Statistical Foundations and Applications of Machine Learning and Artificial Intelligence

A symposium to honor Professor Shaw-Hwa Lo’s 70th Birthday and his collaboration with Professor Herman Chernoff

August 12th, 2022 | Department of Statistics, Columbia University
About

In celebration of Professor Shaw-Hwa Lo’s 70th Birthday and his longtime collaboration with Professor Herman Chernoff, this symposium includes presentations on the latest advances in statistical research. Featuring presentations by Professor Lo’s former students, collaborators, and friends, this symposium will highlight contributions to foundational statistical areas, along with applications in machine learning and artificial intelligence. A special session on collaboration and interdisciplinary research will feature the virtual participation of Professor Chernoff.

Shaw-Hwa Lo began his career as a mathematical statistician and his early contributions to the field were in asymptotics theory, survival analysis, and resampling methods. During his 30+ years career at Columbia, he has continuously expanded his research interests to solve some of the most challenging problems posed by big data. In particular, starting in the 1990s, he started a long-lasting and fruitful collaboration with Professor Herman Chernoff at Harvard that has led to innovative statistical methods for deciphering the genetic basis of complex human traits. Lo’s innovations in statistics have contributed to multiple disciplines such as molecular biology, transportation and medicine. In recognition of his life-time achievements the New England Statistics Society (NESS) recently gave him the inaugural Chernoff Excellence in Statistics Award, its highest honor. Read more about Professor Lo’s life and career in this feature article on his Chernoff award.

Location

The conference will be held in the Department of Statistics at Columbia University (1255 Amsterdam Avenue, New York, NY 10027) on the 9th floor in room 903. The banquet will be served at Bar Boulud, 1900 Broadway (at 64th Street).
Schedule

8:00AM - 8:30AM  Breakfast and Registration

8:30AM - 8:40AM  Opening Remarks

8:40AM - 10:10AM  Session I

**In silico prediction of vaccine effectiveness against new genetic variants**
| Maggie Wang, Chinese University of Hong Kong

**Miniaturizing Data Defect Correlation: A Versatile Strategy for Handling Non-Probability Samples**
| Xiao-Li Meng, Harvard University

**Tweedie's formula for chi-squared distributions**
| Inchi Hu, George Mason University

10:10AM - 10:20AM  Coffee break

10:20AM - 11:50PM  Session II

**Causal Inference in High Dimensions with the Influence Score**
| Jonathan Auerbach, George Mason University

**Distributional Regression Models**
| Hans-Georg Müller, University of California, Davis

**A Bayesian View of Constraints on a 2x2 Table: A preliminary report**
| Joseph B (“Jay”) Kadane, Carnegie Mellon University

11:50PM - 1:00PM  Lunch

1:00PM - 2:30PM  Session III

**Knockoff-based statistics for the identification of putative causal loci in genetic studies**
| Iuliana Ionita-Laza, Columbia University

**Refugee narratives in news and their influence on German native attitudes and behaviors**
| Adeline Lo, University of Wisconsin Madison

**Safe Policy Learning through Extrapolation: Application to Pre-trial Risk Assessment**
| Kosuke Imai, Harvard University
2:30PM - 2:40PM  Coffee break

2:40PM - 4:10PM  Session IV

Test-Fairness Deep Learning with Influence Score
| Henry Horng-Shing Lu, National Yang Ming Chiao Tung University

Deep Learning for the Partial Linear Cox Model
| Jane-Ling Wang, University of California, Davis

Changepoint Like Problems
| David Siegmund, Stanford University

4:10PM - 4:20PM  Coffee Break

4:20PM - 5:20PM  A Tribute to Herman Chernoff
| Moderated by Adeline Lo

6:00PM - 8:00PM  Banquet
Speakers & Abstracts

Maggie Wang

In silico prediction of vaccine effectiveness against new genetic variants

Biography

Dr. Maggie Haitian Wang is an Associate Professor in the JC School of Public Health and Primary Care, Faculty of Medicine, the Chinese University of Hong Kong (CUHK). She received her bachelor's degree (Hon) in Physics in 2005 and PhD degree in ISOM (Statistics) in 2011 from the Hong Kong University of Science and Technology (HKUST). Her research interest is developing statistical and bioinformatic methods towards better vaccine design and human genome interpretation. She serves as a board member of the International Genetic Epidemiology Society (IGES) and a review editor of Human Genetics.

Maggie was a visiting scholar in Prof. Shaw-Hwa Lo’s group in Columbia University in 2009 through the collaboration of her adviser Prof. Inchi Hu and Prof. Lo. She has received the best support and friendship from Tian and Prof. Lo till this day. During her stay in Columbia, Maggie was exposed to a series of wonderful academic activities including the genetic analysis workshop (GAW) and IGES, which have greatly facilitated her later career development and become her academic home today.

Abstract

The effectiveness of vaccines varies with the matching of vaccine strains to circulating strains. Based on the genetic distance of surface proteins of the viruses to vaccine strains, we statistically quantified the relationship between the genetic mismatch and vaccine effectiveness (VE) for influenza A/H1N1pdm09, A/H3N2, influenza B and COVID-19. We also proposed a systematic approach to integrate multiple genes and virus subtypes for overall VE estimation. Evident relationships were identified and validated in independent data. The modelling framework may enable in silico prediction for VE on a real-time basis and inform public health planning.
Xiao-Li Meng

Miniaturizing Data Defect Correlation: A Versatile Strategy for Handling Non-Probability Samples

Biography

Xiao-Li Meng, the Founding Editor-in-Chief of HDSR and the Whipple V. N. Jones Professor of Statistics, and the Founding Editor-in-Chief of Harvard Data Science Review, is well known for his depth and breadth in research, his innovation and passion in pedagogy, his vision and effectiveness in administration, as well as for his engaging and entertaining style as a speaker and writer. Meng was named the best statistician under the age of 40 by COPSS (Committee of Presidents of Statistical Societies) in 2001, and he is the recipient of numerous awards and honors for his more than 150 publications in at least a dozen theoretical and methodological areas, as well as in areas of pedagogy and professional development. In 2020, he was elected to the American Academy of Arts and Sciences. He has delivered more than 400 research presentations and public speeches on these topics, and he is the author of “The XL-Files,” a thought-provoking and entertaining column in the IMS (Institute of Mathematical Statistics) Bulletin. His interests range from the theoretical foundations of statistical inferences (e.g., the interplay among Bayesian, Fiducial, and frequentist perspectives; frameworks for multi-source, multi-phase and multi-resolution inferences) to statistical methods and computation (e.g., posterior predictive p-value; EM algorithm; Markov chain Monte Carlo; bridge and path sampling) to applications in natural, social, and medical sciences and engineering (e.g., complex statistical modeling in astronomy and astrophysics, assessing disparity in mental health services, and quantifying statistical information in genetic studies). Meng received his BS in mathematics from Fudan University in 1982 and his PhD in statistics from Harvard in 1990. He was on the faculty of the University of Chicago from 1991 to 2001 before returning to Harvard, where he served as the Chair of the Department of Statistics (2004-2012) and the Dean of Graduate School of Arts and Sciences (2012-2017).
Abstract

Non-probability samples are deprived of the powerful design probability for randomization-based inference. This deprivation, however, encourages us to take advantage of a natural divine probability that comes with any finite population. A key metric from this perspective is the data defect correlation (ddc), which is the model-free finite-population correlation between the individual's sample inclusion indicator and the individual's attribute being sampled. A data generating mechanism is equivalent to a probability sampling, in terms of design effect, if and only if its corresponding ddc is of $N^{-1/2}$ (stochastic) order, where $N$ is the population size (Meng, 2018, AOAS). Consequently, existing valid linear estimation methods for non-probability samples can be recast as various strategies to miniaturize the ddc down to the order of $N^{-1/2}$. The quasi design-based methods accomplish this task by diminishing the variability among the $N$ inclusion propensities via weighting. The super-population model-based approach achieves the same goal through reducing the variability of the $N$ individual attributes by replacing them with their residuals from a regression model. The double robust estimators enjoy their celebrated property because a correlation is zero whenever one of the variables being correlated is constant, regardless of which one. Understanding the commonality of these methods through ddc reveals the possibility of “double-plus robustness”: a valid estimation without relying on the full validity of either the super-population model or the estimated inclusion propensity, neither of which is guaranteed because both rely on device probability.
Inchi Hu

Tweedie's formula for chi-squared distributions

Biography

Inchi Hu is a professor of Statistics at the College of Engineering and Computing in George Mason University. He received his PhD in statistics from Stanford University. Prior to joining Mason, he was on the faculty of Hong Kong University of Science and Technology (HKUST), University of Pennsylvania, and University of Maryland.

Abstract

The exceptional virtue of Tweedie's formula for the normal distribution, first reported in Robbins (1956), lies in its representation of selection bias as a simple function of the derivative of log marginal likelihood. Since the marginal likelihood and its derivative can be estimated from the data directly without invoking prior information, bias correction can be carried out conveniently. We propose a Bayesian hierarchical model for chi-squared data such that the resulting Tweedie's formula has the same virtue as that of the normal distribution. Because the family of noncentral chi-squared distributions does not constitute an exponential family, our results cannot be obtained by extending existing results. Furthermore, the corresponding Tweedie's formula manifests new phenomena quite different from those of the normal distribution and suggests new ways of analyzing chi-squared data. This is a joint work with Lilun Du.
Jonathan Auerbach

Causal Inference in High Dimensions with the Influence Score

Biography

Jonathan Auerbach is an Assistant Professor in the Department of Statistics at George Mason University. His research covers a wide range of topics at the intersection of statistics and public policy. He has measured selection bias in mortality studies (with Andrew Gelman) and traffic safety studies (with Shaw-Hwa Lo and Tian Zheng) and assessed the quality of the 2020 census (with Paul Biemer and Joseph Salvo). His work also investigates urban myths such as there are not eight million rats in New York City, there are traffic ticket quotas, and the average building is not getting taller (with Phyllis Wan).

Abstract

For many important policies in industry and government, program participants cannot be randomly selected due to ethical or logistical considerations. Adjusting for non-random selection is complicated when researchers do not know which of a large number of potential factors were used for selection. This paper proposes a new measure of association, the influence score, to identify the factors that control for selection bias in high-dimensional panel data. Intuitively, the influence score works by deemphasizing sparsely populated strata, which can distort other, more traditional, measures of association. The paper illustrates the proposed method using data from Vision Zero, a civil-engineering policy aimed at eliminating traffic fatalities. It identifies a variety of effective traffic-safety changes that were overlooked by other analyses, such as speed humps and enhanced crossings.
Hans-Georg Müller
Distributional Regression Models

Biography
Professor of Statistics at UC Davis, Research on Functional Data Analysis, Random Objects (including optimal transport and data on manifolds), Longitudinal Data Analysis, Brain Imaging, Plant Genomics and the Biodemography of Aging. Fellow of IMS, ASA and Am Assoc for the Advancement of Science. IMS Medallion Lecture, Noether Senior Research Award, Humboldt Research Prize, IMS Rietz Lecture.

Abstract
The analysis of samples of random objects that do not lie in a vector space has found increasing attention in statistics in recent years. An important class of such object data are univariate probability measures defined on the real line. Regression models for data that include random distributions as predictors or responses are of interest for various applications. Established approaches including global transformations and Frechet regression for the case where responses are distributional will be briefly reviewed. A recent method is Wasserstein regression that utilizes the geometry of tangent bundles of the Wasserstein metric space. It turns out that distributional regression can also be based on optimal transports in conjunction with a novel transport algebra. This approach is particularly useful for autoregressive distributional modeling in the context of distributional time series. These distributional regression methods are supported by theory and are shown to lead to useful insights for data in demography and climatology.

This talk is based on joint work with Yaqing Chen, Zhenhua Lin and Changbo Zhu.
Joseph B ("Jay") Kadane

A Bayesian View of Constraints on a 2x2 Table: A preliminary report

Biography
Joseph B ("Jay") Kadane is the Leonard J. Savage Professor of Statistics and Social Sciences at Carnegie Mellon University. He has a B.A. in mathematics from Harvard University, and a Ph.D. in statistics from Stanford University. Before coming to Carnegie Mellon University in 1971, he was at Yale and at the Center for Naval Analysis. He served as department head of CMU Statistics 1972–1981. He is an elected fellow of the American Statistical Association, the Institute of Mathematical Statistics, the American Association for the Advancement of Science, the Center for Advanced Study in the Behavioral Sciences, the International Society for Bayesian Analysis and the American Academy of Arts and Sciences. He was awarded a Fulbright Fellowship to Chile in 2004. While he became emeritus in 2006, he continues to be active. In 2014, he was awarded the DeGroot Prize for my book, "Principles of Uncertainty."

Abstract
Sum constraints on the rows and/or columns of a 2x2 table were subject of intense debate among Fisher, Yates, Barnard, Berkson, and others in the 1930's and 40's. This talk takes a Bayesian perspective, which mandates setting aside questions about how to conduct significance tests of sharp null hypotheses. With a common prior, common data, and the common goal of a posterior distribution on the log odds ratio, the talk contrasts three scenarios:

(i) no constraint

(ii) sum constraints on either the rows or the columns, but not both

(iii) sum constraints on both.

When the common prior is a Dirichlet distribution, (i) and (ii) have the same posterior distribution, which differs from that of (iii).

Joint work with Ruobin Gong, Teddy Seidenfeld, Mark Schervish and Rafael Stern.
Iuliana Ionita-Laza

Knockoff-based statistics for the identification of putative causal loci in genetic studies

Biography

Iuliana received her PhD in 2006 from the Courant Institute of Mathematical Sciences at NYU, followed by a postdoctoral fellowship at Harvard School of Public Health. She has been a member of the Department of Biostatistics at Columbia since 2009. Her main research interests lie at the interface of statistics and genomics. She is particularly interested in developing statistical and machine learning methods for the analysis of high-dimensional genetic and functional genomics data. She is also involved in applications of such methods to understand the genetic basis of complex diseases and traits, including autism spectrum disorders, schizophrenia and Alzheimer’s disease. She has been the PI of multiple grants from NIH and NSF, and has served on the editorial boards of Biometrics and Statistics in Biosciences. She is also a regular member of the NIH Genomics, Computation and Technology (GCAT) study section.

She is currently leading a genomics program that brings together an interdisciplinary group of people from multiple departments across Columbia University with diverse research expertise in statistical/computational genomics and other omics, computational biology, biomedical informatics, and interested in understanding biology and human health.

Abstract

Knockoff-based methods are becoming increasingly popular due to their enhanced power for locus discovery and their ability to prioritize putative causal variants in a genome-wide analysis. However, because of the substantial computational cost for generating knockoffs, existing knockoff approaches cannot analyze biobank-scale datasets. I will discuss a scalable knockoff-based method for population-based designs, and related extensions to family-based designs. I will show applications to the UK biobank data, and some family-based studies for autism spectrum disorders.
Adeline Lo

Refugee narratives in news and their influence on German native attitudes and behaviors

Biography

Adeline Lo is an Assistant Professor in the Department of Political Science at the University of Wisconsin-Madison. She researches the factors that motivate or mitigate conflict between groups, with an eye towards investigating migrant inclusionary politics. Adeline also designs statistical tools for prediction and measurement for applied social science data. Her research has led her to work with and construct tools for observational, text, experimental and network data.

Abstract

The forcibly displaced, of which refugees constitute a quarter in proportion, number in over 80 million people worldwide as of the end of 2020 (UNHCR Global Trends 2020). Some of the flows of this population have been due to protracted conflicts and instability in home regions, such as the Syrian refugee crisis, and others from abrupt changes in international politics and regional tensions, such as displacements surrounding the withdrawal of U.S. troops in Afghanistan in 2021. Given the inherent vulnerabilities of refugees, it is even more important to study how national news media have portrayed news related to the population, whether these dynamics have changed over time, and if/how certain frames with which their stories have been presented influence native attitudes and behaviors towards them. This latter question is motivated by the rise in nativist rhetoric in response to immigrants and refugees in many Western host countries (Hangartner et al. 2019). The proposed work takes a within-country sweep of broadcasted television representation and combines it with a causally identified, micro-level randomized survey experiment. We harness a unique broadcast television corpora collected and preprocessed from the ARD Tagesschau, spanning over 40,000 minutes of national broadcast news (daily German national TV 2014-2019) to answer the first question. We find that refugee stories have been increasingly presented a) with more frequency, b) under a topical framing of criminal justice, and c) visually presenting more young faces. Relying on an observational German longitudinal panel survey we demonstrate that watching national broadcast news is correlated with attitudes towards migrants. Finally, we use a randomized survey experiment to evaluate refugee video story frames and their effects on citizen attitudes and policy preferences towards refugees. Beyond contributing to the literatures on the study of high dimensional unstructured data and migration politics, the findings will likely be of interest to policymakers, refugee advocacy organizations and applied researchers working with unstructured data.
Kosuke Imai

Safe Policy Learning through Extrapolation: Application to Pre-trial Risk Assessment

Biography

Kosuke Imai is Professor in the Department of Government and the Department of Statistics at Harvard University. Before moving to Harvard in 2018, Imai taught at Princeton University for 15 years where he was the founding director of the Program in Statistics and Machine Learning. Imai specializes in the development of statistical methods and machine learning algorithms and their applications to social science research.

Abstract

Algorithmic recommendations and decisions have become ubiquitous in today's society. Many of these and other data-driven policies, especially in the realm of public policy, are based on known, deterministic rules to ensure their transparency and interpretability. For example, algorithmic pre-trial risk assessments, which serve as our motivating application, provide relatively simple, deterministic classification scores and recommendations to help judges make release decisions. How can we use the data based on existing deterministic policies to learn new and better policies? Unfortunately, prior methods for policy learning are not applicable because they require existing policies to be stochastic rather than deterministic. We develop a robust optimization approach that partially identifies the expected utility of a policy, and then finds an optimal policy by minimizing the worst-case regret. The resulting policy is conservative but has a statistical safety guarantee, allowing the policy-maker to limit the probability of producing a worse outcome than the existing policy. We extend this approach to common and important settings where humans make decisions with the aid of algorithmic recommendations. Lastly, we apply the proposed methodology to a unique field experiment on pre-trial risk assessment instruments. We derive new classification and recommendation rules that retain the transparency and interpretability of the existing instrument while potentially leading to better overall outcomes at a lower cost.
Henry Horng-Shing Lu received his Ph.D. degree in Statistics from Cornell University in 1994 and his B.S. degree in electrical engineering from National Taiwan University in 1986. He is a Professor in National Chiao Tung University (NCTU) that is merged to National Yang Ming Chiao Tung University (NYCU). He has served as the Vice President for Academic Affairs and Dean for College of Science in NCTU. He is an Elected Member of the International Statistical Institute and Principal Fellow of the High Education Academy. His research interests include statistics, data science, machine learning, image science, biomedical studies and industrial applications.

Abstract

Artificial intelligence is widely used in many situations, so it is important to prevent bias when designing the deep learning methods. However, the training data may contain discrimination and bias when collecting those data from specific populations or groups. The discriminatory behavior of the model will result in distinct prediction performance among different groups. Thus, eliminating the discriminatory effect and making the model fair become an important issue. To mitigate this effect, we develop a new method on the deep learning model to make the model fair while keeping high prediction performance. We adopt Influence Score in the proposed model, which is the feature selection algorithm that takes the interaction between multiple variables into consideration, and the influential features will be included in follow-up predictions. Through this method, the fair model only contains the important features not influenced by the discriminatory factor and keeps its high prediction performance. In the experiment, we apply the method to the ISIC 2019 and Asan skin lesion datasets. The ISIC 2019 dataset is mainly collected from Western, and the Asan dataset is mainly collected from South Korea. Because in the skin lesion literature they show that the probability of diagnosed skin lesion in Western is significantly higher than Asian, we consider the area information as the bias of the prediction model. The result shows that the fair model can fairly and precisely classify the types of skin lesions by eliminating the discriminatory information.
Jane-Ling Wang

Deep Learning for the Partial Linear Cox Model

Biography

Dr. Jane-Ling Wang is Distinguished Professor of Statistics at the University of California, Davis. Her research areas include deep learning, functional data analysis, neuroimaging data analysis, semiparametric and nonparametric modeling approaches and survival analysis. She is interested in theory as well as applications, and has received the Noether Senior Research Award (2016), the ICSA Distinguished Achievement Award (2018) and the Humboldt Research Award (2020). She has been a co-editor of Statistica Sinica (2002-2005) and is currently a co-editor of JASA Theory and Methods (since 2000).

Abstract

While deep learning approaches to survival data have demonstrated empirical success in applications, most of these methods are difficult to interpret and mathematical understanding of them is lacking. This paper studies the partially linear Cox model, where the nonlinear component of the model is implemented using a deep neural network. The proposed approach is flexible and able to circumvent the curse of dimensionality, yet it facilitates interpretability of the effects of treatment covariates on survival. We establish asymptotic theories of maximum partial likelihood estimators and show that our nonparametric deep neural network estimator achieves the minimax optimal rate of convergence (up to a poly-logarithmic factor). Moreover, we prove that the corresponding finite-dimensional estimator for treatment covariate effects is $\sqrt{n}$-consistent, asymptotically normal, and attains semiparametric efficiency. Extensive simulation studies and analyses of two real survival datasets show the proposed estimator produces confidence intervals with superior coverage as well as survival time predictions with superior concordance to actual survival times.
David Siegmund
Changepoint Like Problems

Biography
David Siegmund is the John D. & Sigrid Banks Professor at Stanford University. He was born in 1941 in Webster Groves, Missouri, where he lived for the next seventeen years. He earned a Ph. D. in mathematical statistics from Columbia University. He has taught at Columbia as well as Stanford, where in addition to teaching he served for three years as Associate Dean for the Natural Sciences. He has held visiting positions in Jerusalem, Heidelberg, Oxford, Cambridge, the Free University of Amsterdam, and Singapore. He is an elected member of the National Academy of Sciences and the American Academy of Arts and Sciences. He has served on advisory committees for the National Science Foundation, the National Research Council and for research institutes in Berkeley, Palo Alto, Singapore, and Freiburg im Breisgau, Germany.

Before 1985 he was primarily interested in sequential analysis, especially the design and analysis of sequential clinical trials. A related interest has been “change-point” detection, where one looks for changes in the measurement of a noise corrupted process that indicate changes in the mechanism generating the process. For 15 years beginning in the early 1990s he concentrated on statistical aspects of genetics, specifically gene mapping and detection of different genomic “signals,” most of which have similarities with change-point detection.

Abstract
We consider the problem of segmentation of normal observations according to changes in their mean. Changes can occur continuously, e.g., changes in the slope of a regression line, bumps, or discontinuously, e.g., jump in the mean level. Theoretical results are illustrated by applications to copy number changes, historical temperature records, incidence of COVID-19, and of hate crimes. Confidence regions for the change-points and some difficulties associated with dependent observations are also discussed.

Aspects of this research involve collaboration with Fang Xiao, Li Jian, Liu Yi, Nancy Zhang, Benjamin Yakir, Li (Charlie) Xia, and Keith Worsley.