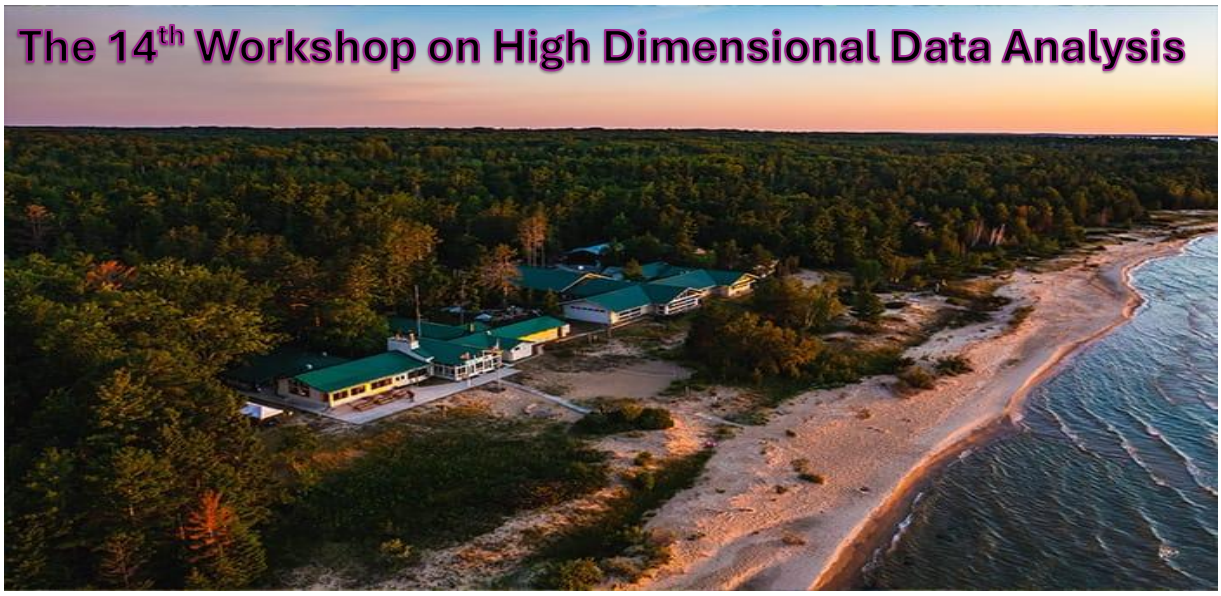
HDDA - XIV Aug 19-22

COLLEGE OF
SCIENCE & ENGINEERING
CENTRAL MICHIGAN UNIVERSITY



PROGRAM ABSTRACTS

(in order of presentation)

Simultaneous Confidence Intervals for Signal Detection and Ascertaining Precision of Adverse Event Rates in Clinical Trials

Authors: Guoqing Diao, Margaret Gamalo, Ram Tiwari

Abstract: The marketing authorization of a medicinal product is contingent upon demonstration of safety and efficacy in support of the product's labeled conditions of use. To demonstrate safety, one group of adverse events that requires detailed consideration is common adverse reactions (ARs). ARs and their frequency are reported in prescription drug labeling in the US and EU. The determination of these adverse reactions generally takes a simple approach - usually, inclusion is through the frequency of reporting and whether the adverse event (AE) rate for the drug exceeds the placebo rate. This standard method does not account for confounders or multiplicity. To overcome these limitations, we propose a Monte-Carlo approach to detect drug safety signals in clinical trials. We fit regression models incorporating covariates to assess the drug effect on the rate of AEs. Adjustment for multiplicity is carried out through the construction of simultaneous confidence intervals accounting for arbitrary correlations. A computationally efficient multiplier bootstrap approach using the Rademacher sequences is developed to generate random samples from the joint distribution of the estimators for all the AE rates. Compared to Bonferroni-based methods, the proposed method leads to narrower simultaneous confidence intervals and is more powerful in detecting potential safety signals.

A Bayesian approach to the consistency for regularized estimators

Rauf Ahmad

Department of Statistics, Uppsala University

Abstract

For high-dimensional data, regularized estimators (henceforth REGs) such as ridge or LASSO replace their classical counterparts, e.g. least-squares estimators. In Bayesian theory, the same estimators follows as posterior MAP estimators (henceforth MAPs) by setting appropriate priors on the parameters, with normal likelihood. In Bayesian context, the consistency pertains to that of the posterior distribution. Since, however, MAPs coincide with REGs, our interest focuses on proving their consistency using Bayesian techniques. We extend Schwartz' theorem to independent, non-identically distributed random variables, and apply this extended version to demonstrate the consistency of the estimators, with special emphasis on ridge and LASSO estimators.

Note It is an ongoing joint work with Silvelyn Zwanzig.

Weighted LAD-Liu-LASSO for robust estimation and sparsity

Murat Genç[1] · Adewale Lukman[2]

Abstract

The Least Absolute Shrinkage and Selection Operator (LASSO) is widely used for parameter estimation and variable selection but can encounter challenges with outliers and heavy-tailed error distributions. Integrating variable selection methods such as LASSO with Weighted Least Absolute Deviation (WLAD) has been explored in limited studies to handle these problems. In this study, we proposed the integration of Weighted Least Absolute Deviation with Liu-LASSO to handle variable selection, parameter estimation, and heavy-tailed error distributions due to the advantages of the Liu-LASSO approach over traditional LASSO methods. This approach is demonstrated through a simple simulation study and real-world application. Our findings showcase the superiority of our method over existing techniques while maintaining the asymptotic efficiency comparable to the unpenalized LAD estimator.

Advancements in Efficient Estimation for Mixed Effects Models with Censored Data

Shakhawat Hossain, University of Winnipeg, MB Canada

Abstract

Longitudinal and repeated measures data are frequently analyzed using mixed model. However, the presence of censored responses, a common occurrence in biomedical research and clinical trials due to detection limits, introduces complexity. These limits can result in left or right censored measurements. While adapted linear mixed effects models are often employed to handle such data, this paper proposes a likelihood-based approach for fitting a linear mixed effects models with normally distributed errors. An expectation-maximization (EM) algorithm is utilized for unrestricted maximum likelihood estimation. Furthermore, the study explores scenarios where model parameters are subject to uncertain linear constraints, leading to the development of a restricted estimator. To improve the estimation of fixed effects, two refined estimator sets are introduced: a pretest estimator, and shrinkage and positive shrinkage estimators. The performance of these proposed estimators is evaluated against the unrestricted maximum likelihood estimator via extensive simulations and an application to longitudinal data from the AIDS Clinical Trials Group protocol A5055 study. Privacy-Constrained Robust Inference for High-dimensional Locations

Estimation and Inference in Tensor Regression Models with Structural Breaks, with Applications to Neuroimaging

Mai Ghannam

University of Toronto

Abstract:

In this talk, we study estimation and inference in a tensor regression model with multiple change-points, under weak dependence assumptions on the covariates and errors, specifically modeled as an L^2-mixingale array. We first establish the asymptotic properties of both the unrestricted estimator (UE) and a restricted estimator (RE) in this framework. Building on this, we introduce a novel class of shrinkage estimators (SEs) tailored for tensor regression and provide sufficient conditions under which these SEs outperform the UE in terms of risk. For inference, we focus on the case of a single possible change-point and formulate a general hypothesis testing problem for the tensor-valued parameter. This setting encompasses, as a special case, the problem of testing for the presence of a change-point. We develop a consistent test for the restriction and characterize its asymptotic power. Simulation studies and neuro-imaging data analysis are presented to support and illustrate the theoretical findings.

FASTER: Feature Alignment and Structured Transfer via Efficient Regularization

Iris Zhang, Yang Feng

Abstract

Most existing transfer learning methods assume identical feature spaces for all domains. However, differences in data collection often create feature variations across domains, making the feature space heterogeneous. To address this, we propose FASTER (**F**eature **A**lignment and **S**tructured **T**ransfer via **E**fficient **R**egularization), a novel two-step transfer learning framework that integrates regularized feature alignment with structured modeling to enhance knowledge transfer. FASTER first aligns source and target domains by learning structured feature mappings through covariance-regularized optimization, ensuring effective information transfer despite feature differences. In the second step, a joint predictive model is trained on the mapped source and target data by minimizing a regularized loss function, followed by an adaptive correction term that refines task-specific differences. Our approach reduces domain disparity while preserving interpretability through structured regularization. Extensive simulations and real-data experiments validate the effectiveness of FASTER in heterogeneous feature adaptation, providing a principled solution for transfer learning across diverse domains.

Note: This work will be presented as an invited talk in the session "*Novel Statistical and Machine Learning Methods for Integrating Multiple Data Sources.*"

Enhancing Variable Selection in High-Dimensional GLMs: Incorporating Weak Signals Through Post-Selection Estimation

Dr. Reza Belaghi and Dr. Abdul A. Hussein

[1]Department of Energy and Technology, Swedish University of Agricultural Sciences, Uppsala, Sweden

[2]Department of Mathematics and Statistics, University of Windsor, Windsor, Ontario, Canada

Abstract

High-dimensional datasets—where the number of predictors exceeds the sample size—are increasingly encountered in diverse fields such as biomedical research, finance, and machine learning. In such contexts, variable selection is essential for building interpretable and predictive models. This study investigates and compares statistical and machine learning approaches for variable selection in the framework of generalized linear models (GLMs), emphasizing the detection of both strong and weak signal covariates. We propose a three- stage post-selection estimation procedure that extends LASSO and Elastic-Net methods to accommodate weak signals often ignored by conventional penalization techniques. This strategy incorporates restricted estimation, weighted ridge regression, and James-Stein-type shrinkage adjustments to improve post-selection inference. Through extensive Monte Carlo simulations, we show that the proposed post-selection shrinkage estimators consistently outperform standard GLM-based LASSO and Elastic-Net models in terms of relative mean squared error and false positive rate. We further validate our approach using a real-world dataset on GDP growth and socioeconomic indicators, where the method achieves more accurate and stable estimates than existing techniques. Additionally, the Boruta algorithm is evaluated as a machine learning-based alternative for variable selection, particularly in scenarios with high multicollinearity. Results suggest that incorporating weak signals and post-selection refinement significantly enhances model performance and reliability in high-dimensional GLM settings. This work underscores the importance of adaptive shrinkage and hybrid strategies for robust and effective variable selection in modern statistical modeling.

Incorporating Additional Information Using Penalized Regression for Gene-Environment Interaction Analysis

Yaqing Xu

Shanghai Jiao Tong University

Abstract:

Gene-environment interactions have important implications for complex disease such as cancer and diabetes. It is challenging to analyze interactions due to the high dimensionality and low signal levels. Given the lack of information, incorporating additional information is desired and can potentially improve the accuracy and interpretability of identification. However, most of existing methods ignore such information and treat genetic factors equally a priori. We propose a penalized approach that is customized to incorporate additional information for identifying important main genetic effects and hierarchical interactions. Under a marginal analysis framework, the proposed method adopts minimax concave penalty for regularized estimation and Laplacian quadratic penalty for additional information. Here, additional information can be the adjacency structure in certain chromosomes, correlation structure among gene expressions, and data extracted from existing literature. The proposed method can be efficiently realized by a coordinate descent algorithm for multiple types of genetic factors. Extensive simulation shows our proposed approach outperforms multiple alternatives. In the analysis of TCGA melanoma, the proposed method demonstrates practical applicability and provides sensible findings.

Enhancing High-Dimensional Clustering with Voting Technique for Most Representative Average (VOMORA)

[1]Kayode Ayinde, [2]Emmanuel Taiwo Adewuyi, and [3]Adewale F. Lukman
[1]Department of Mathematics and Statistics
Northwest Missouri State University, Maryville, Missouri, 64468, USA.
[2]Department of Medical Statistics, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London.
[3]Department of Mathematics, University of North Dakota, Grand Forks, ND 58202, USA. Email Address: ayindek@nwmissouri.edu

Abstract:

The proliferation of high-dimensional data poses significant challenges for clustering algorithms, primarily due to the "curse of dimensionality," which often renders traditional techniques ineffective. To address this issue, we propose a novel clustering approach called Voting Technique for Most Representative Average (VOMORA). This method is specifically designed to tackle the complexities of high-dimensional numerical data sets, where conventional clustering methods often struggle to identify meaningful groupings.

In this study, we applied VOMORA to a benchmark data set from Kaggle, comprising 230 instances and 537 features, along with binary classification labels. Our experimental analysis revealed that the clusters formed by VOMORA exhibit strong agreement with the provided labels, demonstrating its ability to identify robust and meaningful groupings even in the presence of many dimensions.

The results of this study highlight the potential of VOMORA as an effective clustering strategy for high-dimensional data scenarios. By leveraging the voting technique, VOMORA can accurately identify representative averages, leading to more accurate and robust cluster formation. This approach offers a promising solution for various applications where high- dimensional data clustering is crucial, such as data mining, pattern recognition, and machine learning.

The key contributions of this study include:

1. Introduction of VOMORA, a novel clustering approach designed for high-dimensional numerical data sets.
2. Experimental validation of VOMORA's effectiveness using a benchmark data set with binary classification labels.
3. Demonstration of VOMORA's robustness against dimensional complexity, highlighting its potential for various high-dimensional data clustering applications.

Overall, VOMORA presents a promising approach for enhancing cluster formation in high-dimensional data sets, offering both accuracy and robustness against the challenges posed by high dimensionality.

A multivariate extension of size-biased mixtures and its estimation using a penalized EM algorithm

Taehan Bae, University of Regina

Email: Taehan.bae@uregina.ca

Abstract

Finite mixtures have been among the most widely used approaches for modeling complex data drawn from a population composed of heterogeneous sub-populations. Specifically for insurance loss modelling, several mixtures of weighted distributions such as the Erlang mixture, the size-biased Weibull mixture and the size-biased truncated lognormal mixture, have gained popularity due to their desirable modelling features such as the flexibility to fit various distributional shapes including heavy-tailed ones, the stochastic ordering structure and the weak-denseness property. In this talk, an extension of univariate size-biased mixture models to multivariate cases will be discussed. The proposed multivariate model involves a large number of mixture terms, and thus a regularization is essential to avoid overfitting. In this talk, I will discuss about a penalized EM algorithm, driven from Dirichlettype prior for the mixture weight parameters, that facilitates purging of statistically insignificant components, and thus results in a parsimonious model than the one fitted by the classical EM algorithm. Three specific multivariate size-biased models: Erlang, size biased Weibull and size-biased truncated lognormal models, will be illustrated with model fittings on a real-life data set.

Multi-Task Learning in Multiple Imputation by Chained Equation (MTL-MICE)

Authors: Yuyu (Ruby) Chen, Yang Feng

This abstract is dedicated to the invited talk at HDDA workshop in the session "Novel Statistical and Machine Learning Methods for Integrating Multiple Data Sources."

High-dimensional datasets are common in healthcare and public health, where multi-center electronic health records (EHRs) and national surveys pose complex missing data challenges. Traditional imputation methods struggle in these settings, as they handle missing values independently for each task. To address this, we propose Multi-Task Learning via Multiple Imputation by Chained Equations (MTL-MICE), a novel approach that enhances imputation by leveraging shared information across tasks.

MTL-MICE integrates multi-task learning into the MICE framework, capturing correlations among tasks to improve accuracy and robustness. Instead of treating missing data separately, it utilizes shared relationships across features. Additionally, we incorporate a transferable source detection technique to identify informative tasks, refining imputation further.

Through simulations and real-world studies, we show that MTL-MICE significantly reduces imputation error and bias compared to single-task methods while preserving MICE's flexibility. These findings highlight the potential of multi-task learning to improve missing data methodologies for large-scale, high-dimensional studies.

Spacekime Representation, Statistical Inference, and AI prediction using Repeated Measurement Longitudinal Data

Ivo Dinov

University of Michigan

Abstract: Complex-time (kime) representation of repeated measurement longitudinal processes paves the way for advanced spacekime statistical inference and artificial intelligence (AI) applications. Extending time into the complex plane offers a unified framework connecting fundamental quantum mechanics principles, statistical dynamics, and machine learning. Kime representation enhances both model-based statistical inference techniques – utilizing classical probability distributions – and model-free AI prediction and classification algorithms – relying on data and generalized functions. Many open mathematical-physics problems emerge from this formulation, including definition and interpretation of a consistent spacekime-metric tensor and classification of alternative time-series to kime-surfaces transformations. Simulations and observed neuroimaging data demonstrate the utility of complex-time representation and the induced spacekime analytics. These methods enable forward prediction by extrapolating processes beyond their observed timespan and facilitate group comparisons based on corresponding kime surfaces. Additionally, they allow for statistical quantification of differences between experimental groups and conditions, support topological kime surface analysis, and enhance AI prediction for repeated measurement longitudinal data.

MALADY: Multiclass Active Learning with Auction Dynamics on Graphs

Gokul Bhusal

Michigan State University

Abstract—Active learning enhances the performance of machine learning methods, particularly in low-label rate scenarios, by judiciously selecting a limited number of unlabeled data points for labeling, with the goal of improving the performance of an underlying classifier. In this work, we introduce the Multiclass Active Learning with Auction Dynamics on Graphs (MALADY) algorithm, which leverages an auction dynamics technique on similarity graphs for efficient active learning. In particular, the proposed algorithm incorporates an active learning loop using as its underlying semi-supervised procedure an efficient and effective similarity graph-based auction method consisting of upper and lower bound auctions that integrate class size constraints. In addition, we introduce a novel active learning acquisition function that incorporates the dual variable of the auction algorithm to measure the uncertainty in the classifier to prioritize queries near the decision boundaries between different classes. Overall, the proposed method can efficiently obtain accurate results using extremely small labeled sets containing just a few elements per class; this is crucial since labeled data is scarce for many applications. Moreover, the proposed technique can incorporate class size information, which improves accuracy even further. Lastly, using experiments on classification tasks and various data sets, we evaluate the performance of our proposed method and show that it exceeds that of comparison algorithms.

Support vector machine is a powerful statistical machine-learning method that enables us to capture the relationship between a set of covariates and a response variable. On the other side, shrinkage estimation methods play an important role in statistical models for various purposes, such as variable selection and overfitting problems. As biased estimation techniques, shrinking methods improve parameter estimation by constraining the parameter space within a restricted model. In this research, through a probabilistic framework, we propose various shrinkage methods in low and high-dimensional settings for support vector machine using Karush–Kuhn–Tucker conditions. We evaluate the performance of the developed shrinkage support vector methods through extensive numerical studies. Finally, the developed methods are applied to real data examples.

Armin Hatefi

University of Newfoundland

Support vector machine is a powerful statistical machine-learning method that enables us to capture the relationship between a set of covariates and a response variable. On the other side, shrinkage estimation methods play an important role in statistical models for various purposes, such as variable selection and overfitting problems. As biased estimation techniques, shrinking methods improve parameter estimation by constraining the parameter space within a restricted model. In this research, through a probabilistic framework, we propose various shrinkage methods in low and high-dimensional settings for support vector machine using Karush–Kuhn–Tucker conditions. We evaluate the performance of the developed shrinkage support vector methods through extensive numerical studies. Finally, the developed methods are applied to real data examples.

Class-specific Joint Feature Screening in Ultrahigh-dimensional Mixture Regression

Abbas Khalili

Dept. of Math and Stat

McGill University, Canada

Finite mixture of regression (FMR) models are ubiquitous for analyzing complex data. They aim to detect heterogeneity in the effects of a set of features on a response over a finite number of latent classes. When the number of features is large, a direct fitting of FMR can be computationally infeasible and often leads to a poor interpretative value. One practical strategy is to screen out most irrelevant features before an in-depth analysis. In this paper, we propose a novel method for feature screening in ultrahigh-dimensional Gaussian FMR. The new method is built upon a sparsity-restricted expectation-approximation-maximization algorithm, which simultaneously removes varying sets of irrelevant features from multiple latent classes. In the screening process, joint effects between features are naturally accounted and class-specific screening results are produced without ad hoc steps. These merits give the new method an edge to outperform the existing screening methods. The promising performance of the method is supported by both theory and numerical examples including a real data analysis. This talk is based on a joint work with Kaili Jing and Chen Xu.

A Divide-and-Conquer Approach for Testing Joint Significance

Authors: Canyi Chen and Peter X.-K. Song

Abstract: This paper addresses the problem of distributed testing for the presence of a joint significance. The null hypothesis of no joint significance comprises a composite structure of multiple sub-cases. The underlying sub-cases are generally unknown in practice. Classical procedures, including Sobel's test and the MaxP test, exhibit distinct limiting distributions across these sub-cases. Consequently, standard divide-and-conquer strategies based on averaging local statistics fail to adequately control the type I error and exhibit diminished power under the alternative. To address these challenges, we introduce a sequential aggregation approach that separates sub-case determination from parameter estimation. The resulting test statistic provably achieves valid type I error control across the composite null and demonstrates enhanced power. Simulation studies verify the theoretical claims of the proposed method.

A novel high-dimensional model for identifying regional DNA methylation QTLs

KAIQIONG ZHAO
York University

Abstract:

Varying coefficient models offer the flexibility to learn the dynamic changes of regression coefficients. Despite their good interpretability and diverse applications, in high-dimensional settings, existing estimation methods for such models have important limitations. For example, we routinely encounter the need for variable selection when faced with a large collection of covariates with nonlinear/varying effects on outcomes, and no ideal solutions exist. One illustration of this situation could be identifying a subset of genetic variants with local influence on methylation levels in a regulatory region. To address this problem, we propose a composite sparse penalty that encourages both sparsity and smoothness for the varying coefficients. We present an efficient proximal gradient descent algorithm to obtain the penalized estimation of the varying regression coefficients in the model. A comprehensive simulation study has been conducted to evaluate the performance of our approach in terms of estimation, prediction and selection accuracy. We show that the inclusion of smoothness control yields much better results than having the sparsity-regularization only. Using an adaptive version of our penalty function, we can achieve notable additional performance gains. Furthermore, we applied our method to identify genetic variations that affect methylation variability in gene-based regulatory regions using asymptomatic samples drawn from the CARTaGENE cohort. The methodology development has been implemented in R package sparseSOMNiBUS available on GitHub.

Bayesian Complementary Kernelized Learning for Multidimensional Spatiotemporal Data

MENGYING LEI
McGill University

Probabilistic modeling of multidimensional spatiotemporal data is critical to many real-world applications. An important research question in spatial statistics and machine learning is to develop effective and computationally efficient statistical models to accommodate nonstationary/nonseparable processes containing both long-range and short-scale variations for large-scale datasets with various corruption/missing structures. In this paper, we propose a new statistical framework---Bayesian Complementary Kernelized Learning (BCKL)---to achieve scalable probabilistic modeling for multidimensional spatiotemporal data. To effectively characterize complex dependencies, BCKL integrates two complementary approaches---kernelized low-rank tensor factorization and short-range spatiotemporal Gaussian Processes. Specifically, we use a multi-linear low-rank factorization component to capture the global/long-range correlations in the data and introduce an additive short-scale GP based on compactly supported kernel functions to characterize the remaining local variations. We develop an efficient Markov chain Monte Carlo (MCMC) algorithm for model inference and evaluate the proposed BCKL framework on both synthetic and real-world spatiotemporal datasets. Our experiment results show that BCKL offers superior performance in providing accurate posterior mean and high-quality uncertainty estimates, confirming the importance of both global and local components in modeling spatiotemporal data.

Time-aligned Topic Model for Longitudinal Microbiome Data

Pratheepa Jeganathan

Department of Mathematics and Statistics

McMaster University

Latent Dirichlet Allocation (LDA) has been widely used in microbiome data analysis to identify microbial communities, referred to as "topics." Current approaches to applying LDA for longitudinal text data, particularly for long series, involve a two-stage process. First, new cohorts are formed at each time point using subsampling, where samples are chosen based on their temporal distance. Subsequently, the topics are aligned based on the shortest path distance on the learned manifold, which limits the application of this alignment method to data with a limited number of time points.

To address this gap, we propose an alignment based on the log-linear model. We treat aligning topics between consecutive cohorts as a problem of finding the correct permuta- tion. To do this, we create a high-dimensional contingency table, with one dimension for samples and other dimensions for all permutations of topics from two consecutive cohorts that could match up. Using a log-linear model, we analyze this table to identify the most significant topic permutation.
Our simulation and real-world case study findings indicate that the proposed time- aligned LDA approach significantly improves the goodness-of-fit of LDA-based analysis. Thus, it provides researchers with a more accurate tool for examining temporal variations in microbial community compositions.

A regression-based and distribution-free matrix completion framework with side information and row/column correlations

Khaled Fouda
Department of Decision Sciences, HEC Montreal
khaled.fouda@hec.ca

Abstract

Matrix completion models have been studied across different domains (recommendation systems, drug– target interaction prediction, and transportation networks), with models being tailored to specific appli- cations. To the best of our knowledge, no existing framework unifies matrix completion. We introduce a regression-based framework that is distribution-free, incorporates side information, adjusts for spa- tial/temporal correlations, and provides interpretable estimated coefficients for the side information. The response matrix, assumed to be partially observed, is modelled as a linear function of covariate matrices associated with both its rows and columns, along with a latent matrix acting as an element-wise intercept. A nuclear norm penalty is used to capture low-rank property of the latent matrix, and correlations within the rows and columns are accounted for using graph Laplacian regularizers. Finally, a lasso penalty is added on the covariate coefficients to promote sparsity. The model is solved using an alternating least squares algorithm. A key aspect of the implementation is minimizing the computational costs and scaling to high-dimensional matrices by avoiding any operations on large dense matrices. The proposed method demonstrated improved accuracy and computational efficiency over existing regression-based models in simulations with over 90% missingness and with covariates. It was further validated on data from the Montreal bike-sharing system, where it accurately estimated daily station-level ride counts.

Random projection-based response best-subset selector for ultra-high dimensional multivariate data

Jianhua Hu[1]

[a]*School of Statistics and Data Science, Shanghai University of Finance and Economics, Shanghai 200433, China*

Abstract

In this talk, we propose a random projection-based response best-subset selector to perform response variable selection in ultra-high dimensional multivariate data, where both the dimensions of response and predictor variables are substantially greater than the sample size. This method is developed by integrating the response best-subset selector, random projection technique and simultaneous predictor dimensionality reduction around common high-dimensional response variables. Under a multivariate tail eigenvalue condition, such a simultaneous dimensionality reduction on- ly leads to an ignorable error between the original and dimension-reduced models. A computational procedure is presented. The proposed method exhibits model consistency under mild conditions. The efficiency and merit of the proposed method are strongly supported by extensive finite-sample simulation studies. A real breast cancer dataset spanning 22 chromosomes are analyzed to demonstrate the proposed method.

Censored Sliced Inverse Regression based on the Weighted Leverage Score (CSIR-WLS)

TAOUFIK BOUEZMARNI
University of Sherbrooke

Analyzing high-dimensional censored survival data presents significant challenges, as existing methods often focus either on dimension reduction or variable selection, but rarely both. Sliced Inverse Regression (SIR) and Weighted Leverage Scores (WLS) have independently advanced these areas, yet neither fully addresses the complexities of right-censored data, including multicollinearity and feature selection. In this work, we introduce **Censored Sliced Inverse Regression based on the Weighted Leverage Score (CSIR-WLS)**, a novel method that unifies dimension reduction and variable selection for survival analysis. Our approach extends WLS to censored settings by introducing the **Censored Weighted Leverage Score (CWLS)** and integrates it with SIR to identify relevant predictors while mitigating multicollinearity.We establish its theoretical properties and convergence behavior, demonstrating its robustness and efficiency. Through extensive simulations and real-world applications, we show that CSIR-WLS outperforms existing methods in both accuracy and computational scalability.

Illuminant Spectrum Estimation and Inference to Study Animal Vision from Multispectral Camera Images

David Kepplinger

George Mason University

Multispectral images have been playing a crucial role in animal vision research. Reconstructing animal vision from images typically involves camera calibration, spectral reflectance estimation, and linear transformation of camera responses to animal quantum catches — a procedure that lacks transferability across scenes as camera recalibration is required for changing illumination. We propose a statistical framework for reconstructing animal vision in dynamic settings using modified but affordable consumer cameras. Our framework provides estimation and inference for the spectral illumination at a high resolution, thereby requiring only a one-time camera calibration. Unlike most functional data analysis tasks, we observe only camera readings, i.e., the inner products of spectral illumination, sensitivity and reflectance, giving rise to a complicated dependency structure. We therefore propose a penalized constrained likelihood estimator for the coefficients of a tailored basis representation and the variance-covariance function. This framework enables the reconstruction of animal vision with moving objects and dynamic lighting, paving the way for generating animal vision videos in natural habitats.

Statistical Modeling of Fuzzy Counts in High-Dimensional RNA-Seq via Conditional Coarsening

Antonio Calcagnì

University of Padova, Italy

High-dimensional data modeling requires accounting for complex measurement processes that introduce structured uncertainty or partial observability, challenging traditional assumptions of precise and unambiguous measurements. RNA-seq data analysis typically assumes precise, count-based measurements of gene expression. However, in practice, mapping ambiguity —due to multireads aligning to multiple genomic locations and multicovers spanning overlapping genes— creates intrinsic epistemic uncertainty. This uncertainty is not a byproduct of noise or post-processing, but an inherent property of the sequencing and alignment process, suggesting that observed counts should be treated as fuzzy numbers rather than crisp values.

We propose a hierarchical representation of fuzziness within a Bayesian framework. The proposal explicitly models the epistemic mechanism generating fuzzy data through a coarsening not at random (CNAR) paradigm. In this approach, latent gene expression levels remain crisp-valued random variables, while observed fuzzy counts are derived using a conditional schema that captures the mapping uncertainty. While previous approaches have used tailor-made solutions to tackle this problem, our proposal subsumes epistemic fuzziness within the broader framework of statistical coarsening mechanisms.

In our model, once we acknowledge that the sample space is fuzzy rather than crisp, the process generating the fuzziness is no longer independent of the unobserved (latent) true realizations. Instead, the fuzzification mechanism depends on the underlying complete data. To rigorously address this dependence, we adopt a conditional statistical schema akin to those employed in selection models, where the joint distribution is factorized into a latent model for the true values and a conditional model for the observed fuzziness given those true values. Crucially, in this setting, there is no ignorability: the parameters governing the unobserved expression levels are not independent of the parameters governing the fuzzification mechanism. Moreover, if fuzziness is viewed analogously to missingness, in our case all data are effectively missing, namely every observation is fuzzy. This eliminates the operational meaning of partitioning data into observed and missing components. Therefore, our approach —built on this conditional sampling schema for fuzzy data analysis— generalizes and resembles the mechanisms typically assumed in the CNAR framework, providing a principled probabilistic treatment of epistemic uncertainty inherent in the measurement process.

The Bayesian formulation adopted in this study preserves the counting ambiguity usually lost in approaches based on fractional counts and supports integration of heterogeneous data sources by consistently modeling their differing levels of epistemic fuzziness. Overall, our proposal provides an integrated statistical framework for capturing the inherent imprecision of RNA-seq measurements, aligning fuzzy data analysis with high-dimensional and multi-source bioinformatics challenges.

Adaptive Nonconvex Regularization for Toeplitz Covariance Estimation in High-Dimensional Settings

Deliang Dai[1], Yuli Liang[2], Stanislas Muhinyuza[1], and Jianxin Pan[3,4]

[1]Department *of Economics and Statistics, Linnaeus University, Sweden.*
[2]*School of Mathematics and Statistics, Guangxi Normal University, Guilin, 541004, China.*
[3]*Research Center for Mathematics, Advanced Institute of Natural Sciences, Beijing Normal University, Zhuhai, Guangdong 519087, PR China.*
[4]*Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, Guangdong 519087, PR China.*

Abstract

We further develop a novel approach to high-dimensional covariance estimation by employing the smoothing penalty and nonconvex SCAD penalty to regularize Toeplitz parameter estimates by using an efficient Alternating Direction Method of Multipliers (ADMM) based algorithm. This penalty adaptively shrinks small coefficients to zero, encouraging sparsity while mitigating the bias typically introduced by convex penalties such as the lasso. Our method automatically detects the optimal near sparsity structure of the Toeplitz matrix. The convergence rates for the proposed estimator under mild regularity conditions are established and developed with eigenvalue constraints to ensure positive definiteness. The algorithm utilizes local linear approximations of the non-convex penalty, achieving computational efficiency and algorithmic stability. Simulation studies confirm the robustness of our method across various settings, and applications to real-world datasets demonstrate its ability to recover true covariance structures accurately. Our framework offers a theoretically sound and practically effective solution for extreme scenarios when $p \gg n$.

Modeling and clustering of heterogeneous multivariate categorical sequences

Yingying Zhang and Volodymyr Melnykov

Clustering algorithms for quantitative data have been explored in literature extensively. However, many real-life applications involve qualitative data. The range of clustering procedures available in this framework is very limited. Recently, categorical sequences attracted the attention of researchers, and several promising methods were developed for univariate sequences. However, observations are often utilized in the form of multivariate categorical sequences. Currently existing methods either impose restrictive and often unrealistic assumptions or reformulate multivariate sequences as univariate ones with a higher number of states, which is often computationally prohibitive even in low-dimensional cases. The contribution of this paper is twofold. First, we explore the concept of inter-sequence transitions and develop a method for modeling their probabilities. Second, we extend the proposed model to the case of heterogeneity, which allows handling challenging real-life scenarios. As we demonstrate through the series of simulation studies, the developed mixture shows good modeling and clustering performance. The applications of the methodology to stylometry and British Household Panel Survey data produce interesting and meaningful results.

Multiscale Asymptotic Normality in Quantile Regression: Hilbert Matrices and Polynomial Designs

AHMED EL-GHINI
Mohammed V Univ in Rabat, Morocco

(Joint Paper with Azzouz Dermoune and Said Maanan)

Abstract:

This paper investigates the asymptotic properties of quantile regression estimators in linear models, with a particular focus on polynomial regressors and robustness to heavy-tailed noise. Under independent and identically distributed (i.i.d.) errors with continuous density around the quantile of interest, we establish a general Central Limit Theorem (CLT) for the quantile regression estimator under normalization using (Δ_n^{-1}), yielding asymptotic normality with variance $(\tau(1-\tau)/f^2(0)\cdot D_0^{-1})$. In the specific case of polynomial regressors, we show that the design structure induces a Hilbert matrix in the asymptotic covariance, and we derive explicit scaling rates for each coefficient. This generalizes Pollard's and Koenker's earlier results on LAD regression to arbitrary quantile levels $(\tau \in (0,1))$. We also examine the convergence behavior of the estimators and propose a relaxation of the standard CLT-based confidence intervals, motivated by a theoretical inclusion principle. This relaxation replaces the usual $(T^{j+1/2})$ scaling with (T^{α}), for $(\alpha < j+1/2)$, to improve finite-sample coverage. Through extensive simulations under Laplace, Gaussian, and Cauchy noise, we validate this approach and highlight the improved robustness and empirical accuracy of relaxed confidence intervals. This study provides both a unifying theoretical framework and practical inference tools for quantile regression under structured regressors and heavy-tailed disturbances.

Statistical learning does not always entail knowledge

DANIEL ANDRES DIAZ-PACHON
University of Miami

ABSTRACT:

In this talk, we study learning and knowledge acquisition (LKA) of an agent about a proposition that is either true or false. We use a Bayesian approach, where the agent receives data to update his beliefs about the proposition according to a posterior distribution. The LKA is formulated in terms of active information, with data representing external or exogenous information that modifies the agent's beliefs. It is assumed that data provide details about a number of features that are relevant to the proposition. We show that this leads to a Gibbs distribution posterior, which is in maximum entropy relative to the prior, conditioned on the side constraints that the data provide in terms of the features. We demonstrate that full learning is sometimes not possible and full knowledge acquisition (KA) is never possible when the number of extracted features is too small. We also distinguish between primary learning (receiving data about features of relevance for the proposition) and secondary learning (receiving data about the learning of another agent). We argue that this type of secondary learning does not represent true KA. Our results have implications for statistical learning algorithms, and we claim that such algorithms do not always generate true knowledge. The theory is illustrated with several examples.

Design-Based Causal Inference with Missing Outcomes: Missingness Mechanisms, Imputation-Assisted Randomization Tests, and Covariate Adjustment

Siyu Heng (New York University)

Abstract

Design-based causal inference, also known as randomization-based or finite-population causal inference, is one of the most widely used causal inference frameworks, largely due to the merit that its validity can be guaranteed by study design (e.g., randomized experiments) and does not require assuming specific outcome-generating distributions or super-population models. Despite its advantages, design-based causal inference can still suffer from other issues, among which outcome missingness is a prevalent and significant challenge. This work systematically studies the outcome missingness problem in design-based causal inference. First, we propose a general and flexible outcome missingness mechanism that can facilitate finite-population-exact randomization tests of no treatment effect. Second, under this general missingness mechanism, we propose a general framework called "imputation and re-imputation" for conducting randomization tests in design-based causal inference with missing outcomes. We prove that our framework can still ensure finite-population-exact type-I error rate control even when the imputation model was misspecified or when unobserved covariates or interference exists in the missingness mechanism. Third, we extend our framework to conduct covariate adjustment in randomization tests and construct finite-population-valid confidence regions with missing outcomes. Our framework is evaluated via extensive simulation studies and applied to a large-scale randomized experiment.

Analysis on Dynamics of Multi-Layer Neural Network with High-Dimensional Structure

MASAAKI IMAIZUMI
University of Tokyo

We introduce several topics related to the connection between statistics, machine learning, and dynamical systems. The first topic concerns the learning of the XOR function by a neural network with simultaneous training. Feature learning, where the first layer of a multilayer neural network learns important structures from the data, has been recognized as a key advantage of deep networks. However, demonstrating this theoretically requires specific techniques, such as sequential learning algorithms. This study shows that it is possible to learn the XOR function even when both layers of a two-layer neural network are updated simultaneously. To establish this result, we characterize the fine-grained tracking of neuron variability, which differs from conventional dynamical analyses based on optimization. The second topic discusses statistical inference for high-dimensional parameters, specifically the evaluation of the uncertainty of estimators. Inference for high-dimensional parameters often employs a framework that derives distributions using limit theorems for dynamical algorithms. In this study, we extend this approach to GLMs and single-index models, a representative example of nonlinear models, and demonstrate that statistical inference for high-dimensional parameters can be performed within this setting.

On a class of shrinkage estimators in high and ultra-high dimensional data

SEVEREIN NKURUNZIZA

Abstract: In this investigation, we study the risk performance of a class of shrinkage estimators for the mean parameter matrix of a scale mixture of multivariate normal distribution with an unknown variance- covariance matrix. We generalize the existing results in four ways. First, we consider an estimation problem which is encloses as a special case the one about the vector parameter. Second, we weaken the conditions about the distribution of the random sample. Third, we present a class of James-Stein matrix estimators and, we establish a necessary and a sufficient condition for any member of the proposed class to have a finite risk function. Fourth, we derive the conditions for the proposed class of estimators to dominate the classical estimator. Beyond these interesting findings, the additional novelty consists in the fact that, the derived results hold in the classical case as well as in the context of high and ultra-high dimensional data.

Cancer Human Disease Networks (cHDNs) via Deep Learning SEER-Medicare

SHUANGGE MA
Yale School of Public Health

Abstract: In this investigation, we study the risk performance of a class of shrinkage estimators for the mean parameter matrix of a scale mixture of multivariate normal distribution with an unknown variance- covariance matrix. We generalize the existing results in four ways. First, we consider an estimation problem which is encloses as a special case the one about the vector parameter. Second, we weaken the conditions about the distribution of the random sample. Third, we present a class of James-Stein matrix estimators and, we establish a necessary and a sufficient condition for any member of the proposed class to have a finite risk function. Fourth, we derive the conditions for the proposed class of estimators to dominate the classical estimator. Beyond these interesting findings, the additional novelty consists in the fact that, the derived results hold in the classical case as well as in the context of high and ultra-high dimensional data.
Cancer Human Disease Networks (cHDNs) via Deep Learning SEER-Medicare

Cancer patients often also suffer from other disease conditions. For more effective management and treatment, it is crucial to understand the “big picture”. Human disease network (HDN) analysis provides an effective way for describing the interrelationships among diseases. The goal of this study is to mine the SEER-Medicare data and construct the HDNs for major cancer types for the elderly. For network construction, we adopt penalized deep neural networks (pDNNs). The DNNs can be more flexible than the regression-based and other analyses, and penalization can effectively distinguish important disease interconnections from noises. As a “byproduct”, we establish the asymptotic properties of pDNNs. The constructed cHDNs are carefully analyzed in terms of node, module, and network properties.

Nonlinear forecasting with many predictors using mixed data sampling kernel ridge regression models

KRISTOFER MANSSON
Jonkoping University

Abstract:

Policy institutes such as central banks need accurate forecasts of key measures of economic activity to design stabilization policies that reduce the severity of economic fluctuations. Therefore, this paper develops a kernel ridge regression estimator in a mixed data sampling framework. Kernel ridge regression can handle many predictors with a nonlinear relationship to the target variable. Consequently, it has potential to improve the currently used principal component-based methods when the economic data follow a nonlinear factor structure. In a Monte Carlo study, we show that the kernel ridge regression approach is superior in terms of mean square error and is more robust than principal component-based methods to different nonlinear data generating processes. By using a dataset consisting of 24 economic indicators, we forecast Swedish gross domestic production. The results confirm the superiority of the kernel ridge regression approach. Therefore, we suggest that policy institutes consider the use of kernel-based approaches when forecasting key measures of economic activity.

Observing Long-range Dependent data at non regular time intervals

HAYE MOHAMEDOU OULD
Carleton University

We study the effect of observing a long memory stationary process at irregular time points via a renewal process. We establish a sharp difference in the asymptotic behaviour of the self- normalized sample mean of the observed process depending on the renewal process. In particular, we show that if the renewal process has a moderate heavy tail distribution, then the limit is a so- called Normal Variance Mixture (NVM) and we characterize the randomized variance part of the limiting NVM as an integral function of a Levy stable motion. Otherwise, the normalized sample mean will be asymptotically normal.

Estimation of Out-of-Sample Sharpe Ratio for High Dimensional Portfolio Optimization

Abstract

Estimation of Out-of-Sample Sharpe Ratio for High Dimensional Portfolio Optimization

XURAN MENG

University of Michigan

Portfolio optimization aims at constructing a realistic portfolio with significant out-of-sample performance, which is typically measured by the out-of-sample Sharpe ratio. However, due to in-sample optimism, it is inappropriate to use the in-sample estimated covariance to evaluate the out-of-sample Sharpe, especially in the high dimensional settings. In this paper, we propose a novel method to estimate the out- of-sample Sharpe ratio using only in-sample data, based on random matrix theory. Furthermore, portfolio managers can use the estimated out-of-sample Sharpe as a criterion to decide the best tuning for constructing their portfolios. Specifically, we consider the classical framework of Markowits mean-variance portfolio optimization under high dimensional regime of $p/n \to c \in (0,\infty)$, where p is the portfolio dimension and n is the number of samples or time points.

We propose to correct the sample covariance by a regularization matrix and provide a consistent estimator of its Sharpe ratio. The new estimator works well under either of the following conditions: (1) bounded covariance spectrum, (2) arbitrary number of diverging spikes when $c < 1$, and (3) fixed number of diverging spikes with weak requirement on their diverging speed when $c \geq 1$. We can also extend the results to construct global mini- mum variance portfolio and correct out-of-sample efficient frontier. We demonstrate the effectiveness of our approach through comprehensive simulations and real data experiments. Our results highlight the potential of this methodology as a useful tool for portfolio optimization in high dimensional settings.

D-CDLF: Decomposition of Common and Distinctive Latent Factors for Multi-view High Dimensional Data

Hai Shu

New York University

Abstract: Modern biomedical studies often collect multi-view data, that is, multiple types of data measured on the same set of objects. A typical approach to the joint analysis of multiple high dimensional data views is to decompose each view's data matrix into three parts: a low-rank common-source matrix generated by common latent factors of all data views, a low-rank distinctive-source matrix generated by distinctive latent factors of the corresponding data view, and an additive noise matrix. Existing decomposition methods only focus on the uncorrelatedness between the common latent factors and distinctive latent factors but inadequately address the equally necessary uncorrelatedness between distinctive latent factors from different data views. We propose a novel decomposition method, called Decomposition of Common and Distinctive Latent Factors (D-CDLF), to effectively achieve both types of uncorrelatedness. Consistent estimators of our D-CDLF method are established with good finite sample numerical performance. The superiority of D-CDLF over state-of-the-art methods is also corroborated in simulations and real-world data analysis.

FoSeGNN: Cell Type-Specific Protein Prediction via Graph Neural Networks

YUYING XIE
Michigan State University

Abstract: Proteins are functionally critical but costly to measure compared with mRNA. Rare cell types, often key to disease, are hard to model. FoSeGNN leverages cell type relationships and graph neural networks to predict protein abundance from RNA level, improving performance and robustness, especially in rare populations.

Enhancing regression in high dimensions: Adaptive ridge estimators for constrained models with spherically symmetric errors

Idir Ouassou

Cadi Ayyad University, National School of Applied Sciences, Morroco

Corresponding author : i.ouassou@uca.ac.ma

Abstract:

We investigate adaptive ridge regression estimators within a general linear model framework, particu- larly focusing on high-dimensional data where the number of predictors p may exceed the number of observations n. The model is defined as:

$$Y A\beta + \varepsilon,$$

where Y is an $n \times 1$ vector of observations, A is the known $n \times p$ design matrix, β is the $p \times 1$ vector of unknown regression coefficients, and ε is an $n \times 1$ vector of experimental errors with a homogeneous spherically symmetric distribution. In high-dimensional settings, the least squares estimator (LSE) of β is:

$$\hat{\beta} = (A^t A)^{-1} A^t Y,$$

is often unstable or undefined due to the singularity or near-singularity of A^tA. To address these chal- lenges, Hoerl and Kennard (1970) introduced the ridge estimator:

$$\hat{\beta}^{\text{ridge}}(k) = (A^t A + kI_p)^{-1} A^t Y,$$

where k is a positive constant. This estimator stabilizes the estimates by adding a regularization term of A^tA, which is particularly beneficial in high-dimensional scenarios.

In this study, we propose adaptive ridge estimators that outperform the LSE under a general quadratic loss function:

$$L(\delta, \beta) = (\beta - \delta)^t B^{-1} (\beta - \delta),$$

where B is an arbitrary $p \times p$ positive definite matrix. By transforming the model into a canonical form, we derive a class of minimax estimators for the parameter θ, a transformation of β. These estimators are applied to generalized ridge regression when the design matrix A is of full rank or in high-dimensional settings where $p > n$. We focus on scenarios where all components of β are non-negative and demonstrate that our estimators dominate the LSE under specific conditions.

Our findings reveal that the proposed adaptive ridge estimators significantly enhance regression accuracy, particularly in high-dimensional contexts where non-negativity constraints are imposed on the regression parameters. This approach holds great promise for applications in econometrics, genomics, and other fields grappling with high-dimensional data, multicollinearity, and parameter constraints.

Karim Oualkacha[1,*], Abdellah Atanane[1,2], Abdallah Mkhadri[2]

[1]*Department of Mathematics, Université du Québec à Montréal, Montréal, Québec, Canada,*

[2]*Department of Mathematics, Cadi Ayyad University, Marrakesh, Morocco*

[*]*Presenting author (e-mail: oualkacha.karim@uqam.ca*)

Abstract

Large-scale data analysis poses substantial challenges, including storage constraints, the pres-ence of skewed distributions, heteroskedastic variance structures, and intricate dependencies between outcomes and high-dimensional covariates. Distributed storage offers an effective so- lution by alleviating the burden on any single system. The L^k-quantile regression framework, which extends the concepts of both quantile and expectile regression, has gained popularity due to its robustness and effectiveness, especially for $1 < k \leq 2$. In the context of large- scale distributed data, this work investigates penalized L^k-quantile regression incorporating the smoothly clipped absolute deviation (SCAD) and adaptive LASSO (aLASSO) penalties, both of which are effective in identifying heteroskedasticity in high-dimensional settings. To address the distributed nature of the data, a communication-efficient surrogate likelihood—termed the CSL^k (Communication-efficient Surrogate for L^k-quantile regression)—is proposed as a proxy for the global loss function. A distributed optimization algorithm, based on the parallel alternating direction method of multipliers (ADMM), is developed to minimize the proposed penalized CSL^k objective function. Oracle properties of the resulting estimators are established under suitable choices of tuning parameters. The finite-sample performance of the method is evaluated through comprehensive simulation studies and its practical utility is demonstrated through an application to a real dataset.

Latent Space Modeling for Human Disease Network with Temporal Variations: Analysis of Medicare Data

Hao Mei

Renmin University of China

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 been shown as powerful in other network analysis contents, has unique and insightful interpretations, but has been limitedly adopted 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. 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 Medicare data -- one of the most powerful medical claims data, we analyze the admission records of 133 million hospital inpatient treatments from January 2008 to December 2019. 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 understanding and planning medical resources.

Predicting Exoplanetary Atmospheric Absorption Spectra Using High-Dimensional AI-Based Modeling

Vasuda Trehan[1], Kevin H. Knuth[2], and M. J. Way [3,4]

[1] Department of Information Science, University at Albany, Albany NY, USA; vtrehan@albany.edu

[2] Department of Physics, University at Albany, Albany NY, USA; kknuth@albany.edu
[3] NASA Goddard Institute for Space Studies, New York NY, USA; michael.way@nasa.gov
[4] Theoretical Astrophysics, Department of Physics & Astronomy, Uppsala University, Sweden

Abstract:

The research presents a scalable machine learning model and Bayesian inference framework that will predict atmospheric absorption spectra across a high-dimensional space of data with up to 30 planetary parameters such as mass, radius, surface temperature, stellar type, orbital parameters, and more. Each contributes to the complexity of the modeling challenge. Traditional interpolation techniques, such as Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) or splines, quickly become inefficient as dimensionality increases, limiting their applicability in modern exoplanet research.

This research introduces a scalable, non-parametric Bayesian framework using Gaussian Process Regression (GPR) to serve as a forward model for predicting exoplanetary spectra. Our approach integrates various kernel function optimizations to better capture complex, non-linear spectral variations, and Bayesian Adaptive Exploration to prioritize informative sampling, reducing the number of required simulations without sacrificing accuracy. We also incorporate nested sampling to efficiently explore the posterior over kernel hyperparameters and spectral outputs in high-dimensional settings. Using simulated data from NASA's ROCKE-3D General Circulation Model (GCM), we ensure that our model achieves accurate spectral predictions across a broad range of exoplanetary conditions, including varying masses, radii, stellar types, surface temperatures, and orbital characteristics. The framework not only predicts spectra but can also invert observations to estimate underlying planetary properties which can indicate the planet's habitability.

Zheshi Zeng

University of Michigan

Abstract

High-dimensional matrix-valued time series are increasingly prevalent in modern applications, such as asset return matrices in finance or city taxi traffic data. Forecasting future observations in such settings is of great importance, as is providing accurate and robust uncertainty quantification. However, the high dimensionality and complex temporal and cross-sectional dependencies in these data pose significant challenges for both prediction and its uncertainty assessment. Matrix factor models are commonly employed for dimension reduction, yet they often lack predictive power unless the dynamics of the latent factors are explicitly modeled. Moreover, uncertainty quantification becomes even more challenging when prediction models are subject to misspecification or bias.

To improve prediction accuracy, we adopt a two-component dynamic matrix factor model, which extends the standard matrix factor framework by incorporating a matrix autoregressive process for the latent factors. This structure enables effective dimension reduction while equipping the model with forecasting capabilities. It captures the underlying low-rank temporal dynamics while preserving the matrix structure of the observations. This modeling framework is particularly useful for forecasting high-dimensional matrix time series in applications such as regional economic indicators or citywide transportation flows.

To quantify the uncertainty in predictions generated from the two-component dynamic matrix factor model, we develop a robust conformal prediction procedure. Conformal prediction offers prediction intervals with valid coverage guarantees even under model misspecification. This robustness is especially advantageous when nuisance parameters, such as the rank of the factor matrix, are difficult to specify correctly. We propose constructing conformal prediction intervals with **conditional coverage**, focusing on forecasting the current time point—a particularly challenging task due to the violation of exchangeability. To address this, we incorporate the iFusion strategy, calibrating prediction intervals using only "similar" historical instances. Furthermore, we handle the dependence of the time series using weak dependence structures, including mixing conditions and physical dependence measures. We establish theoretical coverage guarantees under these assumptions and demonstrate the robustness and effectiveness of our method through numerical studies and comparisons with existing conformal prediction techniques.

Recent developments in environmental research with high-dimensional data

Professor Azizur Rahman*

*School of Computing, Mathematics and Engineering, Charles Sturt University, Wagga Wagga, NSW 2678, Australia

Email: azrahman@csu.edu.au

Abstract: Climate change, the environment, and human activities have complex interactions. For example, climate change-related events such as droughts, cyclones, and floods have significant effects on water, the environment, and agriculture. Water quality has a direct impact on both human health and the environment. Given the continued technological transformation and advancement in the data-gathering process, the integration of various data sources and large datasets is increasing dynamically. As a result, advanced-level analysis and modelling strategies are essential in multidisciplinary research to understand the individual activities that occur within very complex behavioral, socio-economic and ecological systems. However, the scales at which practical models can be developed and the subsequent issues they can resolve are often restricted by our inability or challenges to effectively understand data that mimic interactions at the finest spatial, temporal, or organizational resolutions. This talk will present recent advancements in environmental data analysis and interpretation, with a particular focus on the selection processes for variables or parameters and validation techniques. It will also report empirical findings from one of our water quality research projects.

A data-driven Bayesian variable selection method for variable selection and estimation of sparsity with applications in biomedical data

Jabed Tomal

Associate Professor, Department of Mathematics and Statistics, Thompson Rivers University, British Columbia, Canada

Abstract: We proposed a flexible and data-driven hierarchical Bayesian variable selection method that simultaneously selects variables and estimates sparsity using a beta-Bernoulli prior. We compared the results of our model to those of a standard Bayesian variable selection method employing a non-hierarchical prior with a fixed selection probability. For our study, we defined sparsity as the proportion of relevant variables among all the variables in a dataset. We applied our method to two datasets: Coronary Artery Disease (CAD) and Breast Cancer (BC). Our method resulted in a smaller Deviance Information Criterion (DIC) than the fixed-prior model, indicating improved fit to the data while penalizing for complexity. Moreover, our proposed method yielded more precise (i.e., narrower) credible intervals, especially for the significant predictors, than the method with fixed-prior selection probability. In CAD, the variables such as asymptomatic chest pain, male sex, elevated fasting blood sugar, exercise-induced ST depression, and high cholesterol were found to be positively associated with CAD. On the other hand, upsloping or downsloping ST slope and higher maximum heart rate were negatively associated. For BC dataset, our method identified 19 genes, compared to 14 selected by the model with fixed prior. Notably, SERPINA1, GSTT2, and RGS4, previously reported in cancer literature, were detected by our model, along with five newly selected genes such as PARP3, ABCG4, CCDC57, SH3GL3 and TMEM105. The estimated sparsity values were 0.88 and 0.73 for CAD and BC, respectively.

Continuous-time Causal Inference with Marked Point Process Weights: An Example on Sodium-Glucose Co-Transporters 2 Inhibitor Medications and Urinary Tract Infection

Sumeet Kalia, PhD

Assistant Professor

Department of Statistics

University of Manitoba

Treatment-confounder feedback is present in time-to-recurrent or longitudinal event analysis when time-dependent confounders are themselves influenced by previous treatments. Conventional models produce misleading statistical inference of causal effects due to conditioning on these factors on the causal pathway. Marginal structural models are often applied to quantify the causal treatment effect, estimated using longitudinal weights which mimic the randomization procedure by balancing the covariate distributions across the treatment groups. The weights are usually constructed in discrete time intervals which is appropriate if the follow-up visits are scheduled and regular. However, in primary care visit times can be irregular and informative, and the treatment history consists of duration and doses. This can be modeled through a continuous-time marked point process. We constructed a continuous-time marginal structural model to estimate the effect of cumulative exposure of Sodium-Glucose co-Transporters 2 Inhibitor (SGLT-2i) medications on time-to-recurrent urinary tract infection (UTI). We used a cohort of type II diabetes patients with chronic kidney disease, and constructed a marked point process which characterized the recurrent flare episodes of primary care visits (i.e., point process) with marks for the multinominal dose (none, low, high) of SGLT-2i medications and recurrent episodes of UTI. We applied the stabilized and optimal treatment weights to estimate the hypothesized causal effect. Our results are concordant with earlier findings in which the recurrent episodes of UTI did not increase when patients were prescribed low dose or high dose of SGLT-2i medications.

Heterogeneous Random Effects Covariance Matrix in High-Dimensional Longitudinal Data

ERFANUL HOQUE
University of Saskatchewan

Longitudinal data often contain missing responses and mismeasured covariates. While generalized linear mixed model (GLMM) frameworks are commonly adopted to simultaneously analyze such data, the random effect covariance matrix in these models is often assumed constant across subjects and is restricted due to its high dimensionality and the positive definite constraint. However, the random effect covariance matrix may vary by measured covariates in many situations, and ignoring this heterogeneity can lead to biased estimates of model parameters.

In this talk, we present an approach to incorporate potential heterogeneity in the random effects covariance matrix. The proposed approach utilizes a modified Cholesky decomposition, allowing the random effects covariance matrix to depend on covariates. This decomposition also guarantees the positive definiteness of the random effects covariance matrix. The performance of the proposed approach is evaluated through simulation studies and demonstrated using longitudinal data from the Framingham Heart Study.