni-Rosalina Andrinopoulou. eleni.r.andrinopoulou@gsk.com

Title: Challenges and Opportunities in Joint Modelling of Multiple Longitudinal and Survival Outcomes

Authors: Eleni-Rosalina Andrinopoulou
Institution: Statistics and Data Science Innovation Hub, GSK

Abstract: Many clinical datasets include multiple longitudinal biomarkers that are repeatedly measured over time, alongside survival outcomes such as time to intervention, infection, dropout, or death. These longitudinal markers often reflect related biological processes and may be interdependent. For example, loss to follow-up due to death can be influenced by disease progression or other clinical factors, making it important to consider these outcomes jointly. Joint modelling approaches have gained significant attention in recent years due to their ability to simultaneously analyse longitudinal and time-to-event data, offering improved inference and predictive performance across a wide range of applications. Examples of such models include the analysis of cardiovascular data, where echocardiographic markers of aortic valve function are studied alongside the risk of intervention and death, or in chronic respiratory diseases such as Chronic Obstructive Pulmonary Disease (COPD) and Cystic Fibrosis (CF), where biomarkers of lung function decline are analysed together with the risk of exacerbation and death. In this presentation, we discuss the challenges and opportunities in the analysis of multiple longitudinal and survival outcomes, with a particular focus on settings where death plays a central role in the modelling process.

- Daniels, Michael J. mdaniels@stat.ufl.edu

Title: A Bayesian Nonparametric Approach for Semi-Competing Risks with Application to Cardiovascular Health
Authors: Gelis-Cadena, Karina, Daniels, Michael J., and Siddique, J.
Institution: University of Florida, University of Florida, Northwestern University
Abstract: We address causal estimation in semi-competing risks settings, where a non-terminal event may be precluded by one or more terminal events. We define a principal-stratification causal estimand for treatment effects on the non-terminal event, conditional on surviving past a specified landmark time. To estimate joint event-time distributions, we use both vine-copula constructions and Bayesian nonparametric enriched Dirichlet-process mixtures (EDPM), enabling inference under minimal parametric assumptions. We index our causal assumptions with sensitivity parameters. We illustrate the proposed method using data from a cardiovascular health study.

- Sebastien Haneuse shaneuse@hsph.harvard.edu

Title: A novel framework for Joint prediction of longitudinal markers and death
Authors: Stephanie Armbruster, Harrison Reeder, Sebastien Haneuse
Institution: Harvard T.H. Chan School of Public Health
Abstract: Clinical decision-making typically involves balancing the prognosis across multiple markers and endpoints, as well as mortality. And yet most risk prediction tools are developed for univariate outcomes, implicitly treating other endpoints at-best as inputs/features. In this work we propose a novel discrete-time framework for jointly modeling one or more longitudinal marker and a terminal event (e.g. mortality) when the underlying goal is prediction. Crucial to the framework is that discretization eliminates implicit extrapolation beyond truncation by the terminal event. The framework accommodates global time-invariant and local time-dependent dependencies between the terminal event and longitudinal marker and captures latent associations via a shared frailty term. The longitudinal submodel incorporates flexible regression structures, including autoregressive patterns. Estimation and inference for model components proceeds through a Bayesian approach, enabling uncertainty quantification while accommodating left truncation, right censoring, and truncation by the terminal event. We obtain posterior predictions of cumulative incidence functions for the terminal event and of future longitudinal trajectories, along with credibility intervals, and benchmark these against oracle predictions. We illustrate the intuitive interpretation of the framework through visualizations. We apply the framework to a study of sudden death and implantable cardioverter defibrillator therapy to emphasize its clinical applicability.

- Sangita Kulathinal sangita.kulathinal@helsinki.fi

Title: Nonparametric estimation of the joint and conditional survival functions of the time to an event of interest and associated integrated covariate processes
Authors: Joshi A, Gasbarra D, Kulathinal S**.**
Institution: Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
Abstract:
During the treatment of chronic diseases, a covariate process can exhibit a response over the treatment period and is associated with the event of interest. Our main aim in this paper is to estimate the joint survival function of the time to event of interest and the integrated covariate process. In real-world settings, the time to event is often censored. Under the assumption that the censoring is independent of the time to event and the covariate process, we propose inverse probability of censoring weighted estimators of the joint survival function. The estimators and their efficiency differ according to the handling of the censoring distribution. For clinical decision making, we study the conditional survival function of the time to event given the history of the covariate process. We illustrate our methodology using two sets of data on age-related macular degeneration of eye; one from a clinical trial and the other from a real-world study.

Session 28: Nonparametric and semiparametric modeling for the analysis of complex data

Organizer: Zhezhen Jin zj7@cumc.columbia.edu
Chairs : Ariane Bercu ariane.bercu@u-bordeaux.fr

- Shanshan Ding sding@udel.edu

Title: Robust causal effect estimation in high dimensional survival analysis using sufficient dimension reduction
Authors: Shanshan Ding and Zhezhen Jin
Institution: University of Delaware, Columbia University
Abstract: Causal inference is an important tool to make inferences about causal relationships between variables. In this talk, we develop a robust nonparametric framework for estimating causal treatment effect in high dimensional survival analysis. The proposed framework is very flexible and alleviates assumptions in existing causal survival analysis. It allows the covariate dimension to grow exponentially fast with the sample size, without imposing stringent model or distributional assumptions for estimating the causal effect. The effectiveness of the methods will be demonstrated through both theoretical and numerical studies.

- Xuewen Lu xlu@ucalgary.ca

Title: Variable Selection in the Joint Frailty Model of Recurrent and Terminal Events Using Broken Adaptive Ridge Regression.
Authors: Christian Chan, Fatemeh Mahmoudi, Chel Hee Lee, Quan Long, and Xuewen Lu
Institution: University of Calgary
Abstract: We introduce a novel method to simultaneously perform variable selection and estimation in the joint frailty model of recurrent and terminal events using the Broken Adaptive Ridge Regression penalty. The BAR penalty can be summarized as an iteratively reweighted squared L2-penalized regression, which approximates the L0-regularization method. Our method allows for the number of covariates to diverge with the sample size. Under certain regularity conditions, we prove that the BAR estimator implemented under the model framework is consistent and asymptotically normally distributed, which are known as the oracle properties in the variable selection literature. In our simulation studies, we compare our proposed method to the Minimum Information Criterion (MIC) method and other competitors. We apply our method on the Medical Information Mart for Intensive Care (MIMIC-III) database, with the aim of investigating which variables affect the risks of repeated hospitalization and death during hospitalization.

- Leilei Zeng lzeng@uwaterloo.ca

Title: Modeling Heterogeneous Disease Processes Using a Mixture of Multistate Hidden Markov Models
Authors: Leilei Zeng, Yidan Shi and Mary E. Thompson
Institution: University of Waterloo
Abstract: Multistate models are widely used for analyzing life history data on disease onset, progression and death over time. Many diseases manifest differently and what appears to be a coherent collection of symptoms is often the expression of a variety of distinct disease subtypes, each with a different rate of onset of symptoms and progression. We propose a mixture of multistate hidden Markov model (MHMM), where the underlying process is characterized by a finite mixture of multiple Markov chains, one for each disease subtype, while the observation process contains states corresponding to the common symptomatic stages of these diseases. Information on type of disease is partially available and reflects the pathway through certain hidden states in the corresponding disease process, facilitating the estimation of parameters involved in the proposed models. The method is demonstrated on a dataset to model the development and progression of dementia caused by Alzheimer's disease and non-AD dementia.

- Hua Shen hua.shen@ucalgary.ca

Title: Integrative Analysis of Multiple Data Sources Involving Misclassification and Missing Data
Authors: Zheng Yu and Hua Shen
Institution: University of Calgary
Abstract: Integrating data from multiple sources is essential for population inference but often challenged by missing and misclassified data. We propose a likelihood-based method that addresses missing outcomes and misclassified covariates in probability samples, and dual

misclassification in nonprobability samples. Simulation studies show substantial bias reduction over naive methods, and a real-world application illustrates the method's practical advantages for improving population estimates.

Session 29: Machine or deep learning for multiple longitudinal and survival data

Organizer: Virginie Rondeau virginie.rondeau@inserm.fr

- Manel RAKEZ rakez.manel@outlook.fr

Title: Joint models and deep learning neural networks for the longitudinal analysis of mammography images in screen-detected breast cancer risk prediction

Authors: Manel Rakez1, Thomas Louis2, Julien Guillaumin2, Foucauld Chamming's3, Pierre Fillard2, Brice Amadeo1, Virginie Rondeau1.
Institution: 1:Bordeaux Population Health U1219, Bordeaux University, ISPED, Bordeaux, France ; 2:Therapixel, Nice, France ; 3:Department of Radiology, Institut Bergonié, Comprehensive Cancer Centre, Bordeaux, France

Abstract: Breast cancer is the most common cancer among women and a leading cause of death worldwide. Early detection and precise risk prediction remain critical for effective prevention and targeted screening. Yet, existing models rarely exploit the Qiitudinal information available from repeated mammograms in population-based screening programs. We aimed to develop and evaluate methods that leverage longitudinal mammographic imaging to enhance individual risk prediction. Two complementary strategies were explored. The first approach combines deep learning with joint modeling to link quantitative imaging biomarkers with cancer risk over time. A convolutional neural network, based on a modified U-Net architecture, was developed to segment breast and dense tissue and estimate validated measures of mammographic density at each screening visit. These quantitative measures, established risk factors for breast cancer, were incorporated into a joint modeling framework to quantify their association with cancer occurrence over time using several latent structures. This approach, developed as the DeepJoint algorithm, generates individualized dynamic risk predictions while preserving statistical interpretability. The second approach, LongiMam, is an end-to-end deep learning model that predicts risk probabilities directly from longitudinal mammogram sequences, with no need for predefined biomarker extraction. By integrating convolutional and recurrent neural networks, it captures spatial and temporal patterns across repeated screenings. This data-driven method aims to maximize predictive accuracy, though at the cost of potentially lower interpretability compared to biomarker-based models. Both methods were applied to a large American screening dataset, addressing complementary aspects of risk prediction. Our findings show that longitudinal modeling improves breast cancer risk prediction and support the use of repeated mammograms in risk assessment. All proposed tools are open source, ensuring transparency and reproducibility for future research.

- Agathe Guilloux agathe.guilloux@inria.fr

Title: Deep Learning for Longitudinal and Survival Data: Toward Realistic Synthetic Patient Generation

Authors: Perrine Chassat, Van Tuan Nguyen, Agathe Guilloux
Institution: HeKA – INRIA – INSERM – Univ. Paris Cité

Abstract: Joint modeling of longitudinal and time-to-event data has become a cornerstone of clinical research, providing a principled framework for understanding disease progression and predicting patient outcomes. Over the past decade, these models have demonstrated remarkable interpretability and clinical relevance, particularly when built upon carefully designed statistical assumptions.
In parallel, the growing complexity and scale of modern health data—characterized by high dimensionality, multimodality, and irregular observation times—have motivated the exploration of complementary approaches based on deep learning. Recent architectures such as variational

autoencoders and Neural Ordinary Differential Equations (Neural ODEs) can flexibly represent dynamic latent processes and capture the intricate dependencies between longitudinal measurements and survival outcomes. These models can also be used to generate realistic synthetic patient trajectories, thereby addressing privacy and data-sharing constraints that frequently limit access to large clinical cohorts.
In this talk, I will provide an overview of recent deep learning methods for longitudinal and survival data, emphasizing their connections and complementarities with traditional joint modeling frameworks. I will then present an enhanced Neural ODE architecture that jointly represents longitudinal dynamics and event-time processes within a unified latent framework. The model introduces an explicit handling of censoring and a dynamic latent representation capturing inter-individual heterogeneity. Applied to real-world clinical datasets, this approach improves the fidelity and temporal coherence of synthetic data while preserving their analytical validity. Such developments open new opportunities for privacy-preserving research, digital twin modeling, and synthetic control arm design in clinical trials.

- Clemens Schaechter clemens.schaechter@uniklinik-freiburg.de

Title: Modeling disease progression and treatment switches in rare disease trials by integrating statistical modeling and synthetic experts

Authors: Clemens Schächter[1], Astrid Pechmann[2], Janbernd Kirschner[2], Harald Binder[1]
Institution: [1]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany ; [2]Department of Neuropediatrics and Muscle Disorders, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg

Abstract: Many rare diseases have limited established treatment options, leading patients to switch therapies when new medications emerge. To analyze the impact of such treatment switches within the small-sample size limitations of rare disease trials, it is important to use all available data sources, which can stem from different multidimensional measurement instruments. Additionaly it can be beneficial to incorporate domain knowledge about the disease progression to further enhance modeling within low sample size trials.
We address these challenges by mapping the measured item values of each instrument with variational autoencoders to a temporal trajectory and aligning them in a unified low-dimensional representation. Disease dynamics and treatment switch effects are then captured through a mixed-effects regression model applied to the latent representations across time points. Subsequently, decoded trajectories are labeled by a LLM acting as a synthetic clinical expert, which guides the training process by injecting knowledge about disease progression.
The demonstrate the methodology on data from a rare disease registry of patients with spinal muscular atrophy to quantify the impact of treatment switches on disease trajectories. In this context, disease progression can be categorized into clinically established disease types, representing prototypical courses of disease progression. We automate classification using LLMs and identify trajectories that change their prototypical type over time. This allows us to enhnace and detect disease altering interventions triggered by treatment switches.

- Cécile Proust-Lima cecile.proust-lima@u-bordeaux.fr

Title: Neural differential equations enhanced linear mixed model

Authors: Aurore Zhe Li, Quentin Clairon, Mélanie Prague, Cécile Proust-Lima
Institution: Univ. Bordeaux, Inserm Bordeaux Population Health Research Center, and Inria Univ. Bordeaux research center, France

Abstract: The linear mixed model remains a cornerstone of longitudinal data analysis in health research. It can naturally handle irregular observation times, is adapted to the hierarchical nature of longitudinal data and provides robust estimates under the 'missing at random' mechanism.
However, when modeling a time-varying outcome according to multiple time-varying exposures, as encountered in life-course epidemiology, its limitations become apparent. The associations between the time-varying exposures and the outcome must be determined in advance, and typically only very specific and simple associations are considered, such as an instantaneous linear effect with no interactions across exposures.
Motivated by the study of the relationship between cardiometabolic health (e.g. blood pressure, adiposity, glycaemia and lipids) and cognitive decline in older adults, we have developed a hybrid mixed model that uses neural ordinary or controlled differential equations to encode multiple time-varying

exposures into latent spaces, which are then used to specify a decoding linear mixed model for the outcome.

Session 30: Causal mediation pathway analysis
Organizer: Peter Song pxsong@umich.edu

- Jiasheng SHI shijiasheng@cuhk.edu.cn

Title: Spillovers Through Time: A Neural Framework for Temporal Network Data and Dynamic Causal Effect Estimation
Authors: Jiasheng SHI
Institution: The Chinese University of Hong Kong (Shenzhen)
Abstract: Spatiotemporal data, characterizing phenomena ranging from pandemic transmission dynamics to environmental conflict, pose fundamental challenges for causal inference due to the dual pressures of spatial spillover and temporal persistence. While recent frameworks have improved exposure mapping, they often struggle to simultaneously account for dynamic interference and high-dimensional treatment strategies. To address these gaps, we propose a novel sequential neural architecture that bridges longitudinal causal inference and network interference. Motivated by encoding generative modeling approaches, our method learns low-dimensional latent representations to mirror the underlying causal data flow. This allows for the estimation of expected potential outcomes under spatiotemporal interference with robust covariate adjustment. The result is a framework that effectively captures evolving dependencies, providing a rigorous tool for policy evaluation in interdependent environments.

- Yen-Tsung Huang ythuang@stat.sinica.edu.tw

Title: Semiparametric mediation analysis using single-index models
Authors: Yen-Tsung Huang
Institution: Institute of Statistical Science, Academia Sinica
Abstract: Mediation analysis is increasingly used to investigate how an exposure Z influences an outcome Y through a set of mediators M. However, most existing methods either rely on parametric assumptions for the mediators and outcome, or are limited to binary exposures and a single mediator. We propose a novel algorithm that accommodates single-index models (SIMs) with non-binary exposures and adjustment for potential confounders, while avoiding parametric assumptions for both mediators and outcome. Specifically, we introduce two SIMs: one modeling the outcome conditional on the exposure, mediators, and confounders, and the other modeling the mediators conditional on the exposure and confounders. We derive a mediation estimand of the form $E\left[\left[\square_{\square}\square_{\square}\square_{\square}\square_{\square}\square_{\square}\right]\square_{\square}\square_{\square}\right]$, representing the expected outcome under a joint intervention setting the exposure to z_a and the mediators to their counterfactual values under z_b. The proposed estimator proceeds in four steps: (1) regress Y on Z and M using an SIM; (2) predict Y given z_a and M; (3) regress the predicted Y on Z using another SIM; (4) predict the outcome at z_b. We establish the asymptotic properties of this estimator and evaluate its finite-sample performance through simulation studies. Finally, we demonstrate the practical utility of our method with two applications. The first examines the effect of socioeconomic status on body mass index mediated by DNA methylation of the *FASN* (fatty acid synthase) gene. The second investigates the effect of obesity on glycemic control through triglyceride, fasting glucose, and urine microalbumin.

- Zhonghua Liu zl2509@cumc.columbia.edu

Title: Mitigating Unmeasured Confounding Bias in Large-Scale Causal Mediation Analysis Via Factor Analysis in Epigenome-Wide Studies
Authors: Zhonghua Liu
Institution: Department of Biostatistics, Columbia University
Abstract: Large-scale causal mediation analysis in epigenome-wide studies investigates whether an exposure affects an outcome through specific DNA methylations. Traditional mediation analysis approaches typically rely on the strong assumption of no unmeasured confounding between the mediators and the outcome. This assumption is often violated in observational studies where neither the exposure nor the mediators are randomized, leading to potentially biased inference on mediation effects.

To address this challenge, we propose a Factor Analysis-based Mediation Analysis (FAMA) framework that corrects for unmeasured confounding in large-scale mediation analysis settings. FAMA integrates factor analysis with methods for handling omitted-variable bias to estimate natural indirect effects in the presence of unmeasured confounding. We establish the theoretical validity of our approach and demonstrate its robust performance in diverse confounding scenarios through extensive simulation studies. We further apply FAMA to the analysis of the epigenome-wide Normative Aging Study to investigate the mediating roles of hundreds of thousands of DNA methylation CpG sites in the causal pathways linking smoking and lung function, showing that FAMA enables robust causal mediation analysis in epigenome-wide studies.

- Wei Zhou zhouwei23@swufe.edu.cn

Title: Statistical Inference for Mediation Models with High Dimensional Exposures and Mediators

Authors: Xinyu Zhang, Wei Zhou, Jingyuan Liu, Jian Kang
Institution: Xiamen University

Abstract: High-dimensional mediation analysis has gained increasing interest in various fields, particularly in genetic and medical research. Compared with existing works that focus mainly on high-dimensional mediators, this paper advocates a new statistical inference framework, named PRIME, when both exposures and mediators are high-dimensional. Estimated direct and indirect effects are established using a group-wise partially penalized least squares method, incorporating a double-layer latent factor structure. F-type and Wald tests for the high-dimensional direct and indirect effects, respectively, are advocated based on the proposed estimators. Both theoretical and numerical performance of PRIME have been carefully studied. PRIME is also applied to investigating direct effects of genetic variants on Alzheimer's disease and indirect effects of them mediated by changes in brain activity intensity.

Session 31: advanced uses of joint modeling techniques with application in longitudinal and time-to-event data

Organizer: Dimitris Rizopoulos d.rizopoulos@erasmusmc.nl

- Jeremy Taylor jmgt@umich.edu

Title: Dynamic Predictions and Predictimands for Salvage Therapy in Recurrent Prostate Cancer Using Joint Models

Authors : Jeremy Taylor
Institution : University of Michigan

Abstract: Prostate cancer patients with biochemical recurrence (BCR) are faced with a decision of whether or not to start salvage therapy (ST), which may reduce the probability of progression to metastatic disease but comes at the cost of side effects. To inform the decision to start ST at BCR, models that make counterfactual predictions with and without treatment are highly desirable. However, estimation of such models using observational data requires care due to time-varying confounding by the longitudinal biomarker prostate-specific antigen (PSA). Moreover, a careful definition of the estimands of interest, referred to as ``predictimands'' in recent work by van Geloven and colleagues, is required due to the possibility of delayed initiation of treatment after biochemical recurrence. In this study, we utilize the framework of joint longitudinal and failure time models to tackle these issues, estimating a joint model for pre-ST PSA trajectories and risk of metastasis that incorporates the effect of ST. We define predictimands of interest under three scenarios: immediately treated, never treated, and treatment under a dynamic regime, where ST is started when PSA is observed to exceed a pre-specified threshold. We propose a Monte Carlo scheme for computing these predictimands, adapted from previous work on dynamic predictions from joint models to account for treatment timing, which we illustrate for an example patient. The methodology proposed could be adapted to a wide variety of applications requiring counterfactual predictions in the presence of time-varying treatments and biomarkers and code to implement such analyses is available in the R package JMbayes2.

- Esra Kurum esra.kurum@ucr.edu

Title: Scalable Time-Dynamic Joint Models for Multilevel Health Data

Abstract: Individuals with end-stage kidney disease (ESKD) on dialysis experience substantial hospitalization burden and high mortality risk over time. Understanding the dynamic effects of patient- and cluster- (region-) level factors on these correlated outcomes is critical for improving care and informing policy. Using data from the United States Renal Data System (USRDS), a national registry comprising repeated observations on patients nested within health service regions (clusters), we propose a scalable Bayesian framework for joint modeling of longitudinal and survival outcomes with time-varying effects. Region-level fixed effects are of direct scientific interest, as quantifying regional differences in risk can reveal disparities and guide targeted interventions. However, as the number of regions and overall data size grow, standard joint modeling approaches become computationally intractable. The proposed Bayesian time-dynamic joint model with cluster-level fixed effects (BJM-CFE) addresses this challenge through a novel two-stage estimation procedure that alternates between cluster-specific and global-level inference. This modular approach enables parallelization and memory efficiency, allowing for high-dimensional temporal modeling in large, multilevel datasets. Application of BJM-CFE to the USRDS data reveals important time-varying patterns in hospitalization and mortality risk and highlights regional disparities across the United States.

- Christos Thomadakis cthomadak@aueb.gr

Title: Shared parameter modeling of longitudinal data allowing for possibly informative visiting process and terminal event

Abstract: Joint modeling of longitudinal and time-to-event data, particularly through shared parameter models (SPMs), is a common approach for handling longitudinal marker data with an informative terminal event. A critical but often neglected assumption in this context is that the visiting/observation process is noninformative, depending solely on past marker values and visit times. When this assumption fails, the visiting process becomes informative, potentially resulting in biased SPM estimates. Existing methods generally rely on a conditional independence assumption, positing that the marker model, visiting process, and time-to-event model are independent given shared or correlated random effects. Moreover, they are typically built on an intensity-based visiting process using calendar time. This study introduces a unified approach for jointly modeling a normally distributed marker, the visiting process, and time-to-event data in the form of competing risks. Our model conditions on the history of observed marker values, prior visit times, the marker's random effects, and possibly a frailty term independent of the random effects. While our approach aligns with the shared-parameter framework, it does not presume conditional independence between the processes. Additionally, the visiting process can be defined on either a gap time scale, via proportional hazard models, or a calendar time scale, via proportional intensity models. Through extensive simulation studies, we assess the performance of our proposed methodology. We demonstrate that disregarding an informative visiting process can yield significantly biased marker estimates. However, misspecification of the visiting process can also lead to biased estimates. The gap time formulation exhibits greater robustness compared to the intensity-based model when the visiting process is misspecified. In general, enriching the visiting process with prior visit history enhances performance. We further apply our methodology to real longitudinal data from HIV, where visit frequency varies substantially among individuals.

- Eleanor Pullenayegum eleanor.pullenayegum@sickkids.ca

Title: Addressing measurement times that are not at random in longitudinal data through joint modelling of outcomes and latent disease processes

Abstract: Longitudinal data collected as part of usual care often feature irregular and informative observation times. In diseases featuring a relapsing-remitting course, observation times are likely to be driven in part by the relapse status; since this is typically unknown prior to the visit observation times are not at random. Furthermore, traditional mixed models are unable to describe disease course at the individual level, since they do not capture fluctuations due to relapses. Both of these issues can be resolved through joint modelling of the outcomes, observation times, and latent relapse status. We show how informative priors can be used to ensure identifiability and illustrate the approach through an analysis of disease course in juvenile dermatomyositis.

Session 32: Tensor-Train Methods, Partial Derivatives, and High-dimensional Inference in Modern Statistics

Organizer: Yong Chen ychen123@pennmedicine.upenn.edu

- Yong Chen ychen123@pennmedicine.upenn.edu

Title: PDA: a unified framework for lossless, one-shot, federated learning algorithms
Authors: Yong Chen
Institution: University of Pennsylvania
Abstract: Multi-site studies are now central to biomedical research, but inference across sites remains challenging due to regulatory limits on sharing individual-level data, heterogeneous data distributions, sparse events in small centers, and the logistical burden of multi-round communication. In this talk, I introduce MOSAiC (Multi-site One-Shot Aggregation of Compressed Risk Functions), a unified framework for modern distributed research networks. Notably, MOSAiC reframes distributed learning as a mathematical problem of compressing and aggregating risk functions, leveraging recent advances in tensor networks—state-of-the-art tools for high-dimensional function approximation in scientific computing. Our MOSAiC enjoys four desirable properties that have never been achieved by any of the existing federated algorithms (except linear regressions): one-shot communication, lossless recovery of pooled-data estimates, inclusiveness of all sites irrespective of size or event rarity, and analytic submodel exploitability without re-querying partners. I will illustrate MOSAiC's validity and efficiency through applications in drug relabeling, drug repurposing, and post-market safety surveillance.

- Jingmei Qiu jingqiu@udel.edu

Title: Projection-Enhanced Interpolation-Based Tensor-Train Decompositions for One-Shot, Privacy-Preserving Distributed Inference
Authors: Jingmei Qiu
Institution: University of Delaware
Abstract: The tensor-train (TT) format is a data-sparse tensor representation that has become a central tool for high-dimensional approximation and scientific computing. To improve interpretability and robustness in data-driven settings, recent work has focused on *skeletonized* low-rank approximations that combine interpolation and projection ideas. While effective, existing skeletonized TT methods may still suffer from accuracy degradation due to information loss from unselected samples, as well as increased computational cost when applied to very high-dimensional data. In this work, we develop projection-enhanced interpolation-based algorithms that systematically improve the accuracy of skeletonized TT approximations while preserving their low-complexity structure. As an application, we demonstrate how the proposed TT approximation framework enables *one-shot* aggregation in privacy-constrained, multi-site inference. By representing locally computed risk functions in compressed TT form, our approach supports the Multi-site One-Shot Aggregation of Compressed Risk Functions (MOSAiC) meta-algorithm for distributed learning, allowing single-round communication, robustness to heterogeneous or small sites, and accuracy comparable to pooled analysis. This application highlights the practical impact of robust and efficient TT approximations in modern distributed research networks.

- Xiaowu Dai dai@stat.ucla.edu

Title: Statistical Learning via Partial Derivatives
Authors: Xiaowu Dai
Institution: University of California, Los Angeles
Abstract: Traditional nonparametric estimation methods often lead to a slow convergence rate in large dimensions and require unrealistically large dataset sizes for reliable conclusions. We develop an approach based on partial derivatives, either observed or estimated, to effectively estimate the function at near-parametric convergence rates. This novel approach and computational algorithm could lead to methods useful to practitioners in many areas of science and engineering. Our theoretical results reveal behavior universal to this class of nonparametric estimation problems. We explore a general setting involving tensor product spaces and build upon the smoothing spline analysis of variance framework. For d-dimensional models under full interaction, the optimal rates with gradient information on p covariates are identical to those for the (d-p)-interaction models without gradients and, therefore, the

models are immune to the curse of interaction. For additive models, the optimal rates using gradient information achieve the surprising parametric rate.

- Yumou Qiu qiuyumou@math.pku.edu.cn

Title: High-dimensional Data Assimilation for Large Scale Physical Systems

Authors: Yumou Qiu

Institution: Peking University

Abstract: Data assimulation is a fundamental method for forecasting in complex nonlinear systems. High-dimensionality of large scale systems poses a challenge for data assimilation in both theoretical property and computation. In this paper, we propose a locally weighted estimator of the analysis states in a data assimilation procedure from the perspective of kriging, through specific prior information on the weights of innovations. The proposed method is shown consistent, and can be applied jointly with other regularization methods to achieve a better performance for forecasting. We also provide an efficient computation algorithm for the proposed method, which allows for high resolution ocean data assimilation. The simulation results in Lorenz 96 and MITgcm show that our method provides the smallest RMSE compared to existing methods, and is less sensitive to model error. A comprehensive real data analysis shows the proposed method reduces the forecast errors of ocean temperature and salinity of the GLO12v4 system (GLORYS) by 17% and 19%, respectively.