University of California, Irvine
Statistics Seminar

Uncover Latents in Heterogeneous Data and Deep Generative Models: Locally Dependent Grade of Membership Estimation and Identifiable Deep Discrete Encoders

Yuqi Gu
Assistant Professor
Department of Statistics
Columbia University

Tuesday, Jan. 7, 2025
6011 Donald Bren Hall

In the era of data science and generative AI, latent structures are ubiquitously employed in various scientific disciplines and machine learning architectures.

The first part of the talk focuses on the mixed membership model for multivariate categorical data widely used for analyzing survey responses and population genetics data. This so-called grade of membership (GoM) model offers rich modeling power but presents significant estimation challenges for high-dimensional polytomous data. Such data take the form of a three-way (quasi-)tensor, with many subjects responding to many items with varying numbers of categories. We introduce a novel approach to flatten the three-way quasi-tensor into a “fat” matrix and then perform SVD to estimate parameters by exploiting the singular subspace geometry. Our fast spectral method can accommodate a broad range of data distributions with arbitrarily locally dependent noise. We establish finite-sample entrywise error bounds for the model parameters. We also develop a new sharp two-to-infinity singular subspace perturbation theory for arbitrary locally dependent and flexibly distributed noise, a contribution of independent interest. Simulations and applications to data in political voting, population genetics, and single-cell genomics demonstrate our method's superior performance.

The second part of this talk focuses on deep generative models (DGMs) with latent representations. Despite DGMs’ impressive empirical performance, the statistical properties for these models remain underexplored. DGMs are often overparametrized, non-identifiable, and uninterpretable black boxes, raising serious concerns when deploying them in high-stakes applications. Motivated by this, we propose an interpretable deep generative modeling framework for rich data types with discrete latent layers, called Deep Discrete Encoders (DDEs). Theoretically, we propose transparent identifiability conditions for DDEs, which imply progressively smaller sizes of the latent layers as they go deeper. Identifiability ensures consistent parameter estimation and inspires an interpretable design of the deep architecture. Computationally, we propose a scalable estimation pipeline of a layerwise nonlinear spectral initialization followed by a penalized stochastic approximation EM algorithm. This procedure can efficiently estimate models with exponentially many latent components. Extensive simulation studies validate our theoretical claims. We apply DDEs to diverse real datasets for hierarchical topic modeling, image representation learning, and response time modeling in educational testing.