# Statistical methods for the deconvolution of bulk tissue RNA sequencing data

Su Xu

Preprint no. 2025-04

**Abstract**

Bulk RNA sequencing (RNA-seq) provides a cost-effective overview of gene expression but lacks resolution to identify cell-type-specific contributions in heterogeneous tissues. Computational deconvolution methods address this by estimating cell-type proportions from bulk data, enabling finer biological insights. This dissertation develops and applies statistical frameworks to improve the accuracy and interpretability of deconvolution results.

We begin by reviewing RNA-seq technologies and the impact of cellular heterogeneity. Deconvolution is then framed as a nonnegative matrix factorization (NMF) problem, with attention to challenges like non-uniqueness and noise sensitivity. Building on recent identifiability theory, we propose a geometric structure-guided NMF (GSNMF) that incorporates biological priors –such as marker genes– and local manifold structure to stabilize estimation.

To further enhance reference-free deconvolution, we introduce pseudo-bulk augmentation: a strategy that synthesizes single-cell-derived mixtures to enrich bulk data. This approach mitigates issues related to underdetermined solutions and improves robustness.

A comprehensive benchmarking study compares reference-based and reference-free methods using metrics like correlation, root mean squared error, and mean absolute deviation. Results show that while high-quality reference data can improve performance, augmented reference-free approaches like GSNMF are highly effective when reference data are scarce. We conclude with future directions and ongoing challenges.

Department of Mathematics, UNC Charlotte, Charlotte, NC 28223