Mathematics, Statistics, and Computer Science **@ UIC**

Statistics and Data Science Seminar

Information-based Optimal Subdata Selection for Clusterwise Linear Regression Model

Yanxi Liu (University of Illinois, Chicago)

Abstract: As the data size increases rapidly, the relationship between input and output variables may not be homogeneous anymore. Conventional statistical models such as generalized linear models (GLMs) may not be well-suited to heterogeneous relationships. Using a Mixture of Expert models is a good solution. The Mixture of Expert models can combine different statistical models to detect heterogeneous patterns while maintaining the benefits of conventional statistical modeling techniques. However, it needs a considerable amount of computer resources, particularly when working with big data. To address this issue, an attractive idea is to analyze a subsample of the data retaining the rich information of the full data. Information-Based Optimal Subdata Strategy (IBOSS), proposed by Wang et al. (2019), is such a strategy. The IBOSS strategy captures most of the relevant information in the full data through a judicious selection of the subdata by "maximizing" the Fisher information matrix. This project aims to develop an algorithm for the Clusterwise Linear Regression model, a type of Mixture of Experts, to select subdata based on IBOSS strategy. However, the Fisher information matrix of the model has no explicit form, which is a major challenge of the work. To overcome this challenge, we propose a surrogate matrix which is proved to be asymptotically equivalent to the Fisher information matrix, and it is used to construct the IBOSS subdata. Further, the proposed subdata selection is proved to be asymptotically optimal, i.e., no other method is statistically more efficient than the proposed one when the full data size is large.

Wednesday, October 20 at 4:00 PM in Zoom