2022 Bohrer Memorial Student Workshop in Statistics

ILLINOIS

I-Hotel & Conference Center
April 23, 2022

Welcome from the Department Chair

Professor Bo Li
Department Chair

Bohrer Workshop Committee:

Dave Zhao – Chair
Daniel Eck
Naveen Narisetty
Xiaohui Chen

Special thanks to:

Melissa Banks
Aaron Thompson
I-Hotel & Conference Center Event Services
University Catering Services

Welcome to the 2022 Robert E. Bohrer Memorial Student Workshop in Statistics! We are proud to feature two outstanding plenary lectures, by Professor Vijay Nair of the University of Michigan and Dawn Woodard, Distinguished Scientist at LinkedIn and a number of invited research presentations by Ph.D. students at the University of Illinois at Urbana-Champaign. Our premier annual event, the Bohrer Workshop, was established in 1994 with support from donations by faculty, students and friends of Professor Robert E. Bohrer to honor his inspiring life and career, and his commitment to students.

Meetings Conduct Policy

The Department of Statistics is committed to providing an atmosphere that encourages the free expression and exchange of ideas. Consistent with this commitment, it is the policy of the department that all participants in University of Illinois activities will enjoy a welcoming environment free from unlawful discrimination, harassment, and retaliation. All participants in university activities also agree to comply with all rules and conditions of the activities, which are subject to change without notice.

Please read the complete Code of Conduct Policy:
https://www.ethics.uillinois.edu/compliance/university_code_of_conduct

2022 Robert Bohrer Memorial Student Workshop in Statistics Schedule
April 23, 2022

8:30 - 9:00	•Breakfast - Alma Mater Room
9:00 - 9:10	•Opening Remarks, Bo Li – Quad Room
9:10 – 9:50	•Speed Session One – Quad Room
9:50 – 10:00	•Break – Alma Mater Room
10:00 – 11:00	•Bohrer Lecture, Vijay Nair (remote)
11:00 – 11:15	•Break – Alma Mater Room
11:15 – 12:00	•Speed Session Two – Quad Room
12:00 – 1:00	•Lunch – Alma Mater Room
1:00 – 1:45	•Paper Session One – Quad Room
1:45 – 2:00	•Break – Alma Mater Room
2:00 – 2:30	•Paper Session Two
2:30 – 2:40	•Break – Alma Mater Room
2:40 – 3:45	•Norton Session
3:45 – 4:00	•Break
4:00 – 5:00	•Wijsman Lecture, Dawn Woodard (remote)
5:00 – 6:00	•Student Award Presentation and Student Poster Presentation
6:00 – 8:00	•Dinner – Alma Mater Room

Wijsman Lecture

Dawn Woodard

- Distinguished Scientist, LinkedIn

Talk Title: Statistical Efficiency of Travel Time Prediction

Talk Abstract: Modern mobile applications such as navigation services and ride-hailing platforms rely heavily on geospatial technologies, most critically predictions of the time required for a vehicle to traverse a route. Two major categories of prediction methods are segment-based approaches, which predict travel time at the level of road segments and then aggregate across the route, and trip-based approaches, which use generic information about the trip such as origin, destination, and route length to predict travel time. Though various forms of these methods have been developed and used, there has been no rigorous theoretical comparison of the accuracy of these two approaches, and empirical studies have in many cases drawn opposite conclusions.

We fill this gap by comparing these two approaches in terms of their statistical efficiency. We study how quickly the prediction error shrinks as the number of trip observations and the size of the road network grow, and prove that segment-based approaches are more efficient for large road networks. We complement our theoretical results with empirical evidence using ride-sharing traffic data. Our work highlights that the accuracy of travel time prediction is driven not just by the sophistication of the model, but also the spatial granularity at which those methods are applied.

Bohrer Lecture

Vijay Nair

- Professor, University of Michigan
- R&D Managing Director, Wells Fargo Bank

Talk
Title: Machine Learning: Applications, Opportunities, and Challenges

Talk Abstract: Machine learning (ML) algorithms are increasingly used in business, industry, and technology over the last few decades. The first part of this presentation will provide an overview of the developments and applications, with a focus on supervised learning methods applied in finance and banking. The second part of the presentation will describe opportunities and challenges as well as some of the ongoing research to mitigate the challenges. In particular, I will discuss the need for interpretability of ML results and recent developments on the use of surrogate models and inherently interpretable ML algorithms. If time permits, I will also describe applications in the analysis of text data using natural language processing techniques.

Norton Lecture

Honorable Mention

Yangfan Zhang

- PhD Student, UIUC

Talk Title: Adaptive Testing for High-Dimensional Data

Talk Abstract: In this article, we propose a class of L_q-norm based U-statistics for a family of global testing problems related to high-dimensional data. This includes testing of mean vector and its spatial sign, simultaneous testing of linear model coefficients, and testing of component-wise independence for high-dimensional observations, among others. Under the null hypothesis, we derive asymptotic normality and independence between L_q-norm based U-statistics for several qs under mild moment and cumulant conditions. A simple combination of two studentized L_q-based test statistics via their p-values is proposed and is shown to attain great power against alternatives of different sparsity. Our work is a substantial extension of He et al. (2021), which is mostly focused on mean and covariance testing, and we manage to provide a general treatment of asymptotic independence of L_q-norm based U-statistics for a wide class of kernels. To alleviate the computation burden, we introduce a variant of the proposed U-statistic by using the monotone indexes in the summation, resulting in a U statistic with asymmetric kernel. A dynamic programming method is introduced to reduce the computational cost from $O(n^{qr})$, which is required for the calculation of full U-statistic, to $O(n^r)$ where r is the order of the kernel. Numerical studies further corroborate the advantage of the proposed test as compared to some existing competitors.

Norton Lecture

Joshua Loyal

- PhD Student, UIUC

Talk Title: An Eigenmodel for Dynamic Multilayer Networks

Talk Abstract: Network (or graph) data is at the heart of many modern data science problems: disease transmission, community dynamics on social media, international relations, and others. In this talk, I will elaborate on my research in statistical inference for complex time-varying networks. I will focus on dynamic multilayer networks, which frequently represent the structure of multiple co-evolving relations. Despite their prevalence, statistical models are not well-developed for this network type.

Here, I propose a new latent space model for dynamic multilayer networks. The key feature of this model is its ability to identify common time-varying structures shared by all layers while also accounting for layer-wise variation and degree heterogeneity. I establish the identifiability of the model's parameters and develop a structured mean-field variational inference approach to estimate the model's posterior, which scales to networks previously intractable to dynamic latent space models. I apply the model to two real-world problems: discerning regional conflicts in a data set of international relations and quantifying infectious disease spread throughout a school based on the student's daily contact patterns.

Bohrer Memorial Student Workshop

Speed Session One Talk Titles

9:10 – 9:50

Anamitra Chaudhuri

- *Sequential Change Detection of a Correlation Structure under a Sampling Constraint*

Robert Garrett

- *Sliced Elastic Distance for Validation of Global Climate Models*

Shishuang He

- *Identifiability and Estimation of Mixed Membership Stochastic Blockmodels*

Sophie Larsen

- *Host heterogeneity and SARS-CoV-2: lessons learned from vaccine uptake across socioeconomic groups*

Zihe Liu

- *Bayesian Selection and Smoothing for Generalized Additive Model*

Adriana Morales Miranda

- *Climate drivers and seasonality of Respiratory Syncytial Virus in the Southern Hemisphere*

Full talk abstracts can be viewed on pages 12-21

Bohrer Memorial Student Workshop

Speed Session Two Talk Titles

11:15 – 12:00

Diptarka Saha

- *Efficient Variational Approach to Sparse BNN for Model Compression*

Rong Tang

- *Minimax Rate of Distribution Estimation on Unknown Submanifold under Adversarial Losses*

Adam Tonks

- *West Nile Virus Forecasting using Graph Neural Networks*

Mengchen Wang

- Asynchronous Changepoint Estimation for Spatially Correlated Functional Time Series

Tianning Xu

- *Variance estimation of random forests*

Rentian Yao

- *Nonparametric estimation problem of a distribution-state dependent drift vector field*

Yubo Zhuang

- *Sketch-and-lift: scalable subsampled semidefinite program for K-means clustering*

Full talk abstracts can be viewed on pages 12-21

Bohrer Memorial Student Workshop

Paper Session Talk Titles

1:00 – 2:30

Alton Barbehenn

- *Asyptotically Optimal Simultaneous Gaussian Mean Estimation with Nonparametric Regression*

Sayan Chakrabarty

- *Subsampling Based Community Detection in Large Networks*

Sarah Formentini

- *Confidence Band Estimation for Survival Random Forests*

Shen Yan

- *Comparing baseball players across eras via the novel Full House Model*

Austin Warner

- *Sequential Change Diagnosis Revisited*

Full talk abstracts can be viewed on pages 12-21

Bohrer Memorial Student Workshop

Norton Session Talk Titles

2:45 – 3:45

Yangfan Zhang

- Honorable Mention
- *Adaptive Testing for High-Dimensional Data*

Joshua Loyal

- Norton Prize
- *An Eigenmodel for Dynamic Multilayer Networks*

Full talk abstracts can be viewed on pages 12-21

Bohrer Memorial Student Workshop

Full Abstracts

Anamitra Chaudhuri

Sequential Change Detection of a Correlation Structure under a Sampling Constraint

- The problem of sequentially detecting a change in the correlation structure of multiple Gaussian information sources is considered when it is possible to sample only two of them at each time instance. It is assumed that all sources are initially independent and that at least two of them become positively correlated after the change. The problem is to stop sampling as quickly as possible after the change while controlling the false alarm rate and without assuming any prior information on the number of sources that become correlated. A joint sampling and change-detection rule is proposed and is shown to achieve the smallest possible worst-case conditional expected detection delay among all processes that satisfy the same constraints, to a first-order approximation as the false alarm rate goes to 0, for any possible number of post-change correlated sources.

Robert Garrett

Sliced Elastic Distance for Validation of Global Climate Models

- Global climate model validation is integral for ensuring climate models produce realistic climatologies. However, many post-hoc statistical evaluation methods rely on simplifying models that discard information and fail to distinguish between different sources of variability. Here, we introduce a functional data analysis approach for computing sliced amplitude and phase distances between spatiotemporal processes, analogous to the sliced Wasserstein distance. Because our method uses time-warping, which respects temporal ordering, we can more precisely quantify differences between climate models than the previous Wasserstein-based approach. Finally, we apply our method to compare the performance of CMIP5 vs. CMIP6 models in representing historical surface temperature and precipitation from 1979-2005.

Shishuang He

Identifiability of the Mixed Membership Stochastic Blockmodel

- The Mixed Membership Stochastic Blockmodel (MMSB) has been widely used in detecting overlapping communities for large network data. In MMSB, each node can have a mixed membership over multiple communities, instead of being associated with only one community. Despite its popularity, MMSB is known to be unidentifiable. Prior work on the identifiability of MMSB relies on the existence of pure nodes (i.e., nodes that belong to one and only one community), which is too strong in practice. In this paper, we establish a new set of identifiability conditions for MMSB, which is weaker than the pure node condition and has a geometric interpretation.

Bohrer Memorial Student Workshop

Full Abstracts

Sophie Larsen

Host heterogeneity and SARS-CoV-2: lessons learned from vaccine uptake across socioeconomic groups

- To understand infectious disease dynamics, one must identify social and biological sources of host heterogeneity. The human host is highly complex, not just at the molecular level but also in terms of social and behavioral dynamics. This has been especially clear for SARS-CoV-2, with disparities in infections, deaths, and vaccination across socioeconomic groups. Using publically available data from several countries, we examine temporal trends of vaccination rates across socioeconomic groups at different spatial levels and identify patterns that are similar across countries. Using these observed patterns, we propose a novel mechanistic model in the form of an ODE system that incorporates socioeconomic information through host heterogeneity, varying the force of infection and vaccination rate across socioeconomic groups. With this model, we can investigate which conditions minimize deaths vs. infections, and compare strategies across countries to optimize vaccination efforts, generating insights that can inform public health policy and human behavior

Zihe Liu

Bayesian Selection and Smoothing for Generalized Additive Model

- We propose a Bayesian procedure to fit a sparse additive model in high dimension. Our method can automatically select relevant features as well as their proper smoothness parameters. For computation, we derive a variational EM algorithm, which is tuning-free and can scale up with large data. We evaluate the empirical performance of our algorithm on both simulated and real examples and compare it with other variable selection alternatives for additive models.

Adriana Morales Miranda

Climate drivers and seasonality of Respiratory Syncytial Virus in the Southern Hemisphere

- Considerable attention has been put forth to further understand the mecha- nisms governing infectious disease dynamics worldwide, but some are still poorly understood. One of them is the role that climate factors play in modulating the seasonality and inter-annual variation of disease incidence, especially in the Southern Hemisphere. Recently available data from Chile provide a unique op- portunity to understand the effect of temperature and humidity on seasonality and inter-annual variation of Respiratory Syncytial Virus (RSV). We examine the association between climatic factors and positive confirmed cases by look- ing at the timing of RSV onset and inter-annual variability by implementing statistical techniques and mechanistic epidemiological models.

Bohrer Memorial Student Workshop

Full Abstracts

Diptarka Saha

Efficient Variational Approach to Sparse BNN for Model Compression

- Model Compression has drawn much attention within the deep learning community recently. Compressing a dense neural network offers many advantages including lower computation cost, deployability to devices of limited storage and memories, and importantly resistance to adversarial attacks. This may be achieved via node pruning or fully discarding certain input features. Here we demonstrate a novel strategy to emulate principles of Bayesian model selection in a deep learning setup. Given a fully connected Bayesian neural network with spike-and-slab priors trained via a variational algorithm we obtain the posterior inclusion probability for every node that typically gets lost. We employ these probabilities for pruning and feature selection on a host of simulated and real world benchmark data, and find evidence of better generalisability of the pruned model in all our experiments.

Rong Tang

Minimax Rate of Distribution Estimation on Unknown Submanifold under Adversarial Losses

- Statistical inference from high-dimensional data with low-dimensional structures has recently attracted lots of attention. In machine learning, deep generative modeling approaches implicitly estimate distributions of complex objects by creating new samples from the underlying distribution, and have achieved great success in generating synthetic realistic-looking images and texts. A key step in these approaches is the extraction of latent features or representations (encoding) that can be used for accurately reconstructing the original data (decoding). In other words, low-dimensional manifold structure is implicitly assumed and utilized in the distribution modeling and estimation. To understand the benefit of low-dimensional manifold structure in generative modeling, we build a general minimax framework for distribution estimation on unknown submanifold under adversarial losses, with suitable smoothness assumptions on the target distribution and the manifold. The established minimax rate elucidates how various problem characteristics, including intrinsic dimensionality of the data and smoothness levels of the target distribution and the manifold, affect the fundamental limit of high-dimensional distribution estimation. To prove the minimax upper bound, we construct an estimator based on a mixture of locally fitted generative models, which is motivated by the partition of unity technique from differential geometry and is necessary to cover cases where the underlying data manifold does not admit a global parametrization. We also propose a data-driven adaptive estimator that is shown to simultaneously attain within a logarithmic factor of the optimal rate over a large collection of distribution classes.

Bohrer Memorial Student Workshop

Full Abstracts

Adam Tonks

West Nile Virus Forecasting using Graph Neural Networks

- Within Illinois, the Illinois Department of Health overseas a program to monitor populations of mosquitoes infected with West Nile virus (WNV). Using the trap data collected from this program to forecast the location and abundance of infected mosquitoes could aid mosquito surveillance and abatement efforts within the state. Although a variety of machine learning methods have been previously utilized for similar problems, none have taken full consideration of the spatial dimension of such data sets. We show that graph neural networks (GNNs) can perform well with geospatial data that has been collected at individual points, without needing to augment the data set via interpolation. Furthermore, we describe a simple, generalisable method to determine the input graph vertices and edges using k-nearest neighbors, and to incorporate weather time series data. A baseline comparison of our model shows that its performance in generating trap-by-trap forecasts exceeds that of a variety of other models, including logistic regression, decision trees and other neural network architectures.

Mengchen Wang

Asynchronous Changepoint Estimation for Spatially Correlated Functional Time Series

- We propose a new solution under the Bayesian framework to simultaneously estimate mean based asynchronous changepoints in spatially correlated functional time series. Unlike previous methods that assume a shared changepoint at all spatial locations or ignore spatial correlation, our method treats changepoints as a spatial process. This allows our model to respect spatial heterogeneity and exploit spatial correlations to improve estimation. Our method is derived from the ubiquitous cumulative sum (CUSUM) statistic that dominates changepoint detection in functional time series. However, instead of directly searching for the maximum of the CUSUM based processes, we build spatially correlated two-piece linear models with appropriate variance structure to locate all changepoints at once. The proposed linear model approach increases the robustness of our method to variability in the CUSUM process, which, combined with our spatial correlation model, improves changepoint estimation near the edges. We demonstrate through extensive simulation studies that our method outperforms existing functional changepoint estimators in terms of both estimation accuracy and uncertainty quantification, under either weak and strong spatial correlation, and weak and strong change signals. Finally, we demonstrate our method using a temperature data set and a coronavirus disease 2019 (COVID-19) study.

Bohrer Memorial Student Workshop

Full Abstracts

Tianning Xu

Variance estimation of random forests

- Ensemble methods, such as random forests, are popular in applications due to their high predictive accuracy. Existing literature views a random forest prediction as an infinite-order incomplete U-statistic to quantify its uncertainty. However, these methods focus on a small subsampling size of each tree, which is theoretically valid but practically limited. This paper develops an unbiased variance estimator based on incomplete U-statistics, which allows the tree size to be comparable with the overall sample size, making statistical inference possible in a broader range of real applications. Simulation results demonstrate that our estimators enjoy lower bias and more accurate confidence interval coverage without additional computational costs. We also propose a local smoothing procedure to reduce the variation of our estimator, which shows improved numerical performance when the number of trees is relatively small. Further, we investigate the ratio consistency of our proposed variance estimator under specific scenarios. In particular, we develop a new "double U-statistic" formulation to analyze the Hoeffding decomposition of the estimator's variance.

Rentian Yao

Nonparametric estimation problem of a distribution-state dependent drift vector field

- This paper concerns the nonparametric estimation problem of the distribution-state dependent drift vector field in an interacting N-particle system. Observing single-trajectory data for each particle, we derive the mean-field rate of convergence for the maximum likelihood estimator (MLE), which depends on both Gaussian complexity and Rademacher complexity of the function class. In particular, when the function class contains α-smooth H{\"o}lder functions, our rate of convergence is minimax optimal on the order of $N^{-\frac{\alpha}{d+2\alpha}}$. Combining with a Fourier analytical deconvolution argument, we derive the consistency of MLE for the external force and interaction kernel in the McKean-Vlasov equation.

Bohrer Memorial Student Workshop

Full Abstracts

Yubo Zhuang

Sketch-and-lift: scalable subsampled semidefinite program for K-means clustering

- Semidefinite programming (SDP) is a powerful tool for tackling a wide range of computationally hard problems such as clustering. Despite the high accuracy, semidefinite programs are often too slow in practice with poor scalability on large (or even moderate) datasets. In this paper, we introduce a linear time complexity algorithm for approximating an SDP relaxed K-means clustering. The proposed sketch-and-lift (SL) approach solves an SDP on a subsampled dataset and then propagates the solution to all data points by a nearest-centroid rounding procedure. It is shown that the SL approach enjoys a similar exact recovery threshold as the K-means SDP on the full dataset, which is known to be information-theoretically tight under the Gaussian mixture model. The SL method can be made adaptive with enhanced theoretic properties when the cluster sizes are unbalanced. Our simulation experiments demonstrate that the statistical accuracy of the proposed method outperforms state-of-the-art fast clustering algorithms without sacrificing too much computational efficiency and is comparable to the original K-means SDP with substantially reduced runtime.

Alton Barbehenn

Asyptotically Optimal Simultaneous Gaussian Mean Estimation with Nonparametric Regression

- Simultaneous estimation of multiple parameters has received a great deal of recent interest, with applications in multiple testing, causal inference, and large-scale data analysis. Most approaches to simultaneous estimation use empirical Bayes methodology. Here we propose an alternative, completely frequentist approach based on nonparametric regression. We show that simultaneous estimation can be viewed as a constrained and penalized least-squares regression problem, so that empirical risk minimization can be used to estimate the optimal estimator within a certain class. We show that under mild conditions, our data-driven decision rules have asymptotically optimal risk that can match the best known convergence rates for this compound estimation problem. Our approach provides another perspective to understand sufficient conditions for asymptotic optimality of simultaneous estimation. Our proposed estimators demonstrate comparable performance to state-of-the-art empirical Bayes methods in a variety of simulation settings and our methodology can be extended to apply to many practically interesting settings.

Bohrer Memorial Student Workshop

Full Abstracts

Sayan Chakrabarty
Subsampling Based Community Detection in Large Networks

- Large networks are increasingly prevalent in scientific applications. Statistical analysis of such large networks become prohibitive due to exorbitant computation cost and high memory requirements. In this project, we develop a subsampling based divide-and-conquer algorithm, SONNET, for community detection in large networks. The algorithm splits the original network into multiple subnetworks with a common overlap and applies a suitable community detection algorithm on each subnetwork. The results from individual subnetworks are aggregated using a label matching method to get the final community labels. This method saves both memory and computation costs significantly as one needs to store and process only the smaller subnetworks. This method is also parallelizable. We derive a theoretical upper bound for the error rate of SONNET applied with any community detection algorithm. We also specialize the bound when SONNET is applied with spectral clustering on a stochastic blockmodel (SBM) and with spherical K-median spectral clustering on a degree corrected blockmodel (DCBM). We demonstrate the effectiveness of SONNET on a real-world network and on networks simulated from SBM and DCBM.

Sarah Formentini
Confidence Band Estimation for Survival Random Forests

- Random survival forest is a popular machine learning tool for modeling censored survival data. However, there is no statistically valid and computationally feasible approach for estimating the confidence band of a survival function currently. In this paper, we proposed an unbiased confidence band estimation by extending recent developments in infinite order incomplete U-statistics. This approach is computationally easy to implement and works well when the subsampling size of a tree is as large as half of the total training sample size. Numerical studies show that this method accurately estimates the confidence band of a survival function and achieves desired coverage rate. We apply this method to veterans' administration lung cancer data.

Bohrer Memorial Student Workshop

Full Abstracts

Yan Shen

Comparing baseball players across eras via the novel Full House Model

- We motivate a new framework that is appropriate for making cross-contextual comparisons of different components in an evolving system at different times. The systems that we consider involve rich data collection on both the quality of components in the system as well as understanding how components enter the system. Our methodology is a crystallization of the conceptual ideas put forward by Stephen Jay Gould. We name this methodology the Full House Model in his honor. The Full House Model works by balancing the performance of components within the system at any given set time and the number of components that are eligible to be included within the system at that time. We assume that each component has a latent ability which can be computed from this balancing act provided the system inclusion mechanism for components of the system is known. The distribution of components can be estimated from nonparametric probability distribution without any assumptions on the distribution of the system components. We demonstrate the utility of the Full House Model in an application of comparing baseball players' statistics across eras. We show that our approach yields defensible era-adjusted baseball statistics which properly balance how players performed against their peers and how many people were eligible to play in Major League Baseball (MLB). We compare our approach with existing approaches which we argue are not properly calibrated for the task of cross-era comparisons of baseball players. Our results reveal a radical reranking of baseball's greatest players that is consistent with what one would expect under a sensible uniform talent generation assumption. Most importantly, we found that the greatest African American and Latino players now sit atop the greatest all-time lists of historical baseball players.

Bohrer Memorial Student Workshop

Full Abstracts

Austin Warner

Sequential Change Diagnosis Revisited

- The problem of sequential change diagnosis is considered, where a sequence of independent random elements is accessed sequentially. At some unknown time there is an abrupt change in its distribution, and there are two main operational goals: to quickly detect the change and to accurately identify the post-change distribution among a finite set of alternatives. Two algorithms - one which we propose and one which we draw from sequential detection literature - are shown to control the worst-case conditional probability of false isolation under some conditions, and at the same time minimize Lorden's criterion for the detection delay, for every possible post-change distribution, to a first-order asymptotic approximation as both the worst-case probability of false isolation and the false alarm rate go to zero, but with the false alarm rate going to zero at a faster rate. Specifically, it is shown that these properties are satisfied under some conditions by a standard change detection algorithm that raises an alarm as soon as the CuSum statistic that corresponds to one of the post-change alternatives exceeds a certain threshold, when the post-change distributions are more distant (in a Kullback-Leibler divergence sense) from one another than from the pre-change distribution. More importantly, a novel, recursive algorithm is proposed and is shown to satisfy the above properties under any pre-change and post-change distributions, subject to similar conditions. The insights of these theoretical results are illustrated by simulation studies under various setups, where these algorithms are compared to each other and to existing procedures in the sequential change diagnosis literature in a framework which we propose for comparison in general.

Bohrer Memorial Student Workshop
Full Abstracts

Yangfan Zhang

Adaptive Testing for High-Dimensional Data

- In this article, we propose a class of L_q-norm based U-statistics for a family of global testing problems related to high-dimensional data. This includes testing of mean vector and its spatial sign, simultaneous testing of linear model coefficients, and testing of component-wise independence for high-dimensional observations, among others. Under the null hypothesis, we derive asymptotic normality and independence between L_q-norm based U-statistics for several qs under mild moment and cumulant conditions. A simple combination of two studentized L_q-based test statistics via their p-values is proposed and is shown to attain great power against alternatives of different sparsity. Our work is a substantial extension of He et al. (2021), which is mostly focused on mean and covariance testing, and we manage to provide a general treatment of asymptotic independence of L_q-norm based U-statistics for a wide class of kernels. To alleviate the computation burden, we introduce a variant of the proposed U-statistic by using the monotone indexes in the summation, resulting in a U statistic with asymmetric kernel. A dynamic programming method is introduced to reduce the computational cost from $O(n^{qr})$, which is required for the calculation of full U-statistic, to $O(n^r)$ where r is the order of the kernel. Numerical studies further corroborate the advantage of the proposed test as compared to some existing competitors.

Joshua Loyal

An Eigenmodel for Dynamic Multilayer Networks

- Network (or graph) data is at the heart of many modern data science problems: disease transmission, community dynamics on social media, international relations, and others. In this talk, I will elaborate on my research in statistical inference for complex time-varying networks. I will focus on dynamic multilayer networks, which frequently represent the structure of multiple co-evolving relations. Despite their prevalence, statistical models are not well-developed for this network type. Here, I propose a new latent space model for dynamic multilayer networks. The key feature of this model is its ability to identify common time-varying structures shared by all layers while also accounting for layer-wise variation and degree heterogeneity. I establish the identifiability of the model's parameters and develop a structured mean-field variational inference approach to estimate the model's posterior, which scales to networks previously intractable to dynamic latent space models. I apply the model to two real-world problems: discerning regional conflicts in a data set of international relations and quantifying infectious disease spread throughout a school based on the student's daily contact patterns.

Bohrer Memorial Student Workshop

Robert E. Bohrer
1939-1993

PhD 1965

Advisor: Wassily Hoeffding

University of North Carolina

Joined University of Illinois faculty 1968

- Simultaneous inference, mathematical statistics, applied statistics and computational algorithms
- Published more than 50 research articles ranging from applied collaborations to Annals of Statistics
- Founded statistical consulting center in the statistics division of the Mathematics Department
- Consulted with Bell Labs, Teepak, State Geological Survey through the early 80's.
- Part of founding of Department of Statistics in 1985
- Suffered from Diabetes and lost his sight in the early 1970's. Eventually required to use a wheel chair in the late 80's.
- Continued to teach and to be active in research for twenty years!
- Friendly and sympathetic mentor for students

Bohrer Memorial Student Workshop

Example of range of publications.

- On Bayes sequential design with two random variables. Biometrika, 1966.
- On sharpening Scheffe bounds. J. Roy. Statist. Soc (Ser. B), 1967.
- An optimality property of Scheffe bounds. Ann. Statist., 1973.
- Multiple three-decision rules for parametric signs. J. Amer. Statist. Assoc., 1979.
- (with M. J. Schervish) An optimal multiple decision rule for signs of parameters. Proc. Nat. Acad. Sci. U.S.A. 1980.
- (with G. G. Judge, T. A. Yancey and M. E. Bock) The nonoptimality of the inequality restricted estimator under squared error loss. J.Econometrics, 1984.
- (with O.D. Sherwood, S.J. Downing, A. J. Rieber, S. W. Fraley and B. C. Richardsnon). Influence of litter size on ante-partum luteolysis and birth in the rat. Proceedings of the 5th Biennial Ovarian Workshop. July 1984.
- (with M. R. Murphy) Tenable assumptions about comminution leading to lognormal and Rosin-Rammler (Weibull) in particle size distributions. Proc. Int. Symp: Control of Digestion and Metabolism in Ruminants, 1984.
- [Algorithm AS 220] Operating characteristics of Jame-Stein and Efron-Morris estimators. J. Roy. Statist. Soc (Ser. C) (1986)
- (with A. Martinsek) On combining Stein estimation problems: an adaptive rule under classical criteria. J. Statist. Plann. Inference, 1986
- (with T. A. Yancey and G.G. Judge) Sampling performance of some joint one-sided preliminary test estimators under squared error loss.
- Econometrica, 1989.

Bohrer Memorial Student Workshop

Robert Wijsman
1920-2005

PhD 1952, Physics

University of California-Berkeley

Joined University of Illinois faculty 1957

- Mathematical Statistics, Multivariate Analysis, Sequential Analysis.
- Published 60 papers, 56 of them are single-authored.
- "Invariant Measures on Groups and Their Use in Statistics", IMS Lecture Notes Monograph Series
- Mentor for many students (14 in Math Genealogy)
- "His work is characterized by its high quality, careful attention to detail, mathematical sophistication, and balance between theory and applications"
- "Bob displayed admirable tenacity in this pursuit of precise, accurate, and useful results, and he solved hard problems of real significance"
- "Bob's colleagues remember well his careful and precise approach to everything he did, his generosity and willingness to work quite hard for the common good, and his dry sense of humor."
- Reference: A. Martinsek (2006) The contributions of Robert A Wijsman to sequential analysis. Sequential Analysis, 25, 3-18.


ILLINOIS