STATISTICAL METHODS FOR THE DECONVOLUTION OF BULK TISSUE RNA SEQUENCING DATA

by

Su Xu

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Applied Mathematics

Charlotte

2025

Approved by:

_____
Dr. Shaoyu Li

_____
Dr. Duan Chen

_____
Dr. Xue Wang

_____
Dr. Daniel Janies

ABSTRACT

SU XU. Statistical methods for the deconvolution of bulk tissue RNA sequencing data. (Under the direction of DR. SHAOYU LI)

Bulk tissue RNA sequencing (RNA-seq) offers a cost-effective and high-throughput snapshot of global gene expression but lacks the resolution to distinguish the contributions of individual cell types within heterogeneous tissues. To overcome this limitation, computational deconvolution methods aim to disentangle bulk RNA-seq signals into cell-type-specific components, enabling more precise biological interpretation of complex samples. This dissertation develops, evaluates, and applies statistical methodologies to enhance both the accuracy and interpretability of deconvolution results.

We begin by reviewing RNA-seq technologies, highlighting key differences between bulk and single-cell approaches, and examining how cellular heterogeneity impacts transcriptomic analyses. Subsequent chapters delve into the mathematical foundations of deconvolution, framing it as a nonnegative matrix factorization (NMF) problem and discussing common challenges such as non-uniqueness and sensitivity to noise. Building on recent theoretical advancements in NMF identifiability, we introduce a geometric structure-guided framework (GSNMF) that integrates biological priors —such as marker gene information —and leverages local manifold structures to improve the stability and accuracy of estimated cell-type proportions.

We further extend GSNMF by introducing pseudo-bulk augmentation, a strategy that fuses statistically simulated single-cell-derived mixtures with existing bulk datasets. This hybrid approach mitigates singularities and the ill-posed nature of reference-free deconvolution, resulting in more robust factorization. The dissertation also presents a comprehensive benchmarking study comparing the performance of both reference-based and reference-free methods across a variety of datasets. Eval-

uation metrics—including correlation, root mean squared error, and mean absolute deviation—show that while high-quality reference data can improve performance, carefully constrained or augmented reference-free methods can be highly effective when reference data are limited or unreliable. We present and interpret the results, and conclude by discussing existing challenges and future research directions.

DEDICATION

To my parents, Yibing Xu and Lina Zhang, my beloved Shanshan Wang, and my dear dogs, PP and Bobby.

ACKNOWLEDGEMENTS

First and foremost, I am deeply grateful to my advisors, Dr. Shaoyu Li and Dr. Chen Duan. Your expertise played a pivotal role in shaping my research questions and methodology. This dissertation would not have been possible without your insightful guidance, unwavering support, and remarkable patience throughout my Ph.D. journey. You consistently went above and beyond to help me balance research and job searching, and I truly cannot thank you enough.

I would also like to express my sincere appreciation to Dr. Shaozhong Deng, Graduate Coordinator of the Mathematics Department, as well as to my committee member Dr. Xue Wang and Graduate Faculty Representative Dr. Daniel Janies. Your valuable feedback, support, and flexibility were instrumental in bringing this work to fruition.

My thanks also go to the Graduate School for their financial support. This research was partially supported by grants from the National Institutes of Health (NIH: R01GM148971) and the UNC Charlotte Summer Seed Grant for Data Science. I am deeply grateful to these institutions for making this research possible.

A special thank you goes to my beloved Shanshan Wang. From our master's studies to the Ph.D. journey, your love, support, and unwavering encouragement have helped me overcome every challenge. I am forever grateful to walk this path with you.

I would also like to thank my two —emotional support animals —dogs PP and Bobby. You helped me establish the "best" daily routine, taught me patience, and brought joy into my life every single day. Your companionship made even the most challenging days brighter.

Lastly, I dedicate this dissertation to my parents. Your unconditional love, unwavering support, and endless encouragement have shaped me into the person I am today. This achievement belongs to you as much as it does to me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 RNA sequencing

1.1.1 Bulk sample RNA Sequencing

Bulk RNA sequencing (RNA-seq) is a high-throughput technique used to quantify gene expression by measuring the complete set of RNA transcripts within a biological sample. Unlike single-cell RNA sequencing (scRNA-seq), which captures gene expression at the resolution of individual cells, bulk RNA-seq provides an aggregate view of transcriptomic activity across an entire population of cells. This approach is widely employed in transcriptome profiling to assess gene expression patterns, identify differentially expressed genes, and explore regulatory mechanisms underlying various biological processes and disease states.

The workflow for bulk RNA-seq begins with a tissue sample or a heterogeneous cell population as the starting material. RNA is extracted, converted into complementary DNA (cDNA), and sequenced to generate a comprehensive dataset representing the combined transcriptional activity of all cells within the sample. While bulk RNA-seq offers valuable insights into overall gene expression trends, its measurements reflect population-level averages, making it unsuitable for capturing cellular heterogeneity within complex tissues. Despite this limitation, bulk RNA-seq remains a fundamental tool for large-scale transcriptomic studies, particularly in cases where single-cell resolution is not required or when working with limited resources.

1.1.2 Cell level RNA sequencing

1.1.2.1 Single-cell RNA sequencing

Single-cell RNA sequencing (scRNA-seq) is a cutting-edge technique that enables high-resolution analysis of gene expression at the level of individual cells. Unlike bulk RNA sequencing—which captures an averaged transcriptomic profile from a heterogeneous cell population —scRNA-seq provides a detailed snapshot of each cell's transcriptome, revealing the complexity, heterogeneity, and functional diversity within tissues. This unprecedented resolution allows researchers to identify distinct cell types, characterize cellular states, and investigate dynamic biological processes in complex systems.

The scRNA-seq workflow begins with isolating single cells from a tissue sample using various methods, including serial dilution, laser capture microdissection, manual picking, or advanced microfluidic platforms. Once isolated, RNA molecules within each cell are tagged with unique cell-specific barcodes, allowing gene expression profiles to be independently tracked. The RNA is then reverse transcribed into complementary DNA (cDNA), followed by amplification and sequencing library preparation for high-throughput analysis.

Since its introduction in 2009, scRNA-seq has revolutionized transcriptomic research by offering new insights into cellular heterogeneity, lineage differentiation, and intercellular communication. Its applications span a wide range of disciplines, including developmental biology, immunology, cancer research, and neuroscience. Recognizing its transformative potential, Nature named single-cell sequencing the "Method of the Year" in 2013, solidifying its status as an essential tool in modern biological research [1]. As scRNA-seq technologies continue to evolve, their integration with multi-omics approaches and spatial transcriptomics further expands our ability to explore biology at unprecedented resolution.

1.1.2.2 Single-nucleus RNA sequencing

Single-nucleus RNA sequencing (snRNA-seq) is a powerful technique for transcriptomic profiling at the single-nucleus level, offering a compelling alternative to single-cell RNA sequencing (scRNA-seq)—especially in cases where whole-cell dissociation is impractical. Unlike scRNA-seq, which requires fresh, viable cells, snRNA-seq enables the isolation and sequencing of nuclei from frozen or archived tissues, effectively decoupling tissue collection from immediate sample processing. This flexibility makes it an invaluable tool for studying rare, fragile, or otherwise difficult-to-dissociate cell types and facilitates retrospective analyses of biobanked specimens.

Moreover, snRNA-seq supports the multiplexed analysis of longitudinal samples from the same individual, making it particularly advantageous for investigating disease progression, aging, and therapeutic responses over time. However, a notable limitation of snRNA-seq is the lower mRNA content in nuclei compared to whole cells, which can affect the sensitivity and resolution of gene expression analyses. Additionally, enriching or depleting specific cell types is more challenging in nuclear preparations, necessitating careful optimization of experimental protocols and bioinformatic pipelines to minimize bias. Despite these challenges, ongoing advances in nuclear isolation and sequencing technologies continue to improve the performance and reliability of snRNA-seq, broadening its applications in fields such as neuroscience, oncology, and developmental biology [2].

1.1.3 Comparative Analysis of Bulk and Cell level RNA Sequencing

Bulk RNA sequencing (bulk RNA-seq) has long served as a foundational technique for transcriptome profiling, valued for its cost-effectiveness, efficiency, and relatively low technical complexity compared to single-cell RNA sequencing (scRNA-seq). In addition to quantifying overall gene expression levels, bulk RNA-seq is instrumental in identifying novel transcripts, detecting alternative splicing events, and analyzing

allele-specific expression. However, a major limitation of this approach is its inability to resolve cellular heterogeneity, as it captures only the average gene expression across a mixed population of cells.

In contrast, single-cell RNA sequencing (scRNA-seq) provides a high-resolution view of the transcriptome by profiling gene expression at the level of individual cells. While more resource-intensive and technically demanding, scRNA-seq enables researchers to identify distinct cell populations, uncover cell-type-specific expression patterns, and track dynamic cellular processes such as differentiation, immune responses, and disease progression. This ability to dissect cellular heterogeneity makes scRNA-seq a powerful tool for studying complex biological systems, advancing precision medicine, and identifying novel therapeutic targets.

By strategically integrating both approaches, researchers can harness bulk RNA-seq for broad transcriptomic insights and scRNA-seq for fine-grained cellular resolution —together offering a comprehensive understanding of gene expression dynamics across diverse biological contexts.

1.2 Cell Composition Heterogeneity

Human tissues comprise a complex and dynamic mosaic of diverse cell types and subtypes, each characterized by distinct gene expression profiles and specialized biological functions. As a result, gene expression measurements obtained from bulk tissue samples reflect not only the transcriptional activity of individual cell types but also their relative abundances within the sample. This cellular composition is shaped by both intrinsic biological factors and technical considerations, making it a critical variable in transcriptomic analyses [3].

Variability in cellular composition—defined as differences in the distribution of cell types within a tissue sample—can arise from biological conditions, disease states, or experimental treatments. This heterogeneity is especially pronounced in pathological contexts. For example, in the brains of individuals with Alzheimer's disease (AD), a

progressive neurodegenerative disorder, neuronal loss is accompanied by a compensatory increase in microglia and astrocytes, resulting in substantial shifts in cellular composition compared to healthy brain tissue [4, 5, 6]. Similarly, pancreatic tissue from individuals with type 2 diabetes (T2D) exhibits a notable reduction in insulin-producing β cells relative to non-diabetic controls [7, 8]. Beyond disease, significant variation in cell composition is also observed across anatomical regions, developmental stages, and physiological conditions, further contributing to tissue heterogeneity [9].

This inherent cellular heterogeneity poses significant challenges for bulk RNA sequencing (RNA-seq) analyses, particularly in differential expression (DE) studies and gene co-expression network construction. The presence of multiple cell types in bulk samples can obscure cell type-specific expression signals, potentially leading to biased interpretations or spurious findings. As such, addressing cellular composition is essential for deriving accurate and biologically meaningful insights from transcriptomic data.

1.3 Deconvolution methods

Deconvolution is a computational technique used to identify the properties and concentrations of individual components within an observed mixture. With a long history spanning multiple disciplines, it was first applied in audio processing to separate speech and music from background noise [10, 11, 12, 13]. It later gained prominence in hyperspectral imaging, where it is used to extract pure spectral signatures from mixed material signals [14, 15, 16, 17, 18]. In biomedical research, deconvolution has become indispensable for analyzing complex biological tissues, particularly in gene expression studies, where it enables the estimation of cell-type-specific signatures from bulk RNA sequencing data. This approach plays a critical role in disease research, aiding in the identification of tumor microenvironments, monitoring immune responses, and investigating neurodegenerative disorders by tracking changes

in cellular populations [19, 20].Over time, deconvolution methodologies have evolved from simple linear models to advanced regression-based and probabilistic approaches, enhancing both accuracy and applicability across diverse scientific domains.

To mitigate the impact of cellular heterogeneity, a variety of computational deconvolution methods have been developed. These methods can be broadly classified into two main categories: reference-based and reference-free approaches.

Reference-based methods for bulk RNA-seq deconvolution utilize expression profiles from single-cell or purified cell types to estimate the proportions of different cell types within bulk RNA-seq data. These methods employ a range of statistical and machine learning techniques to achieve accurate deconvolution. For instance, DeconRNASeq [21] and DCQ [22] use regression-based approaches, while Bseq-SC [23] applies ν-SVR and improves estimation accuracy using scRNA-seq references. Tools such as TIMER [24] and EPIC [25] are specifically designed for deconvolving the tumor microenvironment and incorporate additional biological priors to enhance performance.

CIBERSORTx [26], an extension of CIBERSORT [20], incorporates single-cell data to increase precision and enables high-resolution estimation of cell-type-specific gene expression. Dtangle [27] minimizes bias through the careful selection of marker genes, while MuSiC [28] leverages cross-subject single-cell RNA-seq data for robust estimation across diverse samples. BayesPrism [29] uses Bayesian modeling to enhance inference accuracy, and DAISM-DNN [30] integrates deep learning techniques to improve deconvolution performance.

Together, these reference-based approaches offer diverse strategies for addressing the complexity of cellular heterogeneity in bulk RNA-seq data. In contrast, reference-free deconvolution methods do not rely on external reference datasets. Instead, they infer cell type proportions through matrix factorization, statistical modeling, and optimization techniques. For example, CellDistinguisher [31] employs a clustering-

based method to identify marker genes for each cell type, followed by support vector machines (SVM) to separate gene expression signals into distinct components. DSA (Digital Sorting Algorithm) [32] uses regularized linear regression to estimate cell type proportions based on selected marker genes, offering a fast and interpretable approach without relying on full reference profiles.

TOAST (TOols for the Analysis of heterogeneouS Tissues) [33, 34] enhances reference-free deconvolution by iteratively identifying cell-type-specific features and refining composition estimates. CDSeq (Complete Deconvolution for Sequencing data) [35] employs non-negative matrix factorization (NMF) to simultaneously estimate cell proportions and cell-type-specific gene expression profiles. Linseed [36] applies convex geometry principles to identify extreme points in the expression space and solves the deconvolution problem through convex optimization. GSNMF [37] extends traditional NMF by incorporating geometric and solvability penalty terms to improve interpretability and accuracy.

These reference-free methods provide flexible alternatives when high-quality reference profiles are unavailable. A summary of these methods is presented in Table (1.1). Additionally, enrichment-based deconvolution methods estimate cell-type proportions using predefined marker gene sets rather than full reference expression profiles. These approaches—such as xCell [38] and ESTIMATE [39]—rely on gene set enrichment analysis, rank-based statistics, or regression models to infer relative cell-type proportions. However, a detailed discussion of enrichment-based methods is beyond the scope of this work.

By incorporating these computational strategies, researchers can more accurately account for cell composition variability, thereby improving the robustness of RNA-seq analyses and enhancing the interpretability of transcriptomic findings.

Table 1.1: Overview of different deconvolution algorithms and their characteristics.

Method	Mathematical/ Statistical Foundations	Input(reference)		Output		Reference
		scRNA-seq	Signature matrix	Cell type proportion	Cell-type-specific GEP	
REFERENCE-BASED METHODS						
DeconRNASeq	Non-negative least squares	No	Yes	Yes	No	Gong *et al.* (2013) [21]
DCQ	Regularized constrained least squares	No	Yes	Yes	No	Altboum *et al.* (2014) [22]
Bseq-SC	ν-Support Vector Regression	No	Yes	Yes	No	Baron et al. (2016) [23]
EPIC	Weighted constrained least squares	No	Yes	Yes	No	Racle *et al.* (2017) [25]
Dtangle	Linear mixing model of linear-scale expressions	No	Yes	Yes	No	Hunt *et al.* (2019) [27]
CIBERSORTx	ν-Support Vector Regression	Yes	Yes	Yes	No	Newman *et al.* (2019) [26]
MuSiC	Weighted least squares	Yes	Yes	Yes	No	Wang *et al.* (2019) [28]
TIMER	Regularized linear regression (multivariate normal)	No	Yes	Yes	No	Li *et al.* (2020) [24]
BayesPrism	Bayesian model	Yes	Yes	Yes	No	Chu *et al.* (2022) [29]
DAISM-DNN	Deep neural network	Yes	No	Yes	No	Lin *et al.* (2022) [30]
REFERENCE-FREE METHODS						
DSA	Regularized linear regression	No	Yes	Yes	Yes	Zhong *et al.* (2013) [32]
CellDistinguisher	Non-negative matrix factorization(NMF)	No	No	Yes	Yes	Newberg *et al.* (2018) [31]
CDSeq	Probabilistic model (LDA)	No	No	Yes	Yes	Kang *et al.* (2019) [35]
Linseed	Simplex topology	No	No	Yes	No	Zaitsev *et al.* (2019) [36]
TOAST	NMF and principal component analysis	No	No	Yes	Yes	Li *et al.* (2019) [33] (2020) [34]
GSNMF	Geometric structure guided NMF model	No	No	Yes	Yes	Chen *et al.* (2022) [37]
ENRICHMENT METHODS						
ESTIMATE	Gene set enrichment analysis	No	Yes	Enrichment scores		Yoshihara *et al.* (2013) [39]
xCell	Gene set enrichment analysis	No	Yes	Enrichment scores		Aran *et al.* (2017) [38]

1.4 Mathematical Formulation of Deconvolution

Deconvolution is a computational technique that leverages large-scale bulk tissue RNA sequencing (RNA-seq) data to estimate the cellular composition within a tissue sample. The fundamental assumption underlying deconvolution is that the observed gene expression profile (GEP) of a bulk tissue sample represents a linear combination of the GEPs of its constituent cell types, weighted by their respective proportions. This relationship is mathematically expressed as:

$$g_{ij} = \sum_{l=1}^{k} p_{lj} c_{il}, \quad 1 \leq i \leq N, \quad 1 \leq j \leq n, \tag{1.1}$$

where g_{ij} and c_{il} denote the GEPs of gene i in the j-th sample and the l-th cell type, respectively, while p_{lj} represents the proportion of the l-th cell type in the j-th sample. Typically, the number of genes (N) far exceeds the number of bulk tissue samples (n), and the number of constituent cell types (k) is relatively small, leading to the common setting $N \gg n > k$.

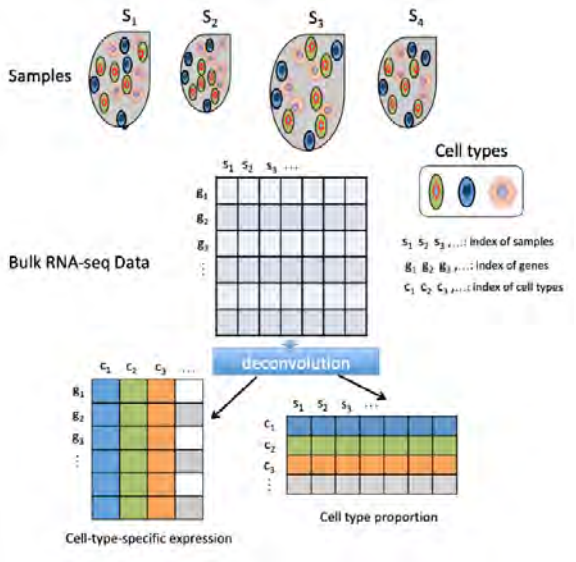


Figure 1.1: Diagram of deconvolution of bulk tissue data

In matrix representation of deconvolution, denote bulk tissue RNA-seq data $\mathbf{G} \in$

$\mathbb{R}^{N\times n}$ with entry g_{ij} being the expression of the i-th gene in the j-th sample; the expected expression levels of genes in these cells $\mathbf{C} \in \mathbb{R}^{N\times k}$ with entry c_{ij} being the reference expression of the i-th gene in the j-th cell type; and the proportion of these cells in the bulk tissue samples $\mathbf{P} \in \mathbb{R}^{k\times n}$ with entry p_{ij} being the proportion of the i-th cell type in the j-th sample. Dimensions $N \gg \max(n, k)$. The following linear relation is assumed:

$$\mathbf{G} = \mathbf{CP} + \epsilon, \tag{1.2}$$

where ϵ is noise.

The problem (1.2) of solving for $\mathbf{P}$, given a known $\mathbf{C}$, is commonly addressed using reference-based deconvolution methods, which have demonstrated remarkable robustness and accuracy. These methods rely on predefined cell type-specific data —such as gene expression profiles obtained from sorted cell populations —to guide the estimation process [20]. With the advent of single-cell RNA sequencing (scRNA-seq) technology, many algorithms now incorporate scRNA-seq reference datasets to further enhance deconvolution accuracy. Notable examples of such methods include CIBERSORTx [20, 26] and MuSiC [28]. However, in more realistic scenarios where little to no information about the underlying cell types is available, the development of reliable complete deconvolution methods remains an open challenge, with only a limited number of models currently established [35, 40, 36].

In contrast, the problem (1.2) of solving for both $\mathbf{C}$ and $\mathbf{P}$, as illustrated in Figure (1.1), is addressed by reference-free deconvolution approaches. These methods infer cell type proportions without relying on predefined reference profiles. Instead, they estimate both the cell type-specific expression profiles and proportions directly from bulk data using matrix decomposition techniques such as singular value decomposition (SVD) or non-negative matrix factorization (NMF). To improve resolution and accuracy, reference-free methods often incorporate additional biological insights

—such as manifold structure or gene co-expression patterns —to aid in identifying distinct cell types. Notable examples include the Digital Sorting Algorithm (DSA) [32], Linseed [36], and GSNMF [37].

1.5 Non-negative Matrix Factorization for Deconvolution

Mathematically, reference-free deconvolution can be formulated as a Non-negative Matrix Factorization (NMF) problem [41, 42, 43]. Extensive research on NMF has been conducted across various disciplines, including spectral unmixing in analytical chemistry [42], remote sensing [44], image processing [45], and topic modeling in machine learning [46].

NMF is a dimensionality reduction technique that decomposes each column of a given input matrix into a non-negative weighted sum of non-negative basis vectors. The number of basis vectors is constrained to be equal to or fewer than the number of columns in the original matrix [47]. In addition to non-negativity, a sum-to-one (STO) constraint is often applied to the columns of matrix $\mathbf{C}$. This constraint restricts the search space, potentially improving the accuracy of the results. In this section, we examine different loss functions and explore the impact of constraint enforcement strategies.

1.5.1 Objective Function Selection

The reference-free deconvolution problem can thus be formulated as follows by (1.2): given data $\mathbf{G} \in \mathbb{R}^{N\times n}$, solve

$$(\mathbf{C}^*, \mathbf{P}^*) = \arg \min_{\mathbf{C}\in\mathbb{R}_+^{N\times k}, \mathbf{P}\in\mathbb{R}_+^{k\times n}} \delta(\mathbf{CP}, \mathbf{G}) \tag{1.3}$$

where $\mathbb{R}_+^{N\times k}$ or $\mathbb{R}_+^{k\times n}$ represent matrices with nonnegative entries and $\delta(\cdot,\cdot)$ is a cost function. The choice of the cost function varies based on prior knowledge of the probability distribution of data noise and its susceptibility to outliers.

1) Choice of loss function: There are various options for suitable loss functions.

The choice of the cost function $\delta(\mathbf{CP}, \mathbf{G})$ is crucial in determining the robustness and accuracy of the solution. Different loss functions can be selected based on assumptions about the noise distribution in $\mathbf{G}$ and the nature of the data.

1. Least Squares Loss (Frobenius Norm):

$$\delta(\mathbf{CP}, \mathbf{G}) = \|\mathbf{CP} - \mathbf{G}\|_F^2 = \sum_{i,j} (\mathbf{CP} - \mathbf{G})_{ij}^2 \tag{1.4}$$

This loss function assumes that the underlying model is perturbed by Gaussian noise. The minimization process serves as a maximum likelihood estimator for additive Gaussian noise. It is computationally efficient but highly sensitive to outliers.

2. L1 Norm Loss (Absolute Error):

$$\delta(\mathbf{CP}, \mathbf{G}) = \sum_{i,j} |(\mathbf{CP} - \mathbf{G})_{ij}| \tag{1.5}$$

The absolute deviation loss is optimal if the noise follows a Laplacian distribution. It is less sensitive to extreme values than the least squares loss but can be more challenging to optimize.

3. Huber Loss (Robust to Outliers):

$$\delta(\mathbf{CP}, \mathbf{G}) = \sum_{i,j} \begin{cases} \frac{1}{2}(\mathbf{CP} - \mathbf{G})_{ij}^2, & \text{if } |(\mathbf{CP} - \mathbf{G})_{ij}| \leq \delta \\ \delta(|(\mathbf{CP} - \mathbf{G})_{ij}| - \frac{1}{2}\delta), & \text{otherwise} \end{cases} \tag{1.6}$$

Huber loss provides a balance between squared error for small deviations and absolute error for large deviations, making it more robust to outliers.

4. Kullback-Leibler (KL) Divergence:

$$\delta(\mathbf{CP}, \mathbf{G}) = \sum_{i,j} \mathbf{G}_{ij} \log \frac{\mathbf{G}_{ij}}{(\mathbf{CP})_{ij}} - \mathbf{G}_{ij} + (\mathbf{CP})_{ij} \tag{1.7}$$

KL divergence is commonly used for Poisson-distributed count data, such as RNA-seq data, where it provides an effective measure of dissimilarity between probability distributions.

5. Support Vector Regression(SVR) Loss Function:

A notable loss function (1.8) used in Support Vector Regression (SVR) is the ϵ-insensitive loss function, which is designed to ignore small errors within a margin ϵ.

$$\delta(\mathbf{CP}, \mathbf{G}) = \sum_{i,j} L_\epsilon \left((\mathbf{CP})_{ij}, \mathbf{G}_{ij} \right) \tag{1.8}$$

where the ϵ-insensitive loss function is defined as:

$$L_\epsilon(y, \hat{y}) = \begin{cases} 0, & \text{if } |y - \hat{y}| \leq \epsilon \\ |y - \hat{y}| - \epsilon, & \text{otherwise} \end{cases} \tag{1.9}$$

where y is the true value, $\hat{y}$ is the predicted value, ϵ is a predefined margin of tolerance. The ϵ-insensitive loss function, similar to Huber loss, it introduces a transition between small and large estimation errors. However, unlike Huber loss, the ϵ-insensitive loss does not penalize errors that fall within the predefined threshold.

2) Choice of Regularizers: Regularization is commonly used to improve the robustness of deconvolution methods by preventing overfitting, enhancing numerical stability, and incorporating prior knowledge about the structure of $\mathbf{C}$ and $\mathbf{P}$. When the reference profile includes many cell types that may not be present in the mixture, or when certain cell types exhibit high correlation, regularization can help sparsify the solution or improve the conditioning of the problem.

1. L2 Regularization (Ridge Penalty):

$$\Omega(\mathbf{C}, \mathbf{P}) = \lambda \left(\|\mathbf{C}\|_F^2 + \|\mathbf{P}\|_F^2 \right) \tag{1.10}$$

L2 regularization discourages large coefficients, resulting in smoother solutions. By penalizing extreme values, it helps control variance and reduces the effects of collinearity.

2. L1 Regularization (Lasso Penalty):

$$\Omega(\mathbf{C}, \mathbf{P}) = \lambda \left(\|\mathbf{C}\|_1 + \|\mathbf{P}\|_1 \right) \tag{1.11}$$

L1 regularization encourages sparsity by driving many elements of $\mathbf{C}$ and $\mathbf{P}$ to be exactly zero. This is particularly useful when only a subset of features or components is expected to contribute significantly.

3. Elastic Net Regularization:

$$\Omega(\mathbf{C}, \mathbf{P}) = \lambda_1 \left(\|\mathbf{C}\|_1 + \|\mathbf{P}\|_1 \right) + \lambda_2 \left(\|\mathbf{C}\|_F^2 + \|\mathbf{P}\|_F^2 \right) \tag{1.12}$$

Elastic Net regularization combines L1 and L2 penalties, balancing feature selection with numerical stability. This approach is particularly effective when dealing with highly correlated features.

4. Group LASSO Regularization:

$$\Omega(\mathbf{P}) = \lambda \sum_{g \in G} \|\mathbf{P}_g\|_2 \tag{1.13}$$

where $\mathbf{G}$ represents a predefined partitioning of the k cell types into groups, $\mathbf{P}_g$ is the submatrix of $\mathbf{P}$ corresponding to group g, $\|\mathbf{P}_g\|_2$ is the L2 norm of the proportions within each group.

Group LASSO extends L1 regularization by enforcing structured sparsity across predefined groups of cell types. Given a grouping $\mathbf{G}$ among cell types, this method encourages similarity within each group while promoting sparsity across different groups. This is particularly useful in biological deconvolution, where certain cell types

are expected to co-occur while others may not be present in a given bulk sample.

5. Non-negative Constraints:

To ensure interpretability in deconvolution, it is common to impose non-negativity constraints:

$$\mathbf{C} \geq 0, \quad \mathbf{P} \geq 0. \tag{1.14}$$

These constraints align with physical and biological interpretations, ensuring that the resulting solutions remain meaningful.

1.5.2 Computational algorithms

Non-negative Matrix Factorization (NMF) is a dimensionality reduction technique widely used in various fields, including bioinformatics, text mining, and image processing. Several numerical methods exist for obtaining a local minimum in NMF, including the Multiplicative Update Algorithm (MUA) [48], Alternating Nonnegativity-Constrained Least Squares (ANLS) [49], and the Alternating Direction Method of Multipliers (ADMM) [50, 51], etc.

Given a data matrix $\mathbf{G} \in \mathbb{R}_{+}^{N \times n}$, the objective of NMF is to approximate $\mathbf{G}$ as the product of two non-negative matrices, $\mathbf{C}$ and $\mathbf{P}$ (1.3). To compute $\mathbf{C}$ and $\mathbf{P}$, various iterative algorithms minimize $\delta(\mathbf{CP}, \mathbf{G})$ while enforcing non-negativity constraints.

1. Multiplicative Update Rules

A widely used algorithm for NMF is the multiplicative update rule, which minimizes the Frobenius norm loss:

$$\delta(\mathbf{CP}, \mathbf{G}) = \|\mathbf{G} - \mathbf{CP}\|_F^2 \tag{1.15}$$

The update rules for $\mathbf{C}$ and $\mathbf{P}$ are:

$$\mathbf{C} \leftarrow \mathbf{C} \odot \frac{\mathbf{GP}^\top}{\mathbf{CPP}^\top} \tag{1.16}$$

$$\mathbf{P} \leftarrow \mathbf{P} \odot \frac{\mathbf{C}^\top \mathbf{G}}{\mathbf{C}^\top \mathbf{C}\mathbf{P}} \tag{1.17}$$

where $\odot$ denotes element-wise multiplication, and division is also performed element-wise.

This method ensures non-negativity without requiring additional constraints and is guaranteed to converge, though not necessarily to a global minimum. However, it is sensitive to initialization, often requiring multiple runs for stability.

2. Alternating Least Squares (ALS)

An alternative approach, *Alternating Least Squares (ALS)*, minimizes the Frobenius norm by solving two subproblems iteratively:

1. Fix $\mathbf{P}$ and solve for $\mathbf{C}$:

$$\mathbf{C} = \arg\min_{\mathbf{C} \geq 0} \|\mathbf{G} - \mathbf{C}\mathbf{P}\|_F^2 \tag{1.18}$$

2. Fix $\mathbf{C}$ and solve for $\mathbf{P}$:

$$\mathbf{P} = \arg\min_{\mathbf{P} \geq 0} \|\mathbf{G} - \mathbf{C}\mathbf{P}\|_F^2 \tag{1.19}$$

This method efficiently solves least squares problems iteratively but requires enforcing non-negativity constraints using optimization techniques such as projected gradient descent. Compared to multiplicative updates, ALS is computationally more expensive.

3. Non-negative Alternating Direction Method of Multipliers (N-ADMM)

A more advanced approach, *Non-negative ADMM (N-ADMM)*, formulates NMF as a constrained optimization problem:

$$\min_{\mathbf{C} \geq 0, \mathbf{P} \geq 0} \|\mathbf{G} - \mathbf{C}\mathbf{P}\|_F^2 + \lambda \|\mathbf{C}\|_1 + \lambda \|\mathbf{P}\|_1 \tag{1.20}$$

where λ controls the sparsity of the factorized matrices.

The ADMM approach enhances NMF by enforcing non-negativity constraints and promoting sparsity via L1 regularization. It achieves this by solving subproblems for $\mathbf{C}$ and $\mathbf{P}$ separately, introducing auxiliary variables for efficient constrained optimization, and updating Lagrange multipliers to ensure feasibility. This method is particularly effective for large-scale NMF problems but is more complex to implement compared to simpler methods.

1.5.3 Ill-posedness of NMF

Solving Eq. (1.3) for $\mathbf{C}$ (or $\mathbf{P}$) while keeping the other variable fixed (reference-free deconvolution) reduces to a convex regression problem. However, solving for both variables simultaneously is inherently non-convex, generally NP-hard, and computational algorithms typically converge only to local minima or stationary points [52]. Furthermore, NMF is an ill-posed problem, meaning its solution is neither unique nor identifiable. Specifically, if $(\mathbf{C}^*, \mathbf{P}^*)$ is a local minimum of Eq. (1.3), then for any $\Omega \in \mathbb{R}^{k\times k}$, the transformed matrices $\hat{\mathbf{C}} = \mathbf{C}^*\Omega$ and $\hat{\mathbf{P}} = \Omega^{-1}\mathbf{P}^*$ are also valid solutions, provided they satisfy the non-negativity constraint.

The non-uniqueness of solutions can significantly impact statistical analyses and influence key decisions in biological applications. Therefore, it is crucial to constrain the search space of variables to enhance the identifiability of solutions, thereby improving interpretability. The uniqueness of the NMF solution is formally defined in the following sense [53]:

Definition 1 (Uniqueness of NMF solution)**.** *The solution* $(\mathbf{C}^*, \mathbf{P}^*)$ *of an NMF is* unique *(or* identifiable*) if and only if, for any other solution* $(\bar{\mathbf{C}}, \bar{\mathbf{P}})$*, there exists a permutation matrix* $\mathbf{\Pi} \in \{0,1\}^{k\times k}$ *and a diagonal scaling matrix* $\mathbf{S}$ *with positive diagonal entries such that*

$$\bar{\mathbf{C}} = \mathbf{C}^*\mathbf{\Pi}\mathbf{S} \quad \textit{and} \quad \bar{\mathbf{P}} = \mathbf{S}^{-1}\mathbf{\Pi}^\intercal\mathbf{P}^*$$

Further results from [52, 53] summarize that under certain conditions [54, 55, 56, 57], NMF solutions can indeed be unique. Two notable results are:

Theorem 1 (Strong identifiability condition)**.** *Let $k = rank(\mathbf{G})$ and assume $\epsilon = 0$. If there exists a solution for problem (1.3) in which both $\mathbf{C}^\intercal$ and $\mathbf{P}$ are separable matrices, then this solution is unique.*

Theorem 2 (Weak identifiability condition)**.** *Let $k = rank(\mathbf{G})$ and assume $\epsilon = 0$. If both $\mathbf{C}^\intercal$ and $\mathbf{P}$ are sufficiently scattered, then the problem (1.3) admits a unique solution.*

Geometric Interpretation

In summary, if the matrices $\mathbf{C}$ and $\mathbf{P}$ satisfy certain identifiability conditions, the NMF problem may yield a unique solution, subject only to the inherent row/column scaling and permutation ambiguities of factorization.

To better understand the strong and weak conditions of uniqueness, we examine the NMF problem from a geometric perspective. For $\mathbf{A} \in \mathbb{R}^{m \times n}$, the notation $cone(\mathbf{A})$ denotes the convex cone generated by the columns of $\mathbf{A}$, defined as

$$cone\{\mathbf{A}\} = \{x \in \mathbb{R}^m \mid x = \mathbf{A}\theta, \quad \text{for some } \theta \in \mathbb{R}^n, \theta \geq 0\}, \tag{1.21}$$

and the convex hull of $\mathbf{A}$ is given by

$$conv\{\mathbf{A}\} = \{x \in \mathbb{R}^m \mid x = \mathbf{A}\theta, \quad \text{for } \theta \in \mathbb{R}^n, \theta \geq 0, \text{ and } \mathbb{1}^\top\theta = 1\}. \tag{1.22}$$

To illustrate this concept, we consider $\mathbf{G}^\top = \mathbf{P}^\top\mathbf{C}^\top$, as gene expression across samples (i.e., the columns of $\mathbf{G}^\top$, corresponding to the rows of $\mathbf{G}$) represents the primary data features of interest. Due to the non-negativity of $\mathbf{C}$ and $\mathbf{P}$, we obtain

$$\mathbf{G}_{(i)} \in cone(\mathbf{P}^\top) \subseteq \mathbb{R}^n_+, \quad 1 \leq i \leq N, \tag{1.23}$$

which implies that $cone(\mathbf{G}^\top) \subseteq cone(\mathbf{P}^\top) \subseteq \mathbb{R}^n_+$. Consequently, problem (1.3) can be interpreted as a nested cone problem: given two nested cones, $cone(\mathbf{G}^\top)$ and $\mathbb{R}^n_+$, the objective is to identify the intermediate nested cone $cone(\mathbf{P}^\top)$ that lies between them.

This geometric interpretation is depicted in the left panel of Figure 1.2, where $N = 8$ and $k = n = 3$. The right panel offers a more intuitive perspective using a convex hull representation, which reduces the dimensional complexity by one compared to the cone view. This simplification is achieved by normalizing the columns of $\mathbf{G}^\top$ and $\mathbf{P}^\top$ to their unit l_1-norm.

From the right panel of Figure 1.2, we observe that the solution for $conv(\mathbf{P}^\top)$ is not unique: the data matrix $\mathbf{G}$ can be enclosed within multiple valid cones (e.g., the red solid and green dashed triangles) formed by different choices of $\mathbf{P}$. This observation suggests that if the rows of $\mathbf{G}$ are sufficiently "spread out" within the nonnegative orthant, then $cone(\mathbf{P}^\top)$ or $conv(\mathbf{P}^\top)$ may be unique. This idea leads to the following theoretical conditions on the identifiability of NMF.

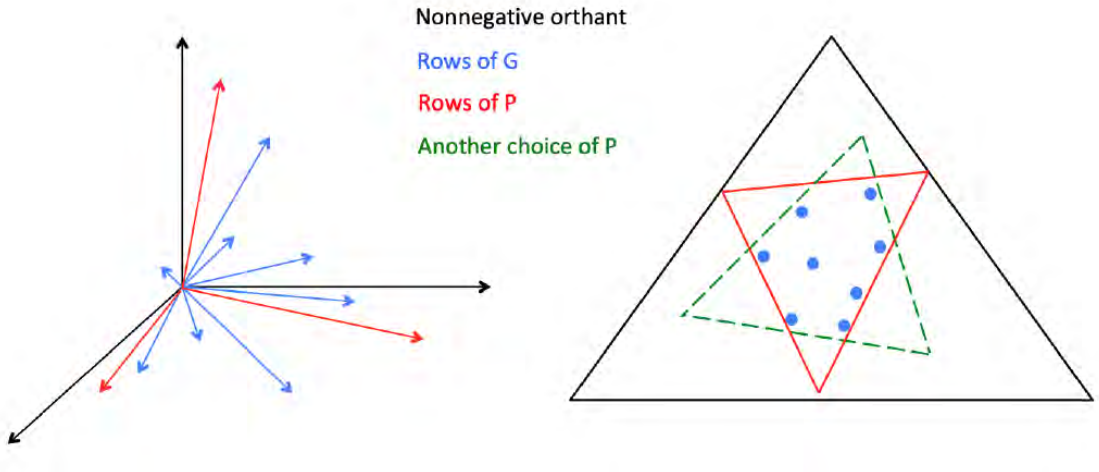


Figure 1.2: Cone (left) and convex hull (right) views of NMF as a nested cone problem.

Strong and weak conditions on identifiability

There are two types of conditions for the uniqueness of the solution to Eq. (1.2) when $\epsilon = 0$. These strong and weak conditions are illustrated above, with the definitions of separable and sufficiently scattered matrices provided as follows:

Definition 2 (Separable Matrix)**.** *A matrix* $\mathbf{A} \in \mathbb{R}_+^{m\times n}$ *is separable if*

$$cone(\mathbf{A}) = \mathbb{R}_+^m. \tag{1.24}$$

Definition 3 (Sufficiently Scattered Matrix)**.** *A matrix* $\mathbf{A} \in \mathbb{R}_+^{m\times n}$ *is sufficiently scattered if:*

1. *The second-order cone in* $\mathbb{R}_+^m$ *is contained within* $cone(\mathbf{A})$*, i.e.,*

$$\mathcal{C} = \{\mathbf{x} \in \mathbb{R}_+^m \mid \mathbf{e}^\top \mathbf{x} \geq \sqrt{m-1}\|\mathbf{x}\|\} \subseteq cone(\mathbf{A}). \tag{1.25}$$

2. *There does not exist any orthogonal matrix* $\mathbf{Q}$ *such that* $cone(\mathbf{A}) \subseteq cone(\mathbf{Q})$*, except for permutation matrices.*

Theorem (1) imposes a strict condition: for a matrix to be separable, $\mathbf{P}$ and $\mathbf{C}^\top$ must contain (scaled) extreme rays of the nonnegative orthant in their respective spaces. That is, for every $r = 1, 2, \ldots, k$, there exists a column index l_r such that

$$\mathbf{P}_{(l_r)} = \alpha_r \mathbf{e}_r, \quad \mathbf{e}_r \in \mathbb{R}_+^k, \tag{1.26}$$

where α_r is a scalar. This condition is illustrated in the left panel of Figure 1.3 for $k = 3$. Since this is a convex hull representation, the columns of $\mathbf{P}$ are shown as blue dots within the unit simplex (red triangle) in $\mathbb{R}_+^3$. As depicted, some columns of $\mathbf{P}$ must be exactly aligned with the unit vectors $\mathbf{e}_r, r = 1, 2, 3$ (overlapping with red dots). A similar condition applies to $\mathbf{C}^\top$. However, such strict assumptions on both variables are often impractical, especially in the presence of noise.

On the other hand, Theorem (2) imposes a much more relaxed condition. The right panel of Figure 1.3 illustrates this scenario: the dashed circle represents the intersection of the second-order cone in $\mathbb{R}^3_+$ with the unit simplex. The blue markers —comprising both dots and pentagons —represent the columns of $\mathbf{P}$. Unlike the stronger condition, none of these columns are required to coincide with the basis vectors $\mathbf{e}_r$; however, some of them (shown as pentagons) must lie outside the dashed circle, that is, outside the second-order cone.

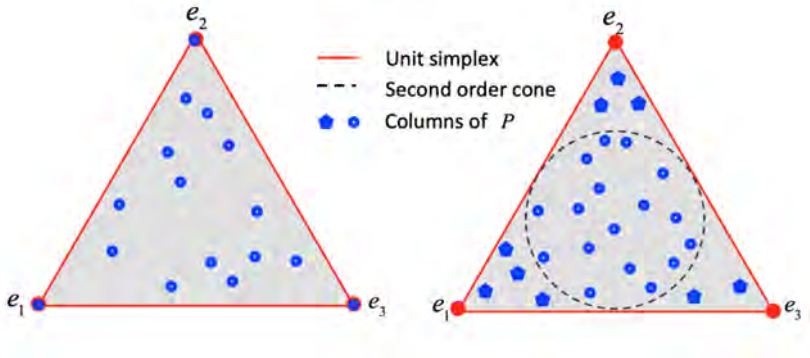


Figure 1.3: Convex hull views for strong (left) and weak (right) identifiability conditions with k = 3

1.6 Statistical Methods for Deconvolution

Deconvolution of bulk RNA sequencing (RNA-seq) data is a crucial computational challenge in transcriptomics, particularly for complex and heterogeneous tissue samples such as tumors. Bulk RNA-seq captures gene expression as an aggregate signal from multiple cell types, complicating downstream analyses. To extract meaningful biological insights, statistical deconvolution methods aim to infer both cell-type proportions and their respective gene expression profiles. Traditional approaches often rely on predefined reference profiles or assume prior knowledge of cellular composition, which may not always be available or reliable.

Two advanced statistical frameworks, CDSeq [29] and BayesPrism [35], have been developed to address these challenges by leveraging different probabilistic models for

robust deconvolution.

CDSeq is a novel probabilistic method based on Latent Dirichlet Allocation (LDA) that performs complete deconvolution without requiring predefined cell-type proportions or cell-type-specific gene expression profiles. By simultaneously estimating both components directly from bulk RNA-seq data, CDSeq provides an unbiased approach to studying heterogeneous tissues. The method has demonstrated superior performance against multiple existing deconvolution techniques across synthetic and real experimental datasets, highlighting its robustness and applicability in complex transcriptomic analyses.

BayesPrism, in contrast, employs a Bayesian hierarchical framework that integrates single-cell RNA sequencing (scRNA-seq) data as prior information to refine cellular composition and gene expression inference. By explicitly modeling discrepancies between bulk RNA-seq and reference scRNA-seq data, BayesPrism improves deconvolution accuracy, particularly in tumor microenvironment (TME) analyses. The method has been successfully applied to various cancer types, enabling integrative analyses across large patient cohorts, refining cancer subtype classifications, and identifying gene expression signatures associated with immune infiltration.

Both CDSeq and BayesPrism represent significant advancements in computational deconvolution, addressing key limitations of conventional methods and expanding the analytical capabilities of bulk RNA-seq data. These approaches enhance our ability to dissect cellular heterogeneity, improve biomarker discovery, and provide deeper biological insights in both healthy and diseased tissues.

1.7 Overview of the Dissertation

The remainder of this dissertation is organized as follows.

In Chapter 2, complete deconvolution of bulk RNA-seq data aims to recover both cell type-specific gene expression profiles (GEPs) and relative cell abundances. However, this remains a challenging task due to the inherently ill-posed nature of models

like nonnegative matrix factorization (NMF). While existing methods have shown promising results in certain cases, a general solution to the ill-posedness problem and improvement in result accuracy is still lacking.

To address these challenges, we investigate the solvability conditions of NMF and analyze how rescaling ambiguity affects solution uniqueness. Building on these insights, we develop a novel NMF-based framework (GSNMF) that integrates marker gene information and geometric structure derived from spectral clustering. This leads to a constrained optimization model with manifold regularization.

Furthermore, we propose a new pipeline—Geometric Structure-Guided NMF with Pseudo-Bulk Augmentation (GSNMF+)—which enhances solvability by augmenting the original bulk RNA-seq data with statistically simulated pseudo-bulk data generated from single-cell RNA-seq profiles. We also introduce a rescaling adjustment step to reduce estimation error. Our method demonstrates substantial performance improvements across several realistic datasets, especially in complex tissue samples with sparse or singular cellular compositions.

In Chapter 3, we conduct a comprehensive benchmarking study to evaluate the performance of various computational deconvolution methods for estimating cell-type proportions in bulk RNA-seq data. The comparison includes both reference-based and reference-free approaches, using in silico pseudo-bulk datasets derived from four distinct single-cell RNA-seq (scRNA-seq) datasets representing diverse tissue types and cell composition scenarios. We assess each method's robustness by comparing its estimates to ground truth using Pearson's correlation (R), root mean squared error (RMSE), and mean absolute deviation (mAD). To evaluate resilience, we introduce distribution shifts—such as mean shifting, truncation, and scaling—into the single-cell profiles.

Our findings reveal that reference-based methods generally perform best when high-quality reference data are available. In contrast, reference-free methods offer more

consistent performance in the absence of reliable references. We also observe that variability in scRNA-seq data and tissue heterogeneity significantly impact method accuracy and stability.

Chapter 4 outlines several directions for future research.

CHAPTER 2: Geometric structure guided NMF model for deconvolution of bulk RNA-seq data

2.1 Introduction

Performing complete deconvolution analysis of bulk RNA-seq data to recover both cell type-specific gene expression profiles (GEPs) and relative cell abundances remains a complex and challenging task. A foundational approach in this domain —nonnegative matrix factorization (NMF)—is inherently ill-posed. While several complete deconvolution methods have been developed and show promising alignment with ground truth in certain datasets, a comprehensive understanding of how to mitigate ill-posedness and improve solution accuracy is still lacking.

In this section, we begin by examining the conditions under which a dataset satisfies the solvability criteria outlined in NMF theory. Even when these criteria are met, the âuniqueâ solutions produced by NMF are typically affected by an unknown rescaling matrix. To address this limitation, we estimate both the local minima to which the model converges and the potential rescaling matrix, leveraging informative initialization strategies.

We propose an NMF-based mathematical model along with corresponding computational algorithms (GSNMF) aimed at enhancing the identifiability of deconvolution solutions for bulk RNA-seq data. Our approach incorporates the biological concept of marker genes into the theoretical framework of NMF solvability, culminating in a geometric structure-guided optimization model. This method first explores the geometric structure of the bulk tissue data using spectral clustering. The identified marker genes are then used as solvability constraints, while a correlation graph constructed from the data provides manifold regularization.

Building on the theoretical insights presented in Chapter 2, we introduce a novel pipeline: Geometric Structure-Guided NMF with Pseudo-Bulk Augmentation (GSNMF+). This approach generates pseudo-bulk tissue data by statistically simulating pseudo-cellular compositions from single-cell RNA sequencing (scRNA-seq) data, which are then integrated with the original bulk dataset. The resulting hybrid dataset is designed to satisfy the weak solvability conditions of NMF. Additionally, we apply an estimated rescaling matrix to adjust the NMF minimizer, with the goal of reducing the mean squared error in the final solutions.

We evaluate our algorithm on multiple realistic bulk tissue datasets and demonstrate significant improvements in performance, particularly in scenarios involving sparse or singular cellular compositions.

2.2 Theoretical foundations in NMF model

The reference-free deconvolution algorithm is crucial for distinguishing whether changes in disease-associated tissue gene expression profiles (GEPs) stem from shifts in cellular composition or alterations in specific cell types, all without relying on a predefined reference. Our group has developed an NMF-based mathematical model and corresponding computational algorithms to enhance the identifiability of solutions when deconvoluting bulk RNA-seq data. In our approach, we integrate the biological concept of marker genes with the solvability conditions of NMF theories and formulate an optimization model guided by geometric structures. First, we explore the geometric structure of bulk tissue data using spectral clustering techniques. Then, we incorporate identified marker gene information as solvability constraints while leveraging the overall correlation graph as manifold regularization.

Mathematical Formulation

Based on the mathematical formulation established in Sections 1.4 and 1.5, the

reference-free deconvolution problem can be formulated as follows: given $\mathbf{G}$, solve

$$(\mathbf{C}^*, \mathbf{P}^*) = \arg \min_{\mathbf{C}\geq 0, \mathbf{P}\geq 0} \frac{1}{2}\|\mathbf{G} - \mathbf{CP}\|_F^2 + \mathbb{1}_T(\mathbf{P}). \tag{2.1}$$

The third term $\mathbb{1}_T(\mathbf{P})$, enforces the sum-to-one constraint on each column of $\mathbf{P}$ (i.e., column stochasticity), reflecting the requirement that the total cellular proportion in each tissue sample must equal one.

The choice of the cost function typically depends on prior knowledge of the noise distribution in the data and its susceptibility to outliers.

In many cases, the minimization process is interpreted as a maximum likelihood estimation problem. For example, when the Frobenius norm is used, the optimization corresponds to a maximum likelihood estimator under additive Gaussian noise. When the noise follows a Poisson process, the Expectation-Maximization (EM) algorithm and maximum likelihood estimation [58] lead to the I-divergence cost function [48]. Alternatively, if the noise follows a Laplace distribution, the cost function can be chosen as the row-wise or column-wise l_1 norm of the difference matrix [47]. As summarized in [52], other possible choices include the Earth Mover's distance metric, α-divergence, β-divergence, γ-divergence, φ-divergence, Bregman divergence, and α-β divergence.

In this work, we focus on addressing data singularity and structure. To simplify the optimization, we use the following cost function: $\frac{1}{2}\|\mathbf{G} - \mathbf{CP}\|_F^2$.

However, as discussed in Section 1.5.3, NMF is inherently an ill-posed problem. Figure 2.1 illustrates various scenarios encountered when solving an NMF problem using the classic Multiplicative Update (MU) rule. Figures 2.1(a) and (b) show two local minimizers of the matrix $\mathbf{C}$ (in blue), neither of which closely approximates the ground truth (in red). Figure 2.1(c) provides an example where the NMF model is applied to real RNA-seq data to estimate cellular composition. Although the computed

results (in blue) are highly correlated with the ground truth (in orange)—yielding a strong correlation score as a performance metric—they still substantially misestimate the proportions of two cell types.

This discrepancy arises because, even when all theoretical assumptions are satisfied, the "unique" solutions of NMF remain subject to permutation and rescaling transformations, as detailed in the following section. While permutation ambiguity can often be resolved using biological prior knowledge, correcting for rescaling requires computational strategies.

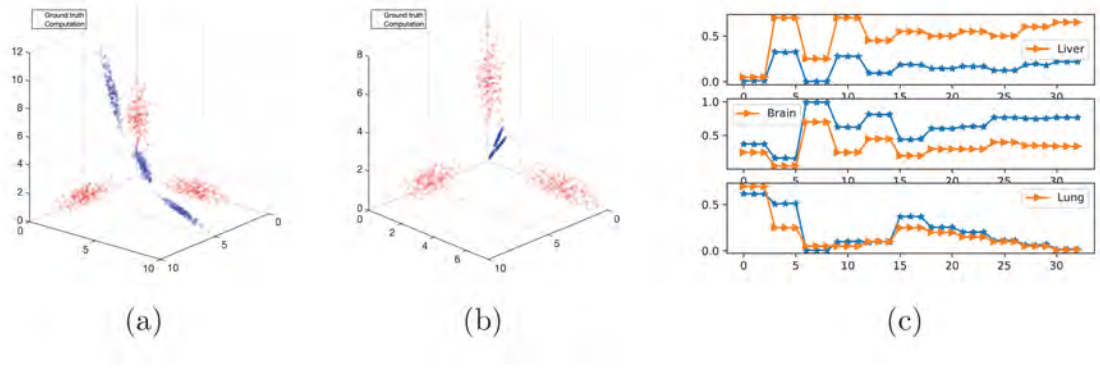


Figure 2.1: Ill-posedness of NMF and its multiple solutions

Unlike in image processing, where solution non-uniqueness may be less consequential, its impact on statistical analysis in biological applications can be substantial, influencing critical decisions. Therefore, it is essential to constrain the search space of variables to enhance solution identifiability, ultimately improving interpretability.

To address this ill-posedness, several guidelines have been proposed in Section 1.5.3. In summary, if the matrices $\mathbf{C}$ and $\mathbf{P}$ satisfy certain identifiability conditions (Theorem 1, Theorem 2), the NMF problem may yield a unique solution, subject only to the inherent row/column scaling and permutation ambiguities of factorization.

Geometric structure-constrained non-negative matrix factorization (GSNMF) is an NMF-based mathematical model that integrates biological features of marker genes to enhance the interpretability of solutions. GSNMF formulates the non-negative

matrix factorization (NMF) problem as a nested cone problem, leveraging the concept of convex cones to understand the geometry of NMF. In this formulation, the columns of the data matrix are interpreted as points within a cone generated by non-negative linear combinations of basis vectors, thereby creating a nested structure of cones that represent the factorization.

Under the strong identifiability condition (depicted in the left panel of Figure 1.3), the factorization must include the extreme rays of the non-negative orthant, leading to overlaps with each unit basis vector $\mathbf{e}$ (shown as red dots). Here, $\mathbf{e}$ denotes the standard unit vectors that span the space. Such assumptions on both variables are typically too restrictive for practical applications, particularly in the presence of noise. Conversely, under the weak identifiability condition (depicted in the right panel of Figure 1.3), approximate uniqueness is achieved by employing second-order cones, which accommodate noise and complexity in the data. In this less stringent scenario, none of the columns are required to overlap with $\mathbf{e}$; however, some columns (shown as pentagons) must lie outside the second-order cone (represented by the circle).

The concept of marker genes helps address challenges encountered when applying the aforementioned NMF-based theories to RNA-seq data. Marker genes associated with a particular cell type are predominantly expressed in that cell type while exhibiting negligible expression in others. Although each cell type may have multiple marker genes, each marker gene is specific to only one cell type. For a given cell type (e.g., the r-th cell type), the expression levels of its marker genes across samples are highly correlated with one another and with $\mathbf{P}$.

Each row of $\mathbf{G}$ can be visualized as a point in a suitable vector space. Due to the strong correlation among marker genes of the same cell type, rows in $\mathbf{G}$ corresponding to these genes naturally form clusters around $\mathbf{P}_{(r)}$. This observation motivates a strategy to first identify marker genes from $\mathbf{G}$ by clustering its rows, thereby enabling a quantitative exploration of the geometric structure within $\mathbf{G}$'s row space (depicted

in the left panel of Figure 2.2).

Furthermore, note that $\mathbf{C}_{(i)}$ represents the coefficient vector of $\mathbf{G}_{(i)}$ with respect to the row-basis vectors of $\mathbf{P}$. Consequently, the second step involves transferring this *geometric structure* of $\mathbf{G}$ to $\mathbf{C}$, thereby enforcing the weak identifiability condition on the variable $\mathbf{C}$ (as illustrated in the right panel of Figure 2.2).

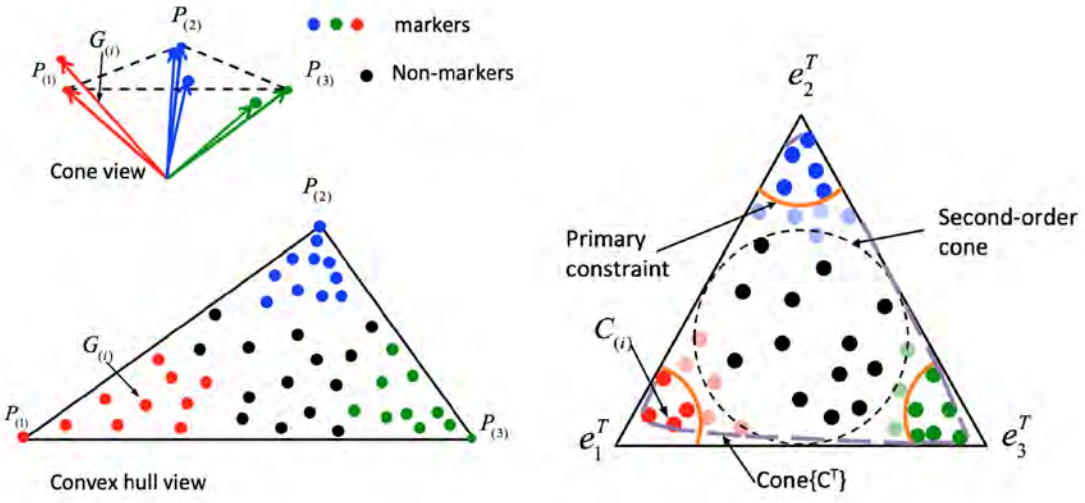


Figure 2.2: RNA-seq data structure (left) and geometric constraints(right).

GSNMF imposes the solvability constraint (2.2) on $\mathbf{C}$ through a clustering approach:

$$\mathcal{F}_1(\mathbf{C}) = \frac{\lambda_1}{2} \sum_{r=1}^{k} \sum_{i \in S_r} d_{\text{eisen}}\big(\mathbf{C}(i), \mathbf{e}_r^{\intercal}\big)^2 \tag{2.2}$$

where λ_1 is a tuning parameter. And in the clustering step, all rows of $\mathbf{G}$ are partitioned into k groups, and their row indices are recorded in $\{\mathbf{G}_r\}_{r=1}^{k}$. Subsequently, within each group, row-wise correlations are ranked, and a subset $\mathbf{S}_r \subset \mathbf{G}_r$ is selected to identify the marker genes for each cell type.

Within each group, a manifold constraint (2.3) is enforced based on the local invariance assumption in manifold regularization.

$$\mathcal{F}_2(\mathbf{C}) = \frac{\lambda_2}{2} \sum_{j=1}^{N} \sum_{i=1}^{N} \omega_{ij} d_{\text{eisen}}(\mathbf{C}(i), \mathbf{C}(j))^2 \tag{2.3}$$

where entry $\omega_{ij} > 0$ in the adjacency matrix $\mathbf{W}$ measures the correlations (larger value represents stronger correlation) between genes i and j in data $\mathbf{G}$.

Specifically, if two data points are close in the intrinsic geometry of the data distribution, their representations in the new basis should also remain close under the same metric. Recall that $\mathbf{C}$ represents the coordinates of the data points $\mathbf{G}$ in the basis defined by $\mathbf{P}$. Therefore, we require that $\mathbf{C}$ preserve the geometric structure of $\mathbf{G}$; in other words, rows of $\mathbf{C}$ belonging to the same cluster should exhibit strong correlations.

Under these constraints, certain rows of $\mathbf{C}$ must lie sufficiently close to each fundamental basis vector such that the convex hull of all rows can span the second-order cone. The distinct groups of rows in $\mathbf{C}$ associated with each basis vector —represented by red, blue, and green dots in the right panel of Figure 2.2—are identified via a clustering algorithm and correspond to sets of biological marker genes. Moreover, we employ the Eisen cosine correlation distance in both geometric constraints, rather than Euclidean distance. This choice captures relative expression patterns and eliminates the need for row normalization.

By combining the solvability condition and the manifold constraint, we arrive at the following geometric-structure-guided nonnegative matrix factorization (GSNMF) model. For convenience, define $T := \left\{ \mathbf{Z} \in \mathbb{R}_+^{k \times n} \;\middle|\; \mathbb{1}^\intercal \mathbf{Z} = \mathbb{1}^\intercal \right\}$, and let $\mathbb{1}_T$ be the corresponding indicator function, where $\mathbb{1}_T(\mathbf{Z}) = 0$ if $\mathbf{Z} \in T$ and $\mathbb{1}_T(\mathbf{Z}) = \infty$ otherwise. With these definitions, solving for $\mathbf{C}$ and $\mathbf{P}$ takes the form of the following optimization problem:

$$\min_{\mathbf{C} \geq 0, \mathbf{P} \geq 0} \frac{1}{2} \|\mathbf{G} - \mathbf{C}\mathbf{P}\|_F^2 + \mathcal{F}(\mathbf{C}) + \mathbb{1}_T(\mathbf{P}), \tag{2.4}$$

where $\mathcal{F}(\mathbf{C}) = \mathcal{F}_1(\mathbf{C}) + \mathcal{F}_2(\mathbf{C})$.

To derive a numerical solution, we employ the Alternating Direction Method of Multipliers (ADMM). First, we rewrite the optimization model (2.4) as:

$$\min_{\mathbf{C},\mathbf{P}} \frac{1}{2}\|\mathbf{G} - \mathbf{C}\mathbf{P}\|_F^2 + \mathcal{F}(\mathbf{C}) + \mathbb{1}_S(\mathbf{C}) + \mathbb{1}_T(\mathbf{P}), \tag{2.5}$$

where $\mathcal{S} := \mathbb{R}_+^{N\times k}$ represents the set of all nonnegative matrices of size $N \times k$.

Introducing auxiliary variables $\mathbf{A}$ and $\mathbf{Q}$, we reformulate (2.5) as:

$$\begin{aligned} &\min_{\mathbf{C},\mathbf{P},\mathbf{A},\mathbf{Q}} \frac{1}{2}\|\mathbf{G} - \mathbf{C}\mathbf{P}\|_F^2 + \mathcal{F}(\mathbf{A}) + \mathbb{1}_S(\mathbf{C}) + \mathbb{1}_T(\mathbf{Q}) \\ &\text{s.t. } \mathbf{C} - \mathbf{A} = 0, \quad \mathbf{P} - \mathbf{Q} = 0, \end{aligned} \tag{2.6}$$

The corresponding augmented Lagrangian function is given by [59]:

$$\mathcal{L} = \frac{1}{2}\|\mathbf{G}-\mathbf{C}\mathbf{P}\|_F^2+\mathcal{F}(\mathbf{A})+\mathbb{1}_S(\mathbf{C})+\mathbb{1}_T(\mathbf{Q})+\frac{\rho}{2}\|\mathbf{C}-\mathbf{A}+\tilde{\mathbf{A}}\|_F^2+\frac{\gamma}{2}\|\mathbf{P}-\mathbf{Q}+\tilde{\mathbf{Q}}\|_F^2 \tag{2.7}$$

where $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{Q}}$ are dual variables of $\mathbf{A}$ and $\mathbf{Q}$, respectively, while $\rho > 0$ and $\gamma > 0$ are penalty parameters.

The ADMM(2.7) iteration from step i to $i+1$ is given in its scaled form [50]:

$$\begin{aligned} \mathbf{C}^{i+1} &:= \arg\min_{\mathbf{C}} \frac{1}{2}\|\mathbf{G} - \mathbf{C}\mathbf{P}^i\|_F^2 + \frac{\rho}{2}\|\mathbf{C} - \mathbf{A}^i + \tilde{\mathbf{A}}^i\|_F^2 + \mathbb{1}_S(\mathbf{C}) \\ \mathbf{P}^{i+1} &:= \arg\min_{\mathbf{P}} \frac{1}{2}\|\mathbf{G} - \mathbf{C}^i\mathbf{P}\|_F^2 + \frac{\gamma}{2}\|\mathbf{P} - \mathbf{Q}^i + \tilde{\mathbf{Q}}^i\|_F^2 + \mathbb{1}_T(\mathbf{P}) \\ \mathbf{A}^{i+1} &:= \arg\min_{\mathbf{A}} \mathcal{F}(\mathbf{A}) + \frac{\rho}{2}\|\mathbf{C}^i - \mathbf{A} + \tilde{\mathbf{A}}^i\|_F^2 \\ \mathbf{Q}^{i+1} &:= \arg\min_{\mathbf{Q}} \frac{\gamma}{2}\|\mathbf{P}^i - \mathbf{Q} + \tilde{\mathbf{Q}}^i\|_F^2 \\ \tilde{\mathbf{A}}^{i+1} &:= \tilde{\mathbf{A}}^i + \mathbf{C}^i - \mathbf{A}^i \\ \tilde{\mathbf{Q}}^{i+1} &:= \tilde{\mathbf{Q}}^i + \mathbf{P}^i - \mathbf{Q}^i \end{aligned} \tag{2.8}$$

Each variable in (2.8) can be solved individually. Specifically, for the $\mathbf{C}$-subproblem,

the Karush-Kuhn-Tucker (KKT) condition [60] yields:

$$\mathbf{C}^{i+1} = [\mathbf{G}\mathbf{P}^{i\top} + \rho(\mathbf{A}^i - \tilde{\mathbf{A}}^i)](\mathbf{P}^i\mathbf{P}^{i\top} + \rho\mathbf{I})^{-1}, \tag{2.9}$$

where $\mathbf{P}^i\mathbf{P}^{i\top}+\rho\mathbf{I}$ is a small $k\times k$ matrix that can be easily inverted. The nonnegativity constraint on $\mathbf{C}$ is enforced using a row-wise active set method.

For the $\mathbf{P}$-subproblem, the KKT condition gives:

$$\mathbf{P}^{i+1} = \Pi\{(\mathbf{C}^{i\top}\mathbf{C}^i + \gamma\mathbf{I})^{-1}[\mathbf{C}^{i\top}\mathbf{G} + \gamma(\mathbf{Q}^i - \tilde{\mathbf{Q}}^i)]\}, \tag{2.10}$$

which is a small-scale problem requiring the inversion of a $k \times k$ matrix, followed by a column-wise probability simplex projection Π [61].

The solution to the $\mathbf{Q}$-subproblem is simply:

$$\mathbf{Q}^{i+1} = \max\{\mathbf{P}^i + \tilde{\mathbf{Q}}^i, 0\}. \tag{2.11}$$

The $\mathbf{A}$-subproblem involves both the solvability condition (2.2) and the manifold constraints (8), making it a nonlinear optimization problem without a closed-form solution. Consequently, we employ the gradient descent method, making this step an inner iteration. Denoting the total objective function of the $\mathbf{A}$-subproblem as:

$$f(\mathbf{A}) = \mathcal{F}(\mathbf{A}) + \frac{\rho}{2}\|\mathbf{C} - \mathbf{A} + \tilde{\mathbf{A}}\|_F^2, \tag{2.12}$$

its gradient, $\nabla f(\mathbf{A})$, is computed based on (2.2) and (2.3). Since $\nabla f(\mathbf{A})$ is nonlinear in terms of $\mathbf{A}$, we improve computational efficiency by utilizing the current outer iteration's result for $\mathbf{C}$, approximating $\nabla f(\mathbf{A}) \approx \nabla f(\mathbf{C})$ without updating it in the inner loop.

Algorithm 1 summarizes the entire GSNMF process, excluding necessary raw data

preprocessing and spectral clustering steps. The stopping criteria are defined by ensuring that both $|\mathbf{C}^{i+1} - \mathbf{C}^i|/|\mathbf{C}^i|_F$ and $|\mathbf{P}^{i+1} - \mathbf{P}^i|/|\mathbf{P}^i|_F$ fall below a specified tolerance.

Algorithm 1 Geometric structure constrained NMF (GSNMF)

Require: Data $\mathbf{G}$, initial guesses $\mathbf{C}_0, \mathbf{P}_0$, structure identifier $\mathbf{C}_g$, graph adjacency matrix $\mathbf{W}$, tolerance ϵ, parameters $\lambda_1, \lambda_2, \rho$ and γ.

Ensure: Matrices $\mathbf{C}$ and $\mathbf{P}$.

for $i = 0, 1, \ldots$ until criteria is satisfied **do** ▷ outer iteration
  Solve $\mathbf{C}$-subproblem in Eq. (2.8) by (2.9);
  Solve $\mathbf{P}$-subproblem in Eq. (2.8) by (2.10);
  for $m = 0, 1, \ldots$ until criteria is satisfied **do** ▷ inner iteration
    Solve $\mathbf{A}$-subproblem in Eq. (2.8) through gradient descent method;
  end for
  Solve $\mathbf{Q}$-subproblem in Eq. (2.8) by (2.11);
  Set $\tilde{\mathbf{A}}^{i+1} := \tilde{\mathbf{A}}^i + \mathbf{C}^i - \mathbf{A}^i$;
  Set $\tilde{\mathbf{Q}}^{i+1} := \tilde{\mathbf{Q}}^i + \mathbf{P}^i - \mathbf{Q}^i$;
end for

2.3 The pseudo-bulk tissue data augmented GSNMF model

With these constraints (Eq. 2.4), GSNMF significantly enhances the interpretability of solutions. However, it still struggles to achieve accurate solutions when the cellular compositions in the bulk tissue data are singular. Recall that both the strong and weak identifiability conditions impose requirements on both matrices $\mathbf{C}^\top$ and $\mathbf{P}$. The weak condition on $\mathbf{C}^\top$ is generally satisfied for a given dataset, making it reasonable to impose constraints on $\mathbf{C}_{(i)}$ due to the biological characteristics of marker genes, which can be identified computationally or inferred from prior knowledge.

However, there is usually no prior information available about the cellular composition. The black star dots in the right panel of Figure (2.3) illustrate a scenario where

the cellular proportions of two cell types are highly correlated across all samples. In such cases, the data points tend to align along a line rather than fully occupying the entire simplex, leading to challenges in accurately estimating cellular compositions.

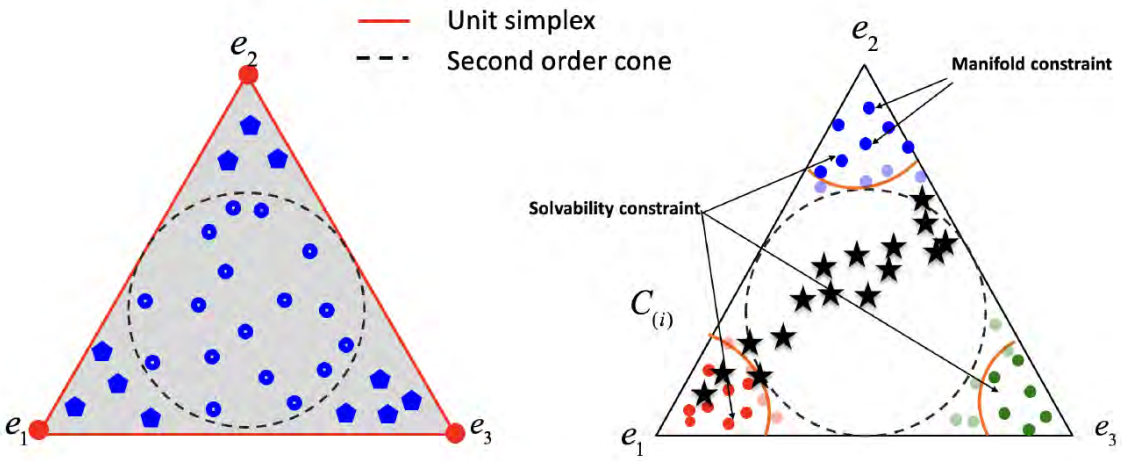


Figure 2.3: Geometric interpretation of solvability condition (left) and constraints of the proposed NMF model (right).

One major contribution of GSNMF+ is its ability to address the challenge of singular data, as mentioned in the introduction. To achieve this, the core idea is to augment the original NMF problem, which involves potentially singular sample data $\mathbf{G}$, by incorporating pseudo-bulk tissue data $\hat{\mathbf{G}}$ with regular cellular proportions. This leads to an augmented NMF formulation:

$$(\mathbf{C}^*, \mathbf{P}^*) = \arg \min_{\mathbf{C}\in\mathbb{R}_+^{N\times k}, \mathbf{P}\in\mathbb{R}_+^{k\times(n+\hat{n})}} \delta(\mathbf{CP}, [\mathbf{G}, \hat{\mathbf{G}}]). \tag{2.13}$$

The key challenge is determining how to construct the regular pseudo-bulk tissue data $\hat{\mathbf{G}}$ in Eq. (2.13).

Constructing Pseudo-Bulk Tissue Data from Single-Cell Data

Single-cell RNA-seq data is represented as a matrix recording the count of a set of N genes across multiple cells of different types and from various samples. In this section, we illustrate how to generate a vector of pseudo-bulk tissue data from this

matrix.

Let $c_{\gamma,l} \in \mathbb{R}^N$ denote the gene expression profile (GEP) of all genes in the l-th cell of the γ-th cell type, where $1 \leq \gamma \leq k$, and the total number of cells depends on the specific dataset. The relationship between the single-cell data and the j-th sample in the pseudo-bulk tissue data is given by:

$$\hat{\mathbf{G}}_{(j)} = \sum_{\gamma=1}^{k} \sum_{l=1}^{m(\gamma,j)} c_{\gamma,l} = \sum_{\gamma=1}^{k} m(\gamma, j)\bar{c}_{\gamma}, \tag{2.14}$$

where $m(\gamma, j)$ represents the total number of cells of the γ-th cell type, or the *cell composition* in the j-th sample. Since the deconvolution problem does not distinguish individual cells, the equation above can be rewritten using the averaged gene expression profile (GEP) $\bar{c}_{\gamma}$ for each cell type. If cell type *proportion* is considered, let $M_j = \sum_{\gamma=1}^{k} m(\gamma, j)$ be the total number of cells in the j-th sample. Then, Eq.(2.14) becomes:

$$\hat{\mathbf{G}}_{(j)} = \sum_{\gamma=1}^{k} \frac{m(\gamma, j)}{\sum_{\gamma=1}^{k} m(\gamma, j)} M_j \bar{c}_{\gamma} = \sum_{\gamma=1}^{k} p_{\gamma,j} M_j \bar{c}_{\gamma}, \tag{2.15}$$

where $\sum_{\gamma=1}^{k} p_{\gamma,j} = 1$ represents the *proportion*.

It is important to highlight the difference between Eq.(2.15) and the j-th column of the NMF model, given by:

$$\mathbf{G}_{(j)} = \sum_{\gamma=1}^{k} p_{\gamma,j} \mathbf{C}^{(\gamma)}. \tag{2.16}$$

Here, the matrix $\mathbf{C} \in \mathbb{R}^{N \times k}$ consists of only k columns $\mathbf{C}^{(\gamma)}$ (for cell type indices $1 \leq \gamma \leq k$) and does not depend on the sample index j. This structure implies an important assumption in bulk tissue data deconvolution: the cell type-specific GEP should be homogeneous across tissue samples. This assumption provides a guideline for constructing pseudo-bulk tissue, ensuring that the total number of cells and the

average GEP, represented by $M_j\bar{c}_\gamma$, remain as consistent as possible across samples.

To generate pseudo-bulk tissue data, the following steps are taken:

- Choose the total number of cells M as a parameter for all pseudo-bulk tissue samples;
- Compute $\bar{c}_\gamma$ from the single-cell RNA-seq data for $1 \leq \gamma \leq k$;
- Select a mechanism to generate random values for $p_{\gamma,j}$, and use Eq. (2.15) to construct the j-th pseudo-bulk sample.

On the other hand, from a linear regression perspective, the matrix $\mathbf{C}$ estimated from bulk tissue data (whether pseudo or real) using the NMF model represents the averaged value:

$$\bar{c}_\gamma = \langle M_j\bar{c}_\gamma \rangle \tag{2.17}$$

across samples. This understanding raises two key concerns when applying the NMF model:

1. The original data must be *well-behaved*, meaning that the GEP across samples should be relatively homogeneous. If prior knowledge suggests that GEP varies significantly between samples, then NMF may not be a suitable model.

2. Proper use of single-cell data: Single-cell RNA-seq (scRNA-seq) data can be directly incorporated into NMF as a reference, reducing complete deconvolution to a standard linear regression problem. However, in practical applications, scRNA-seq data may not align well with the inherent GEP in bulk tissue data, especially if obtained from different sources. Such inconsistencies can significantly reduce the accuracy of both the estimated GEP matrix $\mathbf{C}$ and the inferred cellular proportion matrix $\mathbf{P}$ in the original bulk tissue data.

Pseudo-bulk tissue data with weak identification condition

To generate pseudo-bulk tissue data effectively, it is crucial to employ an appropriate mechanism for constructing pseudo proportions for each sample, as described in Eq. (2.15). Let $\hat{\mathbf{P}}$ denote the pseudo proportion matrix. We impose the weak identification condition on $\hat{\mathbf{P}}$, meaning that it must contain a sufficient number of columns, $\hat{\mathbf{P}}^{(j)}$, that are sufficiently close to $\mathbf{e}_\gamma \in \mathbb{R}^k$. To achieve this, we propose using a multivariate Dirichlet distribution.

Consider the n column vectors in matrix $\hat{\mathbf{P}}$ as n realizations of the random vector $\hat{\mathbf{p}} \in \mathbb{R}^k$. Given parameters $\alpha_1, \alpha_2, \ldots, \alpha_k > 0$, the probability density function with respect to the Lebesgue measure on the Euclidean space $\mathbb{R}^{k-1}$ is given by

$$f(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k; \alpha_1, \alpha_2, \ldots, \alpha_k) = \frac{1}{S(\alpha)} \prod_{\gamma=1}^{k} \hat{p}_\gamma^{\alpha_\gamma - 1}, \quad \text{where } \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k). \tag{2.18}$$

where $\{\hat{p}_\gamma\}$ lie within the standard $k-1$ simplex due to the sum-to-one constraint. The normalizing constant $S(\alpha)$ is defined in terms of the gamma function:

$$S(\alpha) = \frac{\prod_{\gamma=1}^{k} \Gamma(\alpha_\gamma)}{\Gamma\left(\sum_{\gamma=1}^{k} \alpha_\gamma\right)}. \tag{2.19}$$

It is well known that the mean of the variable is $E[\hat{p}_\gamma] = \alpha_\gamma/\alpha_0$, where $\alpha_0 = \sum_{\gamma=1}^{k} \alpha_\gamma$. The variance and covariance of the random vector are given by

$$\mathrm{Var}[\hat{p}_\gamma] = \frac{\alpha_\gamma(\alpha_0 - \alpha_\gamma)}{\alpha_0^2(\alpha_0 + 1)}, \quad \text{and} \quad \mathrm{Cov}[\hat{p}_\gamma, \hat{p}_{\gamma'}] = \frac{-\alpha_\gamma \alpha_{\gamma'}}{\alpha_0^2(\alpha_0 + 1)}. \tag{2.20}$$

By selecting small and equal values for α_γ, one can achieve a smaller α_0, resulting in more widely and evenly scattered vectors within the $k-1$ simplex.

The regularity of matrix $\hat{\mathbf{P}}$ can be examined using the condition number of $\hat{\mathbf{P}}\hat{\mathbf{P}}^T$. If $\hat{\mathbf{P}}$ contains highly correlated or excessively small rows, the condition number of $\hat{\mathbf{P}}\hat{\mathbf{P}}^T$

will be large. Since $\text{Var}[\hat{p}_\gamma] = E[\hat{p}_\gamma^2] - E[\hat{p}_\gamma]^2$ and $\text{Cov}[\hat{p}_\gamma, \hat{p}_{\gamma'}] = E[\hat{p}_\gamma \hat{p}_{\gamma'}] - E[\hat{p}_\gamma]E[\hat{p}_{\gamma'}]$, we derive

$$E[\hat{p}_\gamma^2] = \frac{\alpha_\gamma(\alpha_\gamma + 1)}{\alpha_0(\alpha_0 + 1)}, \quad E[\hat{p}_\gamma \hat{p}_{\gamma'}] = \frac{\alpha_\gamma \alpha_{\gamma'}}{\alpha_0(\alpha_0 + 1)}. \tag{2.21}$$

When the sample size is sufficiently large, the matrix $\hat{\mathbf{P}}\hat{\mathbf{P}}^T$ takes the form

$$\hat{\mathbf{P}}\hat{\mathbf{P}}^T = c_0 \left(\begin{bmatrix} \alpha_1 & & & \\ & \alpha_2 & & \\ & & \ddots & \\ & & & \alpha_k \end{bmatrix} + \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{bmatrix} \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_k \end{bmatrix} \right), \tag{2.22}$$

where $c_0 = n/\alpha_0(\alpha_0 + 1)$ and n is the sample size.

Equation (2.22) provides insight into selecting Dirichlet distribution parameters to ensure a small condition number for $\hat{\mathbf{P}}\hat{\mathbf{P}}^T$. If $\alpha_1 = \alpha_2 = ... = \alpha_k = \alpha$, then

$$\hat{\mathbf{P}}\hat{\mathbf{P}}^T = c_0 \alpha \left(\mathbf{I}_{k \times k} + \alpha \mathbb{1}\mathbb{1}^T \right). \tag{2.23}$$

Since the eigenvalues of $\mathbb{1}\mathbb{1}^T$ are $(k, 0, 0, ...0)$, it follows that $\kappa(\hat{\mathbf{P}}\hat{\mathbf{P}}^T) = k\alpha + 1$. Thus, selecting a uniform vector $\alpha = (\alpha, \alpha, ..., \alpha)$ with a small value of α helps maintain a small condition number. Furthermore, for the mixed bulk tissue data (original and pseudo-bulk), the proportion matrix $\bar{\mathbf{P}} = [\mathbf{P}, \hat{\mathbf{P}}]$ is always regular. Specifically,

$$\bar{\mathbf{P}}\bar{\mathbf{P}}^T = \mathbf{P}\mathbf{P}^T + \hat{\mathbf{P}}\hat{\mathbf{P}}^T \tag{2.24}$$

has full rank, as $\hat{\mathbf{P}}\hat{\mathbf{P}}^T$ contains a diagonal matrix and the remaining matrix is positive semi-definite.

Computational process of the proposed algorithms

In this section, we summarize and assemble all components of the proposed method.

According to Eq. (2.20), we select small and equal values for the parameters α, e.g., $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 1.5$, and use the Dirichlet distribution (2.18) to generate n random vectors $\hat{\mathbf{p}}_j \in \mathbb{R}^k$, where $j = 1, 2, \ldots, n$, each belonging to the standard $k-1$ simplex.

Assume that there are M_γ cells in the single-cell dataset for the γ-th cell type, and let $M = \sum_{\gamma=1}^{k} |M_\gamma|$. Denote $\hat{p}_{\gamma,j}$ as the γ-th entry of vector $\hat{\mathbf{p}}_j$. We then randomly select $\lfloor M\hat{p}_{\gamma,j} \rfloor$ cells from the total M_γ cells of the γ-th type (with replacement). The j-th pseudo-bulk tissue sample, or column of $\hat{\mathbf{G}}$, is then generated as in Eq. (2.14):

$$\hat{\mathbf{G}}^{(j)} = \sum_{\gamma=1}^{k} \sum_{l=1}^{\lfloor M\hat{p}_{\gamma,j} \rfloor} g_{\gamma,l}. \tag{2.25}$$

The subset of marker genes from the augmented data $\tilde{\mathbf{G}} = [\mathbf{G}, \hat{\mathbf{G}}]$ is identified by clustering the rows of $\tilde{\mathbf{G}}$ into k groups. The N rows of $\tilde{\mathbf{G}}$ form the vertex set $V = \{\tilde{\mathbf{G}}_{(i)}\}_{i=1}^{N} \subset \mathbb{R}^n$ in the similarity graph $G = (V, E)$, following the spectral clustering method [62].

The non-negative weights ω_{ij} of edges $E = \{e_{ij}\}$ are computed using a function $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$:

$$\omega_{ij} = \exp\left\{ -\frac{d_{\text{eisen}}\left(\tilde{\mathbf{G}}_{(i)}, \tilde{\mathbf{G}}_{(j)}\right)^2}{\sigma} \right\}, \quad 1 \le i \le N, 1 \le j \le N, \tag{2.26}$$

where

$$d_{\text{eisen}}\left(\tilde{\mathbf{G}}_{(i)}, \tilde{\mathbf{G}}_{(j)}\right) = 1 - \frac{\langle \tilde{\mathbf{G}}_{(i)}, \tilde{\mathbf{G}}_{(j)} \rangle}{|\tilde{\mathbf{G}}_{(i)}||\tilde{\mathbf{G}}_{(j)}|} \tag{2.27}$$

is the Eisen cosine correlation distance, which quantifies the correlation between two vertices, and $\sigma > 0$ is a parameter.

With the adjacency matrix $\mathbf{W} = (\omega_{ij}) \in \mathbb{R}^{N \times N}$ and its degree matrix $\mathbf{D} = \text{diag}(d_1, d_2, \ldots, d_N)$, where $d_i = \sum_{j=1}^{N} \omega_{ij}$, different types of graph Laplacians (gL)

of G are defined:

- Unnormalized gL: $\mathbf{L} = \mathbf{D} - \mathbf{W}$
- Symmetric normalized gL: $\mathbf{L}_{\text{sym}} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$
- Random walk gL: $\mathbf{L}_{\text{rw}} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{W}$

By examining the first few eigenvectors of gL, rows of data $\tilde{\mathbf{G}}$ are clustered into k groups, and the set $\{G_r\}_{r=1}^{k}$ records row indices of $\tilde{\mathbf{G}}$ in the corresponding clusters. The choice of different gLs depends on specific data applications [63]. For our problem, we use the normalized gL $\mathbf{L}_{\text{sym}}$, and the matrix $\mathbf{W}$ is carried forward to the next step. Additionally, the disjoint sets S_γ, $\gamma = 1, 2, \ldots, k$, represent the indices of marker genes for each cell type.

The optimization problem for deconvolution is formulated as:

$$\min_{\mathbf{C}\geq 0, \mathbf{P}\geq 0} \quad \frac{1}{2}\|[\mathbf{G}, \hat{\mathbf{G}}] - \mathbf{C}[\mathbf{P}, \hat{\mathbf{P}}]\|_F^2 + \mathcal{F}(\mathbf{C}) + \mathbb{1}_T([\mathbf{P}, \hat{\mathbf{P}}]). \tag{2.28}$$

For simplicity, we use the Frobenius norm to measure the error between the deconvoluted solutions and the given mixed data. Here, $\mathbf{P} \in \mathbb{R}^{k\times n}$ corresponds to the original n samples, while $\hat{\mathbf{P}} \in \mathbb{R}^{k\times n'}$ represents the pseudo-bulk tissue. Both are subject to sum-to-one constraints on columns, represented by the set

$$T := \{Z \in \mathbb{R}_+^{k\times n} \mid \mathbb{1}^T Z = \mathbb{1}^T\}. \tag{2.29}$$

The indicator function $\mathbb{1}_T$ is defined as:

$$\mathbb{1}_T(Z) = \begin{cases} 0, & \text{if } Z \in T \\ \infty, & \text{otherwise.} \end{cases} \tag{2.30}$$

The regularization function $\mathcal{F}(\mathbf{C})$ is defined in the same way as in Section 2.2, and is illustrated in Equation 2.2 and Equation 2.3.

Computational algorithm and parameter discussion

The Alternating Direction Method of Multipliers (ADMM) [51] is employed to numerically solve this optimization problem. The overall computational process involves two types of parameters.

The first set consists of parameters within the ADMM algorithm itself. A detailed discussion of these parameters can be found in Section 2.2, and the same parameter tuning strategies are adopted in this section.

The second set comprises two key parameters related to the generation and integration of pseudo-bulk tissue data. The first is the vector $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ in the Dirichlet distribution, which determines the pseudo-cellular proportions. As discussed in Section 2.5, choosing small and equal values for α_l $(l = 1, 2, \ldots, k)$ satisfies both the geometric structure requirements and ensures computational stability. The second parameter is the number of pseudo-bulk samples, denoted as n'. If n' is too small, the dataset will lack sufficient information. Conversely, an excessively large n' is unnecessary, as per Definition (3), and would increase computational cost.

Determining the optimal number of pseudo-bulk samples is a nontrivial problem. From Eq. (2.13), we observe that the solution $\mathbf{C}^*$ represents the "average" of the constituent gene expression profiles (GEPs) derived from both the original bulk tissue samples and the reference single-cell RNA sequencing (scRNA-seq) data. If these two datasets are not well-aligned, the generated pseudo-bulk samples may introduce inaccuracies in the estimated GEP, ultimately affecting the inferred cellular proportions. In Section 2.5, we address this issue computationally using a specific dataset, while a more systematic investigation is left for future work.

Evaluation Metrics

In many studies and computational packages, the most commonly used evaluation metric is the row-wise (across samples) correlation between the estimated and ground truth cell proportion matrices, i.e., comparing $\mathbf{P}_{(i)}$ and $\mathbf{P}^*_{(i)}$ for $i = 1, 2, \ldots, k$.

However, this metric has notable limitations.

One issue is that it primarily reflects the *global* similarity of the estimated and true proportions for each cell type across all samples, but provides little insight into the accuracy of estimates for all cell types within a *specific* sample. More importantly, due to the inherent non-uniqueness of NMF solutions, each row of matrix $\mathbf{P}$ may be subject to an arbitrary rescaling factor. Specifically, $\lambda_i \mathbf{P}_{(i)}$ will yield the same correlation coefficient with $\mathbf{P}^*_{(i)}$ as $\mathbf{P}_{(i)}$. Thus, even if all corresponding rows of $\mathbf{P}$ and $\mathbf{P}^*$ are highly correlated, the columns of these matrices could differ significantly, leading to inaccurate estimates of cellular composition for each sample.

To address these shortcomings, in addition to Pearson correlation, we also evaluate simulation accuracy using the normalized mean absolute difference (mAD), which corresponds to the discrete L^1 norm, and the normalized root mean square deviation (RMSD), which corresponds to the discrete L^2 norm. Similar metrics are also applied to the columns of matrix $\mathbf{C}$, which represent the gene expression profiles (GEPs) for each cell type.

2.4 Solution analysis and biological insights

According to the theorem, the solution remains subject to a rescaling factor. That is, if $\mathbf{C}$ and $\mathbf{P}$ represent the "true" solutions of the NMF problem, the best *computational* solution one can obtain is given by

$$\bar{\mathbf{C}} = \mathbf{C}\Omega_\lambda^{-1}, \quad \text{and} \quad \bar{\mathbf{P}} = \Omega_\lambda \mathbf{P}, \tag{2.31}$$

where $\Omega_\lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k)$ is an unknown diagonal matrix with positive entries ($\lambda_k > 0$). This unknown rescaling matrix can lead to challenges in accurately estimating cellular proportions.

A key contribution of this work is the development of posterior error estimates for the NMF solution, providing critical insights into data preparation for realistic

biological applications. First, we investigate how the solution of the non-convex problem depends on initial conditions. Then, we attempt to estimate the rescaling matrix.

It is important to note that the results in this section are based on the matrix decomposition form $\mathbf{G} = \mathbf{CP}$, where $\mathbf{P}$ is assumed to be well-conditioned. Additionally, we do not account for noise or the nonnegativity constraint in this analysis.

Informative initial condition

First, we consider the converged solution.

Theorem 3. *Given initial condition* $\mathbf{P}_{[0]}$ *with* $\mathbf{P}_{[0]}\mathbf{P}_{[0]}^T$ *being invertible and assume the true GEP and cell proportion as* $\mathbf{C}$ *and* $\mathbf{P}$*, respectively, then the solutions of* $\mathbf{G} = \mathbf{CP}$ *converge to* $\mathbf{C}^* = \mathbf{CM}_1\mathbf{M}_2^{-1}$ *and* $\mathbf{P}^* = \mathbf{M}_2\mathbf{M}_1^{-1}\mathbf{P}$*, where* $\mathbf{M}_1 = \mathbf{PP}_{[0]}^T$ *and* $\mathbf{M}_2 = \mathbf{P}_{[0]}\mathbf{P}_{[0]}^T$.

Proof. With the simple iteration,

$$\mathbf{C}_{[0]} = \mathbf{GP}_{[0]}^T(\mathbf{P}_{[0]}\mathbf{P}_{[0]}^T)^{-1} = \mathbf{CPP}_{[0]}^T(\mathbf{P}_{[0]}\mathbf{P}_{[0]}^T)^{-1} = \mathbf{CM}_1\mathbf{M}_2^{-1} \tag{2.32}$$

Then

$$\begin{aligned}
\mathbf{P}_{[1]} &= \left(\mathbf{C}_{[0]}^T\mathbf{C}_{[0]}\right)^{-1}\mathbf{C}_{[0]}^T\mathbf{G} \\
&= \left(\mathbf{C}_{[0]}^T\mathbf{C}_{[0]}\right)^{-1}\mathbf{C}_{[0]}^T\mathbf{CP} \\
&= \left[(\mathbf{M}_2^{-1})^T\mathbf{M}_1^T\mathbf{C}^T\mathbf{CM}_1\mathbf{M}_2^{-1}\right]^{-1}(\mathbf{M}_2^{-1})^T\mathbf{M}_1^T\mathbf{C}^T\mathbf{CP} \\
&= \mathbf{M}_2\mathbf{M}_1^{-1}\mathbf{P}
\end{aligned} \tag{2.33}$$

with such updated $\mathbf{P}_{[1]}$, we compute

$$\begin{aligned}\mathbf{C}_{[1]} &= \mathbf{G}\mathbf{P}_{[1]}^T(\mathbf{P}_{[1]}\mathbf{P}_{[1]}^T)^{-1} \\ &= \mathbf{C}\mathbf{P}\mathbf{P}^T(\mathbf{M}_1^{-1})^T\left(\mathbf{M}_2\mathbf{M}_1^{-1}\mathbf{P}\mathbf{P}^T(\mathbf{M}_1^{-1})^T\mathbf{M}_2^T\right)^{-1} \\ &= \mathbf{C}\mathbf{M}_1\mathbf{M}_2^{-1} \\ &= \mathbf{C}_{[0]}\end{aligned} \tag{2.34}$$

Based on Theorem 3, it is possible to integrate information from marker genes and appropriately select initial conditions to estimate the computational solutions. At this stage, we assume that the marker genes and their associated cell types have been identified. Mathematically, we define the set $\mathcal{M}$ to represent the indices of all marker genes, while the disjoint sets S_γ, for $\gamma = 1, 2, \ldots, k$, represent the indices of marker genes corresponding to the γ-th cell type. Notably, we have $\bigcup_\gamma S_\gamma = \mathcal{M}$, and we denote the complement as $S_\gamma^c = \mathcal{M} \setminus S_\gamma$.

Based on the properties of marker genes, we ideally expect $c_{[S_\gamma^c],\gamma} = 0$, where $c_{[S_\gamma^c],\gamma}$ represents all entries of the γ-th column of $\mathbf{C}$ with row indices in S_γ^c. Then, for certain γ and γ', let $i_1, i_2 \in S_\gamma$ and $i_3 \in S_{\gamma'}$. It follows that:

$$\mathbf{G}_{(i_1)} = c_{i_1,\gamma}\mathbf{P}_{(\gamma)}, \quad \mathbf{G}_{(i_2)} = c_{i_2,\gamma}\mathbf{P}_{(\gamma)}, \quad \text{and} \quad \mathbf{G}_{(i_3)} = c_{i_3,\gamma'}\mathbf{P}_{(\gamma')}. \tag{2.35}$$

From Equation (2.35), we conclude that the rows of $\mathbf{G}$ corresponding to the marker genes of the same cell type (i.e., i_1 and i_2) are linearly dependent since they are scalar multiples of the same γ-th row of $\mathbf{P}$.

Furthermore, for every $1 \leq \gamma \leq k$, we compute the mean of all rows $\mathbf{G}_{(i)}$ over all $i \in S_\gamma$, denoted as $\langle \mathbf{G}_{(i)} \rangle_{S_\gamma}$, and define the result as the γ-th row of a matrix $\tilde{\mathbf{P}}$:

$$\tilde{\mathbf{P}}_{(\gamma)} = \langle \mathbf{G}_{(i)} \rangle_{S_\gamma} = \langle c_{i,\gamma} \rangle_{S_\gamma} \mathbf{P}_{(\gamma)}, \quad \text{or equivalently,} \quad \tilde{\mathbf{P}} = \Omega_c \mathbf{P}, \tag{2.36}$$

where $\Omega_c = \text{diag}(\mathcal{A})$ and $\mathcal{A}^T = [\langle c_{i,1} \rangle_{S_1}, \ldots, \langle c_{i,\gamma} \rangle_{S_\gamma}, \ldots, \langle c_{i,k} \rangle_{S_k}]$.

If a well-defined identification of marker gene sets is achieved, the matrix $\tilde{\mathbf{P}}$ can

be directly computed from the original data $\mathbf{G}$. This matrix is related to the true cellular proportion matrix $\mathbf{P}$ through a diagonal rescaling factor Ω_c. Although this factor remains unknown, $\tilde{\mathbf{P}}$ can still serve as an informative initial condition.

Estimation of computational solutions

First, if initial condition $\mathbf{P}_{[0]}$ as the column normalization of $\tilde{\mathbf{P}}$ with sum-to-one condition, i.e., $\mathbf{P}_{[0]} = \Omega_c \mathbf{P} \Omega_a^{-1}$, where

$$\Omega_a = \text{diag}(\mathcal{A}^T \mathbf{P}^{(1)}, \mathcal{A}^T \mathbf{P}^{(2)}, ..., \mathcal{A}^T \mathbf{P}^{(n)}). \tag{2.37}$$

Consequently, by Theorem (3), solutions converge to

$$\mathbf{C}^* = \mathbf{C}\mathbf{M}_1\mathbf{M}_2^{-1} = \mathbf{C}(\mathbf{P}\Omega_a^{-1}\mathbf{P}^T)(\mathbf{P}\Omega_a^{-2}\mathbf{P}^T)^{-1}\Omega_c^{-1}. \tag{2.38}$$

and

$$\mathbf{P}^* = \mathbf{M}_2\mathbf{M}_1^{-1}\mathbf{P} = \Omega_c(\mathbf{P}\Omega_a^{-2}\mathbf{P}^T)(\mathbf{P}\Omega_a^{-1}\mathbf{P}^T)^{-1}\mathbf{P}. \tag{2.39}$$

Theorem 4. *Assume ideal data* $\mathbf{G}$ *has true solutions* $\mathbf{C}$ *and* $\mathbf{P}$*, i.e.,* $\mathbf{G} = \mathbf{C}\mathbf{P}$*. Additionally, row index set* Ω *of* $\mathbf{C}$ *(also* $\mathbf{G}$*) can be subdivided into* k *disjoint sets* S_γ*, i.e.,* $\Omega = \bigcup_\gamma S_\gamma$*, and entry of* $\mathbf{C}$ *is* $c_{i,\gamma} = 0$ *if* $i \notin S_\gamma$*. Given initial condition* $\mathbf{P}_{[0]} = \Omega_c \mathbf{P} \Omega_a^{-1}$ *and the solutions of* $\mathbf{G} = \mathbf{C}\mathbf{P}$ *converge to* $\mathbf{C}^*$ *and* $\mathbf{P}^*$*, then the following estimations hold for each row of* $\mathbf{C}^*$

$$\beta^{-1}||\mathbf{C}_{(i)}||_2 \leq ||\mathbf{C}^*_{(i)}||_2 \leq \beta||\mathbf{C}_{(i)}||_2 \tag{2.40}$$

and each column of $\mathbf{P}^*$*,*

$$\beta^{-1}||\mathbf{P}^{(j)}||_2 \leq ||\mathbf{P}^{*(j)}||_2 \leq \beta||\mathbf{P}^{(j)}||_2, \tag{2.41}$$

where

$$\beta = \sqrt{\kappa(\mathbf{P}\mathbf{P}^T)} \left(\frac{\lambda^c_{\max}}{\lambda^c_{\min}} \right)^2, \tag{2.42}$$

where $\kappa(\mathbf{P}\mathbf{P}^T)$ *is the condition number of* $\mathbf{P}\mathbf{P}^T$ *and*

$$[\lambda^c_{\max}, \lambda^c_{\min}] = [\max, \min]\{< c_{i,1} >_{S_1}, ... < c_{i,k} >_{S_k}\}. \tag{2.43}$$

Proof. First denote the matrix $\mathbf{R} = (\mathbf{P}\Omega_a^{-1}\mathbf{P}^T)(\mathbf{P}\Omega_a^{-2}\mathbf{P}^T)^{-1}\Omega_c^{-1}$, hence $\mathbf{C}^* = \mathbf{C}\mathbf{R}$ and $\mathbf{P}^* = \mathbf{R}^{-1}\mathbf{P}$. Then we estimate $||\mathbf{R}||_2$ to see how it changes each row of the true GEP matrix $\mathbf{C}^*$. Assuming matrix $\mathbf{P}$ has SVD $\mathbf{P} = \mathbf{U}\Sigma\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{k\times k}$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$ being unitary, while $\Sigma \in \mathbb{R}^{k\times n}$ and has minimum and maximum values as $\sigma_{\min}$ and $\sigma_{\max}$, respectively, then

$$\begin{aligned} ||\mathbf{R}||_2 &= ||\mathbf{U}\Sigma\mathbf{V}^T\Omega_a^{-1}\mathbf{V}\mathbf{U}^T(\mathbf{U}\Sigma\mathbf{V}^T\Omega_a^{-2}\mathbf{V}\mathbf{U}^T)^{-1}\Omega_c^{-1}||_2 \\ &= ||\mathbf{U}\Sigma\mathbf{V}^T\Omega_a^{-1}\mathbf{V}\mathbf{U}^T(\Sigma^T)^{-1}\gamma^T\Omega_a^{-2}\mathbf{V}^{-1}\mathbf{U}^T\Omega_c^{-1}||_2 \\ &= ||\mathbf{U}\Sigma\mathbf{V}^T\Omega_a^{-1}\mathbf{V}\mathbf{I}\mathbf{V}^T\Sigma^T\Omega_a^{-2}\mathbf{V}^{-1}\mathbf{U}^T\Omega_c^{-1}||_2. \end{aligned} \tag{2.44}$$

Note that $\tilde{\mathbf{I}} \in \mathbb{R}^{n\times n}$ and $\tilde{\mathbf{I}} = \begin{pmatrix} \mathbf{I}_{k\times k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$ due to the definition of singular matrix Σ. Hence using the property that unitary matrices preserve spectrum norm, we have the estimate

$$||\mathbf{R}||_2 \leq \frac{1}{\lambda^c_{\min}} ||\mathbf{U}\Sigma\mathbf{V}^T\Omega_a^{-1}\tilde{\mathbf{V}}\mathbf{V}^T\Omega_a^2\Sigma^{-1}\mathbf{U}^T||_2 \tag{2.45}$$

$$= \frac{1}{\lambda^c_{\min}} ||\Sigma\mathbf{V}^T\Omega_a^{-1}\tilde{\mathbf{V}}\mathbf{V}^T\Omega_a^2\mathbf{V}\Sigma^{-1}||_2 \tag{2.46}$$

$$\leq \frac{\sigma_{\max}}{\sigma_{\min}\lambda^c_{\min}} ||\mathbf{V}^T\Omega_a^{-1}\tilde{\mathbf{V}}\mathbf{V}^T\Omega_a^2\mathbf{V}||_2 \tag{2.47}$$

$$= \frac{\sigma_{\max}}{\sigma_{\min}\lambda^c_{\min}} ||\Omega_a^{-1}\tilde{\mathbf{V}}\mathbf{V}^T\Omega_a^2||_2 \tag{2.48}$$

Note that each entry of Ω_a, i.e., $\mathcal{A}^T\mathbf{P}^{(j)}$, is actually a convex linear combination of entries in $\mathcal{A}$ since $\mathbf{P}$ has sum-to-one columns. Then exam the eigenvalues λ_a of Ω_a, we have

$$\lambda^c_{\min} \leq \lambda^a_{\min} \leq \lambda^a_{\max} \leq \lambda^c_{\max}. \tag{2.49}$$

Actually, if $\mathbf{P}$ satisfies weak or strong separability conditions, we have $\lambda^c_{\min} = \lambda^a_{\min}$ and $\lambda^c_{\max} = \lambda^a_{\max}$. Additionally, notice that $\sqrt{\text{cond}(\mathbf{P}\mathbf{P}^T)} = \sigma_{\max}/\sigma_{\min}$, we have the estimate

$$||\mathbf{R}||_2 = ||\mathbf{R}^T||_2 \leq \sqrt{\text{cond}(\mathbf{P}\mathbf{P}^T)} \left(\frac{\lambda^c_{\max}}{\lambda^c_{\min}}\right)^2 \tag{2.50}$$

Finally, by definition of spectral norm of matrix,

$$||\mathbf{C}^{*(i)}||_2 = ||\mathbf{R}^T\mathbf{C}^{(i)}||_2 \leq ||\mathbf{C}^{(i)}||_2\sqrt{\text{cond}(\mathbf{P}\mathbf{P}^T)} \left(\frac{\lambda^c_{\max}}{\lambda^c_{\min}}\right)^2, \tag{2.51}$$

and by the same argument, estimate in (2.50) is also true for $\mathbf{R}^{-1}$. Hence the other side inequality in (2.40) holds since $||\mathbf{C}^{(i)}||_2 = ||(\mathbf{R}^{-1})^T\mathbf{C}^{*(i)}||_2$.

Biological insights

The estimates in Theorem 4 are derived under noiseless conditions, assuming an ideal scenario where marker genes can be easily identified. Despite these simplifications, the results provide valuable qualitative insights into the conditions under which the constituent matrices $\mathbf{C}$ and $\mathbf{P}$ yield numerical solutions that closely approximate the ground truth.

- The quantity β in Eq. (2.42) serves as an estimator of the scaling factor in the "unique" solution, indicating the extent to which the computational results de-

viate from the ground truth. Intuitively, this quantity depends on the regularity of the constituent matrix $\mathbf{P}$, specifically on the condition number of $\mathbf{P}\mathbf{P}^T$.

- The parameters $\lambda^c_{\min}$ and $\lambda^c_{\max}$ represent the minimum and maximum averages of marker gene expression levels within their respective cell types. A smaller ratio of $\lambda^c_{\max}/\lambda^c_{\min}$ leads to a tighter bound on $||\mathbf{R}||_2$, improving numerical accuracy.

- Based on these observations, we conclude that maintaining a homogeneous average gene expression profile (GEP) across different cell types and ensuring regular cellular composition in bulk samples are preferable. For the former, subsets of marker genes within each cell type can be selected to minimize large variations in average GEP. For the latter, the motivation behind introducing pseudo-bulk tissue data is to improve the stability of the decomposition process. As demonstrated in the numerical simulations in the next section, solution accuracy is significantly enhanced when these strategies are applied.

2.5 Numerical Results

In this section, we present the numerical results of the proposed algorithms. First, we use synthetic data to illustrate the geometric structure of pseudo-cellular proportions generated by the Dirichlet distribution under different parameter settings. Properly designed data that satisfy the weak solvability condition significantly enhance the solution quality for singular data.

Next, we apply the proposed algorithms to three different datasets to validate their effectiveness and accuracy. Finally, we perform a computational analysis of the algorithm's resilience. In contrast to the original GSNMF method, the newly proposed pseudo-bulk tissue data augmentation approach is referred to as GSNMF+.

Simulation for synthetic data

Figure 2.4 illustrates the geometric structures of cellular proportions generated from a Dirichlet distribution with various parameter settings. For ease of visualization, the

cell type dimension is set to $k = 3$, allowing all data points to be represented within a 2-simplex (blue triangles). Six different parameter sets, $\boldsymbol{\alpha}$, are considered, including heterogeneous (first row) and homogeneous (second row) components.

The results indicate that when using heterogeneous parameters $\boldsymbol{\alpha}$, the convex hull of the generated data does not enclose the second-order cone, which corresponds to the inscribed circle (not shown) of the triangles, nor does it satisfy the sufficiently scattered condition. Conversely, data generated with homogeneous parameters tend to be more evenly distributed within the simplex. However, for larger values of $\boldsymbol{\alpha}$, the data points become concentrated toward the center of the simplex, preventing their convex hull from covering the second-order cone. This is consistent with the fact that larger values of $\boldsymbol{\alpha}$ lead to a higher condition number of $\mathbf{P}\mathbf{P}^T$.

Based on these observations, we will use homogeneous parameters with small to moderate values in our simulations, specifically $\|\boldsymbol{\alpha}\|_\infty = 0.2$ or $\|\boldsymbol{\alpha}\|_\infty = 1$, where $\|\boldsymbol{\alpha}\|_\infty$ denotes the l_∞ norm of $\boldsymbol{\alpha}$.

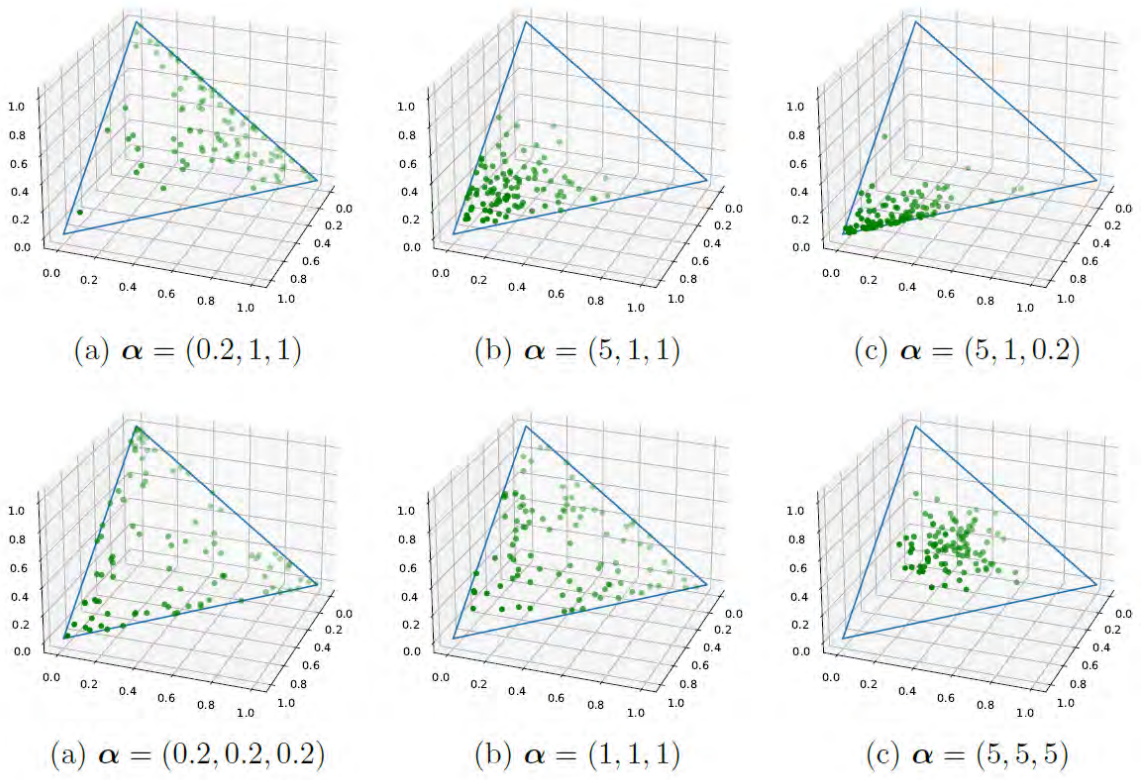


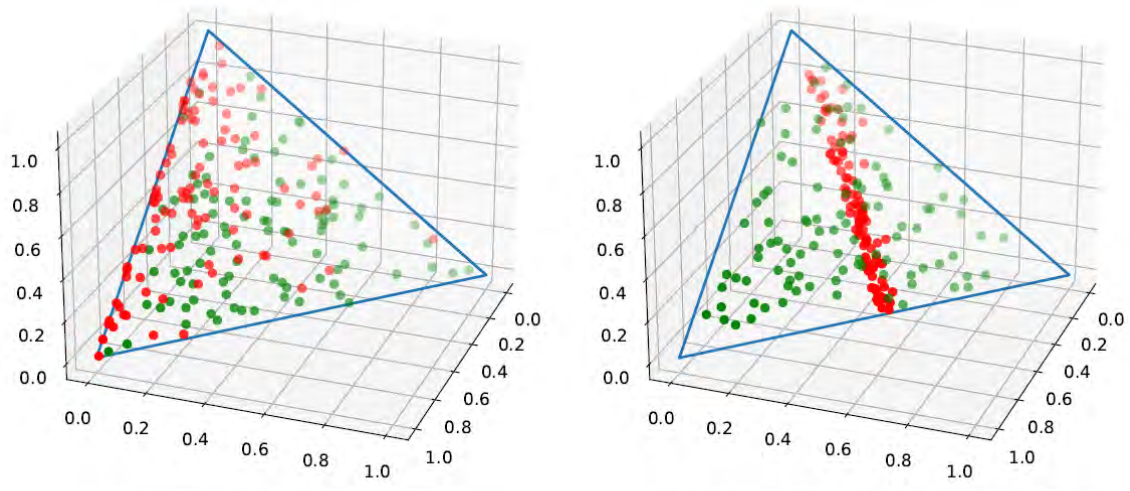
Figure 2.4: Geometric structures of simulated cellular proportions from Dirichlet distribution with various parameters.

Figure 2.5 illustrates how the overall solvability condition improves when singular original cell proportions (red dots) are mixed with pseudo data (green dots). Figure 2.5 (a) depicts a case of data singularity, where one cell type has an extremely small proportion in samples, corresponding to a small row vector in $\mathbf{P}$.

As shown, the red dots are primarily distributed along one edge of the simplex, with one corner notably absent. Another type of singularity occurs when two cell types have highly correlated cellular proportions, as demonstrated in Figure 2.5 (b). In this case, the data points (red dots) align approximately along a single line within the simplex.

In both scenarios, the geometric structures are improved by incorporating pseudo data generated from a Dirichlet distribution (green dots) with $\boldsymbol{\alpha} = (1, 1, 1)$, helping

to mitigate the singularity issues.

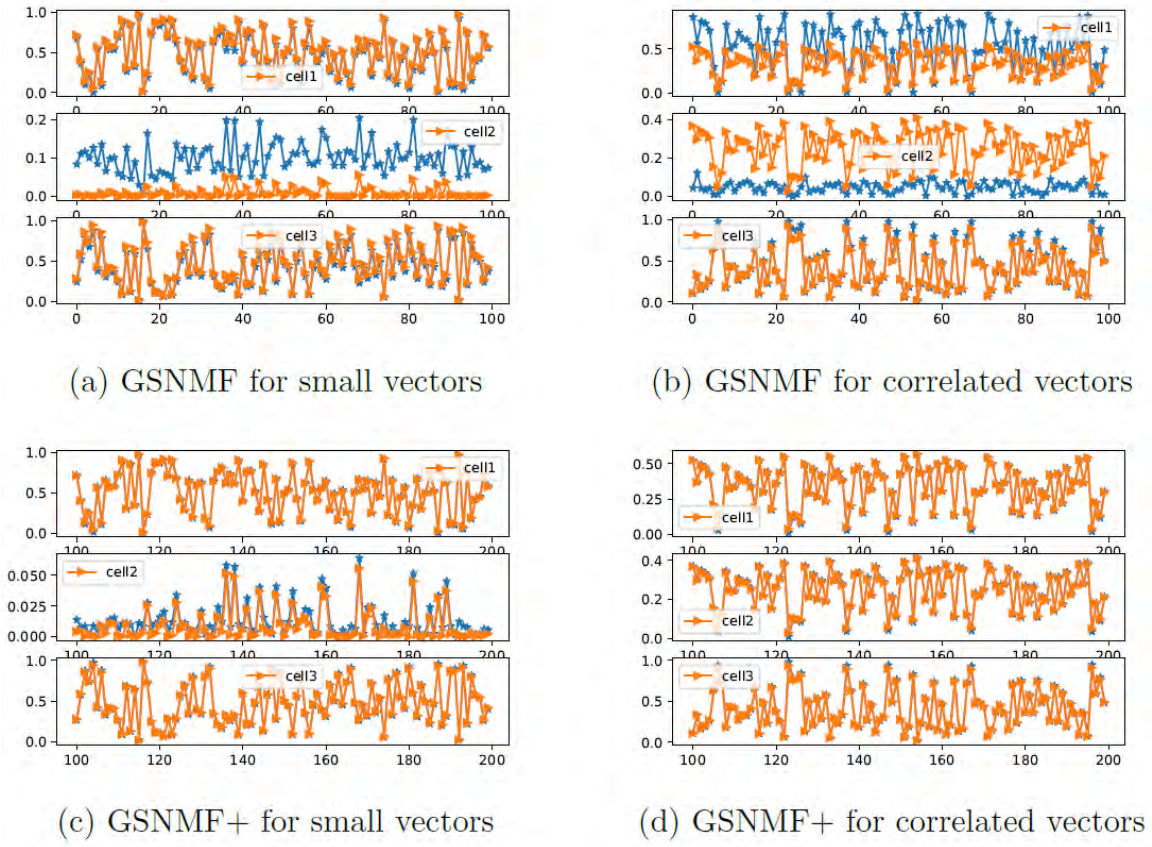


(a) Small row vector (b) Two correlated row vectors

Figure 2.5: Geometric structures of mixed data: original singular cell proportions (red) and simulated regular cell proportions (green).

Figure 2.6 presents the deconvolution results of cellular proportions in a set of synthetic data, where the cellular composition exhibits the aforementioned singularity. The first and second rows correspond to the original GSNMF method and the proposed GSNMF+ method, respectively. The synthetic bulk tissue data, $\mathbf{G}$, is generated by multiplying a predefined matrix $\mathbf{C}$ with the pseudo cellular composition matrix $\mathbf{P}$ and adding noise $\boldsymbol{\epsilon}$. A total of 100 synthetic samples are analyzed. The ground truth of the cellular composition is represented by orange triangles, while the deconvolution results are shown as blue stars. In the left panel, cell type 2 appears infrequently across all samples, resulting in very small proportions (below 0.1). The original GSNMF algorithm fails to produce accurate estimates for this cell type, with a solution correlation of 0.74 and an RMSD of 0.96. Similarly, the correlation for the corresponding column of matrix $\mathbf{C}$ is 0.52, and the RMSD is 0.99. In contrast, the GSNMF+ algorithm significantly improves deconvolution accuracy. For this cell type, the correlation increases to 0.96, while the RMSD drops to 0.06. The accuracy

of matrix $\mathbf{C}$ estimation is also enhanced, with a correlation of 0.98 and an RMSD of 0.13. A similar performance improvement is observed when the data contains correlated cellular compositions, as illustrated in the right panel of Figure 2.6. In this case, cellular proportions for cell types 1 and 2 are intentionally designed to be highly correlated. Consequently, the original GSNMF method produces highly inaccurate deconvolution results. The solution correlation for cell type 1 is 0.99, but the RMSD is 2.21. For cell type 2, both correlation and RMSD are poor, at 0.56 and 2.22, respectively. Moreover, the estimation of matrix $\mathbf{C}$ is entirely inaccurate, with an RMSD of 5.89. When GSNMF+ is applied, the deconvolution accuracy for all cell types improves significantly, with correlations exceeding 0.99 and RMSD values of 0.07, 0.05, and 0.12, respectively. Additionally, the overall RMSD for estimating matrix $\mathbf{C}$ decreases to 0.19.



(a) GSNMF for small vectors

(b) GSNMF for correlated vectors

(c) GSNMF+ for small vectors

(d) GSNMF+ for correlated vectors

Figure 2.6: Deconvolution results of cellular proportion from singular data by GSNMF and GSNMF+.

Table 2.1 presents the performance of GSNMF+ under different noise-to-data ratios (NDRs), showing the root mean square deviation (RMSD) of $\mathbf{C}$ and $\mathbf{P}$, along with the relative residue $\|\mathbf{G}-\mathbf{CP}\|_F/\|\mathbf{G}\|_F$. As indicated in the table, errors in matrix $\mathbf{C}$ become more pronounced as the noise level increases (higher NDR). In contrast, the computation of matrix $\mathbf{P}$ appears to be less affected by noise.

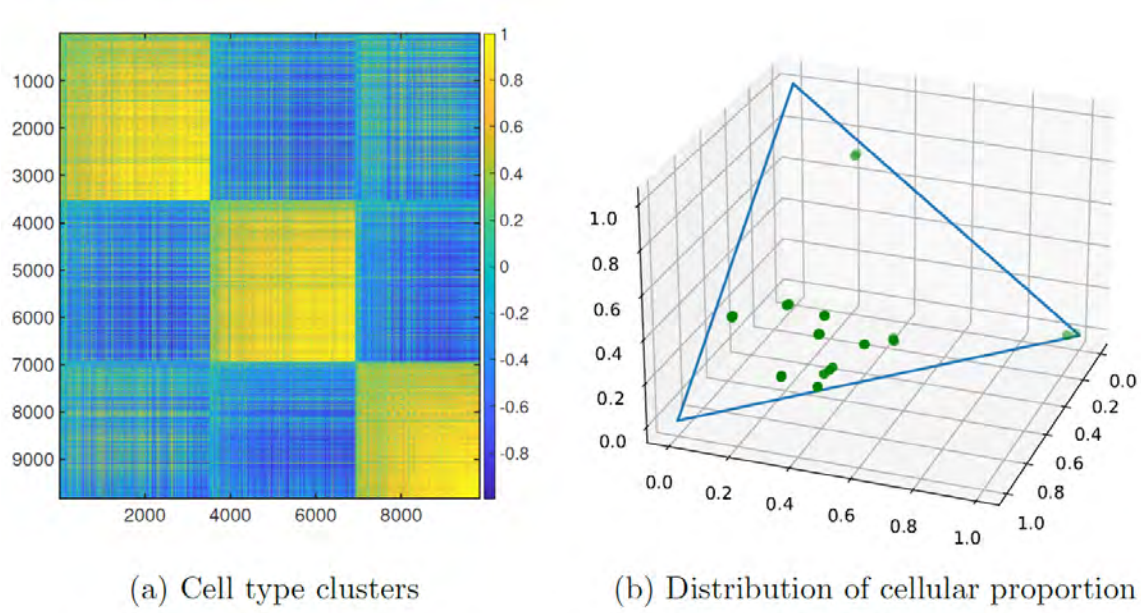
Table 2.1: Quantitative results of GSNMF+ under different noise-to-data ratios (NDRs).

NDR	RMSD of C	RMSD of P	Relative Residue
0.081	0.1802	0.0888	0.0793
0.165	0.2007	0.0940	0.1743
0.346	0.2372	0.1045	0.4024

It is worth noting that noise management is a common challenge across all areas of data science. However, addressing noise is not the primary focus of this work. Extracting meaningful biological insights from the ill-posed NMF model remains an open problem. Therefore, in all synthetic data experiments, we used low to moderate noise levels and considered only standard Gaussian noise.

2.6 Deconvolution Results of Biological Data

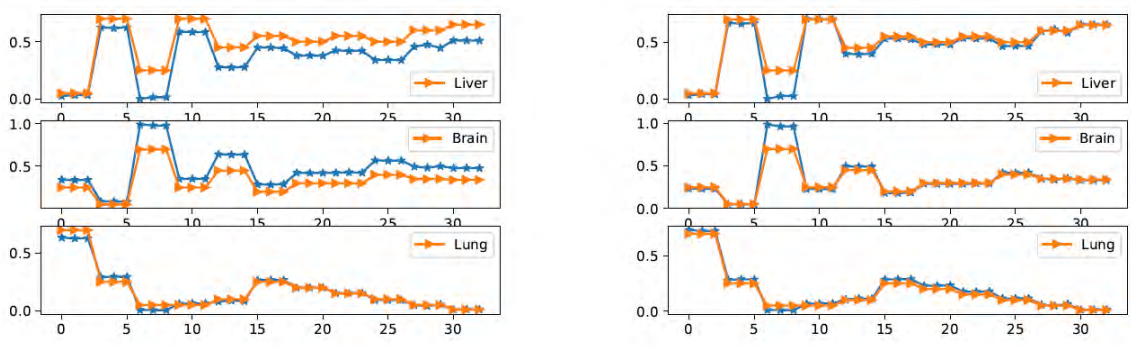
The first biological dataset used to validate the proposed algorithms is GSE19830 [64], which was obtained from tissue samples of the brain, liver, and lung of a single rat using expression arrays (Affymetrix). Homogenates of these three tissue types were mixed at the scRNA homogenate level in known proportions, and the gene expression patterns of each mixed sample were subsequently measured. The GSE19830 dataset mimics a common scenario in heterogeneous biological samples, where both the relative frequencies of the components and the marker genes for each cell type are clearly identifiable. As a result, this dataset has been widely used in the literature [35, 36] for preliminary testing of computational algorithms. In this dataset, the number of cell types is $k = 3$, and the number of tissue samples is $n = 33$. After necessary data preprocessing to exclude obvious outliers (based on row norms, column norms, etc.), we selected $N = 10,000$ genes from the total $\sim 12,000$ genes available.



(a) Cell type clusters (b) Distribution of cellular proportion

Figure 2.7: Visualization of data features for GSE19830.

Since the dimension k is low, the data features are easily visualized, as shown in Figure 2.7. First, the bulk-tissue data is clearly grouped into three clusters corresponding to the three cell types from distinct organs, as shown in Figure 2.7(a), making it straightforward to identify marker genes for each type. Secondly, the constituent (ground truth) cellular proportions were carefully designed to avoid correlation. As shown in Figure 2.7(b), the proportion data is sufficiently scattered within the simplex.



(a) GSNMF results (b) GSNMF+ results

Figure 2.8: Comparisons of simulation results (cellular proportion) to ground truth for dataset GSE19830

Figure 2.8 illustrates this case, where the orange triangle line represents the ground truth of cellular proportions, and the blue stars indicate the corresponding deconvolution results. Our previous GSNMF produced highly accurate solutions, as shown in Figure 2.8(a), with solution correlations of 0.96, 0.98, and 0.99, respectively. However, we observed a consistent pattern where the model tended to overestimate the proportion of one cell type (brain) while underestimating another (liver) across different samples. We hypothesize that this systematic error is due to an unknown rescaling factor.

To address this issue, we applied the proposed GSNMF+ algorithm to the same dataset. By estimating the rescaling factors and incorporating an adjustment strategy, we found that the accuracy was further improved, as shown in Figure 2.8(b). It is important to note that this improvement, or rescaling, is achieved at a global level (across all samples), as the errors in samples $6-8$ remain relatively unchanged.

The qualitative errors, measured as the sample-wise l_2 norm of errors in cellular proportions, are displayed in Figure 2.9. The orange bars represent the original GSNMF algorithm, while the blue bars correspond to the GSNMF+ algorithm developed in this study. Computational errors are expressed in logarithmic form, clearly

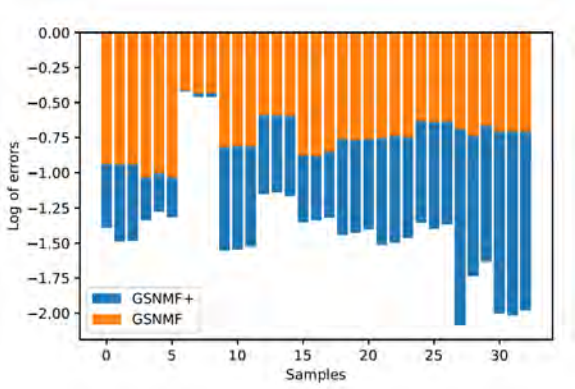


Figure 2.9: Qualitative errors of the deconvolution results in cellular proportions for dataset GSE19830.

demonstrating the improvement in accuracy.

The second dataset, GSE67835, consists of brain tissue samples containing five cell types ($k = 5$): astrocytes, endothelial cells, microglia, neurons, and oligodendrocytes. The dataset includes two distinct sample sets: one with an extremely small number of microglial cells (Case I) and another with a low proportion of endothelial cells (Case II). These imbalanced datasets present significant computational challenges and substantially reduce deconvolution accuracy.

Figure 2.10 compares the deconvolution results for these two cases (top and bottom rows) obtained using GSNMF and GSNMF+, respectively. The orange curves represent the ground truth, while the blue curves depict the deconvolution estimates. GSNMF notably overestimates the cellular proportions of microglia in Case I and endothelial cells in Case II. In contrast, GSNMF+ produces more accurate estimates, demonstrating improved performance in handling these challenging cases.

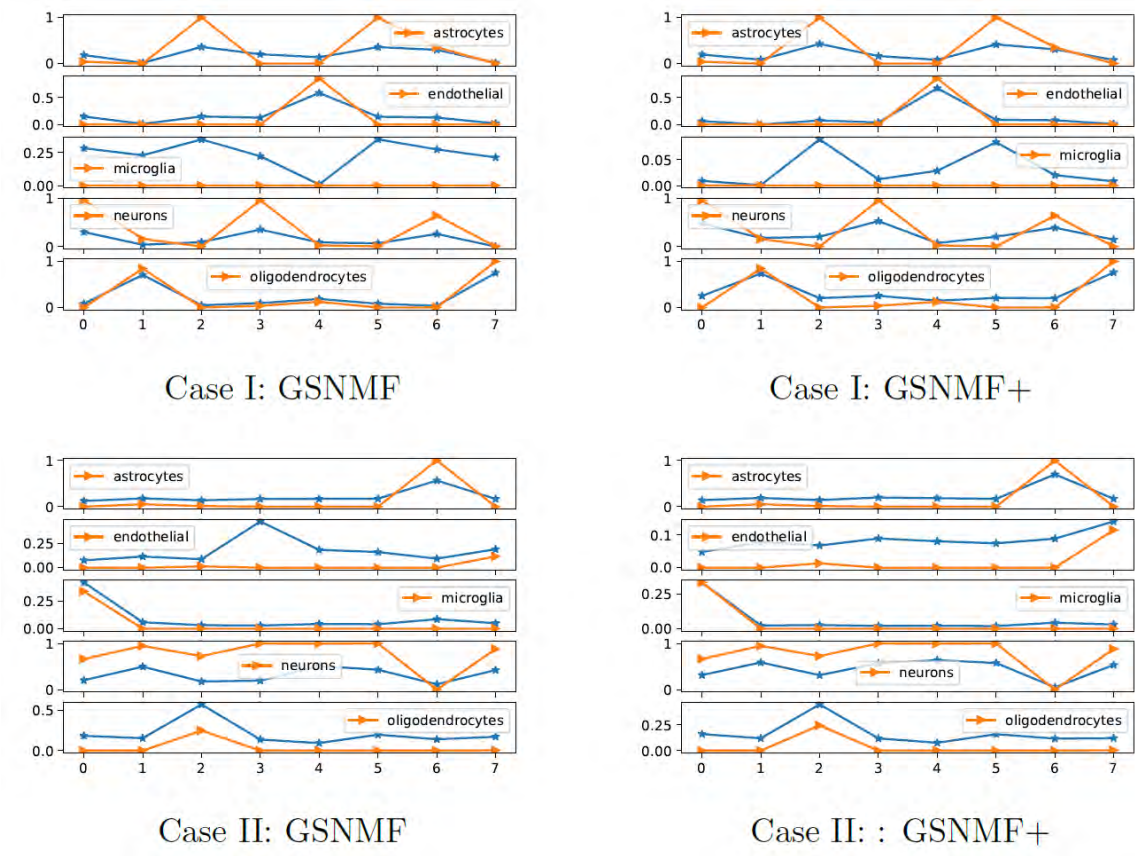


Figure 2.10: Performance comparisons of GSNMF and GSNMF+ for GSE67835.

2.7 Impacts of data variation

It is reasonable to assess the algorithm's performance when varying amounts of pseudo-bulk tissue data are augmented or when scRNA-seq data are perturbed. Figure 2.11 illustrates the solution accuracy for two datasets, GSE67835 and GSE81608, under different levels of pseudo-bulk sample augmentation in the GSNMF+ algorithm.

In these settings, the original bulk tissue data consist of 50 singular samples (denoted as **50S**), characterized by constituent cellular compositions with minimal variation, modeled using a Dirichlet distribution. Additional "regular" samples, which exhibit greater variation in cellular composition, are incrementally introduced into the algorithm, ranging from **10L** to **300L**.

For example, the label **50S+20L** indicates a mixture of 50 original bulk tissue sam-

ples with low variance in cellular proportions and 20 pseudo-bulk tissue samples with high variance in cellular proportions. Solution accuracy is evaluated using Pearson correlation, RMSD, and mAD.

The results indicate that adding "regular" samples significantly improves deconvolution accuracy. However, as the number of "regular" samples increases, solution accuracy fluctuates slightly but does not show a clear trend of improvement or deterioration. This observation aligns with theoretical findings regarding the weak identifiability condition, which suggests that only a few columns of $\mathbf{P}$ need to approximate each fundamental basis for $cone(\mathbf{P})$ to sufficiently cover the second-order cone in $\mathbb{R}^n_+$. Consequently, excessive pseudo-bulk tissue samples do not provide additional benefits.

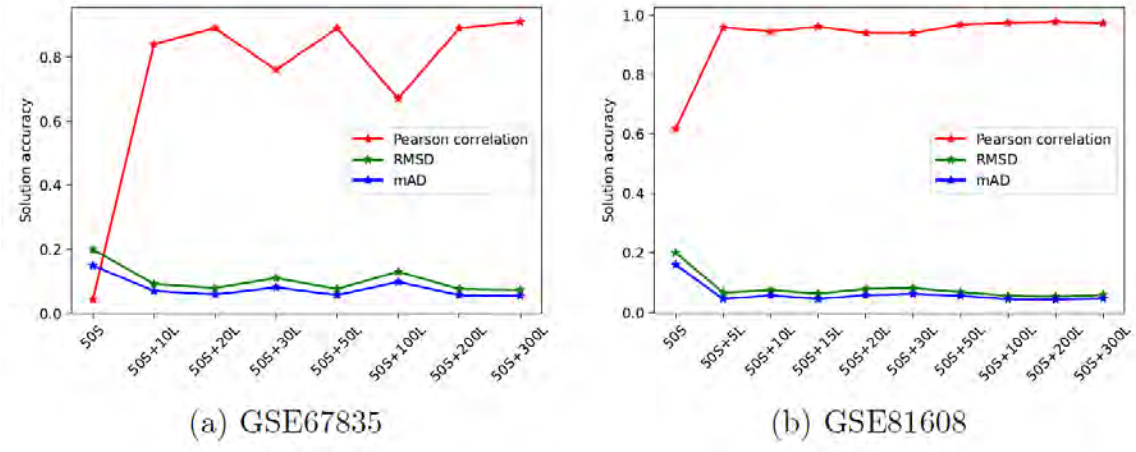


Figure 2.11: Algorithm accuracy for two biological datasets with different amount of pseudo-bulk data.

Another common scenario in real-world applications is that the reference scRNA-seq data used to construct pseudo-bulk tissues may not perfectly align with the constituent cell-specific gene expression profiles (GEP) in bulk tissues. To assess the impact of inconsistent scRNA-seq data on the proposed algorithm, we perform a series of model resilience analyses by generating pseudo-bulk tissue samples using artificially modified reference scRNA-seq data. These modifications include mean

shifting (MS), truncation (TR), and factoring (FA), which are specifically defined as follows:

- **Mean Shifting (MS)**: Each single-cell profile of a given type was modified by adding 10%, 30%, 50%, or 70% of the average expression levels of that cell type.
- **Truncation (TR)**: Single cells were selectively removed from the bulk tissue data generation process by excluding the top or bottom 10% of cells within each cell type.
- **Factoring (FA)**: A scaling factor was applied to all gene expression levels, enabling both upward and downward adjustments. The scaling factors used were 0.4, 0.8, 1.2, and 1.8.

All modifications were applied to the reference set, which consisted of 200 larger samples. As shown in Figure 2.12, these manipulations did not significantly impact the final results. The estimates remained close to the original values, indicating that the algorithm remains robust and effective despite changes to the reference set.

In Table 2.2, we compare GSNMF+ with several popular NMF-based deconvolution algorithms: MuSiC, CIBERSORTx, and Linseed, using data from GSE67835. Among these methods, MuSiC and CIBERSORTx are reference-based (scRNA-seq) approaches, GSNMF+ utilizes reference data only to mitigate data singularity, while Linseed operates without any reference data.

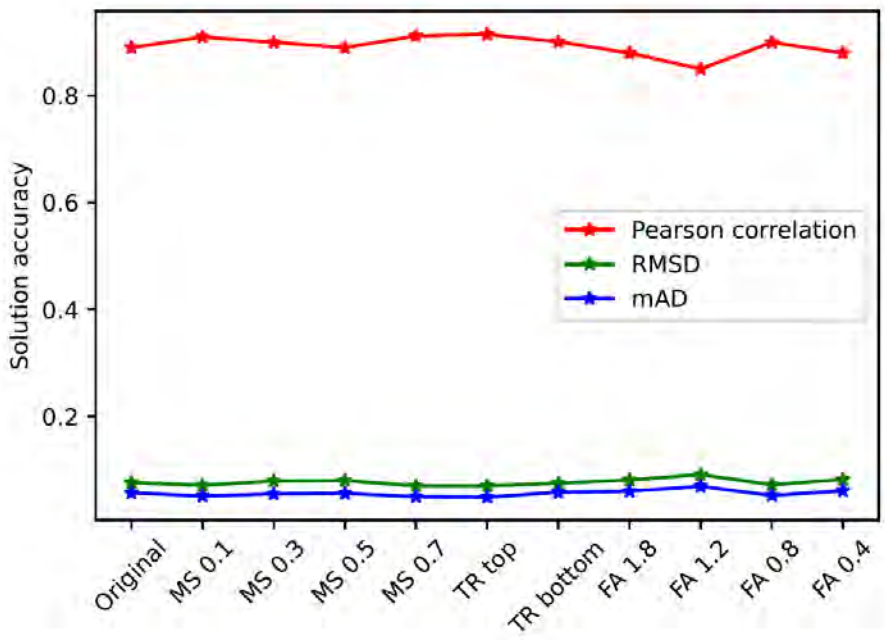


Figure 2.12: Algorithm performance when reference scRNA-seq data is perturbed in pseudo-bulk tissue data.

Table 2.2: Comparison of methods with different reference data.

	Correlation			
Method	**MuSiC**	**CIBERSORTx**	**Linseed**	**GSNMF+**
Original	0.9943	0.9643	0.8777	0.8712
snRNAseq	0.8538	0.7435	0.8888	0.8636
Simulator	0.2727	0.035	0.8749	0.9859
	RMSD			
Original	0.0416	0.1328	0.2025	0.1967
snRNAseq	0.2399	0.2486	0.1942	0.1946
Simulator	0.3806	0.4237	0.2049	0.0606

As previously discussed, reference-based methods (partial deconvolution) generally achieve higher accuracy than reference-free methods (complete deconvolution), assuming that the reference data accurately reflect the constituent gene expression profiles (GEP) in bulk tissue. However, this assumption is often violated in real-world applications. To examine these scenarios, we consider three different references: the

original reference from GSE67835 ("Original"), a single-nucleus RNA-seq ("snRNA-seq") dataset of the human brain from [65], and a reference generated from the scDesign simulator [66], a widely used probabilistic model based on real scRNA-seq data. The top and bottom halves of the table present the correlation and RMSD of cellular compositions obtained from these methods, respectively.

As expected, MuSiC and CIBERSORTx outperform the reference-free methods in both error metrics when the reference data are consistent with the bulk tissue's constituent GEP. This is because, with a well-matched reference, NMF-based deconvolution is effectively reduced to a least squares or linear regression problem. However, when the reference data deviate from the bulk tissue GEP, such as with snRNA-seq or simulated data, the solution accuracy significantly declines. In contrast, Linseed remains largely unaffected, as it does not rely on reference data.

For our proposed GSNMF+, since the reference is used primarily to address data singularity rather than directly drive the deconvolution process, it fundamentally remains an NMF-based method. As a result, its solution accuracy does not degrade with changes in reference data and, in some cases, even improves.

2.8 Discussion

Complete deconvolution is a computational approach used to decompose bulk-tissue RNA-seq data—an experimentally accessible, reliable, and cost-effective resource—into cell type-specific gene expression profiles (GEP) and cellular compositions. These components are critical for differential expression (DE) analysis, enabling the discovery of novel insights into genes and pathways associated with human diseases. While various mathematical and statistical methods have been developed to perform complete deconvolution, nonnegative matrix factorization (NMF) remains a fundamental model in this domain. However, NMF is inherently an ill-posed problem due to its non-separable solutions. Although numerous computational strategies have been proposed to improve solution accuracy, a comprehensive mathematical investi-

gation of the ill-posedness remains underdeveloped.

In this section, we focused on the solvability conditions of NMF in the context of bulk-tissue data. Among the two constituent matrices in NMF, the cell-specific GEP matrix, $\mathbf{C}$, satisfies the weak identifiability conditions due to the presence of marker genes. This allows us to identify and preserve the geometric structures of marker genes when estimating $\mathbf{C}$ from bulk-tissue data. In contrast, the cellular abundance matrix, $\mathbf{P}$, often violates identifiability conditions due to the presence of rare or highly correlated cell proportions. To address this issue, we leveraged the Dirichlet distribution to generate pseudo-cellular proportions and subsequently constructed pseudo-bulk tissue data using available single-cell RNA-seq data.

As a result, our developed GSNMF+ algorithm operates on hybrid bulk-tissue data (a combination of original and pseudo data), ensuring that both constituent matrices satisfy the weak identifiability condition. Despite meeting these solvability conditions, NMF solutions remain subject to an arbitrary rescaling matrix. We addressed this challenge by selecting a specific set of initial conditions and estimating the potential deviation from theoretical solutions. Although this investigation is conducted under idealized, noise-free conditions, it represents the first study explicitly addressing solution ambiguity in NMF. Our findings provide valuable insights into improving the qualitative accuracy of deconvolution results. The proposed algorithm pipeline has been tested on multiple datasets, demonstrating a significant improvement in solution accuracy for bulk-tissue data with singular cellular compositions.

CHAPTER 3: Benchmarking and Evaluation of Computational Deconvolution Methods for Bulk RNA Sequencing Data

3.1 Introduction

In this section, we benchmark the robustness and resilience of computational deconvolution methods for estimating cell type proportions in bulk tissues, focusing on the comparison between reference-based and reference-free approaches. Robustness is assessed by generating in silico pseudo-bulk RNA sequencing data from cell-level gene expression profiles across four distinct tissue types, simulating cellular composition at varying levels of heterogeneity. To evaluate resilience, we introduce intentional alterations to single-cell RNA profiles to create pseudo-bulk RNA-seq data. Deconvolution estimates are then compared with ground truth values using Pearson's correlation coefficient, root mean squared deviation, and mean absolute deviation.

The results demonstrate that reference-based methods perform more robustly when reliable reference data are available, whereas reference-free methods excel in settings where suitable reference data are lacking. Additionally, variations in cell-level transcriptomic profiles and cellular composition emerge as key factors influencing the performance of deconvolution methods. This study provides valuable insights into the factors affecting bulk tissue deconvolution accuracy, offering guidance for users and contributing to the development of more powerful and reliable computational approaches in the future.

3.2 Study Design

We evaluate the robustness and resilience of two reference-based and two reference-free deconvolution methods by assessing their performance across various scenarios

and identifying key risk factors that influence their accuracy. The reference-based methods include CIBERSORTx [31] and MuSiC [28] while the reference-free methods are Linseed[36] and GSNMF [37]. To assess robustness, we use *in silico* pseudo-bulk RNA sequencing data that simulate a wide range of biological scenarios and evaluate the methods' performance. For resilience, we intentionally modify the distribution of single-cell RNA-seq profiles in the pseudo-bulk data while using the original scRNA-seq data as the reference for deconvolution analysis.

To generate pseudo-bulk tissues, we simulate cell proportions using a multivariate Dirichlet distribution with parameter γ, which represents the expected proportions of different cell types. Single-cell profiles are then randomly sampled from an scRNA-seq dataset and summed to create pseudo-bulk RNA-seq data. The four deconvolution methods are applied to estimate cell compositions in each pseudo-bulk sample. Their estimated cell proportions are then compared against the ground truth using three evaluation metrics: Pearson's correlation coefficient (R), root mean squared deviation ($RMSD$), and mean absolute deviation (MAD).

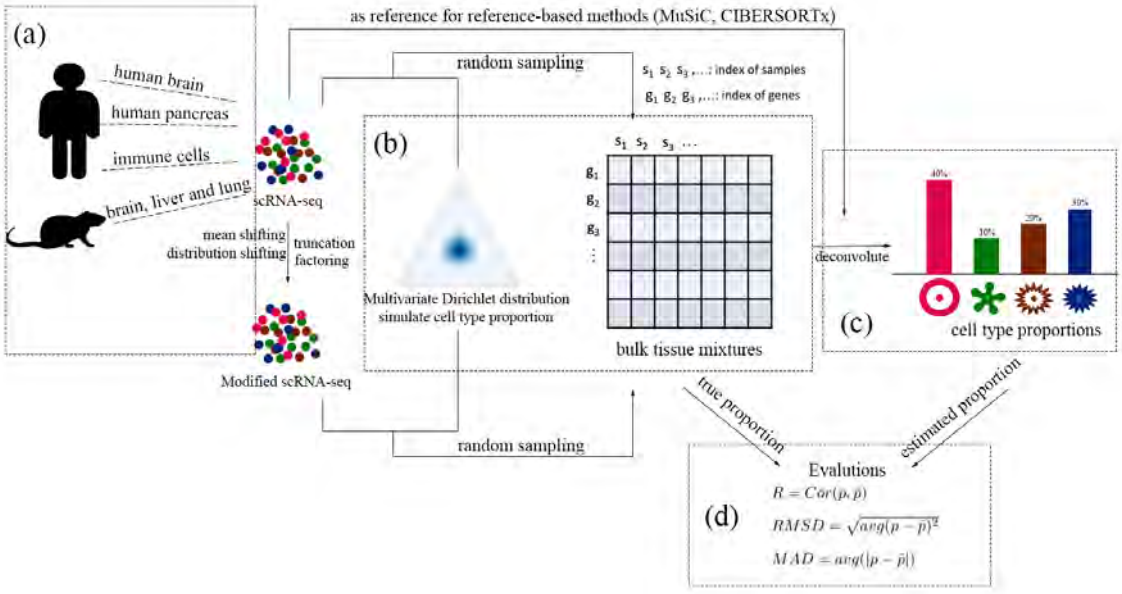


Figure 3.1: Study design. (a) A cell level dataset is used to generated pseudo-bulk tissues. (b) Ideal mixing to generate pseudo bulk tissue RNA-seq data. (c) Deconvolution for cell proportions. (d) Performance evaluation.

Notations and Algorithms

To clarify the notation used in subsequent sections, we introduce the mathematical framework here. Bulk tissue RNA-seq data is represented by a matrix $\mathbf{G}_{M\times n}$, where M denotes the number of genes and n is the number of bulk tissue samples. Assuming the presence of K distinct cell types in the bulk tissue samples, we denote the expected expression levels of these cells by a matrix $\mathbf{C}_{M\times K}$, while their relative proportions are stored in matrix $\mathbf{P}_{K\times n}$. The goal of bulk RNA-seq deconvolution is to estimate the cell type proportions, $\mathbf{P}$, as described in Equation (3.1):

$$\mathbf{G} = \mathbf{CP} + \varepsilon. \tag{3.1}$$

Here, ε represents random error with $E(\varepsilon) = \mathbf{0}$.

Reference-based methods utilize external reference datasets to estimate the expected cell-type-specific expression matrix $\mathbf{C}$ (e.g., the signature matrix in CIBERSORTx) and assume it is known in Equation (3.1), leaving only $\mathbf{P}$ to be estimated. CIBERSORTx, based on a ν-support vector regression model [67], is designed to handle challenges such as noise, unknown mixture content, and closely related cell types. It is widely used across various bulk tissue types, including whole blood and solid tumors [20]. MuSiC, the second reference-based method, selects informative genes based on cross-subject and cross-cell consistency and employs a weighted non-negative least squares (*nnls*) regression model to estimate the cell-type proportions $\mathbf{P}$ [28].

Reference-free deconvolution algorithms are particularly useful when reference datasets are unavailable. In this case, both matrices $\mathbf{C}$ and $\mathbf{P}$ are unknown and must be estimated simultaneously. Computational methods such as LINear Subspace identification for gene Expression Deconvolution (Linseed) [36] and geometric structure-constrained non-negative matrix factorization (GSNMF) [37] have been developed to address this challenge.

Ideally, the expression of cell-type-specific marker genes should be linearly correlated with the corresponding cell-type proportions. Reference-free methods leverage this mutual linearity to uncover the topological structure of bulk RNA-seq data. Linseed, for instance, employs the SISAL algorithm [68] to identify an underlying simplex structure that characterizes cell-type-specific genes and proportions. In contrast, GSNMF improves the identifiability and uniqueness of NMF-based deconvolution by integrating marker gene information as solvability constraints and using a correlation graph for manifold regularization. By incorporating both solvability and manifold constraints, GSNMF aims to enhance the accuracy and robustness of cell proportion estimation.

3.3 Simulation Design

Cell-Level Datasets for Pseudo-Bulk Tissue Generation

Four cell-level expression datasets from different tissue types were used to generate pseudo-bulk tissues in our simulations (Table A.1).

- **GSE19830**: Contains 42 samples, including nine pure rat brain, liver, and lung cell samples, along with 33 mixtures of these three cell types in various proportions [64]. Only the nine pure cell samples were used in our simulations to generate pseudo-bulk tissue samples.

- **Human Leukocyte Subsets (LM22)**: Comprises 22 leukocyte subsets from mature human hematopoietic populations derived from peripheral blood or cultured *in vitro* [20]. This dataset includes detailed cell-type annotations distinguishing subtypes such as naive and memory B cells, CD8 and CD4 T cells (naive, memory resting, memory activated), follicular helper T cells, Tregs, and gamma delta T cells. In this study, we grouped all subtypes into six distinct cell types: B cells, T cells, NK cells, macrophages, dendritic cells, and mast cells.

- **GSE81608**: Contains scRNA-seq data from human pancreatic islet cells ob-

tained from twelve non-diabetic and six type 2 diabetic (T2D) samples [69]. This dataset includes 1,492 cells categorized into pancreatic α, β, δ, and γ cells.

- **GSE67835**: Consists of 466 single cells from healthy human brain samples [70]. It includes nine distinct cell types: astrocytes, endothelial cells, replicating and quiescent fetal neurons, hybrid cells, microglia, regular neurons, oligodendrocytes, and oligodendrocyte precursor cells (OPCs). For our study, we selected five major types—astrocytes, endothelial cells, microglia, neurons, and oligodendrocytes—totaling 267 cells, to generate pseudo-bulk brain tissue RNA-seq data.

Additionally, three more datasets—PBMC8K, E-MTAB-5061, and syn18485175—are listed in Table A.1. These datasets were used either as independent external reference data for deconvolution or to generate bulk tissue RNA-seq data.

Specifically, PBMC8K consists of 8,000 peripheral blood mononuclear cells (PBMCs) from a healthy donor, obtained through scRNA-seq from patient blood samples (https://support.10xgenomics.com/single-cell-gene-expression/datasets). The dataset was sequenced using the Illumina HiSeq 4000, generating approximately 92,000 reads per cell, with a total of 8,381 detected cells. The raw scRNA-seq reads were aligned to the GRCh38 reference genome and quantified using Cell Ranger [71] (10x Genomics). The resulting expression matrix was then processed with Seurat [72] for cell annotation. To annotate PBMC8K, we used Seurat and incorporated prior cell type knowledge from PBMC3K. We identified and annotated five distinct cell types—B cells, T cells, NK cells, macrophages, and dendritic cells—by grouping related subpopulations to construct in silico bulk tissues. We then used the PBMC8K dataset [73] to generate pseudo whole blood tissue samples and employed the LM22 dataset as the external reference.

For the simulated human brain bulk tissue samples, we used the syn18485175 dataset, a single-nucleus RNA-seq (snRNA-seq) dataset of the human brain [65],

as the reference. For the simulated type 2 diabetes (T2D) bulk tissue samples, we used the E-MTAB-5061 dataset [74] as the reference.

In total, 12 pseudo-bulk tissue RNA-seq datasets (four different tissue types × three different levels of variation in cell composition) were generated and used to evaluate the performance of deconvolution methods.

Resilience Analysis: Controlled Modifications to Generating Datasets

To comprehensively analyze resilience, we systematically introduced controlled modifications to the generating datasets. Specifically, we altered the cell-level data to generate pseudo-bulk tissues while using the unmodified original data as the reference for deconvolution analysis. The variation in cell composition was fixed at a medium level during resilience testing. We applied three distinct manipulative techniques:

- **Mean Shifting**: Expression levels were augmented by adding 10%, 30%, 50%, and 70% of the average expression level for each cell type.
- **Truncation**: The top or bottom 10% of expressed cells for each cell type were selectively removed before generating bulk tissue data. The remaining cells were then used for pseudo-bulk tissue generation.
- **Factoring**: Expression levels were scaled using factors of 0.4, 0.8, 1.2, and 1.8 to allow both upward and downward adjustments.

Additionally, we used the scDesign simulator by Li et al. (2019) [66] to generate synthetic scRNA-seq datasets. This simulator estimates the distribution parameters of an assumed probabilistic model based on real scRNA-seq data, enabling the generation of realistic synthetic single-cell expression profiles. In our experiments, we used scDesign to create synthetic datasets for human peripheral blood, brain, and pancreatic tissue cells. The resulting synthetic scRNA-seq data were used to generate pseudo-bulk tissue samples, while the original single-cell datasets served as the reference for deconvolution. In contrast to brain and pancreatic tissues, human

peripheral blood required a different approach: we used the PBMC8k scRNA-seq dataset as input to the scDesign simulator and employed the LM22 dataset as the external reference.

Cell Proportion Generation

We simulated cell proportions in pseudo-bulk tissues using a multivariate Dirichlet distribution with parameter vector $\mathbf{p} = (p_1, p_2, \ldots, p_K)$, where $\gamma_k = p_k / \sum_{k=1}^{K} p_k$ represents the expected proportion of the k^{th} cell type. The sum $p_0 = \sum_{k=1}^{K} p_k$ controls variation in cell composition, with larger values corresponding to smaller variations. Our simulations included three variation levels —small, medium, and large —as detailed in Table A.2.

The expected cell type proportions for different tissue types were set as follows:

- **Brain pseudo-bulk tissues**: $\gamma_{brain} = (0.23, 0.07, 0.06, 0.49, 0.14)$ for astrocytes, endothelial cells, microglia, neurons, and oligodendrocytes.
- **Blood pseudo-bulk tissues**: $\gamma_{blood} = (0.17, 0.43, 0.09, 0.14, 0.09, 0.09)$ for B cells, T cells, NK cells, macrophages, dendritic cells, and mast cells.
- **Pancreatic pseudo-bulk tissues**: $\gamma_{islet} = (0.51, 0.33, 0.1, 0.05)$ for α, β, γ, and δ cells.

Ideal Mixing for Pseudo-Bulk Tissue RNA-seq Data

After simulating cell proportions, single cells were randomly sampled with replacement from the corresponding scRNA-seq dataset. The read counts or expression levels of the selected single-cell RNA profiles were directly summed to generate the total gene expression levels for each pseudo-bulk tissue sample. This method is referred to as the *ideal mixing approach*.

A total of 50 pseudo-bulk mixture samples were generated for each dataset, except for the T2D dataset, for which 100 mixtures were created —50 for the healthy group and 50 for the T2D group. For all simulated bulk tissue samples, genes with fewer

than 10 read counts in at least 80% of the samples were discarded, followed by column normalization. For human brain tissues, the top 1,000 marker genes for each of the five cell types, as identified by McKenzie et al. (2018) [75], were used—resulting in a total of 5,000 genes for brain tissue analysis.

Evaluation Metrics

The deconvolution methods were evaluated using three quantitative metrics:

1. Pearson's Correlation Coefficient: $R = \text{Cor}(p, \hat{p})$
2. Root Mean Squared Deviation (RMSD): $RMSD = \sqrt{\text{avg}(p - \hat{p})^2}$
3. Mean Absolute Deviation (MAD): $MAD = \text{avg}(|p - \hat{p}|)$

where p and $\hat{p}$ represent the ground truth and estimated cell proportions, respectively.

3.4 Results

Through extensive simulations, we gained valuable insights into the performance of various deconvolution methods, particularly in terms of robustness and resilience.

3.4.1 Robustness Results Analysis

Overall, the two reference-based methods, CIBERSORTx and MuSiC, consistently outperformed the two reference-free methods, Linseed and GSNMF, across various tissue types and levels of cell composition variation. This was particularly evident when the reference dataset used was the same dataset used to generate the pseudo-bulk tissues.

CIBERSORTx and MuSiC demonstrated comparable performance across different levels of variation in cell composition, with Pearson's correlation coefficients between estimated and true cell proportions approaching 1, as shown in Figure A.1, A.2, A.3 and A.4. When comparing performance across different bulk tissue types, estimates for pancreatic and brain tissues were slightly less accurate than those for liver, lung,

brain pseudo-bulk tissues, and whole blood pseudo-bulk samples. For example, under medium variation in cell proportions, the observed RMSD for MuSiC was 0.00023, 0.00712, 0.00162, and 0.0472 for liver, lung, brain, whole blood, pancreatic, and brain tissue samples, respectively. The corresponding RMSD for CIBERSORTx was 0.0847, 0.00009, 0.0275, and 0.1037. A similar pattern was observed in Pearson's correlation coefficients and MAD (Figure A.1, A.2, A.3 and A.4).

Linseed and GSNMF struggled to accurately estimate cell proportions when variation in cell composition was minimal. However, their performance improved significantly as variation increased. For instance, in pseudo-bulk tissues for liver, lung, and brain, Pearson's correlation coefficients between estimated cell proportions and ground truth for Linseed were 0.557, 0.760, and 0.751, respectively, under small, medium, and large variation levels (third row in Figure A.1). Similarly, for GSNMF, Pearson's correlation coefficients were 0.126, 0.361, and 0.890 under the same conditions (fourth row in Figure A.1). Compared to GSNMF, Linseed provided more accurate estimates when variation in cell proportions was small to medium, particularly for brain, liver, lung mixtures, and LM22 mixtures. However, as variation increased, GSNMF outperformed Linseed. For example, in pseudo-pancreatic tissues, GSNMF achieved Pearson's correlation coefficients of 0.948 and 0.884 at medium and large variation levels, compared to Linseed's 0.873 and 0.535, respectively(Figure A.3). In pseudo-brain tissues, GSNMF demonstrated performance comparable to Linseed (Figure A.4).

In summary, when pseudo-bulk tissues were generated under ideal mixing conditions, and the same dataset was used as a reference, the two reference-based methods exhibited robust performance in estimating cell proportions across various tissue types and levels of variation in cell composition. Two primary sources of variation in the simulated bulk tissue samples were found to influence method performance: variation in single-cell gene expression profiles and variation in cell compositions.

Variation in Single-Cell Gene Expression Profiles

The four scRNA-seq datasets used in this study differed in the number of cells and tissue types. The liver, lung, and brain dataset contained only nine single cells and exhibited the most distinct gene expression patterns among the three cell types, as illustrated in Figure A.5. The signature genes selected by CIBERSORTx displayed highly distinct expression patterns in liver, lung, and brain cells. Additionally, due to the limited number of cells, random sampling could result in the same cell being selected multiple times, leading to reduced sampling variation compared to other datasets.

The LM22 dataset exhibited similar characteristics. Although LM22 contained more cells, it had been carefully pre-processed and extensively cleaned, resulting in clearer cell-type-specific gene expression patterns, as seen in the heatmap (Figure A.5, b)). In contrast, the T2D and brain cell datasets showed greater variability in gene expression, including less distinct expression patterns for marker genes, increased variability within cell types, and a larger number of cells, contributing to greater overall variation at the cell level.

These findings highlight how factors such as the number of cells, tissue types, and the biological techniques used for cell-level gene expression measurement can significantly affect variation in cell-level data, thereby influencing the clarity and consistency of deconvolution results.

Variation in Cell Proportions

In pseudo-bulk tissues generated under ideal mixing conditions, reference-based methods remained largely unaffected by variations in cell composition, consistently outperforming reference-free methods in terms of Pearson's correlation coefficient, RMSD, and MAD. However, the degree of variation in cell composition had a significant impact on the performance of reference-free methods, with greater variation correlating with more accurate estimations.

This trend can likely be attributed to two key factors:

- Identification of Marker Genes - Reference-free methods assume a linear relationship between cell proportions and the corresponding marker gene expression levels. When cell proportions vary widely, the linearity in expression levels becomes more pronounced, making marker genes easier to identify.
- Mathematical Identifiability - The identifiability condition is better satisfied when deconvolution leverages correctly identified marker genes, leading to more precise results.

In summary, greater variation in cell composition enhances the accuracy of reference-free methods, while reference-based methods remain minimally affected and consistently perform well regardless of variation.

3.4.2 Resilience Results Analysis

We also evaluated the performance of these deconvolution methods under conditions where discrepancies exist between the reference data and the true profiles of individual cells in bulk tissue samples. Specifically, we examined three levels of discrepancy:

1. Artificially manipulated scRNA-seq data used for bulk tissue generation.
2. Synthetic scRNA-seq data generated by a simulator for bulk tissue generation.
3. Three independent real cell-level datasets used separately for bulk tissue generation and as a reference.

The following sections summarize our findings on the resilience of these methods in handling such discrepancies.

Artificial Manipulation by Mean Shifting, Truncation, and Factoring

Our results indicate that mean shifting and factoring have minimal impact on the performance of the four selected deconvolution methods. MuSiC, CIBERSORTx, and GSNMF maintained Pearson correlation coefficients similar to those observed in the unaltered scenario across different tissue types and varying levels of cell composition. However, Linseed exhibited some variability in estimating cell proportions in pseudo-bulk pancreatic tissues. Specifically, its Pearson correlation coefficients were 0.4401, 0.8033, 0.6483, and 0.7422 when 10%, 30%, 50%, and 70% of the mean values were added, respectively (third row in Figure A.12), compared to 0.8725 in the unaltered case (Figure A.3). Interestingly, both Linseed and GSNMF showed improved performance when deconvoluting pseudo-bulk brain tissues with higher percentages of cell-type-specific mean expression values added. Our designed factoring experiments did not result in any significant changes in the performance of the four deconvolution methods (Figures A.7, A.10, A.13, A.16).

Removing the top or bottom 10% of cells based on their overall expression level had a slight impact on the performance of reference-based methods, with Pearson's correlation coefficients remaining close to 1. However, this truncation significantly improved the performance of GSNMF, suggesting its sensitivity to outlier cells with extreme expression levels. Specifically, GSNMF's RMSD values for the four different original pseudo-bulk tissue types were 0.303, 0.230, 0.107, and 0.119. After removing the top 10% of highly expressed cells, these values changed to 0.302, 0.068, 0.103, and 0.106, respectively. Similarly, removing the bottom 10% of cells resulted in RMSD values of 0.301, 0.205, 0.101, and 0.107. (fourth row in Figure A.8,A.11,A.14,A.17)

Linseed displayed a different response to truncation: its performance improved in pseudo whole blood and brain tissues but worsened in pseudo pancreatic tissues, with little change observed in pseudo liver, lung, and brain mixtures. This suggests that Linseed's performance may be tissue-dependent.(third row in Figure A.8,A.11,A.14,A.17)

Impact of Using Synthetic scRNA-seq Data for Bulk Tissue Generation

When artificially simulated scRNA-seq data were used to generate pseudo-bulk tissues while a real scRNA-seq dataset served as the reference, the performance of reference-based methods declined significantly (Figure A.18,A.19,A.20). This effect was particularly evident for CIBERSORTx, where Pearson's correlation coefficients in both types of simulated bulk tissue samples were substantially lower than when the same scRNA-seq dataset was used for both bulk tissue generation and reference.

In contrast, the performance of reference-free methods improved across all levels of variation in cell composition. For example, in pseudo-brain tissue samples, GSNMF achieved Pearson correlation coefficients of 0.6025, 0.9743, and 0.9859 under small, medium, and large variation levels, respectively. Similarly, for pseudo-pancreatic samples, its Pearson correlation coefficients were 0.9695, 0.9884, and 0.9986 across the three variation levels. For pseudo-bulk blood samples, the coefficients were 0.9274, 0.9898, and 0.8649, respectively. Comparable patterns were observed in RMSD and MAD metrics (fourth row in Figure A.18,A.19,A.20).

This improvement can be attributed to the reduced variability of synthetic scRNA-seq data generated by a simulator, as it follows an assumed probability distribution. Conversely, reference-based methods performed worse because this scenario violated a fundamental assumption of regression-based models: that cell-type-specific expression levels are known and observed with minimal error. The substantial drop in CIBERSORTx's Pearson's correlation coefficients, along with increased RMSD and MAD values, further supports this conclusion. Reference-free methods demonstrated greater resilience as they do not rely on external reference datasets.

Impact of Using an Independent Reference Dataset

To further validate these findings, we conducted three additional analyses using real, external, and independent reference datasets for deconvolution.

When we used the PBMC8K dataset to generate pseudo whole blood tissue samples

and employed the LM22 dataset as the external reference, the reference-free methods demonstrated superior performance. Linseed achieved Pearson correlation coefficients of 0.9793, 0.9920, and 0.9972; GSNMF obtained 0.9163, 0.8100, and 0.9005; MuSiC reported 0.8631, 0.8832, and 0.8739; and CIBERSORTx yielded 0.8807, 0.8713, and 0.8149, respectively (Figure A.21).

In bulk pancreatic tissues, MuSiC maintained strong performance across all levels of cell composition variation, with Pearson correlation coefficients of 0.9650, 0.9880, and 0.9990. However, CIBERSORTx exhibited slightly weaker performance compared to the ideally mixed pseudo-bulk tissues discussed earlier, with coefficients of 0.3850, 0.6870, and 0.9650, respectively. As expected, the performance of reference-free methods remained stable, as they do not depend on external reference datasets (Figure A.22).

Furthermore, when pseudo-brain tissue samples were deconvoluted using an external snRNA-seq dataset, the performance of both reference-based methods declined. MuSiCâs Pearson correlation coefficients dropped to -0.1193, 0.3469, and 0.6523 across increasing levels of variation. In contrast, GSNMF achieved coefficients of 0.0431, 0.8894, and 0.8712. Similarly, CIBERSORTx produced 0.5680, 0.6250, and 0.7490, while Linseed yielded 0.1346, 0.7639, and 0.8777, respectively (Figure A.23). Similar patterns were observed in the RMSD and MAD bar plots (Figures A.21, A.22, A.23).

These results suggest that reference-free methods may offer more reliable estimates than reference-based methods when no high-quality reference dataset is available. This advantage becomes especially evident when cell composition variability reaches medium levels or higher.

3.5 Discussion

In this study, we utilized real scRNA-seq datasets to generate in silico pseudo-bulk RNA-seq data and evaluated four computational deconvolution methods: CIBER-

SORTx, MuSiC, Linseed, and GSNMF. These methods were assessed based on their ability to estimate cell proportions using Pearson's correlation coefficient, RMSD, and MAD as evaluation metrics. Our primary objective was to evaluate their robustness across various tissue types and levels of cell composition variation, as well as their resilience when discrepancies existed between the reference dataset and the true transcriptome of cells in bulk tissues.

To examine robustness, we conducted simulations using scRNA-seq datasets from four different tissue types and generated bulk tissue samples with three levels of cell composition variation. To evaluate resilience, we introduced discrepancies by manipulating scRNA-seq data, using simulator-generated synthetic scRNA-seq data, and incorporating independent external scRNA-seq/snRNA-seq datasets as references for pseudo-bulk tissues.

Robustness of Reference-Based and Reference-Free Methods

Both reference-based methods (CIBERSORTx and MuSiC) demonstrated strong robustness across different tissue types and levels of cell composition variation. Across all evaluation metrics, these methods consistently outperformed reference-free approaches, particularly when pseudo-bulk tissues were generated under ideal mixing conditions and true scRNA-seq profiles were used as references. CIBERSORTx and MuSiC produced highly accurate estimates, showing strong correlation with the true cell proportions and lower RMSD and MAD values. Their performance was comparable in our simulations.

In contrast, the reference-free methods (Linseed and GSNMF) showed variable performance depending on tissue type and the level of cell composition variation. Linseed performed best in pseudo-whole-blood tissues, achieving Pearson's correlation coefficients of 0.922, 0.987, and 0.943 at low, medium, and high variation levels, respectively. However, its performance in other tissue types, particularly under low variation conditions, was less satisfactory. GSNMF, on the other hand, performed

best when cell composition variation was high. The differences in performance between these two reference-free methods were largely due to their ability to identify reliable marker genes. In general, variation in cell composition was beneficial for marker gene identification, whereas variation in gene expression levels did not contribute to improved performance. This pattern was consistently observed in RMSD and MAD results as well.

Resilience to Discrepancies Between Reference Data and Bulk Transcriptomes

When discrepancies existed between the reference dataset and the actual single-cell transcriptomes in bulk tissues, reference-based methods were significantly affected. These methods rely on the fundamental assumption of regression models that cell-type-specific gene expression profiles are known and observed with minimal error. When this assumption was violated, their performance declined.

In contrast, reference-free methods, which learn the underlying structure of bulk tissues without depending on external reference datasets, demonstrated greater resilience to such discrepancies. However, they still required some degree of cell proportion variation for optimal performance. GSNMF, as a more mathematical approach, did not explicitly account for scRNA-seq variability when selecting marker genes, whereas Linseed incorporated a permutation-based approach, resulting in smaller errors. Similarly, CIBERSORTx, which employs a support vector regression model, performed better in noisier datasets, while MuSiC, which is based on a regression model, was more effective when data variability was lower.

These findings were further validated in simulations where pseudo-bulk tissue data were generated using one scRNA-seq dataset while an independent external scRNA-seq dataset served as the reference. The results confirmed that reference-free methods exhibited better resilience in these scenarios, particularly when variation in cell composition was at least moderate.

Implications and Future Directions

In summary, we conducted a comprehensive evaluation of two reference-based and two reference-free methods for deconvoluting bulk RNA-seq data. We assessed their performance across different tissue types, cell composition variation levels, and scenarios where discrepancies existed between reference datasets and actual single-cell transcriptomes. Our findings indicate that reference-based methods generally outperform reference-free methods, particularly when the same single-cell profiles used for deconvolution are also used to generate pseudo-bulk tissues. However, when discrepancies between datasets were introduced, reference-free methods sometimes provided more reliable estimates than reference-based approaches. These insights serve as practical guidelines for selecting appropriate deconvolution methods in real-world applications.

We also observed that variation in scRNA-seq profiles and cell composition plays a crucial role in improving and developing more effective computational deconvolution methods. These factors should be carefully considered when designing future deconvolution algorithms.

Study Limitations

Despite its comprehensive scope, this study has some limitations. Although we explored a wide range of scenarios, our simulations were not exhaustive. The tissue types examined were limited to four, whereas real-world applications involve a much broader spectrum. Additionally, there are more than four available deconvolution algorithms, some of which may exhibit different performance characteristics.

Furthermore, our evaluation relied on three commonly used metrics—Pearson's correlation coefficient, RMSD, and MAD—without critically assessing whether these are the most optimal for this type of analysis. Future research could explore alternative evaluation metrics that may provide additional insights into method performance.

Finally, while we introduced large variation in cell composition in our simulations

to investigate its impact on deconvolution, such levels of variation are often unrealistic in real bulk tissue samples. Nevertheless, understanding how variation affects deconvolution methods remains valuable for future algorithm development. Further research should explore ways to leverage variation in a biologically meaningful manner to enhance deconvolution accuracy.

CHAPTER 4: CONCLUSIONS AND FUTURE WORK

In conclusion, we addressed the fundamental challenge of disentangling cell-type composition from bulk RNA sequencing (RNA-seq) data—a problem central to many modern transcriptomic studies. Traditional bulk RNA-seq approaches provide valuable, yet averaged, signals across heterogeneous tissues, thereby obscuring the contributions of individual cell types and confounding downstream analyses.

In Chapter 2, we first highlighted the non-identifiability issues inherent in classical nonnegative matrix factorization (NMF) when reconstructing both cell-type-specific gene expression profiles and their relative proportions. Formal conditions for solution uniqueness, along with geometric interpretations, helped clarify why NMF is often ill-posed in complex biological datasets. Building on recent advances in NMF and leveraging the availability of single-cell data, we introduced theoretical and practical innovations aimed at improving deconvolution methodologies. We proposed a new model—GSNMF —that incorporates biologically plausible priors such as marker-gene knowledge and manifold constraints to improve interpretability and robustness. The use of geometric structure, in particular, provided a valuable framework for enforcing partial identifiability when standard NMF assumptions are relaxed. Expanding on this, the dissertation introduced an enhanced pipeline, GSNMF+, which augments limited bulk data with statistically simulated pseudo-bulk data derived from single-cell RNA-seq profiles. This augmentation helps mitigate "singularity" scenarios where the original bulk dataset alone may not satisfy critical identifiability conditions. By systematically combining pseudo-bulk augmentation with informed initialization and a final rescaling correction, the proposed method achieves more stable and accurate deconvolution across a range of realistic tissue compositions. These methods signifi-

cantly enhance the identifiability and reliability of cell-type-specific gene expression profiles and relative abundances, particularly in complex or sparse tissue contexts.

In Chapter 3, we presented a comprehensive benchmarking study comparing both reference-based and reference-free deconvolution methods across multiple datasets. Using evaluation metrics such as Pearson correlation (R), root mean squared error (RMSE), and mean absolute deviation (MAD), we found that while high-quality reference datasets can substantially improve performance in some scenarios, carefully constrained and augmented reference-free methods often excel in cases where reliable references are unavailable. Furthermore, structured evaluations revealed how variations in data—such as mean shifting, scaling expression values, or gene subsampling—can significantly impact deconvolution accuracy. These results underscore the importance of integrating prior biological knowledge, geometric constraints, and flexible algorithmic strategies to address real-world data variability. Taken together, these contributions advance the field of computational deconvolution in several meaningful ways. By clarifying theoretical foundations and offering empirically validated improvements, this work provides a robust platform for analyzing heterogeneous tissues. Improved estimates of cell-type proportions offer researchers deeper insights into disease progression, immune responses, and tissue-specific regulatory mechanisms. Moreover, the developed methods and evaluations serve as a blueprint for future studies aiming to integrate multi-omic data, adopt spatial transcriptomics, and refine deconvolution techniques under realistic noise conditions.

Based on the findings from this dissertation, we are currently developing a new deconvolution algorithm based on neural networks to address more complex scenarios in future work. This forthcoming method aims to extend the capabilities of current approaches by leveraging deep learning architectures to capture nonlinear relationships, handle higher levels of noise, and experimental conditions.

REFERENCES

[1] D. Jovic, X. Liang, H. Zeng, L. Lin, F. Xu, and Y. Luo, "Single-cell rna sequencing technologies and applications: A brief overview," *Clinical and translational medicine*, vol. 12, no. 3, p. e694, 2022.

[2] M. Slyper, C. B. Porter, O. Ashenberg, J. Waldman, E. Drokhlyansky, I. Wakiro, C. Smillie, G. Smith-Rosario, J. Wu, D. Dionne, *et al.*, "A single-cell and single-nucleus rna-seq toolbox for fresh and frozen human tumors," *Nature medicine*, vol. 26, no. 5, pp. 792–802, 2020.

[3] G. J. Sutton, D. Poppe, R. K. Simmons, K. Walsh, U. Nawaz, R. Lister, J. A. Gagnon-Bartsch, and I. Voineagu, "Comprehensive evaluation of deconvolution methods for human brain gene expression," *Nature Communications*, vol. 13, no. 1, p. 1358, 2022.

[4] L. Crews and E. Masliah, "Molecular mechanisms of neurodegeneration in Alzheimer's disease," *Hum Mol Genet.*, vol. 19, no. R1, pp. R12–20, 2010.

[5] T. Malm, T. Jay, and G. Landreth, "The evolving biology of microglia in Alzheimer's disease," *Neurotherapeutics*, vol. 12, no. 1, pp. 81–89, 2015.

[6] D. Tejera and M. Heneka, "Microglia in Alzheimer's disease: the good, the bad and the ugly," *Curr Alzheimer Res.*, vol. 13, no. 4, pp. 370–80, 2016.

[7] C. Chen, C. Cohrs, J. Stertmann, R. Bozsak, and S. Speier, "Human beta cell mass and function in diabetes: Recent advances in knowledge and technologies to understand disease pathogenesis," *Mol Metab.*, vol. 6, pp. 943–957, Jul 2017.

[8] B. Jew, M. Alvarez, E. Rahmani, Z. Miao, A. Ko, K. M. Garske, J. H. Sul, K. H. Pietiläinen, P. Pajukanta, and E. Halperin, "Accurate estimation of cell composition in bulk expression through robust integration of single-cell information," *Nature communications*, vol. 11, no. 1, p. 1971, 2020.

[9] Juliana M. SÃ , Matthew V. Cannon, Ramoncito L. Caleon, Thomas E. Wellems, David Serre, "Single-cell transcription analysis of Plasmodium vivax blood-stage parasites identifies stage- and species-specific profiles of expression," *PLOS Biology*, vol. 18(5), May 2020.

[10] M. S. Pedersen, U. Kjems, K. B. Rasmussen, and L. K. Hansen, "Semi-blind source separation using head-related transfer functions [speech signal separation]," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. V–713, IEEE, 2004.

[11] M. Yu, W. Ma, J. Xin, and S. Osher, "Multi-channel $l\_\{1\}$ regularized convex speech enhancement model and fast computation by the split bregman method," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 2, pp. 661–675, 2011.

[12] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, 2014.

[13] N. Souviraà-Labastie, A. Olivero, E. Vincent, and F. Bimbot, "Multi-channel audio source separation using multiple deformed references," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1775–1787, 2015.

[14] N. Gillis, "Successive nonnegative projection algorithm for robust nonnegative blind source separation," *SIAM Journal on Imaging Sciences*, vol. 7, no. 2, pp. 1420–1450, 2014.

[15] W.-K. Ma, J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, and T.-H. H. Chan, "A signal processing perspective on hyperspectral unmixing: Insights from remote sensing," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 67–81, 2014.

[16] D. Nuzillard and A. Bijaoui, "Blind source separation and analysis of multispectral astronomical images," *Astronomy and Astrophysics Supplement Series*, vol. 147, no. 1, pp. 129–138, 2000.

[17] V. P. Pauca, J. Piper, and R. J. Plemmons, "Nonnegative matrix factorization for spectral data analysis," *Linear algebra and its applications*, vol. 416, no. 1, pp. 29–47, 2006.

[18] E. Villeneuve and H. Carfantan, "Hyperspectral data deconvolution for galaxy kinematics with mcmc," in *Proc. 20th Eur. Signal Process. Conf.*, pp. 2477–2481, 2012.

[19] A. Kuhn, A. Kumar, A. Beilina, A. Dillman, M. R. Cookson, and A. B. Singleton, "Cell population-specific expression analysis of human cerebellum," *BMC genomics*, vol. 13, pp. 1–15, 2012.

[20] A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, and A. A. Alizadeh, "Robust enumeration of cell subsets from tissue expression profiles," *Nature methods*, vol. 12, no. 5, pp. 453–457, 2015.

[21] T. Gong and J. D. Szustakowski, "Deconrnaseq: a statistical framework for deconvolution of heterogeneous tissue samples based on mrna-seq data," *Bioinformatics*, vol. 29, no. 8, pp. 1083–1085, 2013.

[22] Z. Altboum, Y. Steuerman, E. David, Z. Barnett-Itzhaki, L. Valadarsky, H. Keren-Shaul, T. Meningher, E. Mendelson, M. Mandelboim, I. Gat-Viks, *et al.*, "Digital cell quantification identifies global immune cell dynamics during influenza infection," *Molecular systems biology*, vol. 10, no. 2, p. 720, 2014.

[23] M. Baron, A. Veres, S. L. Wolock, A. L. Faust, R. Gaujoux, A. Vetere, J. H. Ryu, B. K. Wagner, S. S. Shen-Orr, A. M. Klein, *et al.*, "A single-cell transcriptomic

map of the human and mouse pancreas reveals inter-and intra-cell population structure," *Cell systems*, vol. 3, no. 4, pp. 346–360, 2016.

[24] B. Li, T. Li, J. S. Liu, and X. S. Liu, "Computational deconvolution of tumor-infiltrating immune components with bulk tumor gene expression data," *Bioinformatics for Cancer Immunotherapy: Methods and Protocols*, pp. 249–262, 2020.

[25] J. Racle, K. de Jonge, P. Baumgaertner, D. E. Speiser, and D. Gfeller, "Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data," *elife*, vol. 6, p. e26476, 2017.

[26] A. M. Newman, C. Steen, C. L. Liu, A. J. Gentles, A. A. Chaudhuri, F. Scherer, M. S. Khodadoust, M. S. Esfahani, B. A. Luca, D. Steiner, M. Diehn, and A. A. Alizadeh, "Determining cell type abundance and expression from bulk tissues with digital cytometry," *Nature Biotechnology*, vol. 37, no. 7, pp. 773–782, 2019.

[27] G. J. Hunt, S. Freytag, M. Bahlo, and J. A. Gagnon-Bartsch, "Dtangle: accurate and robust cell type deconvolution," *Bioinformatics*, vol. 35, no. 12, pp. 2093–2099, 2019.

[28] X. Wang, J. Park, K. Susztak, N. R. Zhang, and M. Li, "Bulk tissue cell type deconvolution with multi-subject single-cell expression reference," *Nature communications*, vol. 10, no. 1, p. 380, 2019.

[29] T. Chu, Z. Wang, D. Peâer, and C. G. Danko, "Cell type and gene expression deconvolution with bayesprism enables bayesian integrative analysis across bulk and single-cell rna sequencing in oncology," *Nature Cancer*, vol. 3, no. 4, pp. 505–517, 2022.

[30] Y. Lin, H. Li, X. Xiao, L. Zhang, K. Wang, J. Zhao, M. Wang, F. Zheng, M. Zhang, W. Yang, *et al.*, "Daism-dnnxmbd: Highly accurate cell type proportion estimation with in silico data augmentation and deep neural networks," *Patterns*, vol. 3, no. 3, 2022.

[31] L. A. Newberg, X. Chen, C. D. Kodira, and M. I. Zavodszky, "Computational de novo discovery of distinguishing genes for biological processes and cell types in complex tissues," *PloS one*, vol. 13, no. 3, p. e0193067, 2018.

[32] Y. Zhong, Y.-W. Wan, K. Pang, L. M. Chow, and Z. Liu, "Digital sorting of complex tissues for cell type-specific gene expression profiles," *BMC Bioinformatics*, vol. 14, no. 89, 2013.

[33] Z. Li and H. Wu, "Toast: improving reference-free cell composition estimation by cross-cell type differential analysis," *Genome biology*, vol. 20, pp. 1–17, 2019.

[34] Z. Li, Z. Guo, Y. Cheng, P. Jin, and H. Wu, "Robust partial reference-free cell composition estimation from tissue expression," *Bioinformatics*, vol. 36, no. 11, pp. 3431–3438, 2020.

[35] K. Kang, Q. Meng, I. Shats, D. M. Umbach, M. Li, Y. Li, X. Li, and L. Li, "Cdseq: A novel complete deconvolution method for dissecting heterogeneous samples using gene expression data," *PLoS computational biology*, vol. 15, no. 12, p. e1007510, 2019.

[36] K. Zaitsev, M. Bambouskova, A. Swain, and M. N. Artyomov, "Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures," *Nature Communications*, vol. 10, no. 1, p. 2209, 2019.

[37] D. Chen, S. Li, and X. Wang, "Geometric structure guided model and algorithms for complete deconvolution of gene expression data," *Foundations of Data Science*, vol. 4, pp. 441–466, Sep 2022.

[38] D. Aran, Z. Hu, and A. J. Butte, "xcell: digitally portraying the tissue cellular heterogeneity landscape," *Genome biology*, vol. 18, pp. 1–14, 2017.

[39] K. Yoshihara, M. Shahmoradgoli, E. Martínez, R. Vegesna, H. Kim, W. Torres-Garcia, V. Treviño, H. Shen, P. W. Laird, D. A. Levine, *et al.*, "Inferring tumour purity and stromal and immune cell admixture from expression data," *Nature communications*, vol. 4, no. 1, p. 2612, 2013.

[40] D. Repsilber, S. Kern, A. Telaar, G. Walzl, G. F. Black, J. Selbig, S. K. Parida, S. H. Kaufmann, and M. Jacobsen, "Biomarker discovery in heterogeneous tissue samples-taking the in-silico deconfounding approach," *BMC bioinformatics*, vol. 11, pp. 1–15, 2010.

[41] M. D. Craig, "Minimum-volume transforms for remotely sensed data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 3, pp. 542–552, 1994.

[42] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

[43] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[44] W.-K. Ma, J. M. Bioucas-Dias, T.-H. Chan, N. Gillis, P. Gader, A. J. Plaza, A. Ambikapathi, and C.-Y. Chi, "A signal processing perspective on hyperspectral unmixing: Insights from remote sensing," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 67–81, 2013.

[45] X. Fu, W.-K. Ma, T.-H. Chan, and J. M. Bioucas-Dias, "Self-dictionary sparse regression for hyperspectral unmixing: Greedy pursuit and pure pixel search are related," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 6, pp. 1128–1141, 2015.

[46] S. Zhang, W. Wang, J. Ford, and F. Makedon, "Learning from incomplete ratings using non-negative matrix factorization," in *Proceedings of the 2006 SIAM international conference on data mining*, pp. 549–553, SIAM, 2006.

[47] S. Mohammadi, N. Zuckerman, A. Goldsmith, and A. Grama, "A critical survey of deconvolution methods for separating cell types in complex tissues," *Proceedings of the IEEE*, vol. 105, no. 2, pp. 340–366, 2016.

[48] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, 2000.

[49] H. Kim and H. Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," *SIAM journal on matrix analysis and applications*, vol. 30, no. 2, pp. 713–730, 2008.

[50] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[51] J. Eckstein and W. Yao, "Augmented lagrangian and alternating direction methods for convex optimization: A tutorial and some illustrative computational results," *RUTCOR Research Reports*, vol. 32, no. 3, p. 44, 2012.

[52] Y.-X. Wang and Y.-J. Zhang, "Nonnegative matrix factorization: A comprehensive review," *IEEE Transactions on knowledge and data engineering*, vol. 25, no. 6, pp. 1336–1353, 2012.

[53] X. Fu, K. Huang, N. D. Sidiropoulos, and W.-K. Ma, "Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications.," *IEEE Signal Process. Mag.*, vol. 36, no. 2, pp. 59–80, 2019.

[54] N. Gillis *et al.*, "Nonnegative matrix factorization: Complexity, algorithms and applications," *Unpublished doctoral dissertation, Université catholique de Louvain. Louvain-La-Neuve: CORE*, 2011.

[55] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?," *Advances in neural information processing systems*, vol. 16, 2003.

[56] K. Huang, N. D. Sidiropoulos, and A. Swami, "Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition," *IEEE Transactions on Signal Processing*, vol. 62, no. 1, pp. 211–224, 2013.

[57] H. Laurberg, M. G. Christensen, M. D. Plumbley, L. K. Hansen, and S. H. Jensen, "Theorems on positive data: On the uniqueness of nmf," *Computational intelligence and neuroscience*, vol. 2008, no. 1, p. 764206, 2008.

[58] A. Cichocki, R. Zdunek, A. H. Phan, and S. i. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. John Wiley & Sons, 2009.

[59] R. Warren and S. Osher, "Hyperspectral unmixing by the alternating direction method of multipliers," *Inverse Problems and Imaging*, vol. 14, no. 3, pp. 00–00, 2015.

[60] J. Nocedal and S. J. Wright, *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering, Springer Science Business Media, 2 ed., 2006.

[61] W. Wang and M. A. Carreira-Perpinán, "Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application," *arXiv preprint arXiv:1309.1541*, 2013.

[62] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, pp. 395–416, 2007.

[63] J. Qin, H. Lee, J. T. Chi, Y. Lou, J. Chanussot, and A. L. Bertozzi, "Fast blind hyperspectral unmixing based on graph laplacian," in *2019 10th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pp. 1–5, IEEE, 2019.

[64] S. S. Shen-Orr, R. Tibshirani, P. Khatri, D. L. Bodian, F. Staedtler, N. M. Perry, T. Hastie, M. M. Sarwal, M. M. Davis, and A. J. Butte, "Cell type–specific gene expression differences in complex tissues," *Nature methods*, vol. 7, no. 4, pp. 287–289, 2010.

[65] H. Mathys, J. Davila-Velderrain, Z. Peng, F. Gao, S. Mohammadi, J. Z. Young, M. Menon, L. He, F. Abdurrob, X. Jiang, *et al.*, "Single-cell transcriptomic analysis of alzheimerâs disease," *Nature*, vol. 570, no. 7761, pp. 332–337, 2019.

[66] W. V. Li and J. J. Li, "A statistical simulator scdesign for rational scrna-seq experimental design," *Bioinformatics*, vol. 35, no. 14, pp. i41–i50, 2019.

[67] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural computation*, vol. 12, no. 5, pp. 1207–1245, 2000.

[68] J. M. Bioucas-Dias, "A variable splitting augmented lagrangian approach to linear spectral unmixing," in *2009 First workshop on hyperspectral image and signal processing: Evolution in remote sensing*, pp. 1–4, IEEE, 2009.

[69] Y. Xin, J. Kim, H. Okamoto, M. Ni, Y. Wei, C. Adler, A. J. Murphy, G. D. Yancopoulos, C. Lin, and J. Gromada, "Rna sequencing of single human islet cells reveals type 2 diabetes genes," *Cell metabolism*, vol. 24, no. 4, pp. 608–615, 2016.

[70] S. Darmanis, S. A. Sloan, Y. Zhang, M. Enge, C. Caneda, L. M. Shuer, M. G. Hayden Gephart, B. A. Barres, and S. R. Quake, "A survey of human brain transcriptome diversity at the single cell level," *Proceedings of the National Academy of Sciences*, vol. 112, no. 23, pp. 7285–7290, 2015.

[71] G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, *et al.*, "Massively parallel digital transcriptional profiling of single cells," *Nature communications*, vol. 8, no. 1, p. 14049, 2017.

[72] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija, "Integrating single-cell transcriptomic data across different conditions, technologies, and species," *Nature biotechnology*, vol. 36, no. 5, pp. 411–420, 2018.

[73] 10x Genomics, "Datasets - 10x genomics," 2022. Available at: https://www.10xgenomics.com/resources/datasets/.

[74] Å. Segerstolpe, A. Palasantza, P. Eliasson, E.-M. Andersson, A.-C. Andréasson, X. Sun, S. Picelli, A. Sabirsh, M. Clausen, M. K. Bjursell, *et al.*, "Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes," *Cell metabolism*, vol. 24, no. 4, pp. 593–607, 2016.

[75] A. T. McKenzie, M. Wang, M. E. Hauberg, J. F. Fullard, A. Kozlenkov, A. Keenan, Y. L. Hurd, S. Dracheva, P. Casaccia, P. Roussos, *et al.*, "Brain cell type specific gene expression and co-expression network architectures," *Scientific reports*, vol. 8, no. 1, p. 8868, 2018.

APPENDIX A: FIGURES AND TABLES IN CHAPTER 3

Table A.1: Cell level gene expression datasets.

Data	GSE19830	LM22	GSE81608	GSE67835
Tissue Type	rat brain, liver, lung	human blood	human pancreas	human brain
Platform	Affymetrix	Affymetrix	Illumina	Illumina
Num. of genes	13,841	11,845	39,849	22,088
Num. of cells	9	22	1,492	267
Num. of cell types	3	6	4	5
Num. of subjects	NA	NA	18*	NA
Reference	Shen-Orr *et al.* (2010)	Newman *et al.* (2015)	Xin *et al.* (2016)	Darmanis *et al.* (2015)
Data	PBMC8K	E-MTAB-5061	syn18485175	
Tissue Type	human peripheral blood	human pancreas	human brain	
Platform	Illumina	Smart-seq2	Illumina	
Num. of genes	32,738	25,453	17,926	
Num. of cells	8,381	2,209	75,060	
Num. of cell types	10	4	5	
Num. of subjects	NA	10**	48	
Reference	10X Genomics website [73]	Segerstolpe *et al.* (2016)	Mathys *et al.* (2019)	

* 12 healthy and 6 T2D samples.

** 6 healthy and 4 T2D samples.

Table A.2: Values of parameter $\mathbf{p}$ for the Multivariate Dirichlet distribution used for simulations.

Data Set	Small Variation	Medium Variation	Large Variation
GSE19830	(3.3, 3.4, 3.3)	(3.3, 3.4, 3.3)/10	(3.3, 3.4, 3.3)/100
LM22	(1, 2.5, 0.5, 0.8, 0.5, 0.5)	(1, 2.5, 0.5, 0.8, 0.5, 0.5)/10	(1, 2.5, 0.5, 0.8, 0.5, 0.5)/100
GSE81608	(5.4, 3.5, 1.1, 0.5)	(5.4, 3.5, 1.1, 0.5)/10	(5.4, 3.5, 1.1, 0.5)/100
GSE67835	(6.2, 2, 1.6, 13.1, 3.8)	(6.2, 2, 1.6, 13.1, 3.8)/10	(6.2, 2, 1.6, 13.1, 3.8)/100

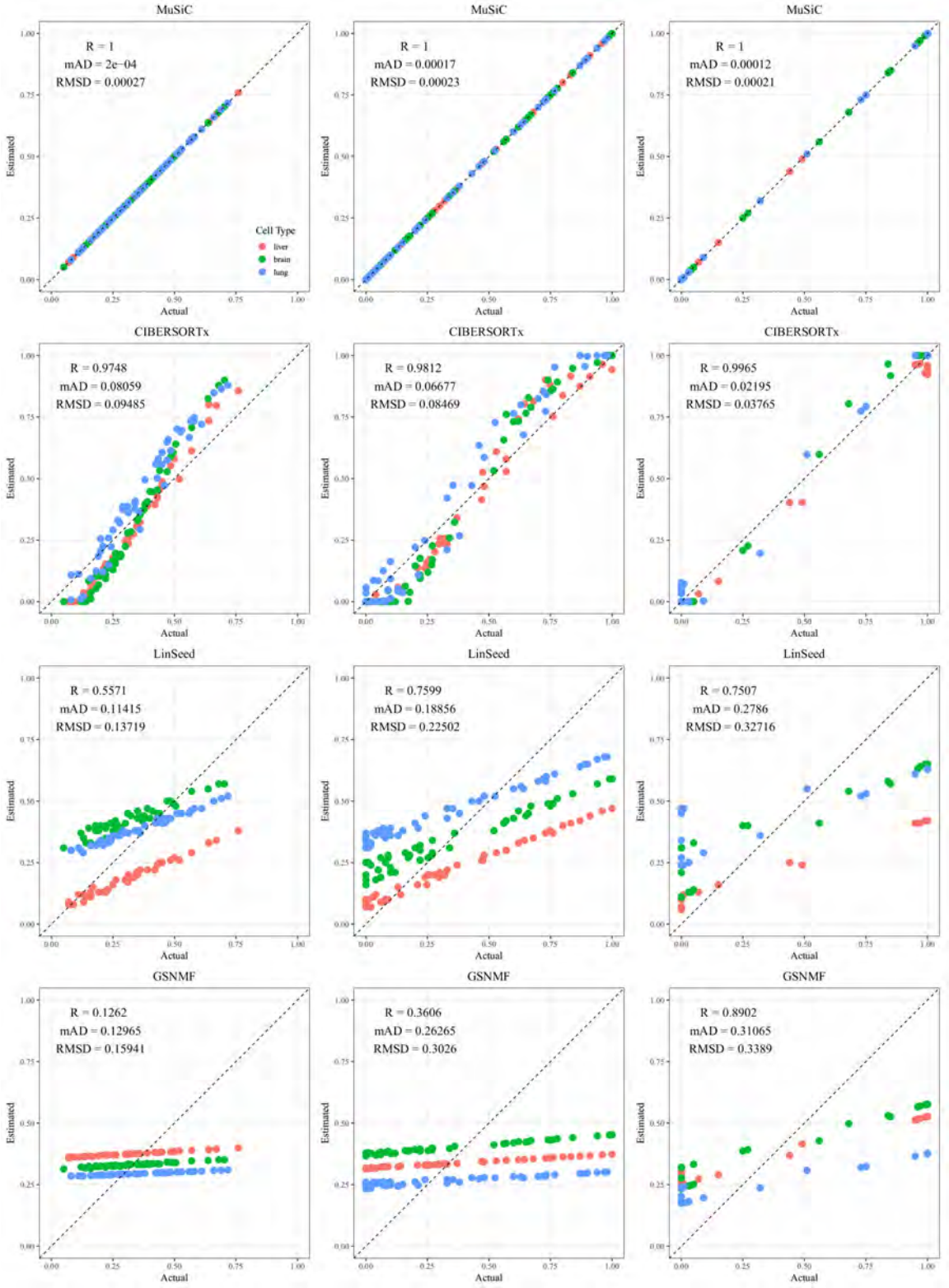


Figure A.1: Comparison of actual and estimated proportions for dataset GSE19830. From left to right, the columns represent small, medium, and large variation. Each row corresponds to a different deconvolution method: MuSiC, CIBERSORTx, LinSeed, and GSNMF. Different colors represent different cell types: liver (red), brain (green), and lung (blue).

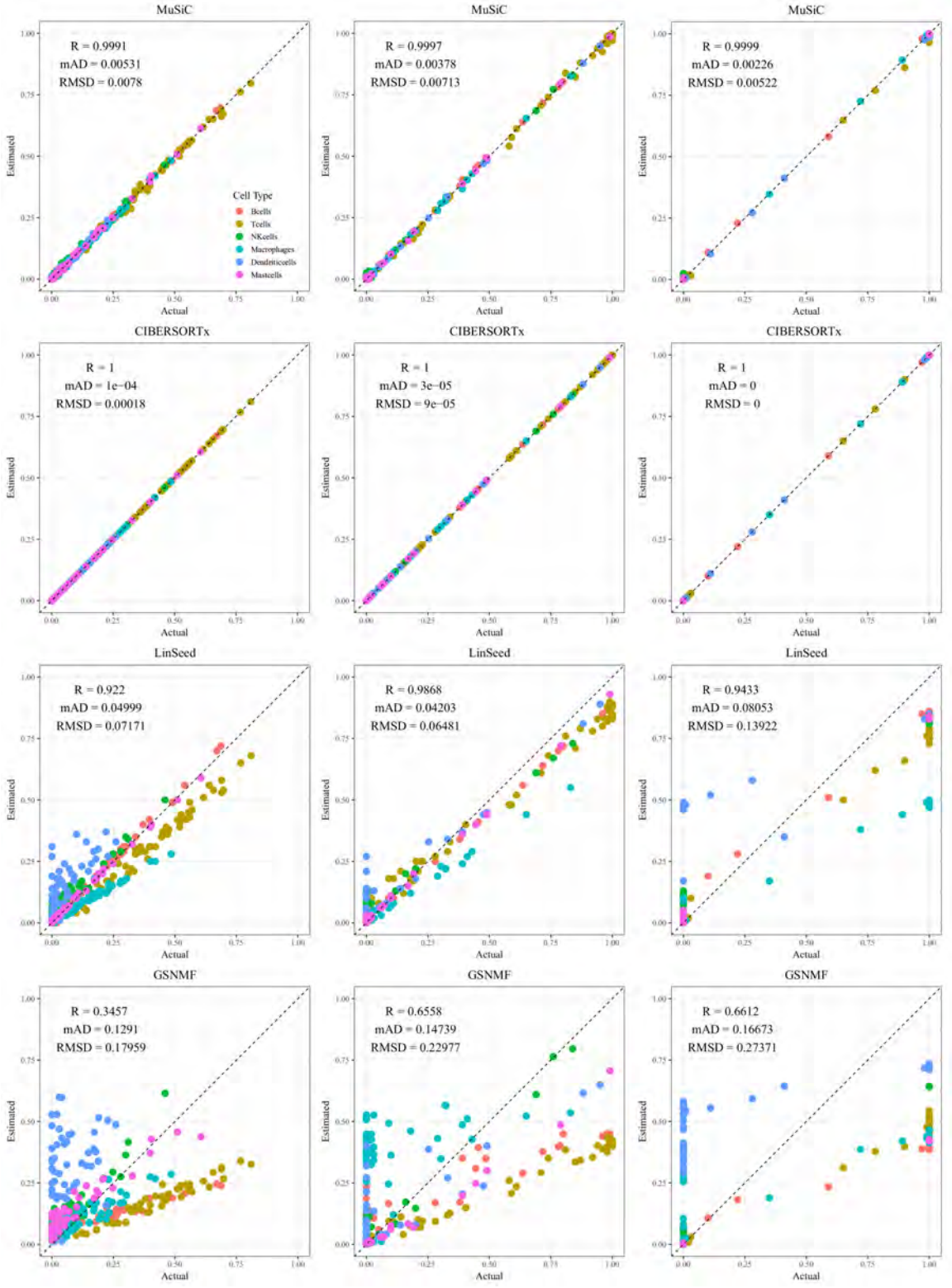


Figure A.2: Comparison of actual and estimated proportions for dataset LM22. From left to right, the columns represent small, medium, and large variation. Each row corresponds to a different deconvolution method: MuSiC, CIBERSORTx, LinSeed, and GSNMF. Different colors represent different cell types: B cells (red), T cells (gold), NK cells (green), Macrophages (cyan), Dendritic cells (blue), and Mast cells (magenta).

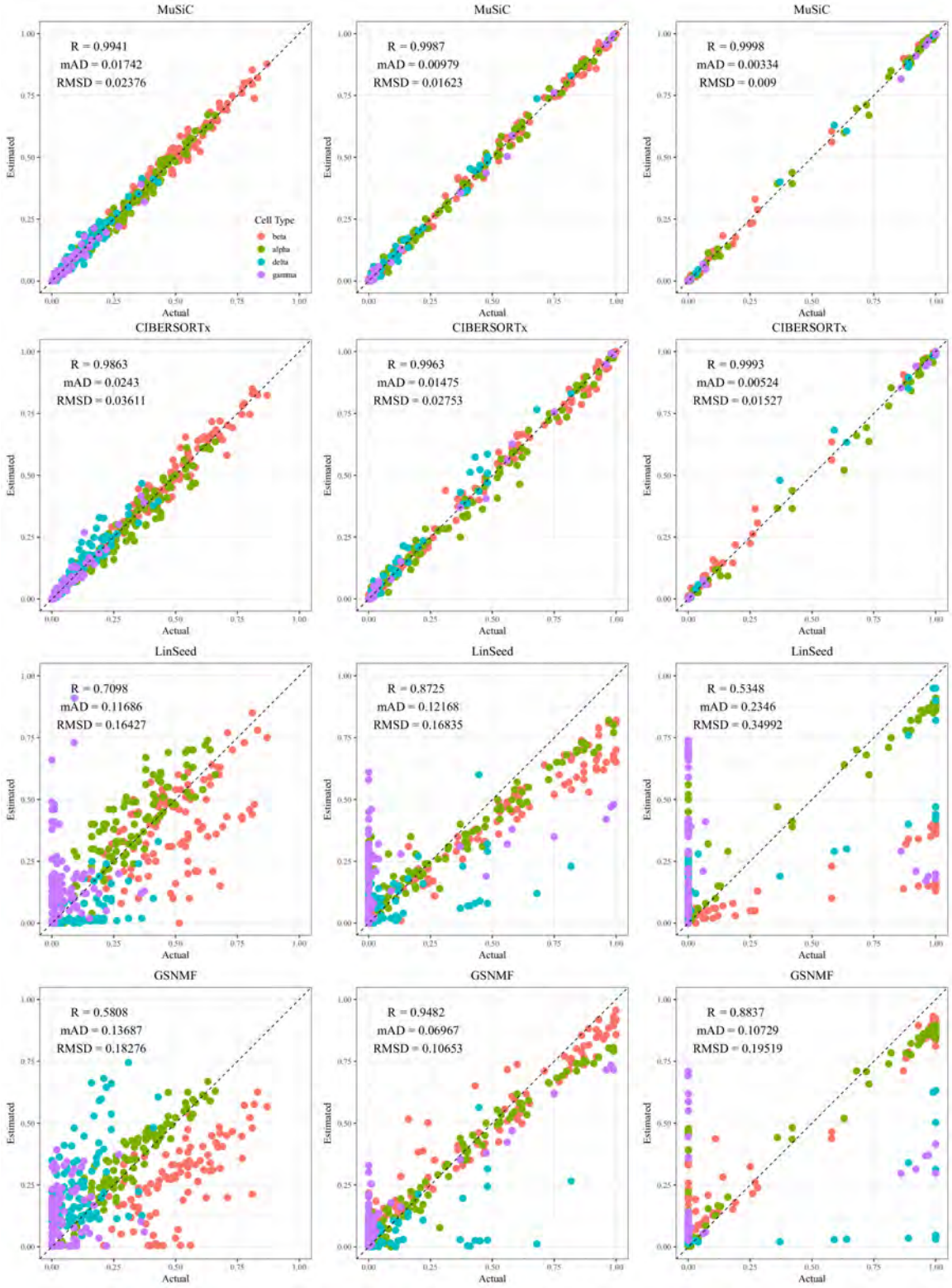


Figure A.3: Comparison of actual and estimated proportions for dataset GSE81608. From left to right, the columns represent small, medium, and large variation. Each row corresponds to a different deconvolution method: MuSiC, CIBERSORTx, LinSeed, and GSNMF. Different colors represent different cell types: beta (red), alpha (green), delta (cyan), and gamma (purple).

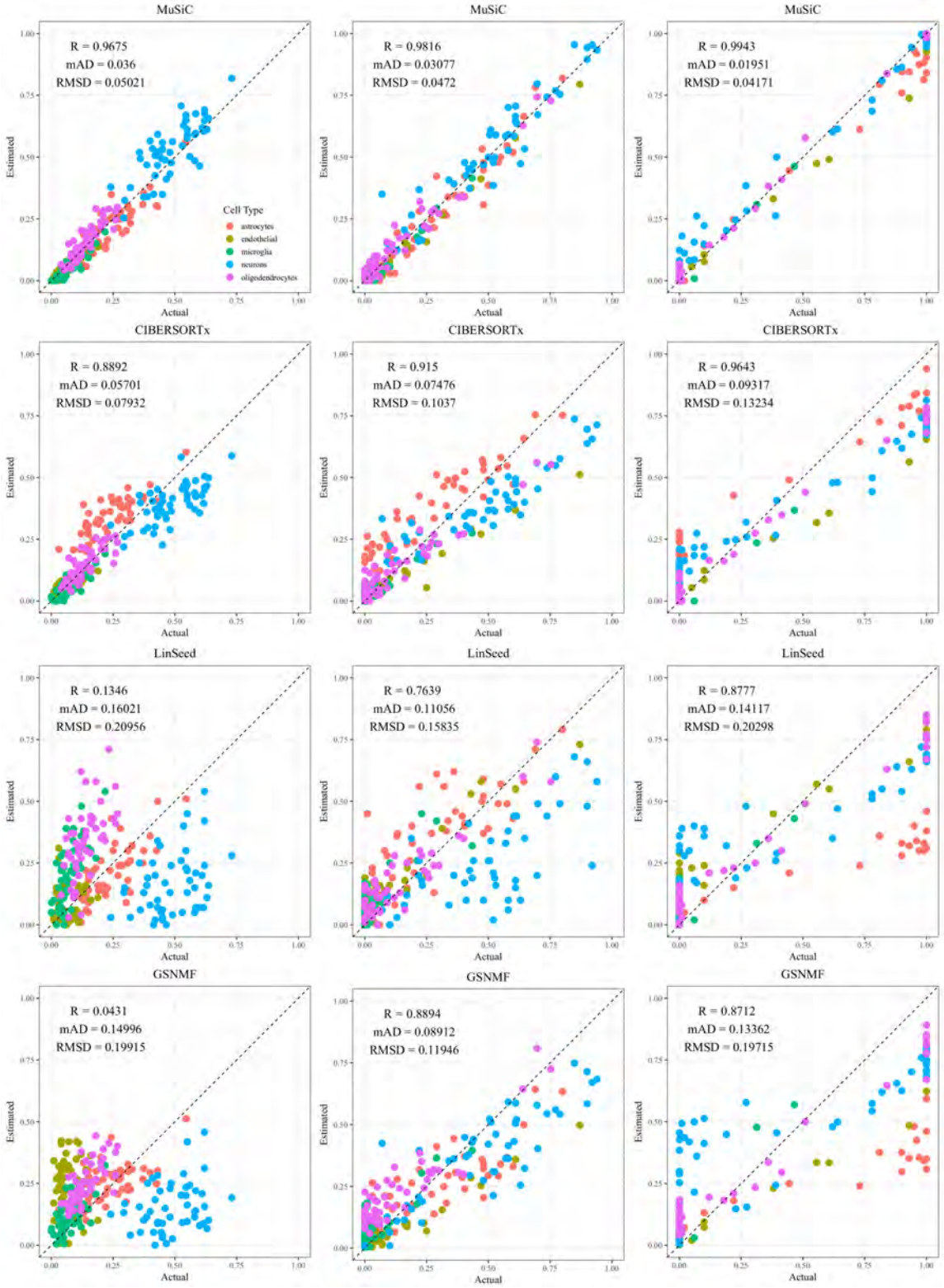


Figure A.4: Comparison of actual and estimated proportions for dataset GSE67835. From left to right, the columns represent small, medium, and large variation. Each row corresponds to a different deconvolution method: MuSiC, CIBERSORTx, LinSeed, and GSNMF. Different colors represent different cell types: astrocytes (red), endothelial (gold), microglia (green), neurons (blue), and oligodendrocytes (magenta).

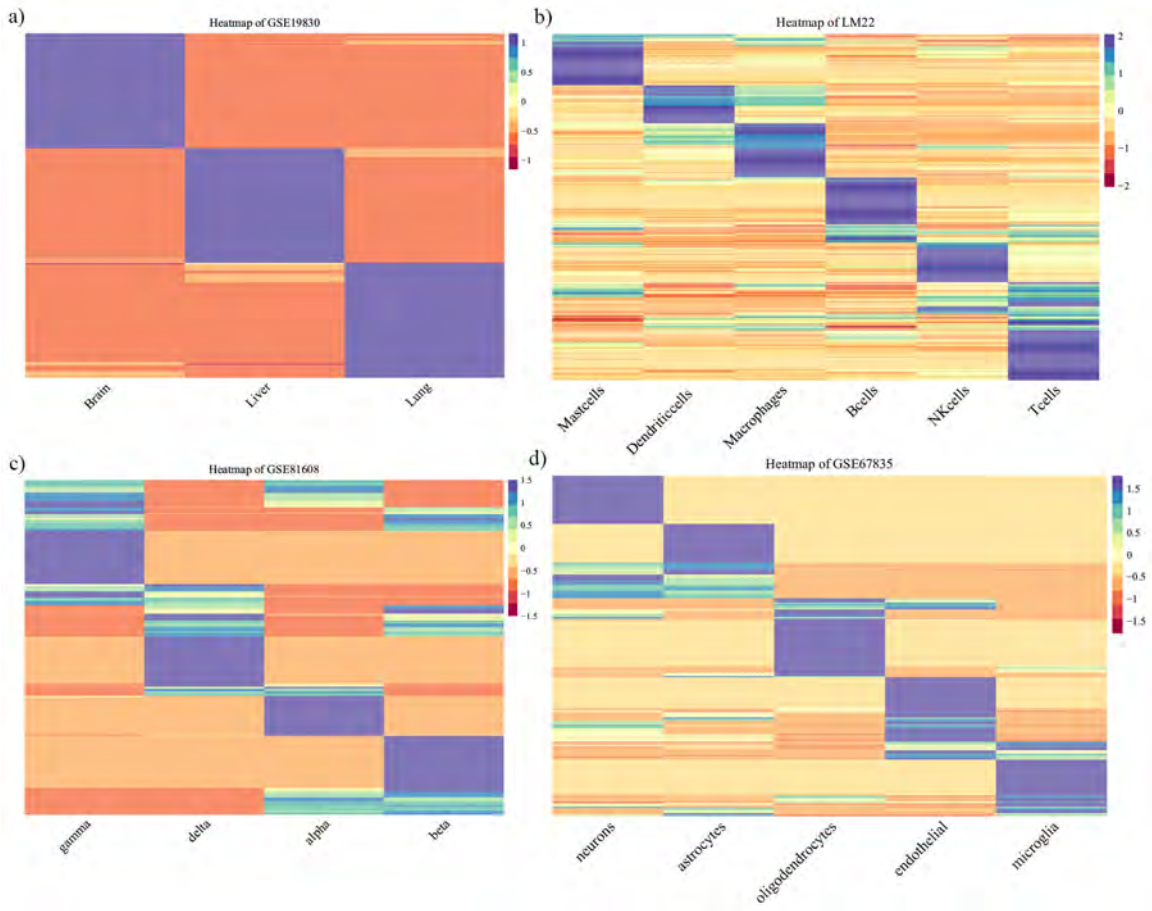


Figure A.5: Heatmap of correlation between signature genes selected by CIBERSORTx: a) GSE19830, b) LM22, c) GSE81608, and d) GSE67835.

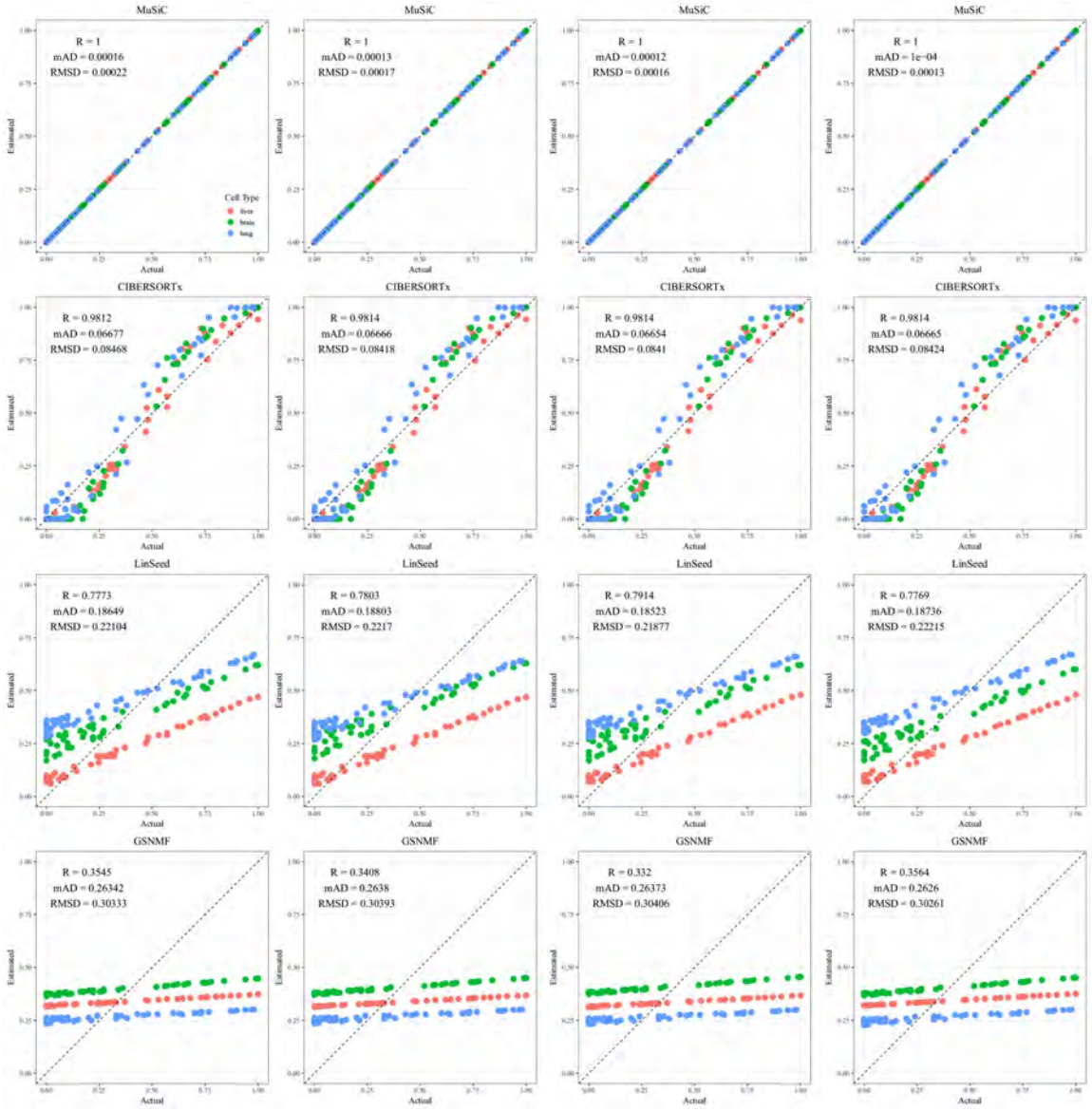


Figure A.6: Comparison of actual and estimated cell-type proportions for dataset GSE19830 under mean shifting conditions. The columns, from left to right, represent 10%, 30%, 50%, and 70% shifts. Each row corresponds to a different deconvolution method: MUSIC, CIBERSORTx, Linseed, and GSNMF. Different colors indicate distinct cell types: liver (red), brain (green), and lung (blue).

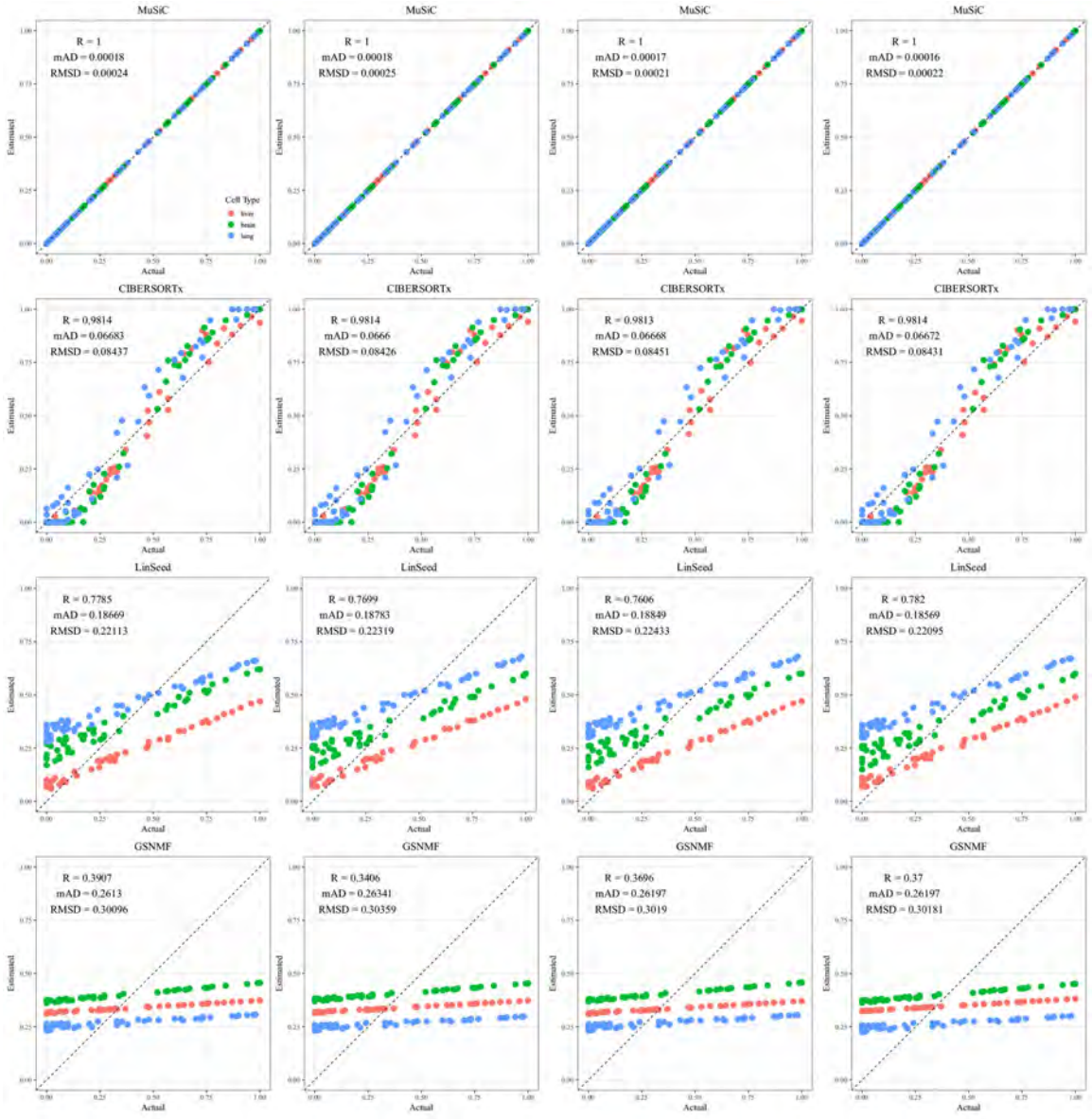


Figure A.7: Comparison of actual and estimated cell-type proportions for dataset GSE19830 under factoring scenarios. The columns, from left to right, represent factoring values of 1.2, 1.8, 0.8, and 0.4. Each row corresponds to a different deconvolution method: MuSiC, CIBERSORTx, LinSeed, and GSNMF. Different colors indicate distinct cell types: liver (red), brain (green), and lung (blue).

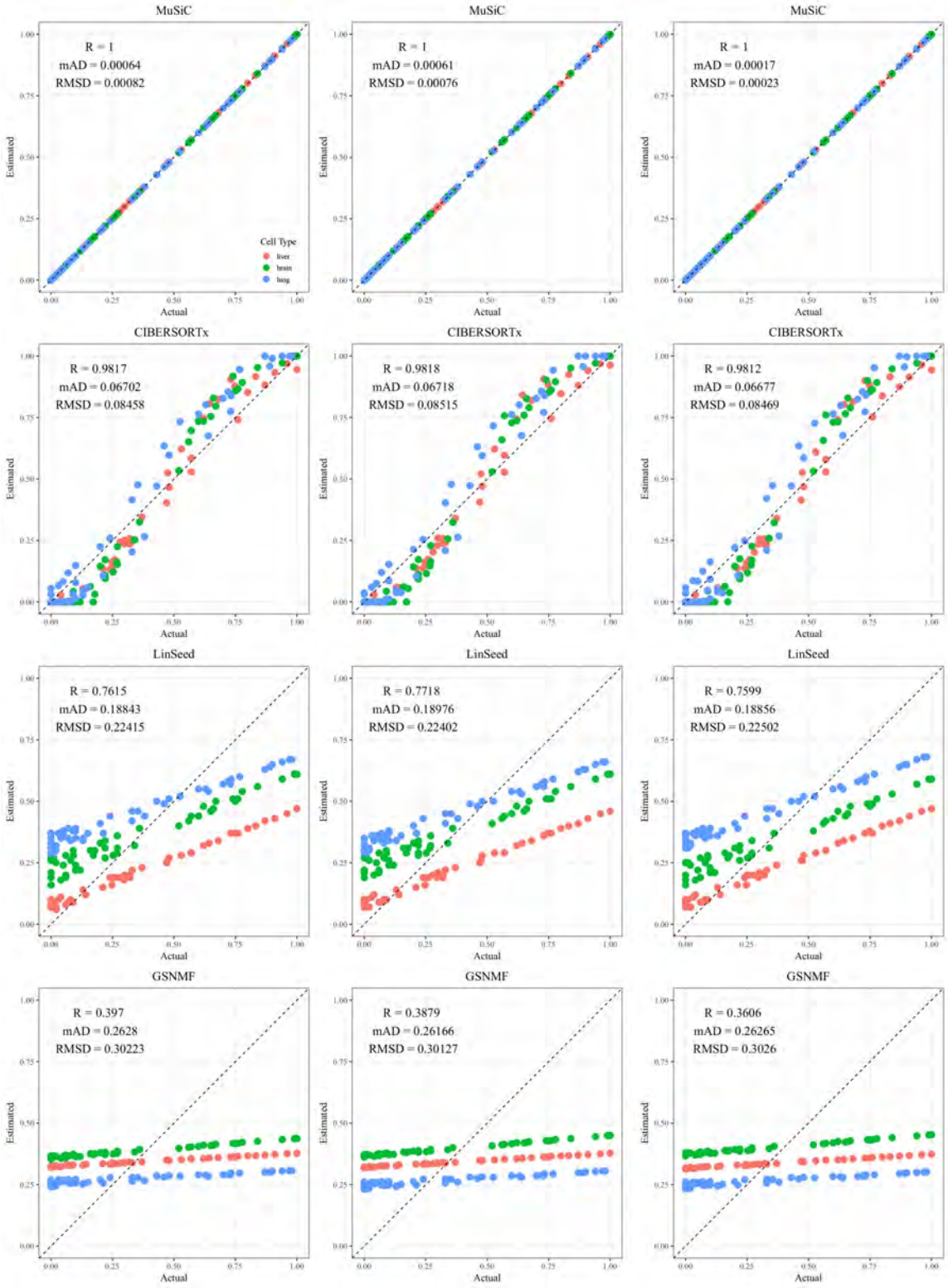


Figure A.8: Comparison of actual and estimated cell-type proportions for dataset GSE19830 under different truncation scenarios. The columns, from left to right, represent truncation of the top 10%, truncation of the bottom 10%, and the original dataset. Each row corresponds to a different deconvolution method: MuSiC, CIBERSORTx, LinSeed, and GSNMF. Distinct colors represent different cell types: liver (red), brain (green), and lung (blue).

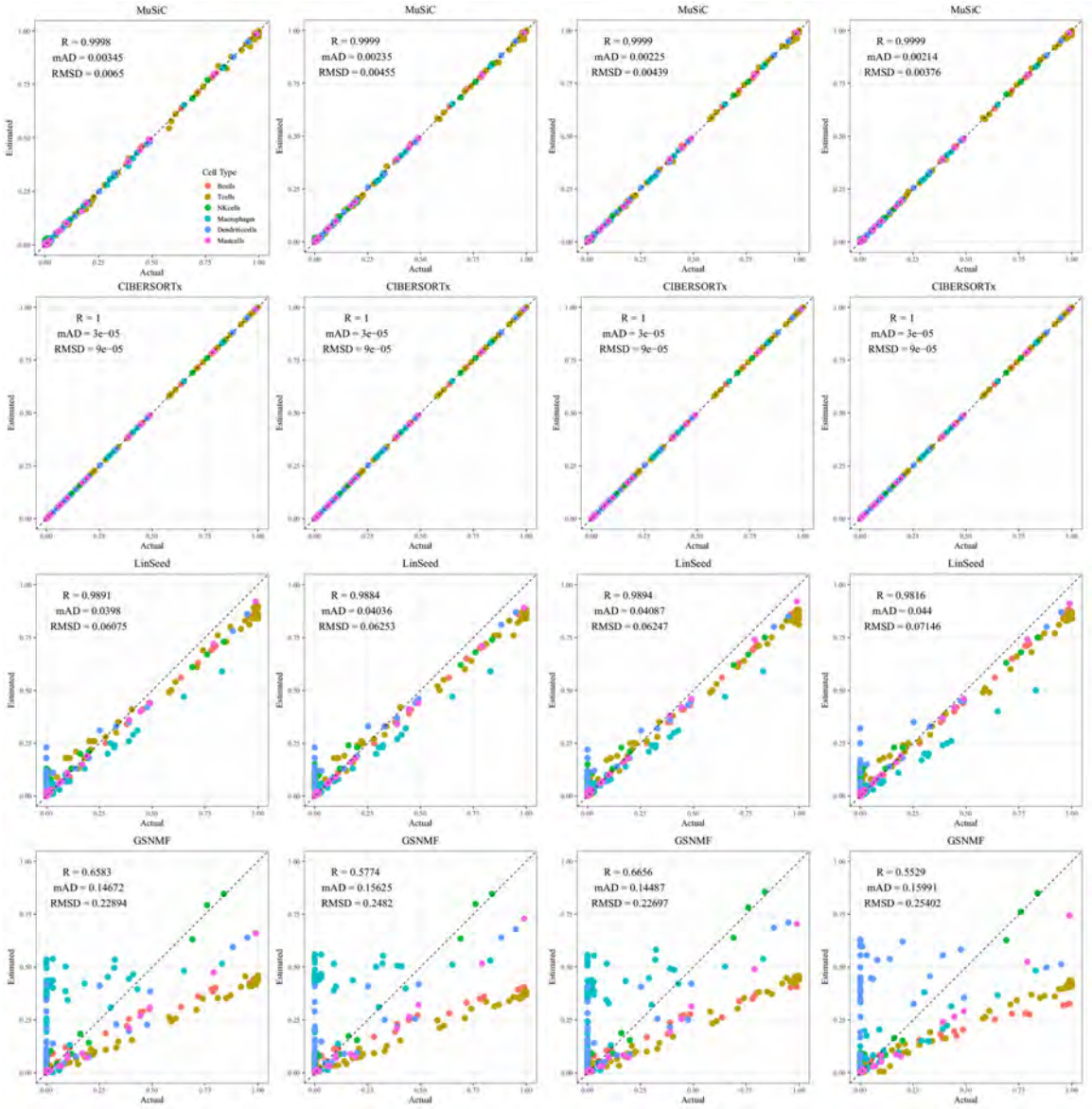


Figure A.9: Comparison of actual and estimated cell-type proportions for dataset LM22 under mean shifting conditions. The columns, from left to right, represent 10%, 30%, 50%, and 70% shifts. Each row corresponds to a different deconvolution method: MUSIC, CIBERSORTx, Linseed, and GSNMF. Different colors represent different cell types: B cells (red), T cells (gold), NK cells (green), Macrophages (cyan), Dendritic cells (blue), and Mast cells (magenta).

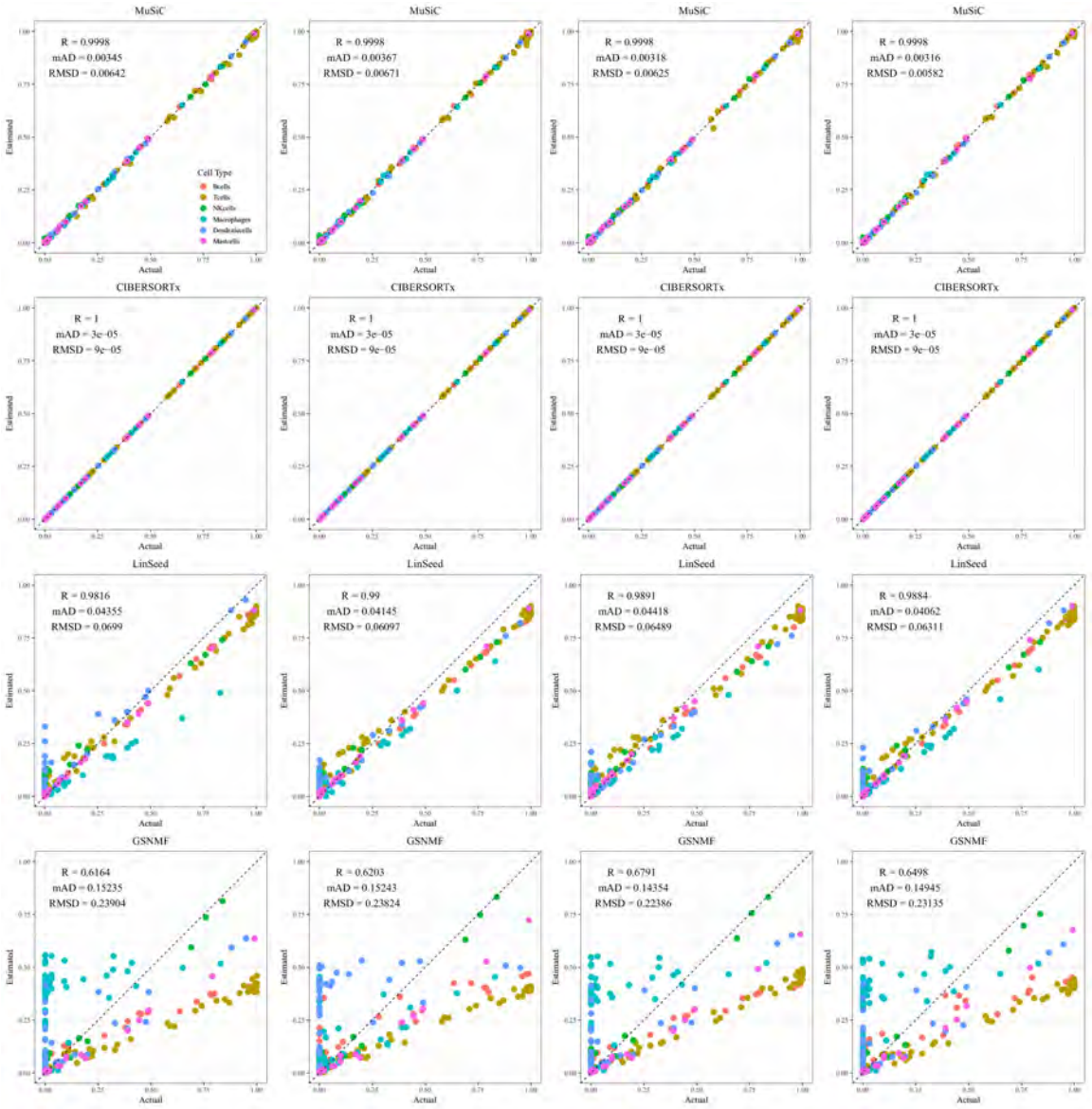


Figure A.10: Comparison of actual and estimated cell-type proportions for dataset LM22 under factoring scenarios. The columns, from left to right, represent factoring values of 1.2, 1.8, 0.8, and 0.4. Each row corresponds to a different deconvolution method: MuSiC, CIBERSORTx, LinSeed, and GSNMF. Different colors represent different cell types: B cells (red), T cells (gold), NK cells (green), Macrophages (cyan), Dendritic cells (blue), and Mast cells (magenta)).

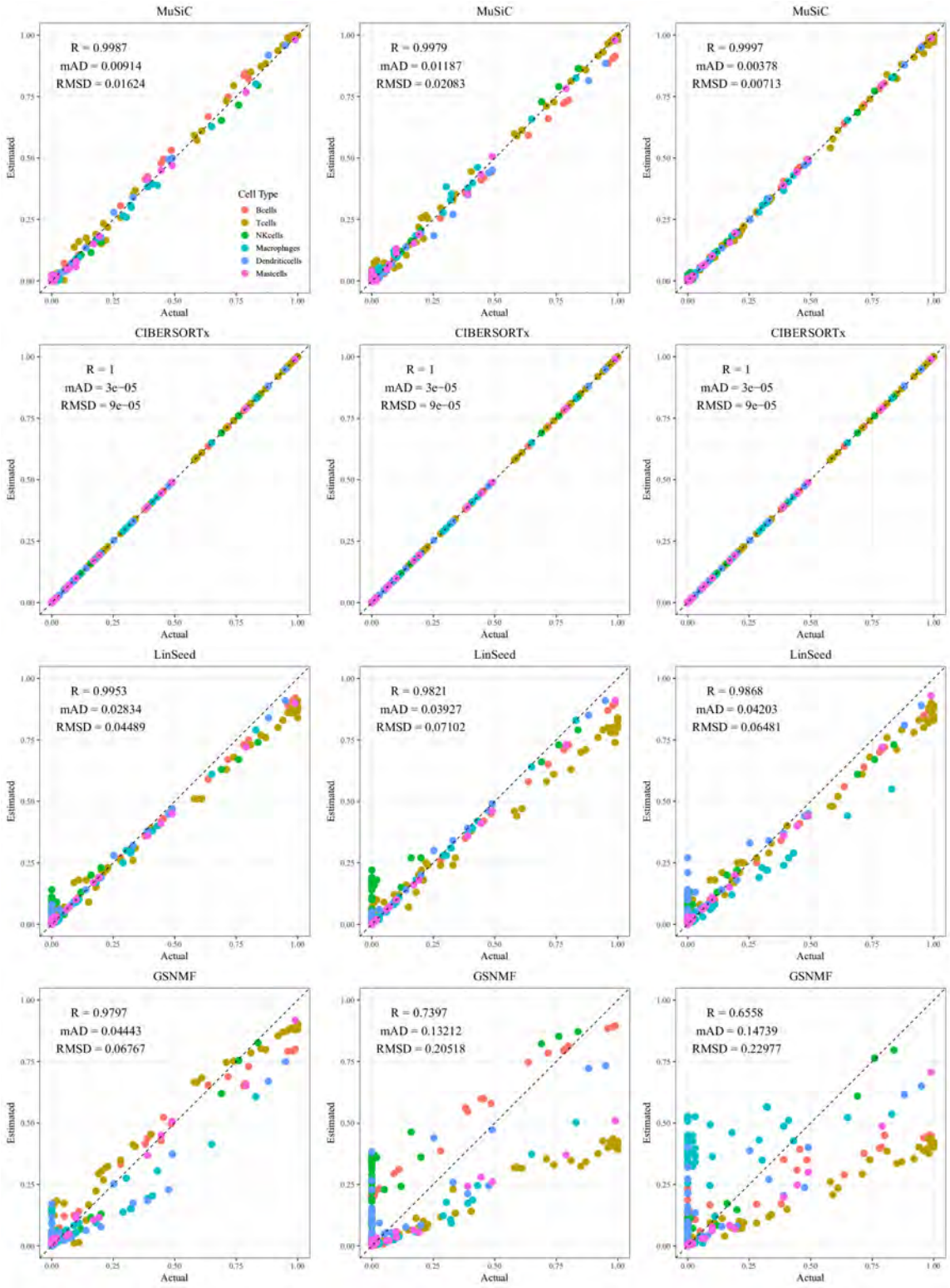


Figure A.11: Comparison of actual and estimated cell-type proportions for dataset LM22 under different truncation scenarios. The columns, from left to right, represent truncation of the top 10%, truncation of the bottom 10%, and the original dataset. Each row corresponds to a different deconvolution method: MuSiC, CIBERSORTx, LinSeed, and GSNMF. Different colors represent different cell types: B cells (red), T cells (gold), NK cells (green), Macrophages (cyan), Dendritic cells (blue), and Mast cells (magenta).

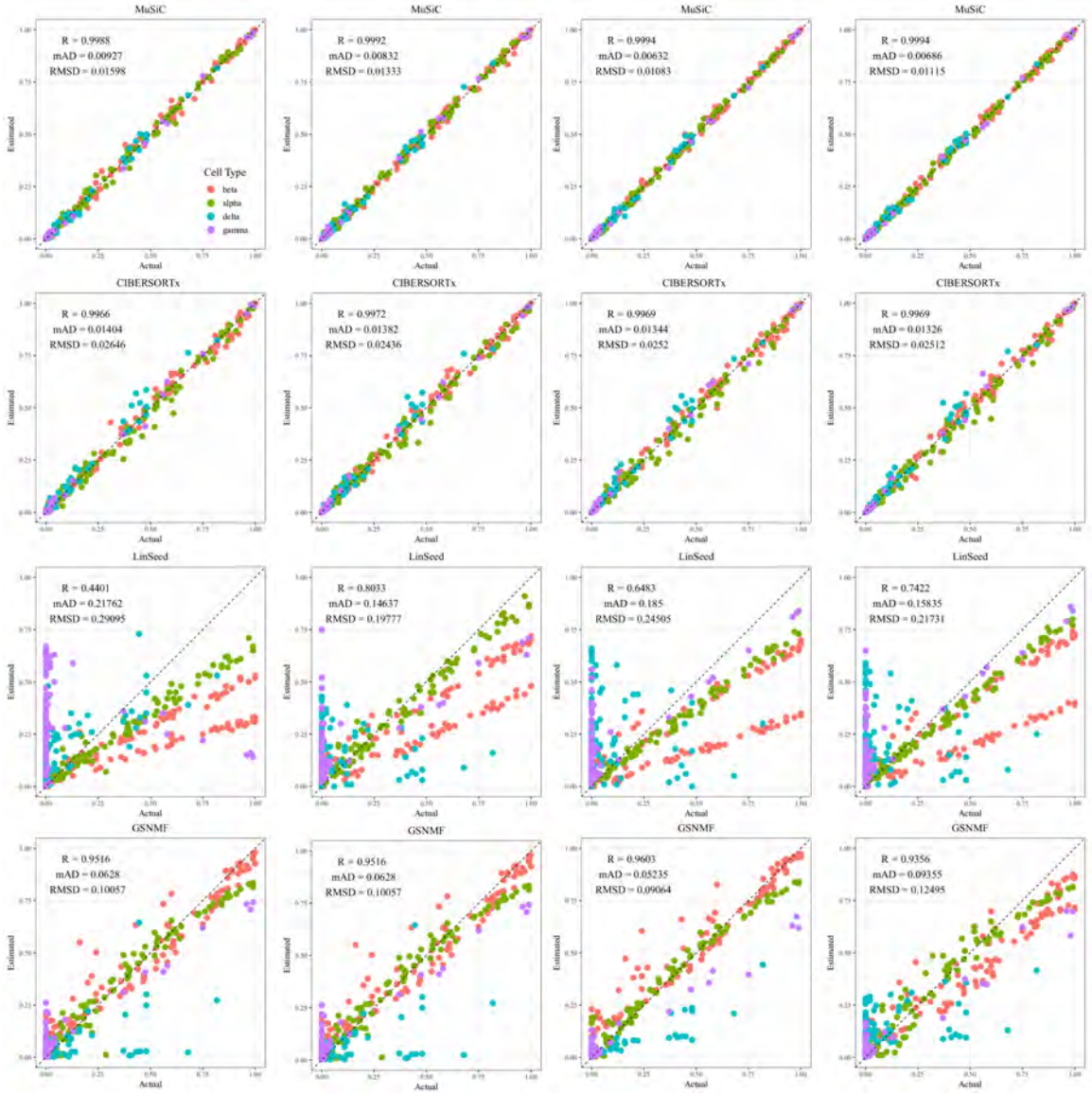


Figure A.12: Comparison of actual and estimated cell-type proportions for dataset GSE81608 under mean shifting conditions. The columns, from left to right, represent 10%, 30%, 50%, and 70% shifts. Each row corresponds to a different deconvolution method: MUSIC, CIBERSORTx, Linseed, and GSNMF. Different colors represent different cell types: beta (red), alpha (green), delta (cyan), and gamma (purple).

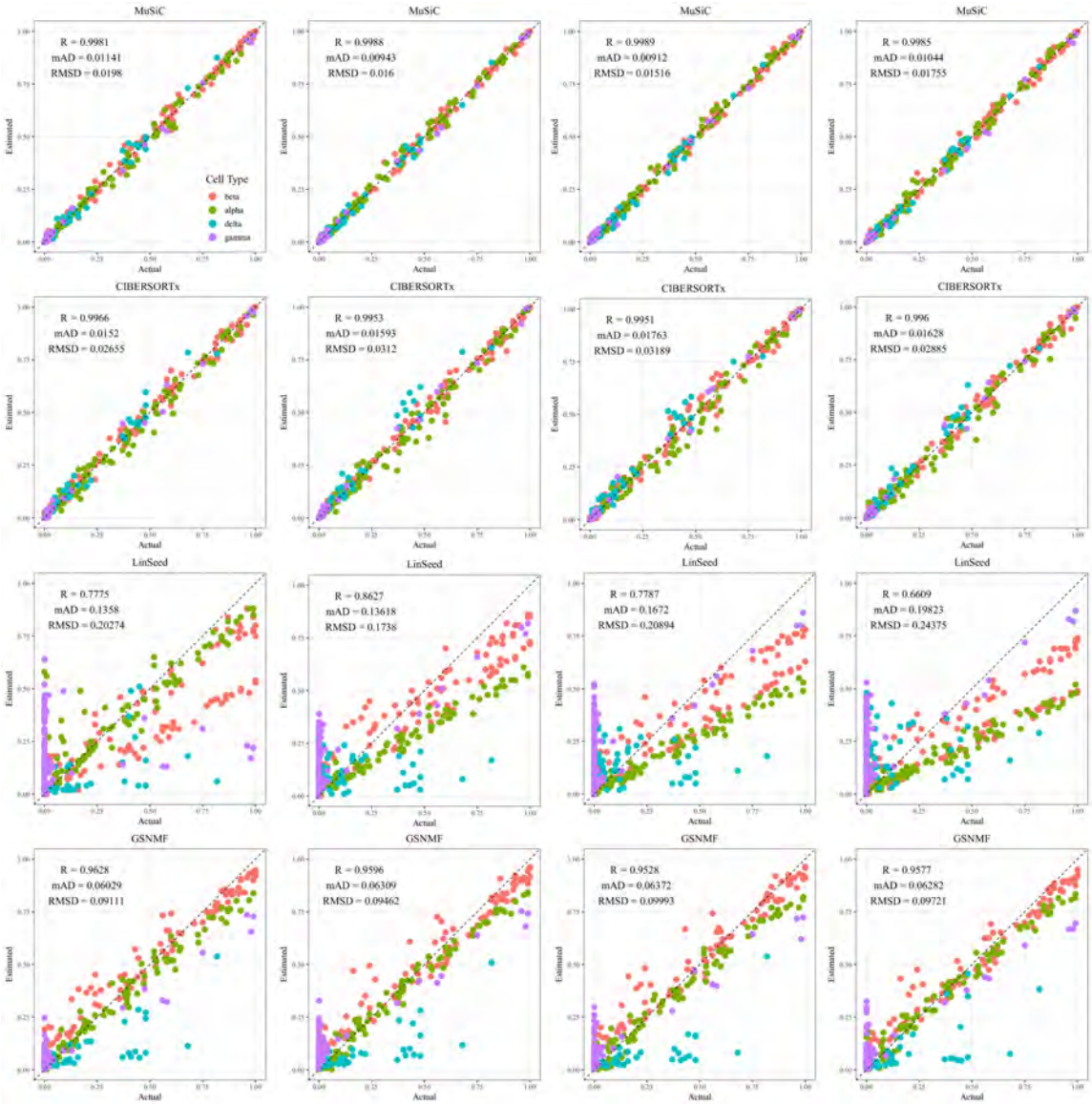


Figure A.13: Comparison of actual and estimated cell-type proportions for dataset GSE81608 under factoring scenarios. The columns, from left to right, represent factoring values of 1.2, 1.8, 0.8, and 0.4. Each row corresponds to a different deconvolution method: MuSiC, CIBERSORTx, LinSeed, and GSNMF. Different colors represent different cell types: beta (red), alpha (green), delta (cyan), and gamma (purple).

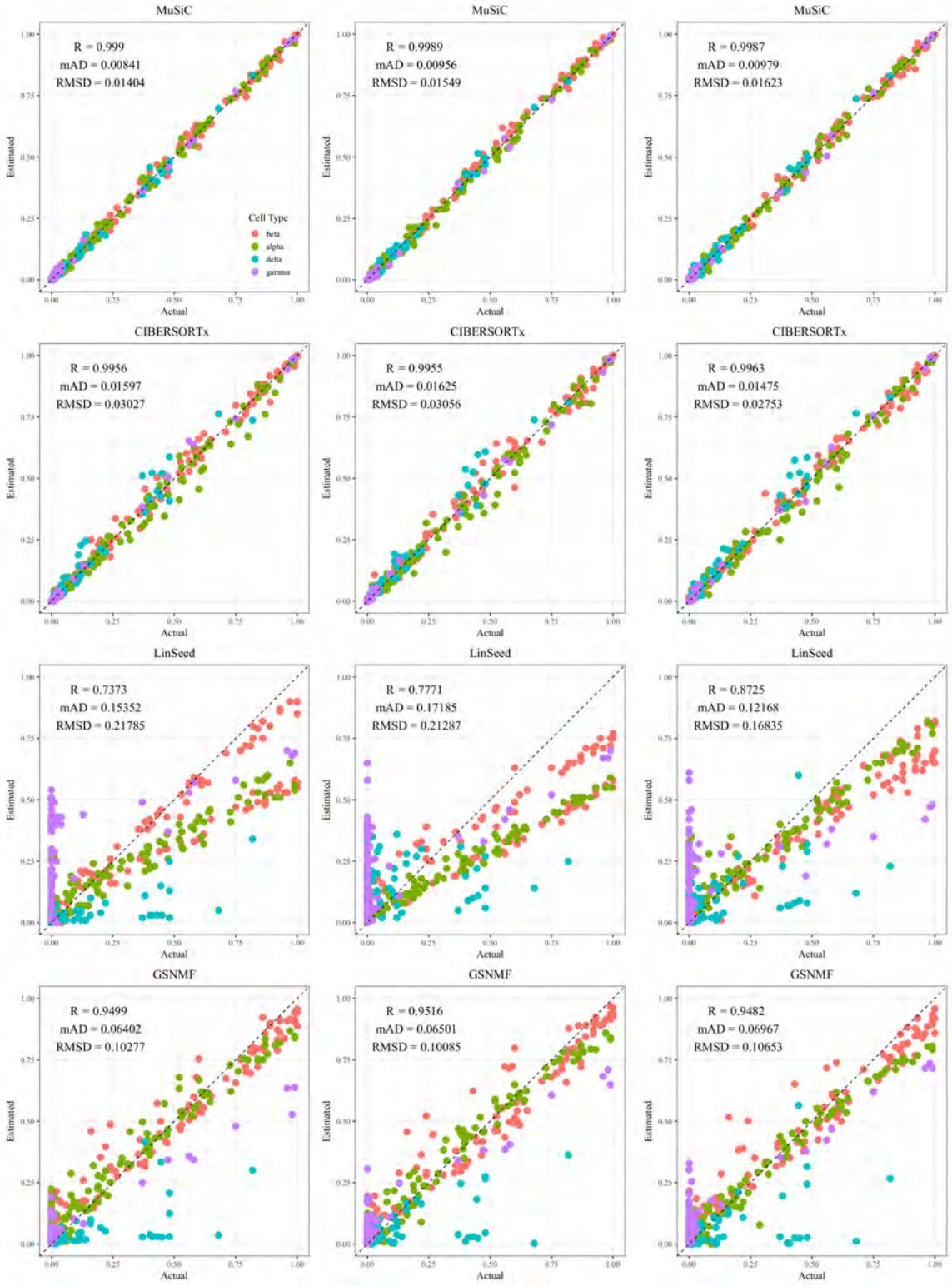


Figure A.14: Comparison of actual and estimated cell-type proportions for dataset GSE81608 under different truncation scenarios. The columns, from left to right, represent truncation of the top 10%, truncation of the bottom 10%, and the original dataset. Each row corresponds to a different deconvolution method: MuSiC, CIBERSORTx, LinSeed, and GSNMF. Different colors represent different cell types: beta (red), alpha (green), delta (cyan), and gamma (purple).

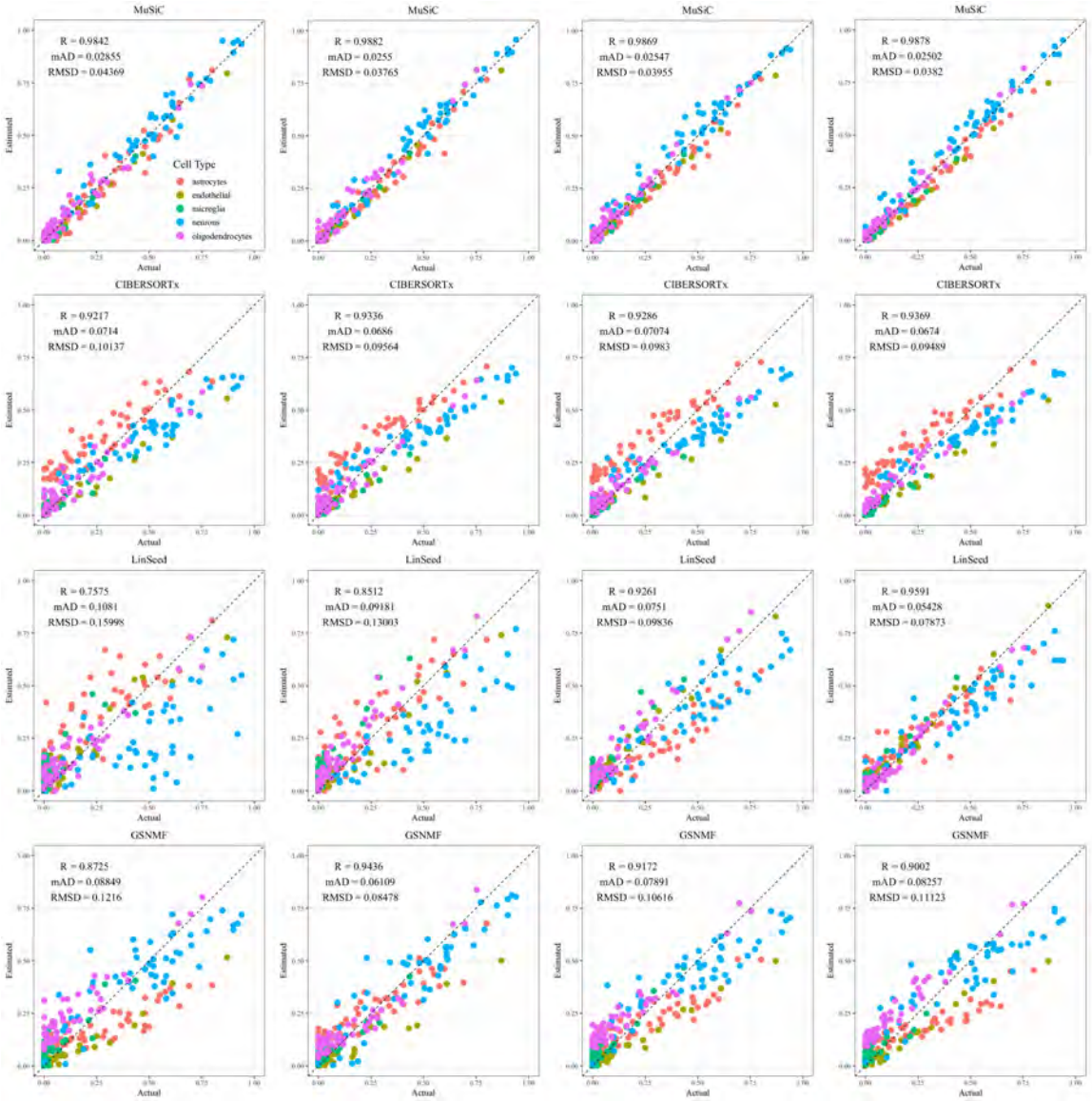


Figure A.15: Comparison of actual and estimated cell-type proportions for dataset GSE67835 under mean shifting conditions. The columns, from left to right, represent 10%, 30%, 50%, and 70% shifts. Each row corresponds to a different deconvolution method: MUSIC, CIBERSORTx, Linseed, and GSNMF. Different colors represent different cell types: astrocytes (red), endothelial (gold), microglia (green), neurons (blue), and oligodendrocytes (magenta).

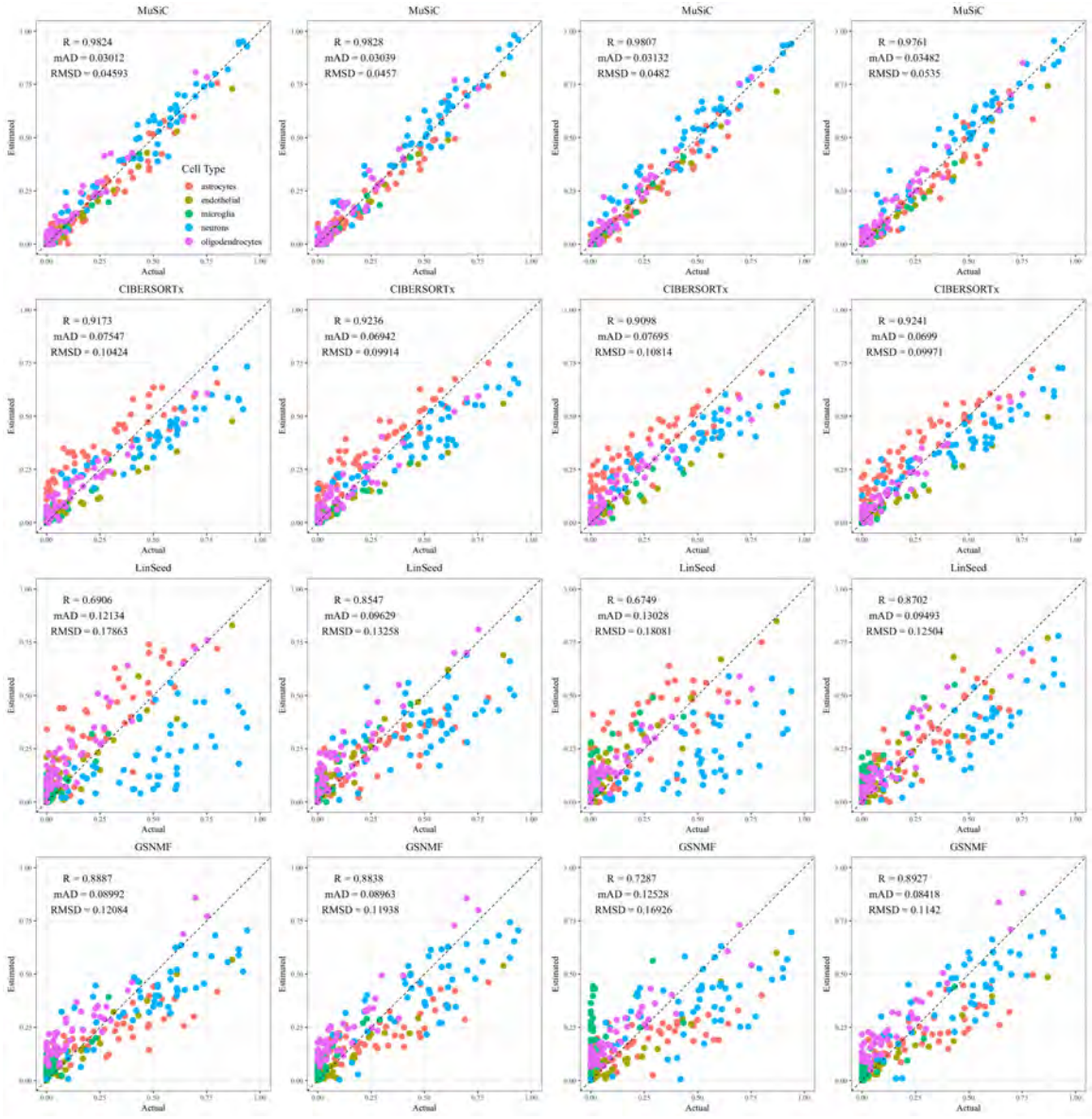


Figure A.16: Comparison of actual and estimated cell-type proportions for dataset GSE67835 under factoring scenarios. The columns, from left to right, represent factoring values of 1.2, 1.8, 0.8, and 0.4. Each row corresponds to a different deconvolution method: MuSiC, CIBERSORTx, LinSeed, and GSNMF. Different colors represent different cell types: astrocytes (red), endothelial (gold), microglia (green), neurons (blue), and oligodendrocytes (magenta).

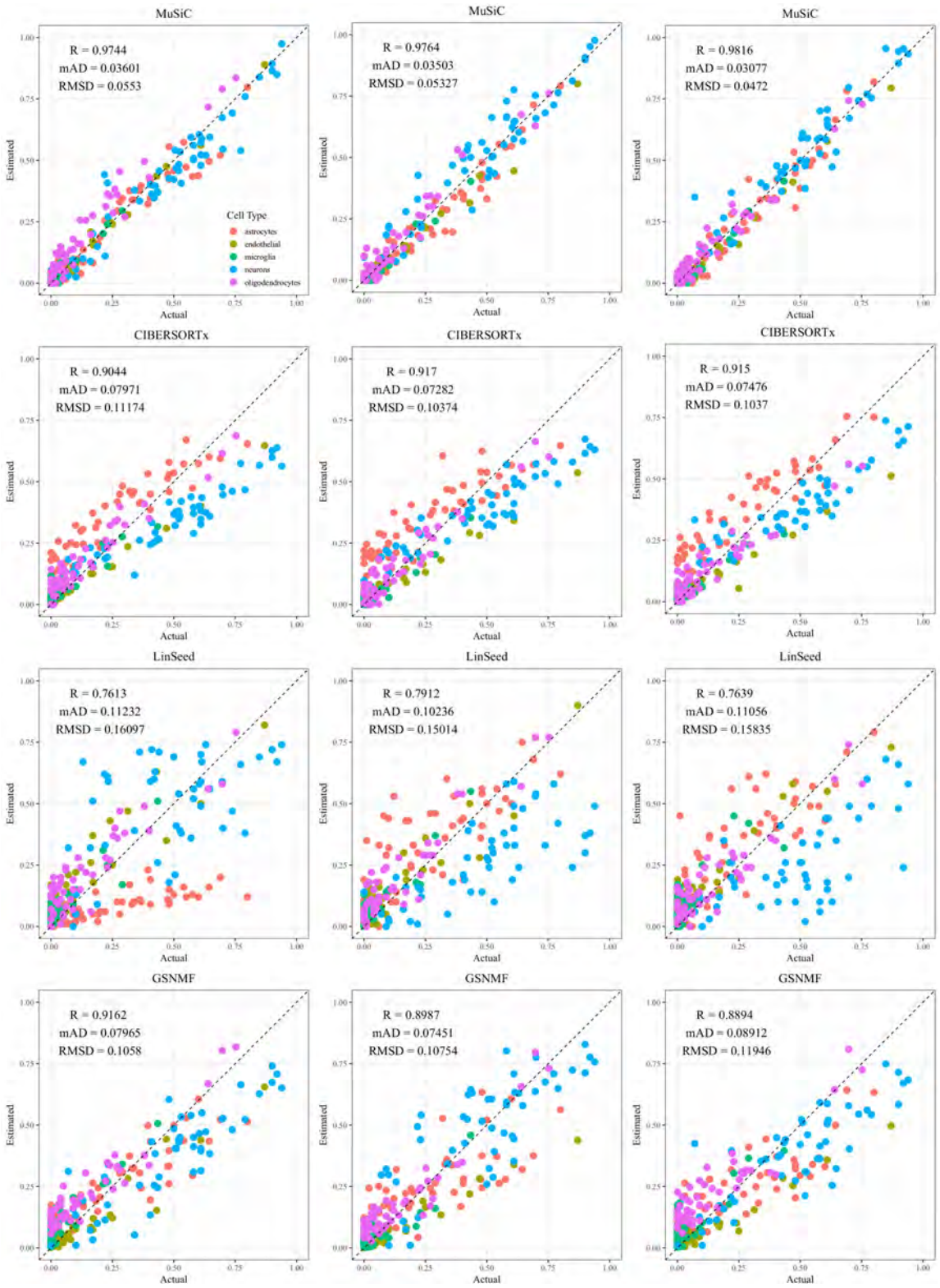


Figure A.17: Comparison of actual and estimated cell-type proportions for dataset GSE67835 under different truncation scenarios. The columns, from left to right, represent truncation of the top 10%, truncation of the bottom 10%, and the original dataset. Each row corresponds to a different deconvolution method: MuSiC, CIBERSORTx, LinSeed, and GSNMF. Different colors represent different cell types: astrocytes (red), endothelial (gold), microglia (green), neurons (blue), and oligodendrocytes (magenta).

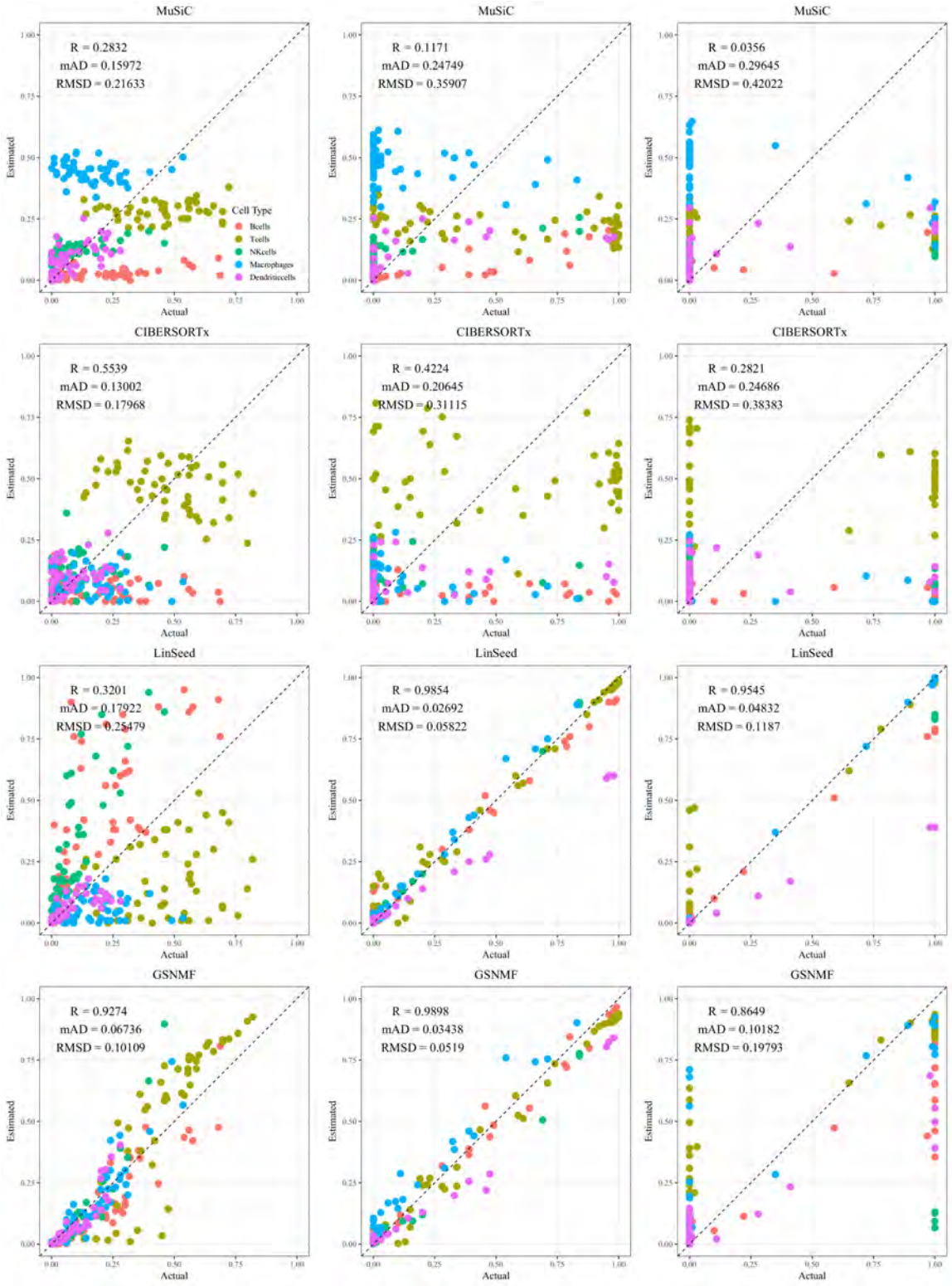


Figure A.18: Comparison of actual and estimated cell-type proportions for dataset PBMC8K under simulator scenarios. From left to right, the columns represent small, medium, and large variation. Each row corresponds to a different deconvolution method: MuSiC, CIBERSORTx, LinSeed, and GSNMF. Different colors represent different cell types: B cells (red), T cells (gold), NK cells (green), Macrophages (blue), and Dendritic cells (magenta).

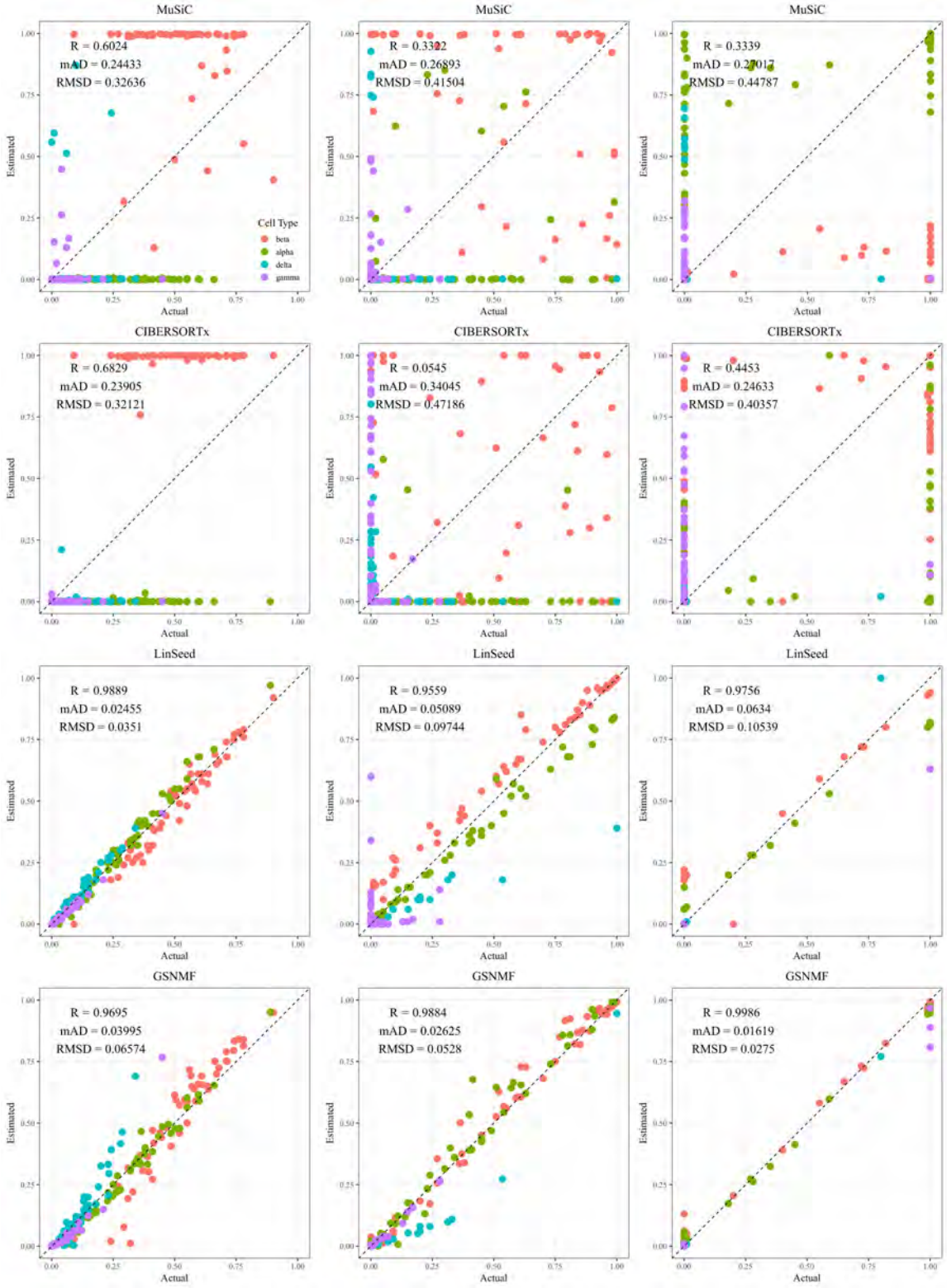


Figure A.19: Comparison of actual and estimated cell-type proportions for dataset GSE81608 under simulator scenarios. From left to right, the columns represent small, medium, and large variation. Each row corresponds to a different deconvolution method: MuSiC, CIBERSORTx, LinSeed, and GSNMF. Different colors represent different cell types: beta (red), alpha (green), delta (cyan), and gamma (purple).

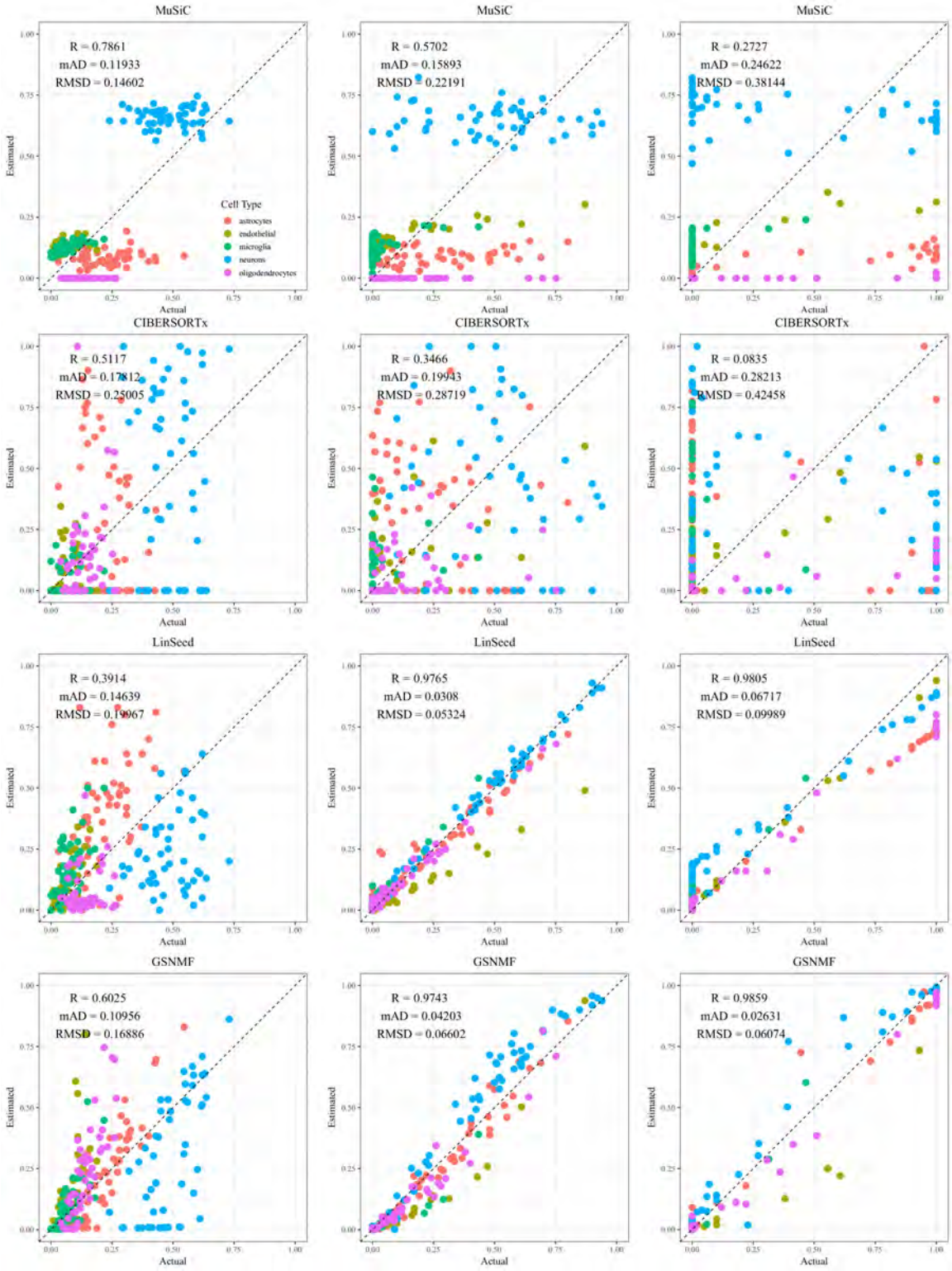


Figure A.20: Comparison of actual and estimated cell-type proportions for dataset GSE67835 under simulator scenarios. From left to right, the columns represent small, medium, and large variation. Each row corresponds to a different deconvolution method: MuSiC, CIBERSORTx, LinSeed, and GSNMF. Different colors represent different cell types: astrocytes (red), endothelial (gold), microglia (green), neurons (blue), and oligodendrocytes (magenta).

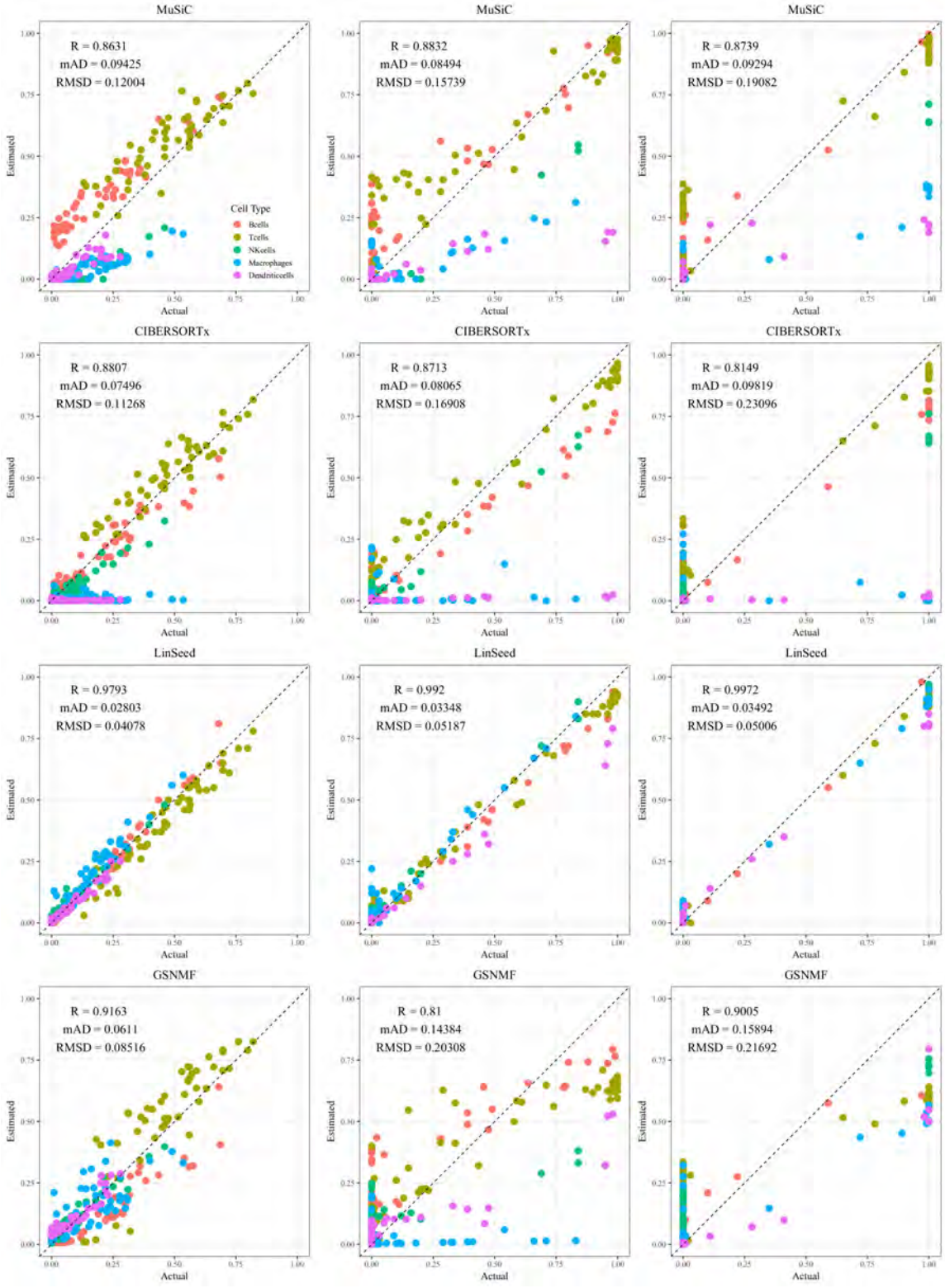


Figure A.21: Comparison of actual and estimated cell-type proportions for the LM22 dataset under real dataset(PBMC8k) scenarios. The columns, from left to right, represent small, medium, and large variation. Each row corresponds to a different deconvolution method: MuSiC, CIBERSORTx, LinSeed, and GSNMF. Different colors represent distinct cell types: B cells (red), T cells (gold), NK cells (green), macrophages (cyan), dendritic cells (blue), and mast cells (magenta).

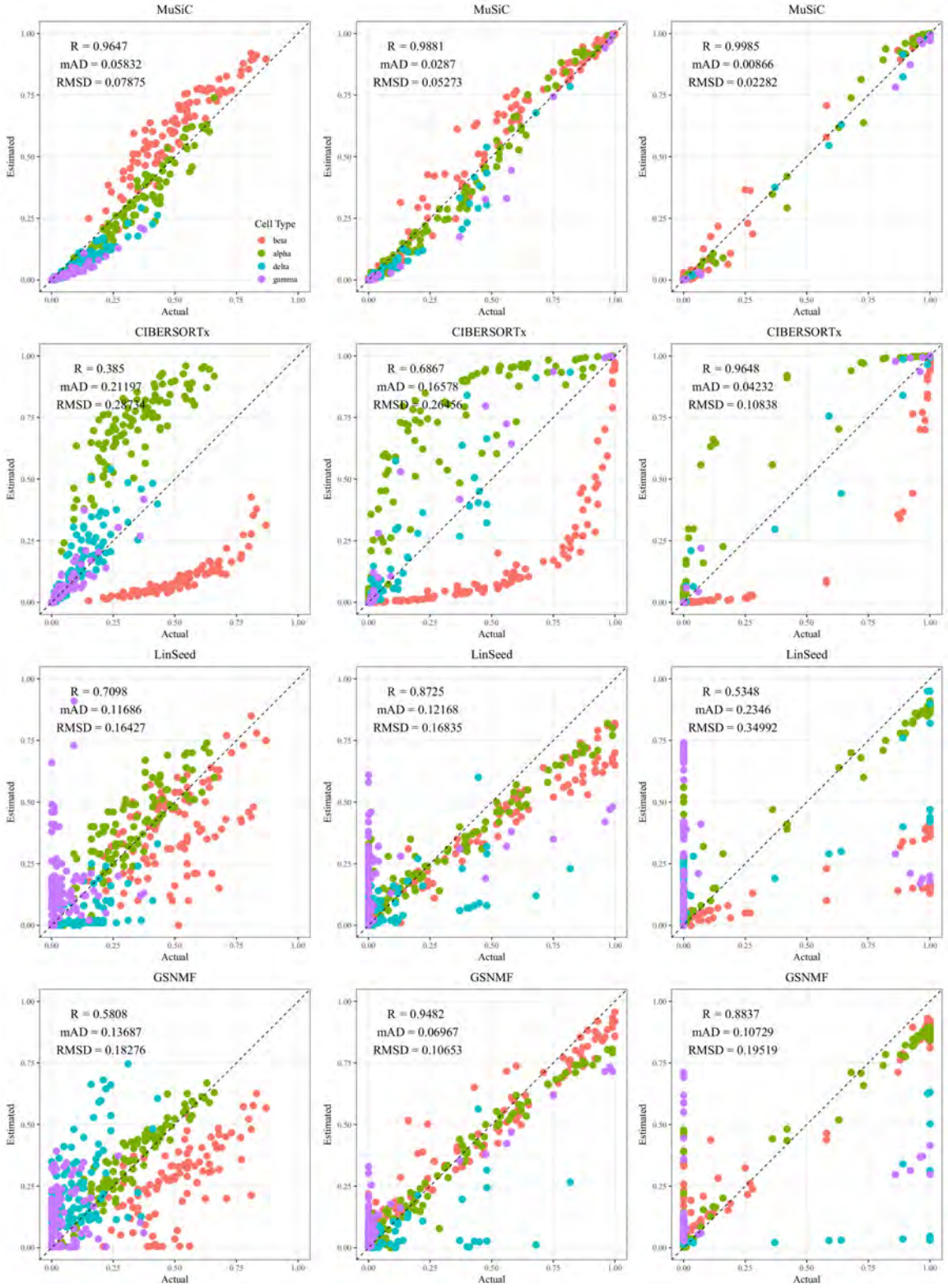


Figure A.22: Comparison of actual and estimated cell-type proportions for the GSE81608 dataset under real dataset(E-MTAB-5061) scenarios. The columns, from left to right, represent small, medium, and large variation. Each row corresponds to a different deconvolution method: MuSiC, CIBERSORTx, LinSeed, and GSNMF. Different colors represent distinct cell types: beta (red), alpha (green), delta (cyan), and gamma (purple).

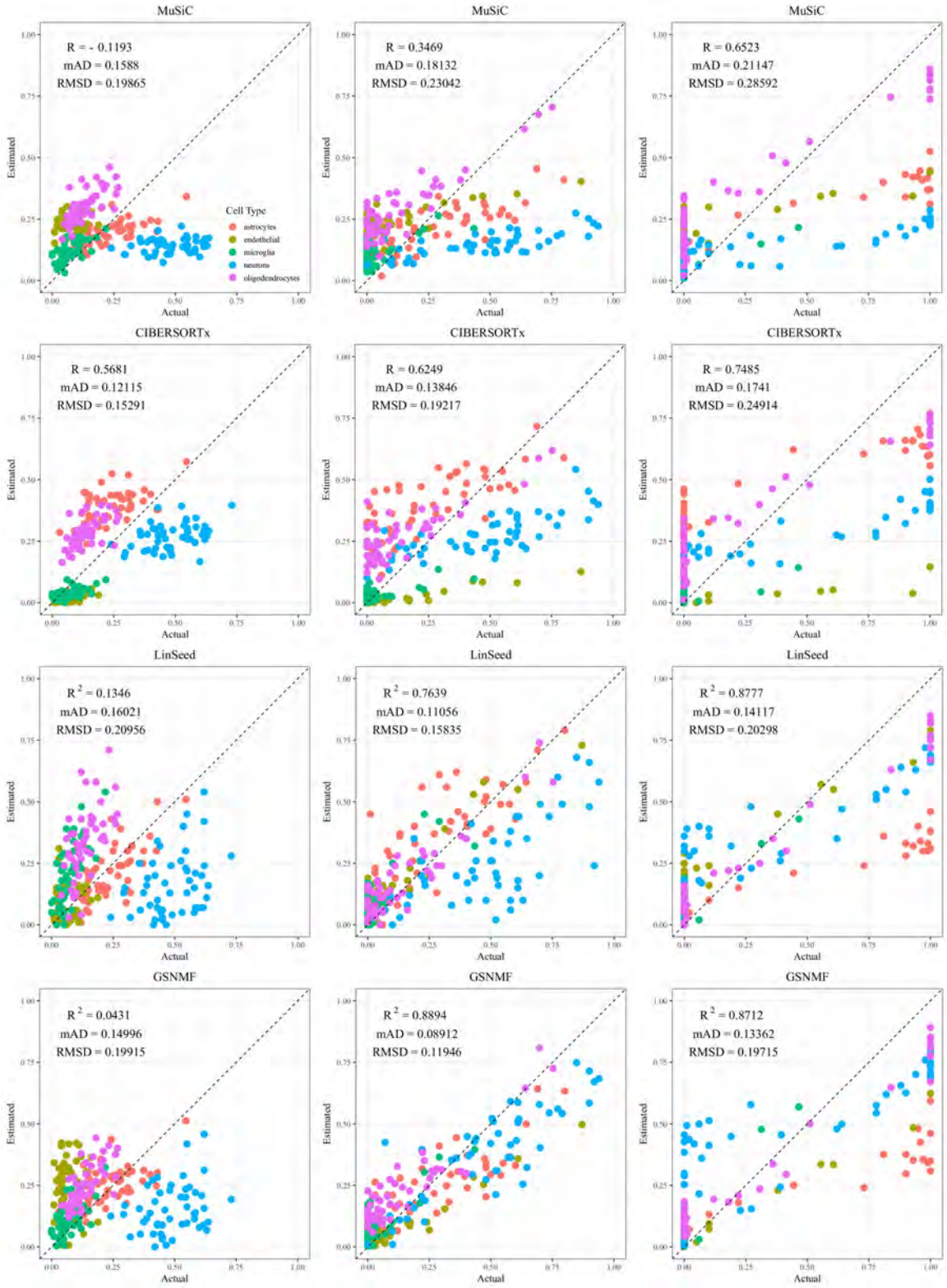


Figure A.23: Comparison of actual and estimated cell-type proportions for the GSE67835 dataset under real dataset (syn18485175) scenarios. The columns, from left to right, represent small, medium, and large variation. Each row corresponds to a different deconvolution method: MuSiC, CIBERSORTx, LinSeed, and GSNMF. Different colors represent distinct cell types: astrocytes (red), endothelial cells (gold), microglia (green), neurons (blue), and oligodendrocytes (magenta).