



THE UNIVERSITY OF
CHICAGO

DEPARTMENT OF STATISTICS

PhD Dissertation Presentation

Yating Liu

Department of Statistics
The University of Chicago

“Title of Talk”

October 27, 2026, at 10:00 AM
Jones 111, 5747 S. Ellis Avenue

Abstract

Biological datasets, such as single-cell gene expression profiles and microbiome abundance measurements, are growing rapidly in size and dimensionality. Identifying latent structure in these data is critical for finding gene programs and microbial communities, characterizing cellular states, and analyzing associations among features. This dissertation studies how these structures can be recovered accurately and used to support trustworthy inference. We examine this problem from three complementary perspectives.

First, we study how microbial subcommunities shared across subjects and time points can be recovered from longitudinal count data. To borrow strength across these dimensions, we develop Tensor Topic Modeling under a nonnegative Tucker decomposition. We establish error bounds for the latent factors and core tensor and assess the coherence of the recovered topics. To support predictive inference with such learned representations, we develop SpeedCP, a conformal prediction method that uses topic proportions or other latent embeddings to construct narrow, calibrated prediction sets for a given predictive model. We establish finite-sample guarantees for approximate conditional coverage, quantify the effect of representation error, and develop an efficient solution-path algorithm.

Second, beyond topic modeling, we consider contextual representations of count data, such as cell embeddings from foundation models. Here, we study how to adapt pretrained models to specific tasks rather than estimate representations from scratch. To address the problem that high-quality target data are limited, we propose SAFE-DRIFT to select examples from a larger candidate pool for supervised fine-tuning. The method uses a shared low-dimensional gradient space to balance target improvement against changes in specified reference behaviors and diagnose target–reference entanglement. Experiments on language models show reduced off-

target drift with competitive target performance.

Third, we complement the study of low-dimensional representations with a network analysis that estimates changes in conditional associations among features and quantifies uncertainty in these changes. Under transelliptical graphical models, we directly estimate the difference between two inverse latent correlation matrices using Kendall's tau and a D-trace loss with weighted ℓ_1 regularization. We then debias the preliminary estimator and establish entrywise asymptotic normality under mild assumptions, enabling asymptotically valid confidence intervals for estimated difference in network parameters.