Synthetic Data–Powered Statistical Inference with Generative AI

Xihong Lin

Department of Biostatistics and Department of Statistics, Harvard University

ABSTRACT

Scalable and robust statistical methods empowered by generative AI offer unprecedent potentials for trustworthy science as they empower statistical analysis, quantify uncertainty, enhance interpretability, and accelerate scientific discovery. In this talk, I will discuss robust and powerful statistical inference by leveraging synthetic data generated by generative AI models, such as diffusion models and transformer, while ensuring valid statistical inference even when generative AI models are misspecified. I will illustrate key points using the analysis of large scale biobanks, such as the analysis of the UK biobank whole genome and electronic health records, and demonstrate the power of scientific discovery by integrating statistics and generative AI using synthetic data.

Keywords: Generative AI; Synthetic data; Power; Statistical genetics; Biobanks

Back to Sessions List

Additive Frechet Regression for Random Objects

Changwon Choi[1], Hans-Georg Mueller[2], **Byeong U. Park**[1], Wookyeong Song[2]

[1]Department of Statistics, Seoul National University

[2]Department of Statistics, University of California, Davis

ABSTRACT

Regression analysis for complex data taking values in a general metric space has gained increasing attention in recent years, particularly in the context of Frechet regression with Euclidean predictors X in R^p. However, nonparametric Frechet regression, while more flexible than global Frechet regression, suffers from the curse of dimensionality when the predictor dimension p > 2. To address this issue while maintaining modelling flexibility, we introduce a novel framework for additive structured nonparametric regression models with responses in general metric spaces. Due to the lack of vector space structure in general metric spaces where the responses reside, we propose a novel formulation that implicitly incorporates the additive structure via projection operators. Our method provides a unified framework for a wide range of response types, including distributions in Wasserstein space, network data represented by graph Laplacians, and spherical data equipped with geodesic distances. We establish consistency and derive convergence rates for the proposed additive \F regression estimators, leveraging smooth backfitting. The practical utility of our approach is demonstrated through applications to brain connectivity network analysis using resting-state fMRI data from Alzheimer's disease and cognitively normal subjects, as well as to distributional physical activity data from the NHANES study.

Keywords: Additive models; metric space valued responses; smooth backfitting; empirical process theory

Goodness-of-Fit and the Best Approximation: An Adversarial Approach

Qiwei Yao[1], Jinyaun Chang, Chengchun Shi, Mingcong Wu, Xinyang Yu

[1]*Department of Statistics, London School of Economics*

ABSTRACT

Diagnostic checking for goodness-of-fit is one of the important and routine steps in building a statistical model. The most frequently used approach for checking the goodness-of-fit is the residual analysis in the context of regression analysis. However for many statistical models there exist no natural residuals, which includes the models for the underlying distributions behind data, or the models for some complex dynamic structures such as the dynamic network models with dependent edges. Furthermore, there are scenarios in which there exist several competing models but none of them are the clear favourite. One then faces a task to choose the best approximation among the wrong models. We propose an adversarial approach in this paper. For checking the goodness-of-fit of a fitted model, we generate a synthetic sample from the fitted model and construct a classifier to classify the original sample and the synthetic sample into two different classes. If the fitted model is adequate, the classifier will have difficulties in distinguish the two samples. For identifying the best model among several candidate models, the classifier will create a distance between the original sample and the synthetic sample generated from each of the candidate model, and the model with the shortest distance is chosen as the best approximation for the truth.

Keywords: Classification, multi-layer perceptron, permutation test, sample-splitting

Back to Sessions List