VC Dimension and Irregular Pairs

Nate Ackerman
Harvard University

Szemeredi's regularity lemma is one of the most important results in graph theory of the last 50 years. This lemma (roughly) says that for every $\epsilon, \delta > 0$ there is a bound $n = n_{\epsilon,\delta}$ such that for every finite graph G there is a partition of the vertex set of G, of size at most n, such that all but a delta-fraction of the pairs of partition elements are "ϵ-regular", i.e. can be "approximated up to ϵ" by a random bipartite graph.

Since its discovery the regularity lemma has been generalized in many directions, including to finite relational structures. There has also been significant work studying the relationship between epsilon, the accuracy of the approximation, and n, the bound on the partition size, with the most notable result being the work of Gowers showing that in the graph case, $n_{\epsilon,\epsilon}$ is at least a tower of 2s of height $1/\epsilon$. There has also been important work showing that under certain model-theoretic tameness assumptions one can greatly improve this relationship. In particular, restricting to graphs with a fixed VC dimension, $n_{\epsilon,\epsilon}$ reduces to being just exponential in $1/\epsilon$.

While there has been extensive work studying the relationship between the size of a regularity partition and its accuracy, there has been significantly less work on the relationship between the partition size and delta, the fraction of irregular pairs. The best known result in the general case by Conlon and Fox who show for any fixed epsilon we have for all k there is graph such that any partition of size k has a $c/log^*(k)$ fraction of irregular pairs, where log^* is the iterative logarithm and c is a constant depending only on epsilon. In this talk I will discuss recent work with Cameron Freer and Rehana Patel where we provide an explicit relationship between the upper bound on the partition size and the fraction of irregular pairs under the assumption that the structure has fixed VC dimension of the structure. In particular we show that if you fix epsilon for any sufficiently large structure and any k there is a partition for which at most a b/k^D fraction of the pairs are irregular, where b and D depend only on the language and the VC dimension.

Area of Specialization: Logic

When Measures Conflict: Towards a Better Understanding of Intergenerational Educational Mobility

Md Nazmul Ahsan[*] Shahe Emran [†] Forhad Shilpi [‡]
Saint Louis University *IPD Columbia* *DECRG, Worldbank*

Hanchen Jiang [§] Orla Murphy [¶]
University of North Texas *Dalhousie University*

October 2022

A large empirical literature on intergenerational educational mobility measures relative mobility by the slope of a conditional expectation function (CEF) relating children's education to parental education. Three measures are widely used: intergenerational regression coefficient (IGRC) with years of schooling as the indicator of educational attainment, intergenerational correlation (IGC) when years of schooling is normalized by its standard deviation, and intergenerational rank-rank slope (IRRS) when schooling ranks in a generation is adopted. The existing evidence suggests that conclusions from IGRC vs. IGC vary substantially, but there is no systematic evidence on whether the IRRS estimates also lead to conflicting conclusions. Using data free of coresidency bias from three developing countries with 42 percent of world population in 2000 (China, India, Indonesia), we provide evidence that the IRRS estimates may lead to dramatically different conclusions about spatial heterogeneity (rural/urban) and evolution across cohorts, especially when the mobility CEF is concave or convex. The rank-rank CEF is consistently more convex (or less concave) compared to the other two CEFs. When different measures lead to conflicting conclusions it is not clear how to interpret the evidence and advise the policymakers. We develop a simple approach to interpret the IGC estimate in terms of the Becker-Tomes model that provides a foundation for a comparative study of IGC vs. IGRC. We find that the idiosyncratic component of children's schooling variance unrelated to the family background plays an important role in IGC. The elasticity of IGC w.r.t IGRC is less than 1 implying that the IGC estimates

[*]Email: nazmul.ahsan@slu.edu.
[†]Email: shahe.emran.econ@gmail.com.
[‡]Email: fshilpi@worldbank.org.
[§]Email: Hanchen.Jiang@unt.edu.
[¶]Email: orla.murphy@dal.ca.

are less responsive to changes in economic forces (such as credit constraint and returns to education) raising questions about the suitability of IGC for understanding the role of changing economic conditions in intergenerational mobility. This also provides an explanation for the puzzle in the literature that IGRC estimates across cohorts show substantial improvements, but the IGC estimates suggest no significant changes. When the mobility CEF is quadratic, by construction, the quadratic coefficient of the CEF for IGC is much larger than that of the CEF for IGRC. This implies that IGC estimates mechanically generate much stronger persistence at the top (for convex) or bottom (for concave) of the distribution. We report evidence that, unlike income, calculating schooling ranks by mid-rank method may fail to neutralize the effects of changing inequality across generations, making IGC a preferable measure for tackling changes in cross-sectional inequality. It is difficult to interpret IRRS in terms of the Becker-Tomes model. The inequality of opportunity approach (Roemer (1998)) suggests that policy advice should focus on the causal effects of policies on the influence of inherited circumstances on children's education which is captured by IGRC. From this perspective, a policy such as school construction or trade liberalization should be considered effective in improving relative educational mobility if the causal effect on IGRC is negative even when a policy fails to affect IRRS significantly.

Area of Specialization: Intergenerational Educational Mobility, Relative Mobility, Alternative Measures, IGRC, IGC, Rank-Rank Slope, Conflicting Evidence, China, India, Indonesia, Policy Advice

Continuity and stability of Fourier series solutions of stochastic wave equations with cubic nonlinearities in 3D

Henri Schurz
Department of Mathematics
Southern Illinois University Carbondale
Qasim Alharbi
Department of Mathematics
Al Qassim University

In this paper, we proved the stability , uniform boundedness and a.s. Hölder continuity of approximate Fourier series solution u to semi-linear stochastic wave equations of the form (in Itô sense)

$$u_{tt} = \sigma^2 \Delta u + (a_1 - a_2 ||u||^2_{\mathbb{L}^2(\mathbb{D})})u - \kappa u_t + (b_0 + b_1 ||u||_{\mathbb{L}^2(\mathbb{D})} + b_2 ||u_t||_{\mathbb{L}^2(\mathbb{D})}) \frac{\partial W}{\partial t}$$

with cubic nonlinearities and homogeneous boundary conditions (HBCs) on general 3D cubes $\mathbb{D} = [0, l_x] \times [0, l_y] \times [0, l_z]$ is studied. The driving Q-regular space-time noise W with linear-growth bounded, state-dependent diffusion intensities $(b_0 + b_1 ||u||_{\mathbb{L}^2(\mathbb{D})} + b_2 ||u_t||_{\mathbb{L}^2(\mathbb{D})})$ is supposed to be general in space $(x, y, z) \in \mathbb{D}$, but white in time $t \geq 0$. The analysis is carried out on an appropriate, separable Hilbert space (i.e. the space of Fourier coefficients) of all solutions which are driven by the eigenfunctions of Laplace operator $\triangle$ subject to HBCs on the 3D domain $\mathbb{D}$. The dynamics of related expected total energy functional $e(t)$ plays a central role in our discussion, depending on diverse parameters such as diffusivity constant σ, damping $\kappa \geq 0$, transport coefficient a_1, length parameters l_x, l_y, l_z, diffusion constants b_r and eigenvalues $\alpha^{i,j,k}_{n,m,l}$ of related covariance operator Q. Several examples illustrate the feasibility of our approach and explain our major results.At the end, we provided some simulations.

Area of Specialization: Stochastic Partial Differential Equations

Minimal dynamical systems with unique probability measure

Dana Bartošová
Department of Mathematics
University of Florida

By a *dynamical system* or a *flow* we mean a continuous action of a topological group on a compact Hausdorff space. The flow is *minimal* if it has no proper non-trivial closed subsets invariant under the action. For every topological group, there is the most complicated minimal flow, so called *universal minimal flow* $M(G)$. For locally compact non-compact groups, the universal minimal flow is non-metrizable, but for other groups it can be metrizable or even trivial. In 2005, Kechris, Pestov, and Todorčević revealed the reason for metrizability of universal minimal flows in the case of groups of automorphisms of countable structures with the pointwise convergence topology. The season lies in the Ramsey theoretic behaviour of the class of finitary substructures of the given structure.

In ergodic theory, one equips dynamical systems with an invariant probability measure, and the analogue of a minimal system is an ergodic system. Often, when a dynamical system admits one invariant measure, it admits infinitely many. However, Angel, Kehris, and Lyons showed using quantitative Ramsey theory that in the case of automorphism groups of countable structures with metrizable universal minimal flows, there may be just one such measure on any minimal system - such groups are called *uniquely ergodic*. In fact,the authors conjectured that this is always the case for such groups. Many new examples appeared since confirming the conjecture in special cases. We will talk about a candidate for a counterexample that Colin Jahel suggested to me.

Area of Specialization: Logic

Probabilistic tools in continuous combinatorics

Anton Bernshteyn
School of Mathematics
Georgia Institute of Technology

In this talk I will describe probabilistic tools that can be used to construct continuous solutions to combinatorial problems on zero-dimensional spaces. I will also discuss some applications of these tools, such as the connection between continuous combinatorics and distributed computing.

Interval Data in the Light of Big Data Sets

L. Billard, University of Georgia

Massively large data sets are becoming routine and ubiquitous given modern computer capabilities. What is not so routine is how to analyze these data. One approach is to aggregate the data sets according to some scientific criteria. The resultant data are perforce lists, intervals, histograms, and so on. Often-times these are then analyzed using some form of classical surrogates (such as averages, variances and the like), expressed most commonly in terms of the midpoints and ranges. While these are intuitively reasonable surrogates, we show through some illustrative examples that these are fraught with mathematical and conceptual difficulties. Instead, it is the entire interval (e.g., likewise for histogram aggregations, etc) that should be used in any analysis.

Efficient Marginalization-based MCMC Methods for Hierarchical Bayesian Inverse Problems

Andrew Brown
School of Mathematical and Statistical Sciences
Clemson University

Hierarchical models in Bayesian inverse problems are characterized by an assumed prior probability distribution for the unknown state and measurement error precision, and hyper-priors for the prior parameters. Combining these probability models using Bayes' law often yields a posterior distribution that cannot be sampled from directly, even for a linear model with Gaussian measurement error and Gaussian prior, both of which we assume in this paper. In such cases, Gibbs sampling can be used to sample from the posterior, but problems arise when the dimension of the state is large. This is because the Gaussian sample required for each iteration can be prohibitively expensive to compute, and because the statistical efficiency of the Markov chain degrades as the dimension of the state increases. The latter problem can be mitigated using marginalization-based techniques, but these can be computationally prohibitive as well. In this paper, we combine the low-rank techniques of Brown et al. (2018) with the marginalization approach of Rue and Held (2005). We consider two variants of this approach: delayed acceptance and pseudo-marginalization. We provide a detailed analysis of the acceptance rates and computational costs associated with our proposed algorithms, and compare their performances on two numerical test cases—image deblurring and inverse heat equation.

Area of Specialization: Statistics

Sufficient Variable Screening with High-Dimensional Controls

Chenlu Ke
Department of Statistics and Operations Research
Virginia Commonwealth University

Variable screening for ultrahigh-dimensional data has attracted extensive attention in the past decade. In many applications, researchers learn from previous studies about certain important predictors or control variables related to the response of interest. Such knowledge should be taken into account in the screening procedure. The development of variable screening conditional on prior information, however, has been less fruitful, compared to the vast literature for generic unconditional screening. In this talk, we introduce a model-free variable screening paradigm that allows for high-dimensional controls and applies to either continuous or categorical responses. The contribution of each individual predictor is quantified marginally and conditionally in the presence of the control variables as well as the other candidates by reproducing-kernel-based R^2 and partial R^2 statistics. As a result, the proposed method enjoys the sure screening property and the rank consistency property in the notion of sufficiency, with which its superiority over existing methods is well-established. The advantages of the proposed method are demonstrated by simulation studies encompassing a variety of regression and classification models, and an application to high-throughput gene expression data.

Area of Specialization: Variable Screening

Bayesian High-Dimensional Bridge-Randomized Quantile Regression

Mai Dao
Department of Mathematics, Statistics, and Physics
Wichita State University
Shen Zhang, Min Wang, Keying Ye
Department of Management Science and Statistics
University of Texas in San Antonio

A bridge-randomized penalization that employs a prior for the shrinkage parameter, as opposed to the conventional bridge penalization with a fixed penalty, often delivers more superior performance compared to many other traditional shrinkage methods. Tian and Song (2020) recently considered a fully Bayesian formulation of the bridge-randomized penalized quantile regression models, which is based on utilizing the asymmetric Laplace distribution as an auxiliary error distribution and the generalized Gaussian distribution prior for the regression coefficients. However, there exist major computational drawbacks in performing posterior sampling associated with this construction in the 'large-p-small-n' or even 'large-p' settings, thus limiting its applicability for high dimensional data analysis. To overcome these issues, we develop an efficient Bayesian computational algorithm via two-block Markov Chain Monte Carlo method for the bridge-randomized penalization in quantile regression. Simulation studies encompassing a wide range of scenarios indicate that the proposed method performs at least as well as, and often better than, other existing procedures in terms of both parameter estimation and variable selection. Finally, a real-data application is provided for illustrative purposes.

Area of Specialization: Bayesian Statistics

Supervised Spatial Regionalization using the Karhunen-Loève Expansion and Minimum Spanning Trees

Ranadeep Daw & Christopher K. Wikle
Department of Statistics
University of Missouri Columbia

The talk presents a methodology for supervised regionalization of data on a spatial domain. Defining a spatial process at multiple scales leads to the infamous ecological fallacy problem. Here, we use the ecological fallacy as the minimization criterion to get the intended regions. The Karhunen-Loève Expansion of the spatial process maintains the relationship between the realizations from multiple resolutions. Specifically, we use the Karhunen-Loève Expansion to define the regionalization error so that the ecological fallacy is minimized. The contiguous regionalization is done using the minimum spanning tree formed from the spatial locations and the data. Then, regionalization becomes similar to pruning edges from the minimum spanning tree. The methodology is demonstrated using simulated and real data examples.

Area of Specialization: Statistics

Dimension Reduction in Matrix-valued Time Series

Tharindu P. De Alwis
Department of Mathematical Sciences
Worcester Polytechnic Institute
S. Yaser Samadi
Department of Mathematics
Southern Illinois University Carbondale

In many fields of science, matrix-valued data are collected over a period of time. For instance, a set of weather characteristics are observed and recorded in different locations every day. The daily observations form a matrix-valued time series data. A limited number of studies exist in the literature for analyzing matrix-valued time series data, however these models suffer from overparametrization. Although, bilinear matrix autoregressive (MAR) models significantly reduce the number of parameters, but they cannot differentiate between relevant and irrelevant information. We propose a new dimension reduction framework for matrix variate time series data by using the minimal reducing subspace that eliminates irrelevant information and achieves efficient estimation by linking the mean function and covariance matrix. We establish the asymptotic properties of the proposed estimators. The efficiency and the accuracy of the proposed method are evaluated and demonstrated with the existing methods via simulation studies and a real data analysis.

Area of Specialization: Statistics

Efficient Online Reinforcement Learning Policies for Continuous Environments

Mohamad Kazem Shirani Faradonbeh
Department of Statistics,
University of Georgia

One of the most popular dynamical models for continuous environments are linear time-invariant control systems that evolve according to stochastic differential equations. An interesting problem in this class of systems is learning to take actions to minimize a quadratic cost function when system matrices are unknown. We discuss implementable online reinforcement learning policies that learn the optimal control actions fast. In fact, the proposed policy efficiently balances exploration versus exploitation by carefully randomizing the parameter estimates such that the regret grows as the *square-root of time* multiplied by the *number of parameters*. Mathematical performance analysis as well as simulations for learning to control an airplane will be presented to show efficiency. To obtain the results, we develop novel techniques that provide a useful cornerstone for reinforcement learning problems in continuous environments.
Area of Specialization: Data Science

A Sufficient reductions in regression with mixed predictors

Liliana Forzani
Department of Mathematics
Universidad Nacional del Litoral
Argentina

Most data sets comprise of measurements on continuous and categorical variables. In regression and classification, modeling high-dimensional mixed predictors has received limited attention in the Statistics literature.
In this talk we present a general regression problem of inferring on a variable of interest based on high dimensional mixed continuous and binary predictors. The aim is to find a lower dimensional function of the mixed predictor vector that contains all the modeling information in the mixed predictors for the response, which can be either continuous or categorical. The approach we propose identifies sufficient reductions by reversing the regression and modeling the mixed predictors conditional on the response.
We derive the maximum likelihood estimator of the sufficient reductions, asymptotic tests for dimension, and a regularized estimator, which simultaneously achieves variable (feature) selection and dimension reduction (feature extraction).
We study the performance of the proposed method and compare it with other approaches through simulations and real data examples.
Area of Specialization: Statistics. Regression analysis

Computable PAC Learning of Continuous Features

Cameron Freer
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

We introduce definitions of computable PAC learning for binary classification over computable metric spaces. We provide sufficient conditions on a hypothesis class to ensure that an empirical risk minimizer (ERM) is computable, and bound the strong Weihrauch degree of an ERM under more general conditions. We also give a presentation of a hypothesis class that does not admit any proper computable PAC learner with computable sample function, despite the underlying class being PAC learnable.

Joint work with Nathanael Ackerman, Julian Asilis, Jieqi Di, and Jean-Baptiste Tristan.

Area of Specialization: Logic & Probability

Random counterexamples in query learning

James Frietag
University of Illinois at Chicago
Mathematics, Statistics, and Computer Science

Traditionally, query learning analyzes worst case mistake bounds for a given concept class. In this work we will consider what happens in the case of random counterexamples. We will show Littlestone dimension characterizes the expected mistake bound.

Incorporation of multiple sources of information in statistical regularization

Jaroslaw Harezlak
Department of Epidemiology and Biostatistics
Indiana University Bloomington
and multiple co-authors

Prior information use in principled manner can improve the quality of the regression coefficient estimation. Our proposal incorporates structural connectivity derived from the Diffusion Weighted Brain Imaging and cortical spatial distance in the penalized approach. Extending previously developed methods informing the estimation of the regression coefficients, we incorporate such information via a Laplacian matrix based on the proximity measures. The penalty term is constructed as a weighted sum of a function of structural connectivity and proximity between cortical areas. Simulation studies show improved estimation accuracy. We apply our approach to the data collected in the Human Connectome Project, where the cortical properties of the left hemisphere are found to be associated with vocabulary comprehension.

Area of Specialization: Biostatistics

Densely computable structures and isomorphisms

Valentina Harizanov
Department of Mathematics
George Washington University

In recent years, computability theorists have investigated dense computability of sets such as generic and coarse computability. These notions of approximate computability have been motivated by asymptotic density problems in combinatorial group theory. We extend the notions from sets to arbitrary countable structures by introducing generically and coarsely computable structures and further notions of densely computable structures. There are two directions in which these notions could potentially trivialize: either all structures in a certain class could have densely computable isomorphic copies, or only those having computable or computably enumerable isomorphic copies. We also investigate the notions of generically and coarsely computable isomorphisms and their weaker variants. This is joint work with Wesley Calvert and Douglas Cenzer.

References

(1) W. Calvert, D. Cenzer and V. Harizanov, Generically and coarsely computable isomorphisms, to appear in *Computability*.

(2) W. Calvert, D. Cenzer and V. Harizanov, Densely computable structures, *Journal of Logic and Computation* 32 (2022), pp. 581–607.

Scaled Envelope Models for Multivariate Time Series

Wiranthe Herath
Department of Information Management and Business Analytics
Drake University
Yaser Samadi
School of Mathematical and Statistical Sciences
Southern Illinois University Carbondale

Vector autoregressive (VAR) models have been extensively used to model multivariate time series data. Efficient estimation of the VAR coefficients is an important problem. The envelope technique for VAR models is demonstrated to have the ability to yield significant gains in efficiency and accuracy by incorporating linear combinations of the response vector that are basically immaterial to the estimation of the VAR coefficients. When VAR responses are measured on different scales, the efficiency improvements promised by envelopes are not always guaranteed. In this talk, we address this issue by introducing the scaled envelope VAR model that retains the ability of the envelopes to boost efficiency while remaining invariant to scale changes. Some simulation studies and a real data analysis will be presented to demonstrate the efficiency of the proposed model.

Area of Specialization: Statistics

Free structures and limiting density

Johanna Franklin
Department of Mathematics
Hofstra University
Meng-Che "Turbo" Ho
Department of Mathematics
California State University, Northridge
Julia Knight
Department of Mathematics
University of Notre Dame

Gromov asked what a typical group looks like, and he suggested a way to make the question precise in terms of *limiting density*. The typical finitely presented group is known to share some important properties with the non-abelian free groups. To make this precise, Knight conjectured that the typical group satisfies a zero-one law and has the same first-order theory as the free group.

We ask Gromov's question more generally, for structures in an arbitrary *algebraic variety* (in the sense of universal algebra), with presentations of a specific form. We focus on elementary properties. We give examples illustrating different behaviors of the limiting density. Based on the examples, we identify sufficient conditions for the elementary first-order theory of the free structure to match that of the typical structure; i.e., a sentence is true in the free structure iff it has limiting density 1.

Area of Specialization: Logic

An Improved Accelerated Sequential Sampling Scheme with Illustrations

Jun Hu
Department of Mathematics and Statistics
Oakland University, Rochester

In this talk, we present a k-at-a-time improved accelerated sequential sampling scheme, which also incorporates three other types of sampling procedures: i) the classic Anscombe-Chow-Robbins purely sequential sampling procedure; ii) the ordinary accelerated sequential sampling procedure; and iii) the relatively new k-at-a-time purely sequential sampling procedure. The asymptotic efficiency of this newly-proposed sequential sampling scheme is fully investigated with three illustrations on minimum risk point estimation, bounded variance point estimation, and bounded risk point estimation in a linear model, respectively.

Area of Specialization: Statistics

Dynamically adjusted federated learning for heterogeneous data

Yilun Huang, Sounak Chakraborty,
Department of Statistics
University of Missouri-Columbia
Anjishnu Banerjee
Division of Biostatistics
Medical College of Wisconsin
Tanujit Dey
Center for Surgery and Public Health
Harvard Medical School

Collaborative machine learning, or more specifically federated learning, is a machine learning paradigm where many models are trained on piece wise data silos, while the inference parameters are combined using a central node. The unique feature of federated learning is that data is never shared between silos, maintaining privacy, thus keeping the model training decentralized. Meaningful inference in federated learning therefore depends critically on the combination step at the central server. We investigate the effect of heterogeneity on the resultant inference. While federated learning itself is a relatively new concept, there is limited literature on how to handle heterogeneity, with most approaches assuming exchangeable data distributions. We study the bias when estimating federated learning models across variables and present a novel method to dynamically flag heterogeneous variables during federated training and testing, and maintain validity of inference. Additionally, we also propose a statistical algorithm using bootstrap for approximate inference for those heterogeneous variables. Finally, we demonstrate the performance of our proposed method in simulated and real data settings.

Area of Specialization: Variable Selection, Distributed Statistical Inference, Federated Learning

Spatio-temporal Matrix Autoregression via Tensor Decomposition

Rukayya Ibrahim
Department of Computer Science and Mathematics
Penn State Harrisburg
S Yaser Samadi
Department of Mathematics
Southern Illinois University Carbondale
Tharindu DeAlwis
Department of Mathematical Science
Worcester Polytechnic Institute

Massive, highly interactive datasets, like time dependent big data and especially spatio-temporal data, are constantly arising in many modern applications. The increasing popularity of such data and data dimensionality made it crucial to seek methods that efficiently handle their high order. To efficiently handle the increasing size and complexity of such datasets in analysis, prediction and forecast, tensor decomposition techniques are employed. Tensor decomposition methods provide the advantage of latent structure and information extraction, data imputation, complexity control, etc., giving way to a highly active research area in statistics, applied mathematics, and data science.

In this work, we model matrix valued spatio-temporal data using matrix autoregression expressed as tensor regression model by folding the autoregression matrices into a four dimensional tensor. We compare our model to other successful models previously used for such data. In addition, we propose two estimators for the transition tensor when dimension are low and when dimension are high respectively. The estimator when data is low dimensional is a least square type estimator while for high dimension, we employ the use of regularization. We use nuclear norm regularized estimator to jointly impose sparsity and low rankness of the estimator to achieve further dimension reduction. We derive some asymptotic and non asymptotic properties of our estimator. Finally, we perform simulations and real data analysis to illustrate the advantages of our model.

Area of Specialization: Statistics

A preservation property for products of metric structures

Mary Leah Karker
Department of Mathematics and Computer Science
Providence College

This talk will focus on analogues, for continuous logic, of results of Feferman and Vaught on model-theoretic properties of various kinds of product. In particular, we will show that a natural continuous-logic version of direct product enjoys the following preservation property: if a sentence θ is true in $\prod_{i=0}^{k} \mathcal{M}_i$ for every $k \in \mathbb{N}$, then θ is true in $\prod_{i \in \mathbb{N}} \mathcal{M}_i$.

Area of Specialization: Logic

Sample Size Requirements for Confirmatory Factor Analysis

Jennifer Koran
Quantitative Methods
Southern Illinois University Carbondale

Minimum sample size recommendations for confirmatory factor analysis generate considerable confusion. This is attributable to several factors. Different criteria for what defines an adequate sample for confirmatory factor analysis are applied to make sample size recommendations. Further, the relevant outcomes are influenced by several characteristics of the analysis.

In this talk results are presented that demonstrate minimum adequate sample size meeting multiple criteria. The variation in minimum adequate sample size according to several characteristics of the analysis is shown.

Area of Specialization: Psychometrics and Statistics

Estimating Gravity Coefficients with Multiple Layers of Heterogeneity

Erick Kitenge
Sajal Lahiri
Vandeveer Chair Professor of Economics and Distinguished Scholar
School of Analytics, Finance and Economics-SIUC

We estimate a gravity model for 205 countries over the period 1954-2014, allowing for multiple layers of heterogeneity. The first layer arises from the interactions between the gravity variables The second one comes from country pairs that differ in values of binary gravity variables. Further layers come from different income groups and regions. Our results show the importance of heterogeneities at various levels. For instance, landlockedness is less restrictive when trading partners are contiguous; language elasticity of trade is lower when contiguous countries trade, and when they are colonially linked. We also find, for example, that the positive interaction between contiguity and landlockedness is more pronounced when the importing countries are not high-income ones; the negative interaction between language and contiguity is more pronounced when the importing countries are from South Asia; the negative interaction between contiguity and colonial links is stronger when the importing countries are from Sub-Saharan Africa.

Design Considerations for Clinical Trials with Survival Endpoints

Jennifer Le-Rademacher
Division of Clinical Trials & Biostatistics,
Department of Quantitative Health Sciences
Mayo Clinic, Rochester MN

Survival is a common and important outcome in medical research. Specifically in cancer, regulatory approval of a new therapy relies on demonstration of its survival benefit compared to a control (whether a placebo or a current standard of care) in randomized clinical trials. Survival benefit can be quantified using one of the following measures (endpoints): a) the mortality hazard rate, b) the survival probability at a pre-specified time point, or c) the mean survival time (restricted to a pre-specified time point). These endpoints have different interpretations and are analyzed using different statistical methods. Their performance depends on the assumptions about the underlying survival distributions of the experimental treatment and the control. Which endpoint is optimal for a clinical trial depends on multitude of factors, including the mechanism of action of the therapy, the expected pattern of the treatment effect, and the clinical relevance of that endpoint in the specific disease setting.

In this talk, we will define the three summary measures. We will describe the clinical trial design features and assumptions associated with each of these endpoints. We will discuss the different points to consider when choosing the optimal survival endpoint for a clinical trial.

Area of Specialization: Biostatistics

Multiple change Point Detection in High Dimensional Low Rank Models

George Michailidis
Department of Statistics
University of Florida

We study the problem of detecting and locating change points in high-dimensional models with low rank structure. We develop a simple two step algorithm for the problem at hand and establish performance guarantees in the form of finite sample bounds for the accuracy of the estimated locations of the change points and the underlying model parameters. We illustrate the detection strategy on data from three different domains that employ different statistical models: macroeconomics, neuroimaging and political science.

Area of Specialization: Data Science

Spatial Envelope

Hossein Moradi Rekabdarkolaee
Department of Mathematics and Statistics
Southern Dakota State University

The envelope is a parsimonious version of the classical multivariate regression model that identifies a minimal reducing subspace of the responses. However, existing envelope methods assume an independent error structure in the model. While the assumption of independence is convenient, it does not address the additional complications associated with spatial or temporal correlations in the data. Therefore, we propose a Spatial Envelope method for dimension reduction in the presence of dependencies across space. The covariance function for the multivariate spatial data can be separable and non-separable. We show that for cases where the marginal spatial correlations are different from each other, the non-separable model provides better estimation and inference than the related separable model, and provides tighter inference than a non-separable spatial model without dimension reduction when there is immaterial variation in the data. We study the asymptotic properties of the proposed estimators and show that the asymptotic variance of the estimated regression coefficients under the spatial envelope model is smaller than that of the traditional maximum likelihood estimation. Furthermore, we present a computationally efficient approach for inferences. The efficacy of the proposed method is investigated through simulation studies and data analysis.

Area of Specialization: Statistics

Improved detection of allelic imbalance using biologically informed priors

Sally Paganin
Department of Biostatistics
Harvard T.H. Chan School of Public Health

There is growing interest in developing tools for cancer screening and monitoring based on the analysis of DNA sequencing data derived from non-invasive procedures such as blood samples. At early cancer stages, such samples contain DNA from a majority of normal cells and a low fraction of tumor cells. Cancer presence can be assessed measuring allelic imbalance: since a person inherits one allele from each parents, the allele proportion at heterozygous loci is close to 0.5 in normal cells, whereas significant deviations from 0.5 are indicative of the presence of cancer. To efficiently and sensitively detect such deviations, we model the allele proportions over the genome via a novel Bayesian hierarchical Hidden Markov Model. We leverage prior knowledge from population genome databases while borrowing information across multiple samples from the same subject. Hypothesis testing for cancer presence is embedded in the model via a spike and slab prior.
Area of Specialization: Biostatistics

Participant Retention: Factors associated with sustained device use in the Electronic Framingham Heart Study

Chathurangi H. Pathiravasan
Department of Biostatistics
Boston University School of Public Health

mHealth is a promising tool which enables clinicians and researchers to monitor patient health and improve daily care. Long-term use of digital devices is critical for successful clinical or research use, but digital health studies are challenged by a rapid drop-off in participation. Since few studies have deployed integrated solutions for digital and mobile devices for cardiovascular disease (CVD) phenotyping among cohort study participants, little is known about factors related to long-term use of system components, and inter-individual variation of wearable use patterns.

We designed an eCohort embedded in the Framingham Heart Study (eFHS) to address these challenges and to compare the digital data to traditional data collection. The eFHS system contains three system components: a smartphone app, and two devices (a smartwatch and digital blood pressure (BP) cuff). Participants were asked to complete app-based surveys every 3 months, send watch data daily and BP data weekly over the 1-year period after enrollment. To better understand the facilitators and barriers of system use, we investigated factors that are associated with long-term use of individual eFHS system components. We further examined inter-individual variability of smartwatch use patterns and its association with physical activity levels. Understanding these patterns and factors associated with smartphone app and wearable device use may help clinicians, health advocates, and developers to support individuals in self-managing and improving their cardiovascular health.

Area of Specialization: Biostatistics

Wald Type Tests with the Wrong Dispersion Matrix

Kosman W Rajapaksha
Department of Mathematics and Statistics
Washburn University
David J. Olive
School of Mathematical & Statistical Sciences
Southern Illinois University

A Wald type test with the wrong dispersion matrix is used when the dispersion matrix is not a consistent estimator of the asymptotic covariance matrix of the test statistic. One class of such tests occurs when there are p groups and it is assumed that the population covariance matrices from the p groups are equal, but the common covariance matrix assumption does not hold. The pooled t test, one-way ANOVA F test, and one-way MANOVA F test are examples of this class. Another class of such tests is used for weighted least squares. Two bootstrap confidence regions are modified to obtain large sample Wald type tests with the wrong dispersion matrix.

Area of Specialization: Bootstrap; confidence region; MANOVA; weighted least squares.

Inference in response-adaptive trials when the patient population varies during time

Masimilliano Russo
Harvard University

A common assumption of data analysis in clinical trials is that the patient population, as well as treatment effects, do not vary during the study. However, when trials enroll patients over several years, this hypothesis may be violated. Ignoring variations of the outcome distributions over time, under the control and experimental treatments, can lead to biased treatment effect estimates and poor control of false positive results. We propose and compare two procedures that account for possible variations of the outcome distributions over time, to correct treatment effect estimates, and to control type-I error rates. The first procedure models trends of patient outcomes with splines. The second leverages conditional inference principles, which have been introduced to analyze randomized trials when patient prognostic profiles are unbalanced across arms. These two procedures are applicable in response-adaptive clinical trials. We illustrate the consequences of trends in the outcome distributions in response-adaptive designs and in platform trials and investigate the proposed methods in the analysis of a glioblastoma study.

Tempered functional time series

Farzad Sabzikar
Department of Statistics
Iowa State University
Piotr Kokoszka
Department of Statistics
Colorado State University

We propose a broad class of models for time series of curves (functions) that can be used to quantify near long-range dependence or near unit root behavior. We establish fundamental properties of these models and rates of consistency for the sample mean function and the sample covariance operator. The latter plays a role analogous to sample crosscovariances for multivariate time series, but is far more important in the functional setting because its eigenfunctions are used in principal component analysis, which is a major tool in functional data analysis. It is used for dimension reduction of feature extraction. We also establish a central limit theorem for functions following our model. Both the consistency rates and the normalizations in the CLT are nonstandard. They reflect the local unit root behavior and the long memory structure at moderate lags.

Area of Specialization: Stochastic processes

Bayesian density estimation under parameter constraints

Nasser Sadeghkhani
Department of Statistics
Ohio State University

This talk discusses the problem of Bayesian density estimation under restricted parameter space. We present some explicit closed-form density estimators which are computationally inexpensive. More specifically, we show that the proposed density estimators dominate other existing density estimators along with the plug-in type density estimators under the Kullback Leibler Loss function. We also study some applied real-world examples to illustrate the proposed methods.

Area of Specialization: Bayesian Statistics

Bayesian Cluster Analysis for High-dimensional Time Series Data

Hadi Safari Katesari
Department of Mathematical Sciences
Stevens Institute of Technology

This talk investigates how Bayesian model-based conducts clustering and captures uncertainty in ultra high-dimensional time series data while avoiding slow mixing and slow computation problems. Firstly, I investigate the suitability of Bayesian model-based for dynamic clustering as a mechanism for dimension reduction in the context of ultra high-dimensional time series data. Secondly, I propose a dependency model to capture the correlation available in the dataset and between augmented marginal distributions of high-dimensional mixed data. Thirdly, I rigorously examine the potential of the above for fast computation and mixing.

Area of Specialization: Statistics

Integrative Discriminant Analysis Methods for Multi-view Data

Sandra Safo
University of Minnesota
Division of Biostatistics

Many diseases are complex heterogeneous conditions that affect multiple organs in the body and depend on the interplay between several factors that include molecular and environmental factors, thus requiring a holistic approach in understanding the complexity and heterogeneity. In this talk, I will present some of our current statistical and machine learning methods for integrating data from multiple sources while simultaneously classifying units or individuals into one of multiple classes or disease groups. The proposed methods are tested using both simulated data and real-world datasets, including RNA sequencing, metabolomics, and proteomics data pertaining to COVID-19 severity. We identified signatures that better discriminated COVID-19 patient groups, and related to neurological conditions, cancer, and metabolic diseases, corroborating current research findings and heightening the need to study the post sequelae effects of COVID-19 to devise effective treatments and to improve patient care.

Big Ramsey degrees for internal colorings

Lynn Scow
California State University, San Bernardino
Mathematics
Dana Bartošová, Mirna Džamonja, and Rehana Patel.

In this talk, I will define what it means for a coloring of substructures of an ultraproduct structure to be “internal”, and a notion of finite big Ramsey degree for internal colorings. I will also present a certain Ramsey degree transfer theorem from countable sequences of finite structures to their ultraproducts, assuming AC and some additional mild assumptions. The version of this theorem for sequences of arbitrary length makes essential use of a certain assumption on the ultrafilter: that it be a finitely additive 0-1 valued measure that is, however, not countably additive, and I will expand on this topic in the talk.

Title: Anticoncentration of Random walks on o-minimal sets

Stanford University
Szego Assistant Professor
Mathematics

Abstract: If we take n independent steps in a Lie group G, how well can we upper bound the probability we land on a submanifold S? If S contains lots of additive structure, then we can always rig the walk to end up on S with quite high (constant) probability so we can't even say the probability tends to 0 as the number of steps tends to infinity! I'll describe a special class of manifolds coming from model theory which have the property that if S doesn't contain any ``obvious'' additive structure (namely exponential arcs in G), then the probability tends to 0 at a rate 1/n^C with C depending only on the dimension of S.

Fréchet Sufficient Variable Selection with Graphical Structure Among Predictors

Jiaying Weng
Mathematical Sciences
Bentley University

Fréchet regression has received considerable scholarly attention to encounter metric-space valued responses that are complex and non-Euclidean data, such as matrices, graphs, and probability functions. However, several unresolved questions remain about the development of Fréchet sufficient dimension reduction and variable selection in ultra-high dimensions. This paper studies Fréchet sufficient variable selection with graphical structure among multivariate Euclidean predictors. The ultimate goal is to find several linear combinations of predictors containing almost all regression information and select active variables when predictors contain additional graphical information. In the ultra-high dimension, a long-standing issue is that the covariance matrix is ill-conditioned, so estimating its precision matrix is challenging. We propose a penalized deference of trace loss to avoid directly computing the inverse of a large covariance matrix. Our proposed penalization can be easily applied to high-dimensional predictors while utilizing the prior graphical information among predictors to improve accuracy and consistency. Theoretically, we derive the asymptotic estimation consistency and the variable selection consistency of the proposed estimator. We demonstrate the superior finite-sample performance of our proposals over existing methods through comprehensive simulations and data analysis.

Fréchet regression; Graphical structure; Sufficient dimension reduction; Sufficient variable selection; Weighted inverse regression ensemble.

Sequential Gradient Descent for Change-Point Analysis

Xianyang Zhang & Trisha Dawn
Department of Statistics
Texas A&M University

One common approach to detecting change-points is minimizing a cost function over possible numbers and locations of change-points. The framework includes several well-established procedures, such as the penalized likelihood and minimum description length. Such an approach requires finding the cost value repeatedly over different segments of the data set, which can be time-consuming when (i) the data sequence is long and (ii) obtaining the cost value involves solving a non-trivial optimization problem. This paper introduces a new method based on sequential gradient descent (SeGD) to find the cost value effectively. The core idea is to update the cost value using the information from previous steps without re-optimizing the objective function. The new method is applied to change-point detection in generalized linear models and penalized regression. Numerical studies show that the new approach can be orders of magnitude faster than the Pruned Exact Linear Time (PELT) method without sacrificing estimation accuracy.

Area of Specialization: Statistics

Merging of opinions for computable Bayesian agents

Francesca Zaffora Blando
Carnegie Mellon University
Department of Philosophy

Abstract: In this talk, I will discuss the phenomenon of merging of opinions with increasing information in the setting of computable probability spaces. We will see that the theory of algorithmic randomness can be used to both (1) characterize the data streams on which computable priors are guaranteed to merge and (2) define notions of agreement between computable priors that are strong enough to entail merging of opinions with probability one.