The Israel Statistics and Data Science Association
الرابطة الإسرائيلية للإحصاء وعلم البيانات
האיגוד הישראלי לסטטיסטיקה ומדע הנתונים

Annual conference - 1 June, 2023

Abstracts

MORNING PLENARY SESSION

Susan Murphy, Harvard University

We used Reinforcement Learning; but did it work?

Reinforcement Learning provides an attractive suite of online learning methods for personalizing interventions in Digital Behavioral Health. However after an reinforcement learning algorithm has been run in a clinical study, how do we assess whether personalization occurred? We might find users for whom it appears that the algorithm has indeed learned in which contexts the user is more responsive to a particular intervention. But could this have happened completely by chance? We discuss some first approaches to addressing these questions.

I A: STATISTICAL THEORY (Chair: Itai Dattner, Haifa University)

Valentin Vancak, Karolinska Institute

Sensitivity Analysis of G-estimators to Invalid Instrumental Variables

Instrumental variables regression is a tool that is commonly used in the analysis of observational data. The instrumental variables are used to make causal inference about the effect of a certain exposure in the presence of unmeasured confounders. A valid instrumental variable is a variable that is associated with the exposure, affects the outcome only through the exposure (exclusion), and is not confounded with the outcome (exogeneity). Unlike the first assumption, the other two are generally untestable and rely on subject-matter knowledge. Therefore, a sensitivity analysis is desirable to assess the impact of assumptions' violation on the estimated parameters. In this paper, we propose and demonstrate a new method of sensitivity analysis for G-estimators in causal linear and non-linear models. We introduce two novel aspects of sensitivity analysis in instrumental variables studies. The first is a single sensitivity parameter that captures violations of exclusion and exogeneity assumptions. The second is an application of the method to non-linear models. The introduced framework is theoretically justified and is illustrated via a simulation study. Finally, we illustrate the method by application to ..real-world data and provide guidelines on conducting sensitivity analysis

Ilan Livne, Technion

A zero-estimator approach for estimating the signal level in a high-dimensional regression setting

We study a high-dimensional linear regression model in a semi-supervised setting, where for many observations only the vector of covariates X is given with no responses Y. We do not make any sparsity assumptions on the vector of coefficients, nor do we assume normality of the covariates. We aim at estimating the signal level, i.e., the amount of variation in the response that can be explained by the set of covariates. We propose an estimator, which is unbiased, consistent, and asymptotically normal. This estimator can be improved by using a zero-estimator approach, where a zero-estimator is a statistic arising from the unlabeled data, whose expected value is zero. More generally, we present an algorithm based on the zero-estimator approach that in principle can improve any given estimator. We further relax the linearity assumption, study some asymptotic properties of the proposed estimators, and demonstrate their finite sample performance in simulated and real datasets.

Havi Murad, Gertner Institute

Missing time-dependent covariate values in a Cox model – Joint Models approach versus combination of Multiple Imputation and Joint Models

We present a novel combination of two approaches used when estimating the association between a time-dependent covariate (marker), measured with missing values, and a survival outcome: multiple imputation (MI) and jointly modeling longitudinal and survival data (JM).
We have previously developed a procedure for imputing missing values for time-dependent covariates in a discrete time Cox model using the chained equations method. This time-sequential MI procedure multiply imputes the missing values for each time-period in a time-sequential manner, using completed covariates (imputed and observed) from previous time-periods, but not from future ones, as well as the survival outcome. Recently, we have developed a Fully Conditional Specification MI version that is compatible with the substantive model, in our case the discrete Cox model. Following Bartlett et al., we term it the Substantive Model Compatible FCS (SMC-FCS). In this method, the missing values are imputed in a chain over all time-periods. In each time-period, the imputation model includes past-imputed values of the marker, as well as future ones in time-periods before the event has occurred, and the survival outcome. Both versions of MI can be applied using the MI procedure in SAS with FCS statement or using similar packages in other software, e.g. the mice package in R.
In this presentation, we demonstrate a novel two-step approach, which: (i) multiply imputes the missing values and then (ii) applies JM to each completed data file and combines the estimates using Rubin's rule. In the first step, we present two versions: Time-sequential MI and SMC-FCS MI. We therefore compare three methods: (i) Time-sequential MI + JM (ii) SMC-FCS MI + JM and (iii) a one-stage JM (simple JM). The JM can be executed using the packages JointModel or JointModelBayes in R.
We use simulations based on data of glucose control variables (plasma glucose and %HbA1c) among diabetic patients, from a large Israeli population-based cohort database (n=546,000) [4], using these methods to evaluate the association of glucose control with the risk of cancer. We examine different patterns of missing data in the glucose control variables (completely missing at random, missing at random and non-missing at random) and the impact of these patterns on the performance of the three methods.

Barak Sober, HUJI

Estimation of Manifolds and Manifold-Based Estimation from Noisy Samples

A common observation in data-driven applications is that data has a low intrinsic dimension, at least locally. Thus, when one wishes to work with data that is not governed by a clear set of equations but still wishes to perform statistical or other scientific analysis, an optional model is the assumption of an underlying manifold from which the data was sampled. Accordingly, the manifold is not given explicitly, but we can obtain samples of it (i.e., the individual data points).

In this talk, we will consider the problem of estimating manifolds from a finite set of samples, possibly recorded with noise. We provide an algorithm to approximate the manifold and prove its convergence properties. We will further explain how one could use this framework to perform optimization of functions whose domains are such manifolds.
The motivation for this work is based on the analysis of the evolution of shapes through time (e.g., handwriting or primates' teeth).

I B: DATA SCIENCE (Chair: Tal Sarig, Meta)

Tamir Bendory, TAU

Signal estimation under algebraic groups

In this talk, I will introduce a family of estimation problems, in which each observation of the signal is corrupted not only by noise but also by an unknown element of a known group. These models are mainly motivated by structural biology technologies, such as single-particle cryo-electron microscopy and X-ray free-electron lasers. In particular, I will discuss recent results on the sample complexity of these models, for generic and sparse signals, and their relation to the algebraic structure of the statistical model.

Tom Hope, HUJI

NLP for Scientific & Clinical Predictive Models

With over one million papers added every year to the PubMed biomedical index alone — the explosion of scholarly knowledge presents tremendous opportunities for accelerating research across the sciences. In this talk, I will present our recent work toward helping researchers and clinicians make use of knowledge embedded in the literature. In particular, I will focus on methods that use information in the literature for training predictive models. This includes models that predict (1) clinical outcomes of hospital patients, (2) new links in biomedical knowledge graphs, and (3) hypotheses in AI research.

Nir Rosenfeld, Technion

Strategic Classification: Learning with Data that “Behaves”

The growing success of machine learning has made it appealing as a tool for informing decisions about humans. But humans are not your conventional input: they have goals, beliefs, and aspirations, and take action to promote their own interests. Given that standard learning methods are not designed to handle inputs that "behave"---a natural question is: how should we design learning systems when we know they will be deployed and used in social settings?

As a starting point, I will present the problem of strategic classification, in which users can modify their features (at a cost) in response to a learned classifier in order to obtain favorable predictions. I will then describe some of our work in this field, demonstrating how even mild forms of strategic behavior can dramatically transform the learning problem. Finally, I will argue for strategic classification as a useful formal framework for reasoning about learning under strategic user behavior in general, and which holds potential for applying more elaborate forms of economic modeling.

Daniel Yekutieli. TAU

Empirical Bayes for Big problems

I will discuss Big data problems that their inherent invariance to a group of actions make them amenable to empirical Bayes analysis. I will explain this property and show how it occurs in the Normal linear model, Normal covariance matrix estimation and detection of correlated databases. I will show that this property implies the existence of “empirical” Oracle Bayes rules that minimize frequentist risks and yield optimal tests for non-null discovery. I will demonstrate how these Oracle Bayes rules may be approximated by hierarchical Bayes procedures.

II A: STATISTICAL MODELS & APPLICATION (Chair: Jonathan Rosenblatt, Fairmatic)

Wiessam Abu Ahmad, HUJI

Utilizing diagnostic analytics in meta-analysis: Lessons from combining results of individual studies addressing the association between $PM_{2.5}$ and birth weight

Meta-analysis, used to combine the results of individual studies, has become an important statistical analysis in life sciences, including public health and environmental epidemiology. The fixed-effect model used to combine studies when they share the same true effect size (homogeneous population) but results deviate due to sampling error. The random-effect model is used when studies have different true effect sizes, due to different study characteristics or target populations, suggesting another source of variability. However, when the between-studies heterogeneity is substantial, the recommendation is not to combine the studies.

In this talk we review analytic and graphic tools to explore sources of heterogeneity and identify potential outlying studies. These tools include Cochran's Q, I2 statistics and GOSH plots to evaluate heterogeneity, as well as the Viechtbauer and Cheung approach and Baujat plots to investigate the influence of individual studies. Additionally, we evaluate publication bias using funnel plot, p-curve, and Egger’s test.

We demonstrate these methods for 57 studies that investigated the relationship between exposure to particulate matter with aerodynamic diameter ≤2.5 μm (PM2.5) during pregnancy and continuous birth weight (BW) or binary indicator for low birth weight (LBW, BW<2500g).

The review protocol was registered on the PROSPERO website (CRD42020188996) and followed PRISMA guidelines.

Our diagnostic analysis demonstrated substantial heterogeneity for both outcomes and we identified two outlying studies for LBW that affect the results. The sources of heterogeneity between studies included study region and time period. Therefore, we suggest not to pool summary measures of the associations between PM2.5 and birth outcomes, and that policy would be informed by local evidence.

Yuval Nov, Haifa University

Modeling and mitigating taxonomic bias in citizen-science biodiversity data

Global internet platforms such as iNaturalist and eBird allow "citizens" (i.e., non-experts) to document and share wildlife sightings, and currently hold hundreds of millions of observations. Such data, however, is not as reliable as data collected through traditional scientific protocols.

A common problem is “taxonomic bias”, whereby the personal preferences of people toward certain species affect their documentation patterns.

We have devised a statistical learning methodology that mitigates taxonomic bias in citizen-science biodiversity data. In one inference approach, we assume that nothing is known a priori about the preferences of the observers or about the encounter rates with a target species; under this assumption, we can estimate only the ratios of the unknown encounter rates across locations and times. In another approach, we assume that a small sub-group of observers have known preferences, or that the true encounter rates in a small portion of the domain have been reliably estimated; under this assumption, we are able to estimate the absolute encounter rates for the entire domain considered.

Inbal Goldshtein, KI - Computational Health Institute

Causality under fire: Covid19 vaccines during pregnancy & delivery

ניסויים קליניים לחיסון קורונה החריגו נשים הרות, ובהיעדר מידע אימהות רבות בעולם חששו להתחסן והתקשו לחזור לשגרה. בעקבות גל תחלואה ודיווחים על סיבוכים בקרב נדבקות הרות, ישראל הייתה הראשונה להרחיב את מדיניות חיסוני קורונה להכללת נשים הרות. מידע תצפיתי נדרש בדחיפות כדי להשלים את התמונה של יעילות ובטיחות החיסון לאם והילוד. נדבר על חילוץ ושימוש משני בדטה דינמי וטרי, הרכבת פאזל מריבוי מקורות נתונים, מקדמים תלויי זמן במודל הישרדות, ושיטות חישוביות המדמות תנאי ניסוי אקראי.

Tal Galili and Tal Sarig, Meta

Balancing biased data samples with the 'balance' Python package

The “balance” Python package is a new open-source software by Meta (released in late 2022). The package offers a simple workflow and methods for dealing with biased data samples when looking to infer from them to a population of interest.

Bias in survey data is often the result of survey non-response or when the data collection suffers from sampling bias. Directly inferring insights from data with such biases can result in erroneous estimates. Hence, it is important for practitioners to understand if and how data is biased and, when possible, use statistical methods to minimize such biases.

The “balance” package addresses this issue by providing a simple, easy-to-use, framework for weighing data and evaluating its biases. The package is designed to provide best practices for weight fitting and offers several modeling approaches. The methodology in “balance” can support ongoing automated survey data processing, as well as ad-hoc analyses of survey data.

The main workflow API of balance includes three steps:

(1) understanding the initial bias in the data relative to a target we would like to infer,
(2) adjusting the data to correct for the bias by producing weights for each unit in the sample based on propensity scores, and
(3) evaluating the final biases and the variance inflation after applying the fitted weights.

The adjustment step provides a few alternatives for the researcher to choose from: Inverse propensity weighting using logistic regression model based on LASSO (Least Absolute Shrinkage and Selection Operator), Covariate Balancing Propensity Scores, and post-stratification. The focus is on providing a simple to use API, based on Pandas data-frame structure, which can be used by researchers from a wide spectrum of fields.

In this talk, we present the capabilities of the balance package, demonstrating the flow of the package and its ease of use.

II B: STATISTICS & MACHINE LEARNING (Chair: Amichai Painsky, TAU)

Amit Moscovich, Stat. TAU

Data preprocessing can break cross-validation

Regression and classification methods are typically evaluated by cross-validation: repeatedly splitting the data into a training set and a validation set, learning a predictive model on the training set, then averaging its loss on the validation set. Under the i.i.d. assumption, this gives an unbiased and consistent estimator for the risk of the trained model. However, in practice, many data sets go through various stages of preprocessing, such as rescaling, dimensionality reduction, and outlier removal. Such “unsupervised” preprocessing procedures, that do not involve the responses or class labels, are often considered harmless. However, they introduce a subtle leakage of information from the validation set to the trained model. This breaks the assumptions of cross-validation, potentially leading to biased risk estimates and sub-optimal model selection. In this talk, we make the case that this subtle error should receive more attention since it is prevalent in scientific research, potentially harmful, and typically easy to fix. We will explain where the bias is coming from and how to eliminate it.

Joint work with Saharon Rosset.

Aryeh Kontorovich, BGU

Local Glivenko-Cantelli (or: estimating the mean in infinite dimensions)

If μ is a distribution over the d-dimensional Boolean cube $\{0,1\}^d$, our goal is to estimate its mean $p \in [0,1]^d$ based on n iid draws from μ. Specifically, we consider the empirical mean estimator $\hat{p}_n$ and study the maximal deviation $M=\max_{j\in[d]} |\hat{p}_n(j)-p(j)|$. In the classical Universal Glivenko-Cantelli setting, we seek distribution-free (i.e., independent of μ) bounds on M. This regime is well-understood: for all μ, we have $\mathbb{E}[M] \lesssim \sqrt{\log(d)/n}$ up to universal constants, and the bound is tight. Our present work seeks to establish dimension-free (i.e., without an explicit dependence on d) estimates on M, including those that hold for $d=\infty$. As such bounds must necessarily depend on μ, we refer to this regime as *Local* Glivenko-Cantelli, and are aware of very few previous bounds of this type — which are quite sub-optimal. Already the special case of product measures μ is quite non-trivial. We give necessary and sufficient conditions on μ for $\mathbb{E}[M] \to 0$, and discover a novel sub-Gamma-type maximal inequality for shifted Bernoullis. A number of challenging open problems are posed for future research. Joint work with Doron Cohen.

Wasim Huleihel, TAU

Statistical-Computational Gaps in Modern Statistics

The ever-increasing size and complexity of modern data sets pose many challenges for statistical inference, while classical statistical analysis has lagged behind. Among these challenges are fundamental questions related to the interplay and tradeoffs between statistical and computational requirements, which emerge due to the presence of complex underlying combinatorial structures in modern high-dimensional problems. In this talk, we will discuss some of our recent novel developments in this research area.

Giora Simchoni, TAU

Integrating Random Effects in Deep Neural Networks

Modern approaches to supervised learning like deep neural networks typically implicitly assume that observed responses are statistically independent. In contrast, correlated data are prevalent in real-life large-scale applications, with typical sources of correlation including spatial, temporal and clustering structures. These correlations are either ignored by DNNs, or ad-hoc solutions are developed for specific use cases. We propose to use the mixed models framework to handle correlated data in DNNs. The key to combining mixed models and DNNs is using the Gaussian negative log-likelihood (NLL) as a natural loss function that is minimized with DNN machinery

including stochastic gradient descent (SGD). Since NLL does not decompose like standard DNN loss functions, the use of SGD with NLL presents some theoretical and implementation challenges, which we address. We present excellent results in a regression setting for tabular datasets, as well as preliminary results for image and text datasets, and other settings such as classification and dimensionality reduction.

AFTERNOON PLENARY SESSION

סמדר לב , הוועדה לתכנון וניתוח סקרים, למ"ס

האם סקר טוב הוא המצאה?

תקציר: מה הוא סקר טוב? איך ניתן להבחין בין סקר טוב לרע? מה נדרש כדי לתכנן ולהפיק סקר טוב? הנחיות לסקרים מפותחות בימים אלה בלשכה המרכזית לסטטיסטיקה, בשיתוף עם בעלי עניין ומומחים שונים. ההנחיות נולדו מתוך צורך, הן של משרדי ממשלה והן של גופי מחקר ומכונים פרטיים, להפיק סטטיסטיקה איכותית. בהרצאה תינתן סקירה קצרה על ההנחיות שפותחו.

אליסף ורמן, ראש אגף תכנון ומדיניות, למ"ס

הלשכה המרכזית לסטטיסטיקה ומהפכת הדאטה

תקציר: מהפכת הדאטה מגיעה לממשלה, ובעלי התפקידים השונים מבקשים לצרוך יותר ויותר נתונים לטובת מדיניות, קבלת החלטות, בקרה על ביצועי הממשלה ועוד. מה תפקידה של הלמ"ס בתוך התהליך, האם הלמ"ס ערוכה לתת מענה לצרכים השונים? ננסה לגעת בסוגיות השונות ולשרטט את תמונת העתיד, מתוך שולחן העבודה האסטרטגי של הלשכה, כפי שדחף וקידם אותן דני בשנות כהונתו.