Statistical methods for transcriptomic deconvolution and temporal gene expression modeling

Suxian Zhou

Preprint no. 2026-05

Abstract

High-throughput RNA sequencing has advanced the study of cellular heterogeneity and dynamic gene expression. This dissertation develops two statistical learning frameworks that integrate bulk and single-cell RNA-seq data.

The first framework, GSNMF+, addresses bulk RNA-seq deconvolution using a geometry-guided nonnegative matrix factorization model. By incorporating single-cell reference information, augmented pseudo-bulk mixtures, solvability-guided regularization, and a manifold-based penalty, GSNMF+ improves the stability and interpretability of estimated cellular compositions. Simulation studies show improved accuracy, and real data analyses demonstrate consistent estimates across independent single-cell references.

The second framework, BetaDE, models pseudotime-dependent gene expression in single-cell RNA-seq data. It uses beta-shaped basis functions to capture diverse temporal patterns and accommodates overdispersion and excess zeros through multiple count-based models. Model selection, hypothesis testing, and functional clustering are used to identify genes with significant and biologically interpretable dynamic expression patterns.

Together, these frameworks provide statistical tools for estimating hidden cellular compositions and characterizing temporal gene expression dynamics from transcriptomic data.

Department of Mathematics, UNC Charlotte, Charlotte, NC 28223