

ISDSA Annual Conference

May 30, 2024, National Library of Israel, Jerusalem

Abstracts

I A: Statistical Methods in Medical Research

Chair: Malka Gorfine, TAU

Bella Vakulenko-Lagun, Haifa U

Regression analysis with censored covariates in the presence of a cured fraction

Many health surveys collect data in a cross-sectional way and record only a current value of a patient reported outcome. The problem is that at the time of data collection, some of the important events, related to the studied outcome, had happened for some of the survey participants, but not for others. In addition, the incomplete observation of this time-to-event covariate might be complicated by the presence of a cured fraction (i.e., those patients who belong to the target population but will never experience this event). There are only few available methods for regression analysis with censored covariates, and none of them accounts for a cured fraction. We propose a 2-stage estimation approach that allows to disentangle, identify and estimate the effects of baseline covariates on cured and uncured subjects and the effects of a post-baseline time-to-event covariate on the uncured subjects. We establish asymptotic properties of the proposed estimator. Finally, we apply our approach to the data on long-term outcomes of Perthes' disease, a rare childhood disorder in a hip.

Rachel Axelrod, TAU

A Sensitivity Analysis Approach for the Causal Hazard Ratio in Randomized and Observational Studies

The Hazard Ratio (HR) is often reported as the main causal effect when studying survival data. Despite its popularity, the HR suffers from an unclear causal interpretation due to a built-in selection bias. While alternative approaches exist, the HR remains the most popular measure used by practitioners, and therefore, analysis approaches directly targeting a causally interpretable HR are of interest. A recently proposed alternative is the causal HR, defined as the ratio between hazards across treatment groups among the study participants that would have survived regardless of the assigned study group. We discuss the challenge in identifying the causal HR from the observed data and present a sensitivity analysis

approach for identification in randomized controlled trials utilizing a working frailty model. We further extend our framework to adjust for potential confounders using inverse probability of treatment weighting. We present a Cox-based and non-parametric kernel-based estimation under right censoring. We study the finite-sample properties of the proposed estimation methods through simulations and illustrate the utility of our framework using two real-data examples.

Jonathan Woodbridge, *HUJI*

Estimating Mean Viral Load Trajectory from Intermittent Longitudinal Data and Unknown Time Origin

Viral load (VL) in the respiratory tract is the leading proxy for assessing infectiousness potential. Understanding the dynamics of disease-related VL within the host is very important and help to determine different policy and health recommendations. However, often only partial followup data are available with unknown infection date. In this paper we introduce a discrete time likelihood-based approach to modeling and estimating partial observed longitudinal samples. We model the VL trajectory by a multivariate normal distribution that accounts for possible correlation between measurements within individuals. We derive an expectation-maximization (EM) algorithm which treats the unknown time origins and the missing measurements as latent variables. Our main motivation is the reconstruction of the daily mean SARS-Cov-2 VL, given measurements performed on random patients, whose VL was measured multiple times on different days. The method is applied to SARS-Cov-2 cycle-threshold-value data collected in Israel.

Yael Travis-Lumer, *HUJI*

Pseudo-Observations for Bivariate Survival Data

The pseudo-observations approach has been gaining popularity as a method to estimate covariate effects on censored survival data. It is used regularly to estimate covariate effects on quantities such as survival probabilities, restricted mean life, cumulative incidence, and others. In this work, we propose to generalize the pseudo-observations approach to situations where a bivariate failure-time variable is observed, subject to right censoring. The idea is to first estimate the joint survival function of both failure times and then use it to define the relevant pseudo-observations. Once the pseudo-observations are calculated, they are used as the response in a generalized linear model. We consider two common nonparametric estimators of the joint survival function: the estimator of Lin and Ying (1993) and the Dabrowska estimator (Dabrowska, 1988). For the former, we show that our bivariate pseudo-observations approach produces regression estimates that are consistent and asymptotically normal. Our proposed method enables estimation of covariate effects on quantities such as joint survival probabilities and conditional survival probabilities. We demonstrate the method using simulations and an analysis of two real-world datasets.

Bar Weinstein, TAU

Causal inference with misspecified network interference structure

Under interference, the potential outcomes of a unit depend on treatments assigned to other units. The interference structure between units is typically represented by a network, which is assumed to be accurate but is often only partially or incorrectly measured. This incomplete measurement can result in the misspecification of the interference structure. We study the problems resulting from misspecifying these networks. First, we derive bounds on the bias arising from estimating causal effects under a misspecified network. Then, we propose a novel estimator that leverages multiple networks simultaneously and is unbiased if one of the networks is correct, thus providing robustness to network specification. Additionally, we develop a probabilistic bias analysis that quantifies the impact of a postulated misspecification mechanism on the causal estimates.

We also discuss how a model-based Bayesian analysis can assist in estimating causal effects while accounting for uncertainty in the network. This approach involves a substantial computational burden, which we alleviate using recent advancements in efficient computations of MCMC sampling. We illustrate key issues in simulations and demonstrate the utility of the proposed methods in a social network field experiment and a cluster-randomized trial with suspected cross-clusters contamination.

Yakir Berchenko, BGU

Simplicity Bias in Overparameterized Machine Learning

Traditional learning theory finds itself at odds with the surprising generalization capabilities observed in deep networks and other overparameterized models. Despite their ability to represent highly complex functions that perfectly fit training data (while performing poorly out of sample), these networks actually demonstrate an unexpected proficiency in generalizing to new, unseen data. Here we resolve this by observing that the random construction of these functions inherently biases outcomes towards simpler functions, offering an explanation for their generalization performance. We explore this concept through examples such as learning Boolean functions, analyzing wide (and shallow) networks and deep networks, and examining a surprisingly efficient “trial and error” algorithm.

Ayelet Benjamini, Google

Global Flood Prediction

Floods are one of the most common natural disasters, with a disproportionate impact in developing countries. Accurate and timely warnings are critical for mitigating flood risks, but hydrological simulation models typically must be calibrated to long data records in each watershed. We present an artificial intelligence-based forecasting system that achieves reliability in predicting extreme riverine events at up to a five-day lead time. With these developments, artificial intelligence can provide flood warnings earlier and over larger and

more impactful events in ungauged basins. The model was incorporated into an operational early warning system that produces publicly available forecasts in real time in over 80 countries.

Yuval Nov, *Haifa U*

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.

II A: Statistical Methodology

Chair: Boaz Nadler, Weizmann

Amitai Eldar, *TAU*

Object Detection Under the Linear Subspace Model

Detecting unknown objects in noisy data is a key problem in many fields of science, such as electron microscopy imaging. A common model for the unknown objects is the linear subspace model, which assumes that the objects can be expanded in some known basis (such as the Fourier basis). In this talk, I will present an object detection algorithm that under the linear subspace model is asymptotically guaranteed to find all objects while making only a small percentage of false discoveries.

Shira Yoffe, *HUJI*

Spectral Extraction of Unique Latent Variable

Multimodal datasets contain observations generated by multiple types of sensors. Most works to date focus on uncovering latent structures in the data that appear in all modalities. However, important aspects of the data may appear in only one modality due to the differences between the sensors. Uncovering modality-specific attributes may provide insights into the sources of the variability of the data. For example, certain clusters may appear in the analysis of genetics but not in epigenetic markers. Another example is hyper-spectral satellite imaging, where various atmospheric and ground phenomena are

detectable using different parts of the spectrum. In this paper, we address the problem of uncovering latent structures that are unique to a single modality. Our approach is based on computing a graph representation of datasets from two modalities and analyzing the differences between their connectivity patterns. We provide an asymptotic analysis of the convergence of our approach based on a product manifold model. To evaluate the performance of our method, we test its ability to uncover latent structures in multiple types of artificial and real datasets.

Amichai Painsky, *TAU*

Distribution Estimation under the Infinity Norm

We present novel bounds for estimating discrete distributions under the infinity norm. We provide data-dependent convergence guarantees for the maximum likelihood estimator which significantly improve upon currently known results. Further, we show that our results are nearly optimal in various precise senses. Our proposed scheme utilizes a variety of techniques, from Chernoff-like inequalities to Bernstein empirical bounds. Finally, we demonstrate our suggested framework in a basic selective inference problem, as we infer the most frequent events in the sample.

Joint work with Aryeh Kontorovich.

Yuli Slavutsky, *HUJI*

Robust Zero-Shot Representations for Class Distribution Shifts

In zero-shot learning, training data includes only a subset of the classes that may be encountered when using the model in practice. In many real-life applications, the distribution of classes might shift in deployment, due to changes in class attributes (for example, a shift in gender or race distribution in the task of person identification). Moreover, during training it is usually unknown which attribute is expected to cause the shift. In such cases, the performance on training classes is no longer indicative of the performance on test classes, presenting a challenge: how to learn data representations that are robust to unknown attribute shifts? To address this, we present a new framework that combines hierarchical sampling with out-of-distribution generalization techniques, and demonstrate its effectiveness in achieving improved performance on diverse class distributions.

II B: Large Language Models

Chair: Uri Shalit, Technion

Yoav Goldberg, *BIU & Allen Institute for AI*

Title and abstract TBA.

Ariel Goldstein, *HUJI*

Temporal Structure of Natural Language Processing in the Human Brain Corresponds to Layered Hierarchy of Deep Language Models

Large Language Models (LLMs) provide a novel computational paradigm for understanding the mechanisms of natural language processing in the human brain. Unlike traditional psycholinguistic models, LLMs use layered sequences of continuous numerical vectors to represent words and context, allowing many emerging applications. Our results reveal a connection between human language processing and LLMs, with the LLM's layer-by-layer accumulation of contextual information mirroring the timing of neural activity in high-order language areas.

Ronen Eldan, *Weizmann & Microsoft*

The power of synthetic datasets: From TinyStories to the Phi models

This talk presents a line of research that demonstrates how synthetic datasets generated by large language models can enable training smaller and more efficient models for specific tasks. We first discuss TinyStories, a dataset of short stories using only very simple words, generated by GPT-3.5/4. TinyStories attempts to preserve the essential elements of natural language, such as grammar, vocabulary, facts, and reasoning, while being more compact and focused than typical corpora. While language models as big as 1B parameters often struggle to produce coherent text beyond one or two sentences, we show that TinyStories can be used to train language models that are much smaller than the state-of-the-art models (below 10 million parameters), or have much simpler architectures (with only one transformer block), yet still produce fluent and consistent stories with several paragraphs that are diverse and have almost perfect grammar, and demonstrate certain reasoning capabilities. However, most attempts to create synthetic data using LLMs usually end up in datasets which are very repetitive and seem to lack the diversity which is needed so that a model trained on them would exhibit any ability beyond the memorization of these repeating patterns. The generation of TinyStories relies on the (new) idea of attaining this diversity by injecting randomness into the prompt. Following the same paradigm, we went on to train a series of models, starting with Phi-1 - a new large language model for code, trained using a combination of "textbook quality" data from the web and a dataset of synthetically generated textbooks and exercises. Despite having only 1.3B parameters, it achieves pass@1 accuracy 50.6% on HumanEval and 55.5% on MBPP, surpassing models more than 10 times its size. We discuss the implications of these results for the development, analysis and research of language models, especially for low-resource or specialized domains, and the potential of synthetic datasets to improve the performance and efficiency of LLMs.

Yoav Levine, *Stanford*

Assessing the Economic Rationality of Large Language Models

There is increasing interest in using LLMs as decision-making "agents." Doing so includes many degrees of freedom: which model should be used; how should it be prompted, etc? Settling these questions – and more broadly, determining whether an LLM agent is reliable enough to be trusted – requires a methodology for assessing such an agent's economic rationality. We taxonomize a large set of fine-grained rational decision making "elements" that an agent should exhibit, along with dependencies between them, and propose a benchmark distribution that quantitatively scores an LLMs performance on these elements. Finally, we describe the results of a large-scale empirical experiment with 14 different LLMs, characterizing the current state of the art and the impact of different model sizes on models' ability to exhibit rational behavior.