Indian Statistical Institute

Applied Statistics Unit

SEMINAR NOTICE

Speaker: Abhishek Chakrabortty, Texas A & M University

Title: Semi-supervised Inference with Big Data: Robustness, Efficiency and the Challenges Beyond

Date: 21 June, 2022

Time: 16:15 PM

Venue: ASU Seminar Room

Online Platform: Google Meet (meet.google.com/bkw-estr-sds)

Abstract:

Semi-supervised (SS) settings are of increasing relevance in the 'big data' era. In SS settings, apart from a usual labeled (supervised) data *L*, one also has a much larger sized unlabeled (unsupervised) data *U* available with |*U*| >> |*L*|, which makes them unique and different from standard missing data problems. Such data arise naturally when the response, unlike the covariates, is difficult and/or expensive to obtain – a frequent scenario in modern studies involving large databases. In this talk, I will discuss the subtleties and associated challenges of SS inference in two different aspects.

In the first part, we consider SS inference in a more 'traditional' sense, where the natural goal is to investigate whether/how the information from *U* can be exploited to improve efficiency over a supervised approach. For a broad class of *Z*-estimation problems, we demonstrate a family of SS *Z*-estimators that are robust and adaptive, ensuring they are always as efficient as the supervised estimator and more efficient (often optimal) when the information from *U* actually relates to the parameter. These properties are crucial for advocating 'safe' use of unlabeled data and are often unaddressed. Our framework provides a unified understanding of these problems.

In the second part, we move away from the traditional SS literature (that implicitly assumes *L* and *U* to be equally distributed) to situations where there is inherent selection/labeling bias in *L*. There is hardly any work under missing at random (MAR) labeling with selection bias, which is more realistic but also challenging due to the decaying nature of the labeling probability here. To address this gap, we consider SS mean estimation as a prototype problem under such MAR settings, and develop a 'double robust' SS mean estimator. Our results reveal novel insights into

the problem and its non-standard behavior, extending the traditional MAR literature to this extreme setting.

All are invited to attend.
Please write to SOMENATH DAS somenath1011@isical.ac.in in case you do not receive the invitation link 48 hours before the seminar time.