Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Statistical Inference of Cell-Type Proportions Estimated from Bulk Expression Data

Biao Cai, Jingfei Zhang, Hongyu Li, Chang Su & Hongyu Zhao

To cite this article: Biao Cai, Jingfei Zhang, Hongyu Li, Chang Su & Hongyu Zhao (2024) Statistical Inference of Cell-Type Proportions Estimated from Bulk Expression Data, Journal of the American Statistical Association, 119:548, 2521-2532, DOI: 10.1080/01621459.2024.2382435

To link to this article: https://doi.org/10.1080/01621459.2024.2382435

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 119, NO. 548, 2521–2532: Applications and Case Studies
https://doi.org/10.1080/01621459.2024.2382435

Statistical Inference of Cell-Type Proportions Estimated from Bulk Expression Data

Biao Cai[a], Jingfei Zhang[b], Hongyu Li[c], Chang Su[d], and Hongyu Zhao[c]

[a]Department of Management Sciences, City University of Hong Kong, Kowloon, Hong Kong; [b]Goizueta Business School, Emory University, Atlanta, GA; [c]Department of Biostatistics, Yale University, New Haven, CT; [d]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA

ABSTRACT

There is a growing interest in cell-type-specific analysis from bulk samples with a mixture of different cell types. A critical first step in such analyses is the accurate estimation of cell-type proportions in a bulk sample. Although many methods have been proposed recently, quantifying the uncertainties associated with the estimated cell-type proportions has not been well studied. Lack of consideration of these uncertainties can lead to missed or false findings in downstream analyses. In this article, we introduce a flexible statistical deconvolution framework that allows a general and subject-specific covariance of bulk gene expressions. Under this framework, we propose a decorrelated constrained least squares method called DECALS that estimates cell-type proportions as well as the sampling distribution of the estimates. Simulation studies demonstrate that DECALS can accurately quantify the uncertainties in the estimated proportions whereas other methods fail. Applying DECALS to analyze bulk gene expression data of post mortem brain samples from the ROSMAP and GTEx projects, we show that taking into account the uncertainties in the estimated cell-type proportions can lead to more accurate identifications of cell-type-specific differentially expressed genes and transcripts between different subject groups, such as between Alzheimer's disease patients and controls and between males and females. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received September 2022
Accepted June 2024

KEYWORDS
Cell type deconvolution; Cell-type-specific analysis; Cell-type proportions; Decorrelation; Uncertainty quantification

1. Introduction

The need to analyze gene expression data collected either through microarrays or sequencing to answer different biological questions has motivated the developments of a great number of statistical methods in the last two decades. The applications of these methods have gained novel biological insights on disease mechanisms, identified informative biomarkers, and led to novel treatments for some diseases (e.g., Zhang and Horvath 2005; Barabási, Gulbahce, and Loscalzo 2011; Trapnell et al. 2012; Zhang et al. 2013; Mostafavi et al. 2018; Zhang and Li 2022). While the literature on gene expression analysis is steadily growing, most gene expression data gathered to date are from bulk samples which consist of distinct cell types. For example, a brain sample usually has astrocytes, endothelial cells, microglia, neurons, oligodendrocytes, and oligodendrocyte precursor cells (Darmanis et al. 2015). Therefore, even if two samples have the same gene expression profiles at the cell type level, their aggregated bulk expression profiles may differ if their cell-type proportions are different. Due to the heterogeneous cell-type proportions across samples, the analysis of gene expression data at the bulk level may lead to false positive findings and miss true biological signals. Moreover, such an analysis only offers an aggregated view of the biological mechanisms in different cell types, while most disease etiologies are cell-type-specific (Hekselman and Yeger-Lotem 2020; Li et al. 2021; Zhu et al. 2022). To gain a more accurate and comprehensive view of the underlying biological mechanisms, a desirable approach is to analyze gene expressions in specific cell types. While such cell-type-specific gene expressions are not directly available from bulk sample data, it can be inferred if the cell-type proportions for bulk samples are given. This task of inferring cell-type proportions and/or expressions from bulk samples is often referred as *deconvolution*.

In recent years, many deconvolution methods have been proposed (Abbas et al. 2009; Newman et al. 2015; Wang et al. 2019; Jew et al. 2020; Tang, Park, and Zhao 2020; Yang et al. 2021). These methods rely on the availability of signature genes for different cell types with their expressions usually gathered from single-cell RNA sequencing (scRNA-seq) data, and they differ in the details on how the information from these signature genes is used. The estimated cell-type proportions from these methods together with the bulk expression data have made it possible to address a number of important cell-type-specific (CTS) biological questions, For example, based on gene expression data collected from two groups of bulk samples, it may be possible to infer genes having different CTS expression levels between the two groups (Jin et al. 2021; Wang, Roeder, and Devlin 2021; Tang, Park, and Zhao 2022). It is also possible to infer CTS co-expression patterns (Su, Zhang, and Zhao 2021) and CTS expression quantitative trait loci (eQTLs) (Patel et al. 2021; Little et al. 2022) from bulk samples. Furthermore, instead of making group-level CTS inference, methods have also been proposed

CONTACT Jingfei Zhang *jingfei.zhang@emory.edu* Goizueta Business School, Emory University, Atlanta, GA 30322; Hongyu Zhao *hongyu.zhao@yale.edu* Department of Biostatistics, Yale University, New Haven, CT 06511.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

to infer CTS expression levels at the individual sample level (Newman et al. 2019; Jaakkola and Elo 2022). These sample-level inferred CTS expressions have been used to infer CTS differentially expressed genes between groups, genetic variants that have CTS effects on gene expressions, and CTS co-expressions (Jin et al. 2021; Wang, Roeder, and Devlin 2021; Jaakkola and Elo 2022).

1.1. The ROSMAP Study on Alzheimer's Disease

Alzheimer's disease is a neurodegenerative disorder that causes progressive and irreversible loss of neurons in the brain (Winblad et al. 2016), and is estimated to affect 5.8 million people in the United States. Genetic factors are known to be important in Alzheimer's disease, with an estimated heritability of 58–79% for late-onset Alzheimer's disease (Sims, Hill, and Williams 2020). In recent years, increasing evidence suggests cell-type-specific pathogenesis of Alzheimer's disease (De Strooper and Karran 2016). Our work considers the bulk RNA-seq data collected from the Religious Orders Study and Rush Memory and Aging Project (ROSMAP; Bennett et al. 2018), a clinical-pathologic cohort study of Alzheimer's disease. In the ROSMAP study, postmortem brain samples from $n = 541$ subjects were collected from the dorsolateral prefrontal cortex, a brain region that is strongly associated with Alzheimer's disease pathology (Salat, Kaye, and Janowsky 2001; Montembeault et al. 2016). Among the 541 subjects, 219 were Alzheimer's disease patients and 322 were controls. The bulk gene expression levels for each subject were collected in Mostafavi et al. (2018)[1] and measured in units of FPKM (Trapnell et al. 2010). Single cell data from the same study (Mathys et al. 2019) found several major brain cell types in these tissue samples including neuron, astrocyte, oligodendrocyte, microglia and endothelial cell. See more details in Section 4.2.

The large sample size of the ROSMAP study and the growing literature on deconvolution methods together enable the inference of genes having different expression levels between Alzheimer's disease patients and controls in specific cell types (Jin et al. 2021; Wang, Roeder, and Devlin 2021; Tang, Park, and Zhao 2022), providing a better understanding of the CTS etiology of Alzheimer's disease. However, existing methods for this purpose often ignore the uncertainty in estimated cell-type proportions and treat them as known quantities. This can lead to missed or false findings in identifying CTS differentially expressed genes. To overcome this challenge, our study aims to develop a statistical deconvolution framework that estimates both the cell-type proportions and their sampling distributions, and can incorporate uncertainties from estimated cell-type proportions when inferring CTS differentially expressed genes.

1.2. Existing Methods and Our Approach

The majority of the existing CTS analysis methods implicitly assume that the true cell-type proportions for bulk samples are available, even though they are often estimated with errors from deconvolution models. Limited efforts have been made to investigate and quantify the impacts of uncertainties in estimated cell-type proportions on downstream CTS analysis methods, even though not considering such uncertainties in estimated cell-type proportions can lead to missed or false findings in downstream CTS analyses. Two recent methods have been proposed in the literature to quantify the uncertainties in estimated cell-type proportions. Erdmann-Pham et al. (2021) proposed a likelihood-based deconvolution method using single-cell reference data, referred to as `RNA-Sieve`, and confidence intervals of the estimated proportions can be calculated as a by-product of the estimation procedure. Their approach assumes that the error terms from modeling the signature gene expressions are independent and Gaussian. However, this assumption will likely fail for real data because there are correlations among genes and RNA-seq data are more appropriately modeled by nonnormal distributions, for example negative binomial distributions. Xie and Wang (2022) developed a method based on a measurement error model that incorporates the errors in inferring signature gene expression levels from single cell data in the estimates of cell-type proportions in bulk samples, referred to as `MEAD`. The estimated proportions are shown to be asymptotically normal and the covariance is estimated through a sandwich type estimator with an estimated gene-gene dependence set. However, this covariance estimator may be biased as the subject-specific covariance among signature genes is not consistently estimated in `MEAD`, and this can reduce the accuracy of inferential tasks such as constructing confidence intervals. Specifically, in our simulation studies in Section 3.3, we show that the confidence intervals for cell-type proportions calculated using `RNA-Sieve` and `MEAD` both suffer from under-coverage, sometimes substantially.

In this article, we use a new statistical deconvolution framework to estimate the cell-type proportions and their sampling distributions, and to incorporate the uncertainties in downstream CTS analysis methods. Our approach does not impose parametric assumptions on the distributions of bulk expressions and allows a general covariance among the signature genes that can be cell-type- and subject- specific. Specifically, we consider a decorrelated constrained least squares framework (`DECALS`) to estimate the cell-type proportions, such that the estimated proportions are nonnegative and add up to 1, and the distribution of the estimated proportions is derived by decorrelating the signature gene expressions in each bulk sample via their sample-specific covariance. One major challenge in estimating the distribution of estimated proportions in a sample, say i denoted as $\boldsymbol{\pi}_i$, is the need to characterize the covariance among signature gene expressions in this sample, denoted as $\boldsymbol{\Sigma}_i$. As bulk expressions are aggregated over different cell types, covariance $\boldsymbol{\Sigma}_i$ is a function of $\boldsymbol{\pi}_i$ and the unknown CTS covariances. To consistently estimate the CTS covariances, we consider a novel moment-based estimator that borrows information across all bulk samples and further consider a finite sample bias correction to improve accuracy. We demonstrate in simulation studies that `DECALS` can accurately quantify the uncertainties in the estimated proportions whereas other methods fail to offer accurate uncertainty estimates. In Section 4, we apply `DECALS` to analyze bulk gene expression data from post mortem brain samples from the ROSMAP and GTEx projects and show that taking into account the uncertainties in the estimated cell-type proportions can lead to more accurate identifications of cell-type-specific

[1] *https://www.synapse.org/#!Synapse:syn3388564*

differentially expressed genes and transcripts between different subject groups, such as between Alzheimer's disease patients and controls and between males and females. As DECALS is flexible, easy to compute and free from parametric assumptions, it can be easily combined with most CTS analysis methods based on bulk samples to incorporate the uncertainties of estimated cell-type proportions and improve the accuracy and interpretability of the biological findings.

The rest of the article is organized as follows. Section 2 introduces the cell type convolution model, the estimation of cell-type proportions and their sampling distributions. Section 3 reports the simulation results. Section 4 performs downstream analysis to identify CTS differentially expressed genes and transcripts between groups of samples for two real studies, demonstrating that taking into account the uncertainties in the cell-type proportions can lead to more enriched and interpretable biological findings. The article is concluded with a discussion section.

2. Estimation and Inference of Cell-Type Proportions

2.1. Cell Type Deconvolution Model

Suppose we have gene expression data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \mathbb{R}^p$ collected from n bulk RNA-seq samples across p signature genes. We assume that there are K cell types, and the bulk level expression for sample i is the sum of these K cell types written as

$$\boldsymbol{y}_i = \sum_{k=1}^{K} \pi_{ik} \boldsymbol{x}_i^{(k)}, \tag{1}$$

where π_{ik} and $\boldsymbol{x}_i^{(k)}$ are the proportion and expression profile of cell type k in sample i, respectively, and $\sum_{k=1}^{K} \pi_{ik} = 1$. This deconvolution model has been commonly considered (Abbas et al. 2009; Newman et al. 2015; Wang et al. 2019; Jew et al. 2020; Tang, Park, and Zhao 2020; Yang et al. 2021). See more discussions in Section A3.3. In this article, we do not make any parametric assumptions on the distributions of CTS expression profile $\boldsymbol{x}_i^{(k)}$ and bulk expression $\boldsymbol{y}_i$. Denoting $\mathbb{E}(\boldsymbol{x}_i^{(k)}) = \boldsymbol{w}_k$, where $\boldsymbol{w}_k$ represents the signature gene expression profile for the kth cell type, we may write

$$\boldsymbol{y}_i = \sum_{k=1}^{K} \pi_{ik} \boldsymbol{w}_k + \boldsymbol{\epsilon}_i, \quad \mathbb{E}(\boldsymbol{\epsilon}_i) = 0, \tag{2}$$

where $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{ip})$ is a vector of random variables with mean zero. In cell type deconvolution analysis, the CTS mean expressions $\{\boldsymbol{w}_k\}_{1 \le i \le K}$ are usually gathered from pure cell types (Newman et al. 2015; Li et al. 2016) or scRNA-seq data (Wang et al. 2019; Newman et al. 2019; Jew et al. 2020). For our model, we recommend using mRNA abundance measure FPKM (Trapnell et al. 2010) and without log transformation (see discussions in Section A1). Given bulk expressions $\{\boldsymbol{y}_i\}_{1 \le i \le n}$ and CTS mean expressions $\{\boldsymbol{w}_k\}_{1 \le i \le K}$, we focus on the inference of $\{\boldsymbol{\pi}_i\}_{1 \le i \le n}$, where $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK})$ denotes the vector of cell-type proportions in sample i.

Before we proceed, we first highlight some important differences between (2) and a standard linear regression problem. *First*, model (2) estimates $\boldsymbol{\pi}_i$ with p observations $(y_{i1}, \ldots, y_{ip})$ representing the bulk expressions of p signature genes in sample i. The statistical units in (2) are the p signature genes, rather than the n bulk samples. Hence, the estimation accuracy of $\boldsymbol{\pi}_i$ is expected to be more closely related to p, the number of signature genes, than n, the number of samples. *Second*, the error terms $(\epsilon_{i1}, \ldots, \epsilon_{ip})$ in (2) are not independent. Specifically, $\mathrm{cov}(\boldsymbol{\epsilon}_i)$ can be written as a sum of CTS covariances between the signature genes weighted by cell-type proportions $(\pi_{i1}, \ldots, \pi_{iK})$; see (5). As a result, drawing inference on $\boldsymbol{\pi}_i$ via (2) demands estimating $\mathrm{cov}(\boldsymbol{\epsilon}_i)$, termed *subject-specific covariance* in this article. *Third*, as π_{ik}'s are cell-type proportions in sample i, they must satisfy the constraints that $\pi_{ik} \ge 0$ and $\sum_{k=1}^{K} \pi_{ik} = 1$. The above unique aspects in (2) pose new and significant challenges in the statistical inference of cell-type proportions, which we will address in the ensuing development.

2.2. Estimation of Cell-Type Proportions

From (2), we estimate the cell-type proportion vector $\boldsymbol{\pi}_i$ in sample i via solving the following constrained least-squares problem:

$$\begin{aligned} &\min_{\boldsymbol{\pi}_i \in \mathbb{R}^K} \sum_{j=1}^{p} \left(y_{ij} - \sum_{k=1}^{K} \pi_{ik} w_{kj} \right)^2, \\ &\text{s.t. } \pi_{ik} \ge 0 \text{ and } \sum_{k=1}^{K} \pi_{ik} = 1. \end{aligned} \tag{3}$$

The solution to (3) is denoted as $\hat{\boldsymbol{\pi}}_i$. Note that $\pi_{ik} \le 1$ is implied by the constraints in (3). In (3), we consider a constrained ordinary least squares. Alternatively, one may wish to consider a constrained generalized least squares that multiplies the regression equation (2) by $\mathrm{cov}(\boldsymbol{\epsilon}_i)^{-1/2}$. While the generalized least squares estimator can be more efficient, we demonstrate in Section 2.4 that it may suffer from large biases in practice, due to the uncertainty in estimating $\mathrm{cov}(\boldsymbol{\epsilon}_i)^{-1/2}$ for each sample i. On the other hand, our empirical investigations show that $\hat{\boldsymbol{\pi}}_i$ is more robust and computationally efficient. See detailed discussions and comparisons in Section 2.4. The nonnegative constraint on cell type proportions in (3) is important and leads to more biologically interpretable results in real data analysis. See Section A3.6 in the supplement for details.

The optimization problem in (3) is a quadratic programming problem and we solve it via the standard dual method (Goldfarb and Idnani 1982, 1983). Writing $\boldsymbol{W} = [\boldsymbol{w}_1^\top, \ldots, \boldsymbol{w}_K^\top] \in \mathbb{R}^{p \times K}$, the dual function of (3) can be written as

$$\begin{aligned} &\max_{\boldsymbol{\lambda}} \boldsymbol{c}^\top \boldsymbol{\lambda} + \frac{1}{2} (\boldsymbol{y}_i^\top \boldsymbol{y}_i - \boldsymbol{\pi}_i^\top \boldsymbol{W}^\top \boldsymbol{W} \boldsymbol{\pi}_i), \\ &\text{s.t. } \boldsymbol{A}^\top \boldsymbol{\lambda} + \boldsymbol{W}^\top \boldsymbol{y}_i = (\boldsymbol{W}^\top \boldsymbol{W}) \boldsymbol{\pi}_i, \end{aligned} \tag{4}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^{K+1}$ is the dual vector, $\boldsymbol{A} = (\boldsymbol{I}_K, \boldsymbol{1}_K)^\top \in \mathbb{R}^{(K+1) \times K}$, $\boldsymbol{1}_K = (1, \ldots, 1) \in \mathbb{R}^K$, $\boldsymbol{c} = (0, \ldots, 0, 1)^\top \in \mathbb{R}^{K+1}$ and $\boldsymbol{\pi}_i$ is a solution to (3). Given (3) and (4), the standard dual method (Goldfarb and Idnani 1982, 1983) that uses the unconstrained least squares estimator as the initial value and the Cholesky and QR factorizations for parameter updates is applied to calculate $\hat{\boldsymbol{\pi}}_i$.

To quantify the uncertainties in the estimated proportions, we establish the asymptotic distribution in Theorem 1 in the supplementary materials. According to this theorem,

$\text{cov}(\sqrt{p}\hat{\boldsymbol{\pi}}_i)$ converges to $\boldsymbol{V}_i$ where $\boldsymbol{V}_i = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$, $\text{cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_i$ and

$$\boldsymbol{D} = \left(\frac{1}{p}\boldsymbol{W}^\top\boldsymbol{W}\right)^{-1}\left(\frac{1}{p}\boldsymbol{W}^\top\boldsymbol{\Sigma}_i\boldsymbol{W}\right)\left(\frac{1}{p}\boldsymbol{W}^\top\boldsymbol{W}\right)^{-1},$$

$$\boldsymbol{U} = \boldsymbol{I} - \left(\frac{1}{p}\boldsymbol{W}^\top\boldsymbol{W}\right)^{-1}\mathbf{1}_K^\top\left\{\mathbf{1}_K\left(\frac{1}{p}\boldsymbol{W}^\top\boldsymbol{W}\right)^{-1}\mathbf{1}_K^\top\right\}^{-1}\mathbf{1}_K.$$

This result is useful for uncertainty quantification in downstream analyses that require cell-type proportions, such as CTS differential expression analysis and CTS eQTL analysis. We will demonstrate two such real data examples in Section 4.2. As previously commented, one major challenge in deriving the asymptotic properties of $\hat{\boldsymbol{\pi}}_i$ is that the random variables $(y_{i1},\ldots,y_{ip})$ in (2) are not independent and need to be *decorrelated* using the subject-specific covariance $\boldsymbol{\Sigma}_i$. In practice, covariance $\boldsymbol{\Sigma}_i$ is unknown and we discuss its estimation in the next section.

2.3. Estimation of Subject-Specific Covariances

In this section, we discuss the estimation of subject-specific covariance $\boldsymbol{\Sigma}_i$ in (S6). Assuming that the CTS expression profiles $\boldsymbol{x}_i^{(1)},\ldots,\boldsymbol{x}_i^{(K)}$ are independent, the covariance between bulk expressions reduces to a weighted sum of cell-type-specific covariances and it is written as

$$\boldsymbol{\Sigma}_i = \text{cov}\left(\sum_{k=1}^K \pi_{ik}\boldsymbol{x}_i^{(k)}\right) = \sum_{k=1}^K \pi_{ik}^2\boldsymbol{\Sigma}^{(k)}. \tag{5}$$

To evaluate the plausibility of (5), we leveraged single-cell RNA-seq data (Fujita et al. 2022) to examine the magnitude of $\text{cov}(\boldsymbol{x}_i^{(k_1)},\boldsymbol{x}_i^{(k_2)})$'s in Section A3.7 and found they are small when compared to $\boldsymbol{\Sigma}^{(k)}$'s.

In order to estimate $\boldsymbol{\Sigma}_i$, we first focus on the estimation of $\boldsymbol{\Sigma}^{(k)}$. Centering by $z_{ij} = y_{ij} - \sum_{k=1}^K \pi_{ik}w_{kj}$, it is easy to see that

$$\mathbb{E}(z_{ij}z_{ij'}) = \sum_{k=1}^K \pi_{ik}^2\boldsymbol{\Sigma}_{jj'}^{(k)}, \quad 1 \le j,j' \le p.$$

The above observation facilities an efficient least squares estimation of $(\boldsymbol{\Sigma}_{jj'}^{(1)},\ldots,\boldsymbol{\Sigma}_{jj'}^{(K)})$ by taking $z_{ij}z_{ij'}$ as the response and $(\pi_{i1}^2,\ldots,\pi_{iK}^2)$ as the vector of predictors. Writing $\boldsymbol{z}_j = (z_{1j},\ldots,z_{nj})$ and $\boldsymbol{H} = \left(\pi_{ik}^2\right)_{n\times K}$, the CTS covariances between genes j and j', that is, $(\boldsymbol{\Sigma}_{jj'}^{(1)},\ldots,\boldsymbol{\Sigma}_{jj'}^{(K)})$, can be consistently estimated with

$$\boldsymbol{b}_{jj'} = (\boldsymbol{H}^\top\boldsymbol{H})^{-1}\boldsymbol{H}^\top(\boldsymbol{z}_j \circ \boldsymbol{z}_{j'}),$$

where $\circ$ denotes the element-wise product. The above CTS covariance estimation was first considered by Su, Zhang, and Zhao (2021), where true cell-type proportions are assumed to be available.

In our setting, the true cell-type proportions are unknown and we only have access to $\hat{\boldsymbol{H}} = \left(\hat{\pi}_{ik}^2\right)_{n\times K}$ and $\hat{\boldsymbol{z}}_j$'s, where $\hat{z}_{ij} = y_{ij} - \sum_{k=1}^K \hat{\pi}_{ik}w_{kj}$. In this case, a natural estimator to consider is

$$\hat{\boldsymbol{b}}_{jj'} = (\hat{\boldsymbol{H}}^\top\hat{\boldsymbol{H}})^{-1}\hat{\boldsymbol{H}}^\top(\hat{\boldsymbol{z}}_j \circ \hat{\boldsymbol{z}}_{j'}). \tag{6}$$

Our result in Theorem 1 suggests that $\boldsymbol{b}_{jj'} - \hat{\boldsymbol{b}}_{jj'} = O_p(1/\sqrt{p})$. Hence, $\hat{\boldsymbol{b}}_{jj'}$ is also a consistent estimator for $(\boldsymbol{\Sigma}_{jj'}^{(1)},\ldots,\boldsymbol{\Sigma}_{jj'}^{(K)})$ as p increases.

In our empirical studies, we find that the finite-sample bias in $\hat{\boldsymbol{b}}_{jj'}$ often leads to a deflated estimation of $\boldsymbol{V}_i = \text{cov}(\hat{\boldsymbol{\pi}}_i)$ and correspondingly, an under-coverage of the confidence intervals calculated for $\boldsymbol{\pi}_i$. As an example, Figure 1(a) shows the coverage probabilities of 95% confidence intervals calculated with (6) under the simulation setting in Section 3.1 and some under-coverage is seen. A further investigation shows that the bias is majorly caused by the finite-sample difference between $\hat{\boldsymbol{H}}$ and $\boldsymbol{H}$ and that between $\hat{\boldsymbol{H}}^\top\hat{\boldsymbol{H}}$ and $\boldsymbol{H}^\top\boldsymbol{H}$. To address this issue, we consider a finite-sample bias-corrected estimator

$$\hat{\boldsymbol{b}}_{jj'}^{\text{correct}} = \left\{\hat{\boldsymbol{H}}^\top\hat{\boldsymbol{H}} - \boldsymbol{B}_1\right\}^{-1}(\hat{\boldsymbol{H}} - \boldsymbol{B}_2)^\top(\hat{\boldsymbol{z}}_j \circ \hat{\boldsymbol{z}}_{j'}), \tag{7}$$

where $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are calculated by explicitly quantifying $\mathbb{E}(\hat{\boldsymbol{H}}^\top\hat{\boldsymbol{H}}) - \boldsymbol{H}^\top\boldsymbol{H}$ and $\mathbb{E}(\hat{\boldsymbol{H}}) - \boldsymbol{H}$, respectively, and given in Proposition 1 below. The proof is given in Section A8.

Proposition 1. Letting $\boldsymbol{\pi}_i^{\circ 2} = (\pi_{i1}^2,\ldots,\pi_{iK}^2)$. If $\sqrt{p}(\hat{\boldsymbol{\pi}}_i - \boldsymbol{\pi}_i) \sim \mathcal{N}(\mathbf{0},\boldsymbol{V}_i)$, it holds that

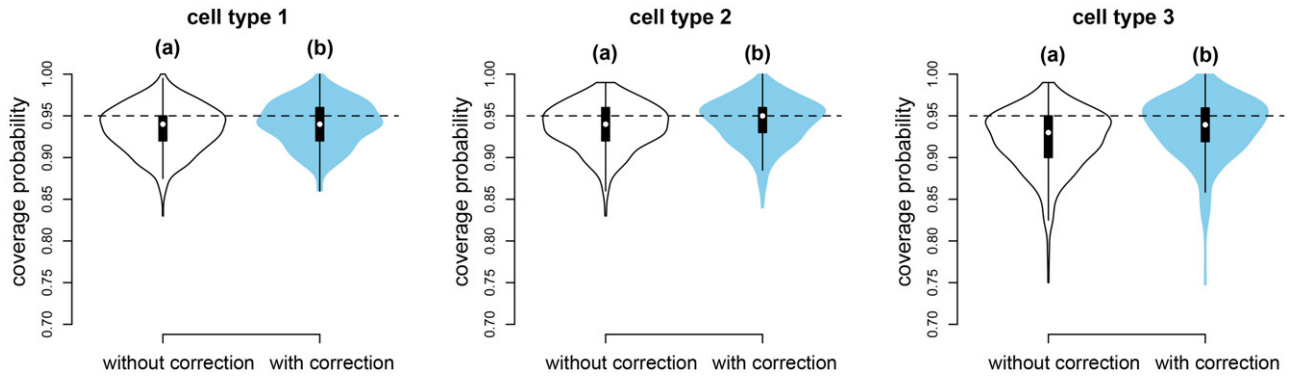

Figure 1. The coverage probabilities of 95% confidence intervals (CI) in three cell types under the simulation setting in Section 3.1. The (a) `without correction` CIs are calculated with (6) and the (b) `with correction` CIs are calculated with (7).

$$\begin{aligned}\boldsymbol{B}_1 &= \frac{1}{p}\sum_{i=1}^{n}\boldsymbol{\pi}_i^{\circ 2}\boldsymbol{u}_i^\top + \frac{1}{p}\sum_{i=1}^{n}\boldsymbol{u}_i\boldsymbol{\pi}_i^{\circ 2\top} \\ &+ \frac{4}{p}\sum_{i=1}^{n}(\boldsymbol{\pi}_i^{\circ 2\top}\boldsymbol{\pi}_i^{\circ 2})\circ\boldsymbol{V}_i + \frac{1}{p^2}\sum_{i=1}^{n}\boldsymbol{T}_i, \\ \boldsymbol{B}_2 &= \frac{1}{p}[\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n]^\top,\end{aligned} \tag{8}$$

where $\boldsymbol{u}_i = (\boldsymbol{V}_{i,11},\ldots,\boldsymbol{V}_{i,KK})^\top$ and $\boldsymbol{T}_i$ is a $K\times K$ matrix with $T_{i,jj'} = 2V_{i,jj'}^2 + V_{i,jj}V_{i,j'j'}$.

Based on Proposition 1 and given $\boldsymbol{V}_i$, we can estimate $\boldsymbol{B}_1$ by

$$\begin{aligned}\hat{\boldsymbol{B}}_1 &= \frac{1}{p}\sum_{i=1}^{n}\hat{\boldsymbol{\pi}}_i^{\circ 2}\boldsymbol{u}_i^\top + \frac{1}{p}\sum_{i=1}^{n}\boldsymbol{u}_i\hat{\boldsymbol{\pi}}_i^{\circ 2\top} \\ &+ \frac{4}{p}\sum_{i=1}^{n}(\hat{\boldsymbol{\pi}}_i^{\circ 2\top}\hat{\boldsymbol{\pi}}_i^{\circ 2})\circ\boldsymbol{V}_i + \frac{1}{p^2}\sum_{i=1}^{n}\boldsymbol{T}_i.\end{aligned} \tag{9}$$

As $\boldsymbol{V}_i$ is unknown in practice, we propose to iteratively update $\boldsymbol{\Sigma}_i$ and $\boldsymbol{V}_i$ in the estimation procedure. The details are summarized in Algorithm 1.

In Step 2 of Algorithm 1, we initialize $\boldsymbol{V}_i^{[0]}$ by estimating σ_i^2 with $\frac{1}{p-1}\sum_{j=1}^{p}\left(y_{ij} - \sum_{k=1}^{K}\pi_{ik}w_{kj}\right)^2$. When p is large, the accumulated errors across $O(p^2)$ entries in $\hat{\boldsymbol{\Sigma}}^{(k)}$ can be excessive, especially when p, the number of signature genes, exceeds n, the number of bulk samples. Hence, in Step 3.3, we consider a sparse estimation of $\boldsymbol{\Sigma}^{(k)}$, which is plausible as gene co-expressions are expected to be sparse when p is large (Zhang and Horvath 2005). Specifically, after calculating $(\tilde{\boldsymbol{\Sigma}}^{(k)})^{[t]}$ in Step 3.2, we consider a SCAD thresholding procedure (Rothman, Levina, and Zhu 2009) with the tuning parameter selected using cross validation (see Section A2.1). Our results in Sections 3–4 are calculated with the SCAD thresholding procedure.

Algorithm 1 The DEcorrelated ConstrAined Least Squares (DECALS) algorithm

Input: Bulk expressions $\{\boldsymbol{y}_i\}_{1\le i\le n}$ and the signature gene matrix $\boldsymbol{W}$.

Step 1: Calculate the constrained least squares estimator $\hat{\boldsymbol{\pi}}_i$ from (3) for $1\le i\le n$.

Step 2: Initialize $\boldsymbol{V}_i^{[0]}$ for $1\le i\le n$.

Repeat the following steps for $t = 0, 1, \ldots$ until convergence.

Step 3.1: Calculate $\boldsymbol{B}_1^{[t]}$ and $\boldsymbol{B}_2^{[t]}$ with (9), $\hat{\boldsymbol{\pi}}_i$ and $\boldsymbol{V}_i^{[t]}$.

Step 3.2: Calculate $(\tilde{\boldsymbol{\Sigma}}^{(k)})^{[t]}$ with (7), $\hat{\boldsymbol{H}}$, $\hat{\boldsymbol{z}}_j$, $\boldsymbol{B}_1^{[t]}$ and $\boldsymbol{B}_2^{[t]}$.

Step 3.3: Calculate $(\boldsymbol{\Sigma}^{(k)})^{[t]}$ by applying SCAD thresholding to $(\tilde{\boldsymbol{\Sigma}}^{(k)})^{[t]}$.

Step 3.4: Calculate $\boldsymbol{\Sigma}_i^{[t]}$ with (5), $\hat{\boldsymbol{\pi}}_i$ and $(\boldsymbol{\Sigma}^{(k)})^{[t]}$.

Step 3.5: Calculate $\boldsymbol{V}_i^{[t+1]}$ with (S6), $\boldsymbol{W}$ and $\boldsymbol{\Sigma}_i^{[t]}$.

Output: The estimated proportions $\{\hat{\boldsymbol{\pi}}_i\}_{1\le i\le n}$ and covariances $\{\hat{\boldsymbol{V}}_i\}_{1\le i\le n}$.

2.4. The Constrained Generalized Least Squares

In our approach, we estimate $\boldsymbol{\pi}_i$ via the constrained least squares in (3). Recalling $\mathrm{cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_i$ and assuming $\boldsymbol{\Sigma}_i$ is positive definite, one may prefer to estimate $\boldsymbol{\pi}_i$ via the following constrained generalized least squares (GLS):

$$\begin{aligned}&\min_{\boldsymbol{\pi}_i\in\mathbb{R}^k}\left\|\boldsymbol{\Sigma}_i^{-1/2}\boldsymbol{y}_i - \boldsymbol{\Sigma}_i^{-1/2}\boldsymbol{W}\boldsymbol{\pi}_i\right\|_2^2, \\ &\text{s.t. } \pi_{ik}\ge 0 \text{ and } \sum_{k=1}^{K}\pi_{ik} = 1.\end{aligned} \tag{10}$$

The solution to (10), denoted as $\hat{\boldsymbol{\pi}}_i^{\mathrm{GLS}}$, is expected to be more efficient than $\hat{\boldsymbol{\pi}}_i$ (Greene 2003). Specifically, denoting $\mathrm{cov}(\hat{\boldsymbol{\pi}}_i^{\mathrm{GLS}})$, we have that

$$\begin{aligned}\boldsymbol{V}_i^{\mathrm{GLS}} &= (\boldsymbol{W}^\top\boldsymbol{\Sigma}_i^{-1}\boldsymbol{W})^{-1} \\ &\left\{\boldsymbol{I} - \mathbf{1}\{\mathbf{1}^\top(\boldsymbol{W}^\top\boldsymbol{\Sigma}_i^{-1}\boldsymbol{W})^{-1}\mathbf{1}^\top\}^{-1}\mathbf{1}^\top(\boldsymbol{W}^\top\boldsymbol{\Sigma}_i^{-1}\boldsymbol{W})^{-1}\right\}.\end{aligned} \tag{11}$$

As demonstrated in Section 2.3, the estimation of the subject-specific covariance $\boldsymbol{\Sigma}_i$ in our problem is nontrivial. When $\boldsymbol{\Sigma}_i$ is unknown but estimated with potentially high noise, the estimate of $\boldsymbol{\pi}_i^{\mathrm{GLS}}$ from (10) and its variance from (11), which further requires $\boldsymbol{\Sigma}_i^{-1}$, can much deteriorate. As an example, Figure 2 shows the coverage probabilities of 95% confidence intervals calculated with $\hat{\boldsymbol{\pi}}_i$ and $\hat{\boldsymbol{\pi}}_i^{\mathrm{GLS}}$, respectively, with the true $\boldsymbol{\Sigma}_i$, referred to as *oracle*, and estimated $\hat{\boldsymbol{\Sigma}}_i$ (see details in Section A2.2) in the simulation setting in Section 3.1. By comparing plots (a) and (b), it is seen that the GLS estimator is slightly more efficient than DECALS when $\boldsymbol{\Sigma}_i$ is known. However, when $\boldsymbol{\Sigma}_i$ is unknown and needs to be estimated from data, $\hat{\boldsymbol{\pi}}_i^{\mathrm{GLS}}$ and its estimated sampling variance can be biased and the coverage probabilities of the 95% confidence intervals are unsatisfactory. We also considered different constraint weighted least squares estimation and the results are presented in Section A2.3 of the supplement.

3. Simulation Studies

We conduct simulations to evaluate the performance of DECALS in two types of settings. In Section 3.1, we generate both the signature gene matrix $\boldsymbol{W}$ and cell-type proportions $\boldsymbol{\pi}_i$'s from pre-specified parametric distributions. In Section 3.3, we use the signature gene matrix $\boldsymbol{W}$ and cell-type proportions $\boldsymbol{\pi}_i$'s inferred from real data (see more details in Section 3.3). Additionally, in Section 3.2, we conduct a sensitivity analysis that examines the performance of DECALS when $\boldsymbol{W}$ is observed with errors.

We compare DECALS with three alternative inferential methods including a naive OLS method, referred to as OLS, RNA-Sieve from Erdmann-Pham et al. (2021) and MEAD from Xie and Wang (2022). In OLS, the cell-type proportions are estimated from ordinary least squares with no constraints; the proportion estimates are taken to be approximately normal and the covariance is calculated assuming the error terms in (3) are iid. RNA-Sieve (Erdmann-Pham et al. 2021) is a likelihood-based deconvolution method that estimates the cell-type proportions from bulk samples and uses single-cell reference data to infer the distribution of the signature

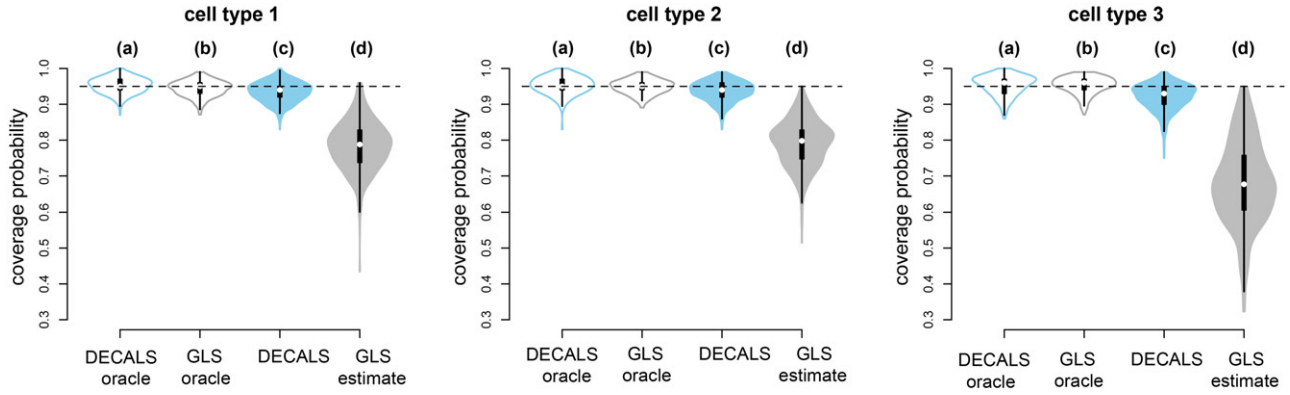

Figure 2. The coverage probabilities of 95% confidence intervals (CI) for three cell types under the simulation setting in Section 3.1. The (a) `DECALS oracle` and (c) `DECALS` CIs are calculated as in Section 2.2 with the true and estimated covariance $\boldsymbol{V}_i$, respectively; the (b) `GLS oracle` and (d) `GLS estimate` CIs are calculated with the true and estimated covariance $\boldsymbol{V}_i^{\text{GLS}}$, respectively.

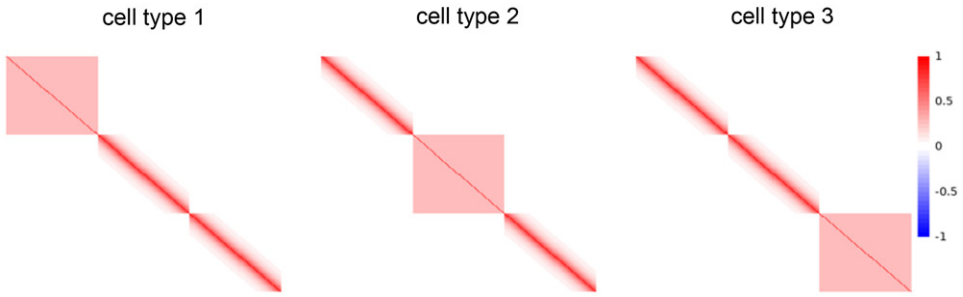

Figure 3. An illustration of the CTS correlation matrices in Section 3.1.

gene matrix. In `RNA-Sieve`, confidence intervals of the estimated proportions can be calculated as a by-product of the estimation procedure. Note that `RNA-Sieve` requires names of the signature genes to link to single cell reference data and hence, it is only implemented in Section 3.3, where such gene information is available. `MEAD` (Xie and Wang 2022) uses an error-in-variable regression framework, where the signature gene matrix is estimated from single cell data with noise, to make inference on cell-type proportions. The estimated proportions are shown to be asymptotically normal and the subject-specific covariance is estimated through a sandwich type estimator with an estimated gene-gene dependence set, though this subject-specific covariance estimator may not be consistent. In our implementation of `MEAD`, we supply the true signature gene matrix without measurement errors and the true gene-gene dependence set under each simulation setting. We compare the performance of these methods through evaluating the coverage probabilities of the confidence intervals constructed by these methods in our experiments. In Section A3.4, we also compare with `Bisque` (Jew et al. 2020) on the accuracy of estimating bulk expressions and cell type proportions.

3.1. Experiments with Simulated W and π_i's

We consider three cell types $K = 3$ and sample $\boldsymbol{\pi}_i$, the cell-type proportions in sample i, from $\boldsymbol{\pi}_i \sim \text{Dirichlet}(3, 2, 1)$. Under this setting, the three cell types have average proportions of 1/2, 1/3, and 1/6, respectively. The bulk gene expression for sample i is calculated as $\boldsymbol{y}_i = \sum_{k=1}^{K} \pi_{ik} \boldsymbol{x}_i^{(k)}$, where the expression profile $\boldsymbol{x}_i^{(k)}$ is simulated from $\boldsymbol{x}_i^{(k)} \sim \mathcal{N}(\boldsymbol{w}_k, \boldsymbol{\Sigma}^{(k)})$. Non-Gaussian distributions are considered in Section 3.3. Entries in $\boldsymbol{w}_k$ are iid from $\mathcal{N}(0, 1^2)$ and $\boldsymbol{\Sigma}^{(k)} = 10 \times \boldsymbol{R}^{(k)}$, where $\boldsymbol{R}^{(k)}$ is the correlation matrix in cell type k. We let

$$\boldsymbol{R}^{(1)} = \text{diag}(\boldsymbol{R}_1, \boldsymbol{R}_2, \boldsymbol{R}_2); \quad \boldsymbol{R}^{(2)} = \text{diag}(\boldsymbol{R}_2, \boldsymbol{R}_1, \boldsymbol{R}_2);$$
$$\boldsymbol{R}^{(3)} = \text{diag}(\boldsymbol{R}_2, \boldsymbol{R}_2, \boldsymbol{R}_1),$$

where $\boldsymbol{R}_1 \in \mathbb{R}^{\frac{p}{3} \times \frac{p}{3}}$ with $\boldsymbol{R}_{1,jj'} = 0.3$ and $\boldsymbol{R}_2 \in \mathbb{R}^{\frac{p}{3} \times \frac{p}{3}}$ with $\boldsymbol{R}_{2,jj'} = 0.7 \times 0.9^{|j-j'-1|}, j \neq j'$; see Figure 3 for an illustration. We let the number of signature genes $p = 300$ and the number of samples $n = 500$.

We apply `OLS`, `MEAD`, and `DECALS` to infer cell-type proportions for each subject. Specifically, we construct 95% confidence intervals for π_{ik}'s using each method and estimate the coverage probabilities using 100 data replicates. The results are summarized in Figure 4. It is seen that `DECALS` has the best performance for all three cell types, with coverage probabilities close to the nominal level of 95%. `OLS` tends to overestimate the CTS proportion variances and the resulting coverage probabilities for the 95% confidence intervals are consistently greater than 95%. This is majorly because the correlations among signature genes are ignored in `OLS`. As all CTS correlations are positive in this simulation setting, ignoring these positive correlations inflates the variance estimates of `OLS`, leading to an over-coverage of the `OLS` confidence intervals. `MEAD` tends to underestimate the variances for the estimated cell-type proportions, likely because the subject-specific covariance is not consistently estimated using the sandwich estimator. Additionally, we

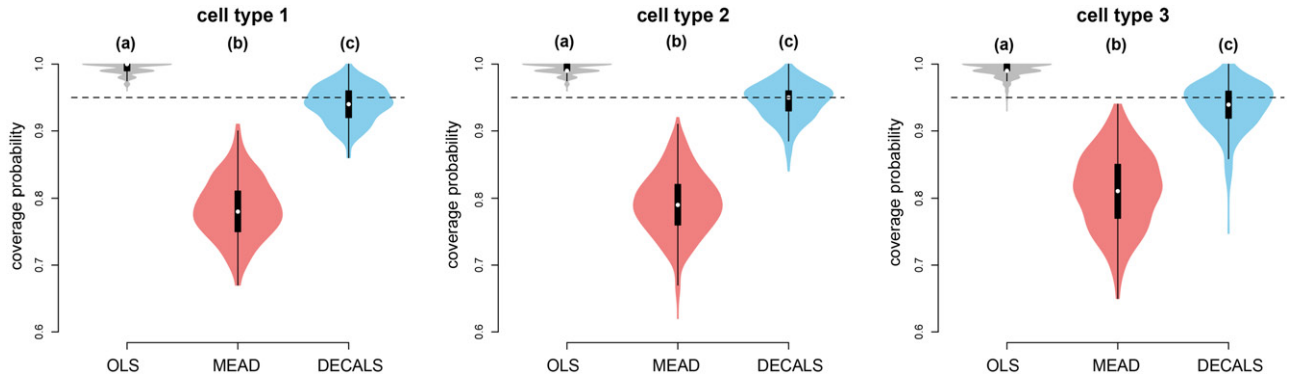

Figure 4. The coverage probabilities of 95% confidence intervals in three cell types with (a) OLS, (b) MEAD, and (c) DECALS.

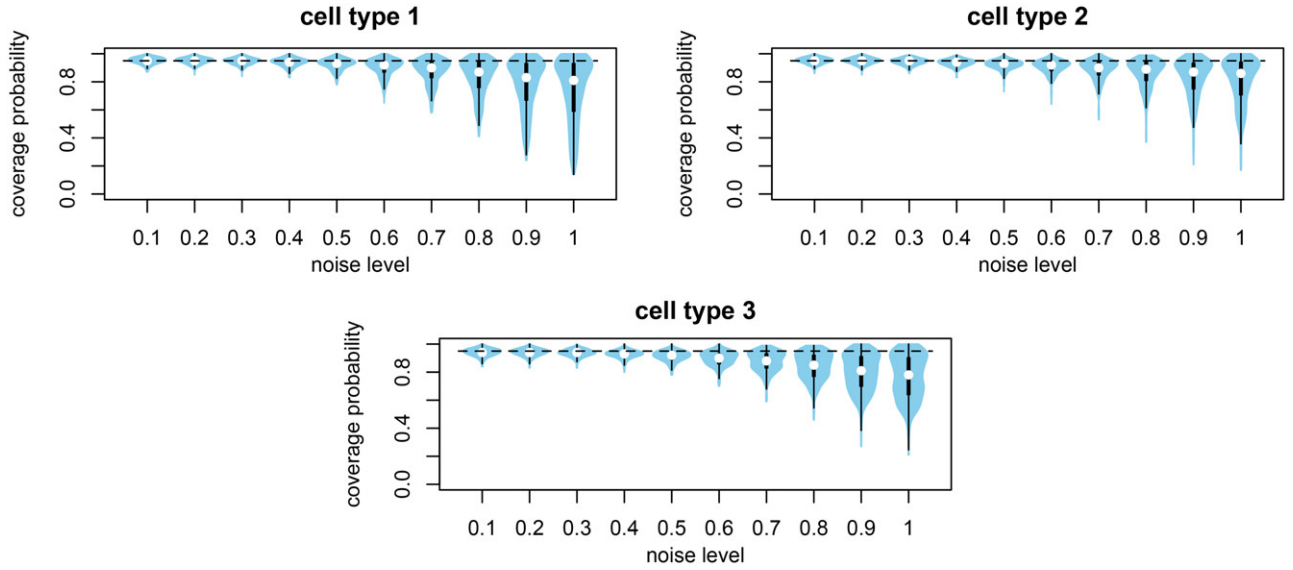

Figure 5. The coverage probabilities in three cell types as the noise level a_0 varies.

investigated how the sparsity levels of $\boldsymbol{\Sigma}_i$'s affect the performance of DECALS and the results are discussed in Section A2.7 in the supplement. Finally, we also investigate the estimation accuracy of $\boldsymbol{V}_i$ with varying p and $\boldsymbol{W}$ and the results are shown in Table S1 in the supplementary materials. The estimation accuracy of DECALS is satisfactory and it improves with the number of signature genes and the variance of signature gene expressions.

3.2. Sensitivity Analysis

In this section, we conduct a sensitivity analysis to examine the performance of DECALS when the signature gene expression matrix $\boldsymbol{W}$ is inaccurate and observed with errors. Consider the simulation settings in Section 3.1, where the mean signature gene expression is generated using $w_{kj} \overset{\text{iid}}{\sim} N(0, 1^2)$. In this sensitivity analysis, we assume that instead of w_{kj}, we observe $\tilde{w}_{kj} = w_{kj} + e_{kj}$, where $e_{kj} \overset{\text{iid}}{\sim} N(0, a_0^2)$ and a_0 varies between 0.1 and 1. Figure 5 reports the coverage probabilities of 95% confidence intervals with DECALS under various noise levels. It is seen that under this inaccurate signature gene matrix setting, DECALS still performs reasonably well, with the coverage probabilities close to the nominal level of 95% when a_0 is as large as 0.6.

3.3. Experiments with W and π_i's Inferred from Real Data

For experiments in this section, we use the signature gene matrix $\boldsymbol{W}$, cell-type proportions $\boldsymbol{\pi}_i$'s and CTS covariances $\boldsymbol{\Sigma}^{(k)}$'s inferred from the real data analysis in Section 4.2. There are $K = 5$ five cell types in this dataset, $n = 541$ bulk samples and $p = 159$ signature genes.

As the dataset in Section 4.2 uses expression unit FPKM (Trapnell et al. 2010) to measure gene expression, which is continuous and positive, we generate CTS expression profile $x_{ij}^{(k)}$'s from Gamma distributions. Specifically, given the mean $\boldsymbol{w}_k$ and target covariance $\boldsymbol{\Sigma}^{(k)}$ inferred from real data, we simulate $\boldsymbol{x}_i^{(k)}$ using a copula approach (see Section A2.6 in the supplement), similar to that in Tian, Wang, and Roeder (2021). We apply OLS, MEAD, RNA-Sieve and DECALS to infer cell-type proportions for each subject. Specifically, we construct 95% confidence intervals for π_{ik}'s by each method and estimate the coverage probabilities using 100 data replicates. The results are summarized in Figure 6. It is seen that DECALS has the best performance in all five cell types, with coverage probabilities close to the nominal level of 95%. Similar as before, MEAD tends to underestimate the variance for the estimated cell-type proportions, which results in confidence intervals with under-coverage. It is seen that RNA-Sieve also suffers from under-coverage, which

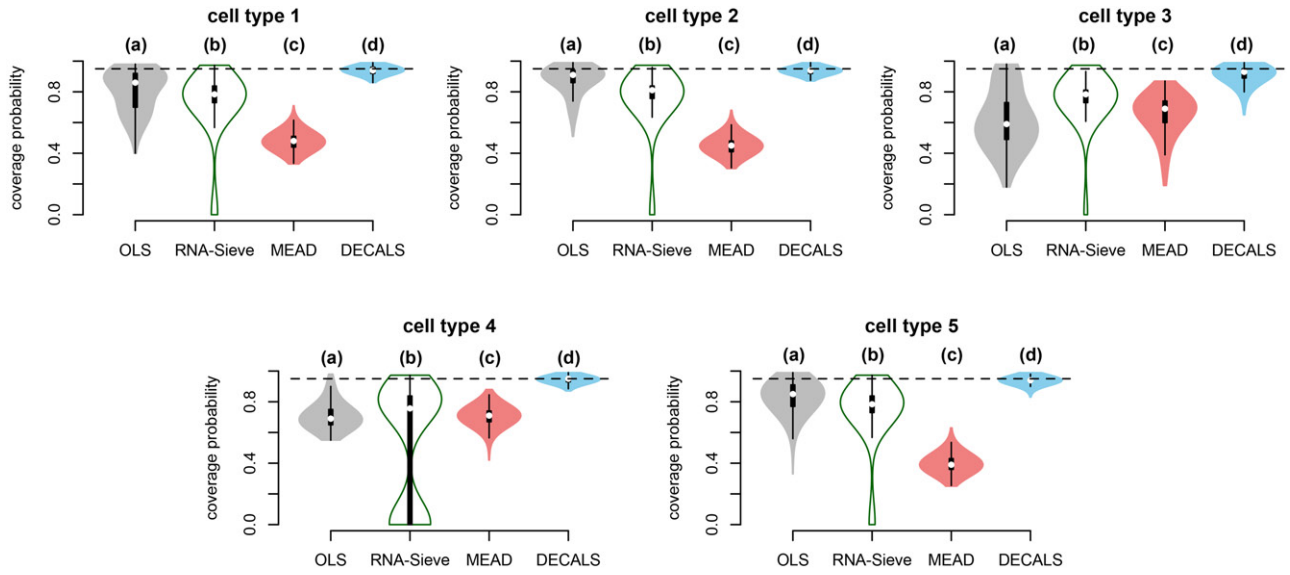

Figure 6. The coverage probabilities of 95% confidence intervals with (a) `OLS`, (b) `RNA-Sieve`, (c) `MEAD`, and (d) `DECALS`.

could be due to violations of key `RNA-Sieve` model assumptions. For example, `RNA-Sieve` assumes that the expression levels in different genes are independent and Gaussian distributed, which does not hold under this simulation setting. We also simulate gene expression data from Gaussian distributions. The results are similar and can be found in Section A2.5 in the supplement.

4. Using DECALS in CTS Analysis from Bulk Samples

4.1. A Sampling Approach to Incorporating Uncertainties

As mentioned above, most CTS analysis methods using bulk samples require cell-type proportions across samples as input, including methods that infer CTS gene expressions (Wang, Roeder, and Devlin 2021), CTS differentially expressed genes (Jin et al. 2021; Wang, Roeder, and Devlin 2021), CTS eQTLs (Patel et al. 2021; Little et al. 2022) and CTS co-expressions (Su, Zhang, and Zhao 2021). As the cell-type proportions used in these methods are not known but estimated from bulk sample data, incorporating the uncertainties in the estimated proportions with `DECALS` can mitigate the potential bias resulting from treating the proportions as known, a common assumption made in the existing CTS methods, and lead to more accurate and biologically more interpretable findings.

One possible approach to incorporating the inferred uncertainties for a specific CTS analysis method is to repeatedly sample the cell-type proportions from the distributions of $\hat{\boldsymbol{\pi}}_i$'s inferred from `DECALS` and perform analysis for each set of these sampled proportions. We can then summarize the results across these repeats. More specifically, we sample M sets of proportions denoted as $\{\hat{\boldsymbol{\pi}}_i^{[m]}\}_{1\le i\le n}$ for $m = 1,\ldots,M$. For each set of sampled proportions $\{\hat{\boldsymbol{\pi}}_i^{[m]}\}_{1\le i\le n}$, we apply the CTS analysis method and get an output, denoted as $\mathcal{S}^{[m]}$. Here $\mathcal{S}^{[m]}$ can be CTS gene expression estimates or a set of CTS differentially expresssed genes. With the results $\mathcal{S}^{[1]},\ldots,\mathcal{S}^{[M]}$ from M sets of sampled proportions, we can incorporate uncertainty from cell-type proportion estimates in the CTS analysis method via, for example, computing confidence intervals.

In Sections 4.2 and 4.3, we implement the above procedure by applying `DECALS` to two real datasets to infer uncertainties associated with cell-type proportion estimates, and incorporate these uncertainties in downstream analysis that identifies differentially expressed genes/transcripts in a specific cell type. Specifically, we combine `DECALS` with a downstream CTS analysis method in Wang, Roeder, and Devlin (2021), referred to as `bMIND`, that uses bulk sample data and cell-type proportions to identify CTS differentially expressed genes/transcripts. `bMIND` adopts a Bayesian approach to estimating CTS expressions from bulk RNA-seq data, which are then used to detect CTS differentially expressed genes/transcripts. The method takes the bulk RNA-seq data and cell-type proportions across samples as the input and outputs the set of genes/transcripts inferred to be differentially expressed in each cell type. We show that, by considering uncertainties of the estimated proportions in `bMIND`, we can get results that are more enriched for biologically relevant functions and more interpretable. Note that when estimating cell type proportions, `DECALS` uses signature genes and assumes they share the same mean expressions for subjects in different groups. This assumption on signature genes is evaluated in Section A3.5 of the supplement.

4.2. ROSMAP Data

We consider the bulk RNA-seq data collected from the ROSMAP study. Using the single-nucleus RNA-seq data from Mathys et al. (2019),[2] we applied the CIBERSORTx S-mode (Newman et al. 2019) to correct for batch effects and to obtain a candidate signature gene matrix for five major cell types, including neurons (Neu), oligodendrocytes (Oli), astrocytes (Ast), microglia (Mic), and endothelials (End). To ensure that the final selected signature genes had strong differential signals across these five cell types, we further took the intersection of this candidate gene set

[2] *https://www.synapse.org/#!Synapse:syn21261143*

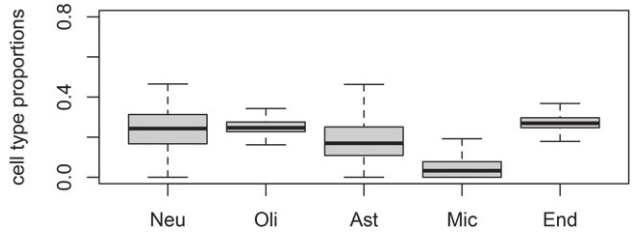

Figure 7. Estimated cell-type proportions across all samples in the ROSMAP data.

with the differentially expressed marker genes for each cell type from Mathys et al. (2019), which finally gave a signature matrix with $p = 159$ genes for the five cell types.

Given the bulk sample expressions and the signature gene matrix, we applied DECALS to estimate the cell-type proportions and their sampling distributions. Figure 7 presents the estimated proportions across 541 samples, which shows a good agreement with the cell-type abundances reported in Patrick et al. (2020) on a subset of ROSMAP samples. More discussions about CTS proportions can be found in Section A3 of the supplement.

Next, we focused a set of 9328 protein coding genes with FPKM$\geq$ 4 in at least half of the samples, and applied bMIND to identify CTS differentially expressed (DE) genes between Alzheimer's disease patients and controls. We considered two different approaches to inferring DE genes. The first approach directly applied bMIND with $\hat{\boldsymbol{\pi}}_i$'s estimated from (3) and calculated a p-value for each gene j in each cell type k, denoted as p_{jk}. In bMIND, gene j is considered a DE gene in cell type k if $p_{jk} < 0.05$. The second approach combines bMIND with DECALS as described in Section 4.1. More specifically, given the estimated sampling distributions of $\hat{\boldsymbol{\pi}}_i$'s from DECALS, we sampled 100 sets of proportions denoted as $\{\hat{\boldsymbol{\pi}}_i^{[m]}\}_{1\leq i\leq n}$ for $m = 1, \ldots, 100$. For each set of sampled proportions $\{\hat{\boldsymbol{\pi}}_i^{[m]}\}_{1\leq i\leq n}$, we applied bMIND and calculated the p-value for gene j in cell type k, denoted as $p_{jk}^{[m]}$. After 100 repeats, gene j was considered a DE gene in cell type k if $\sum_{m=1}^{100} 1\{p_{jk}^{(m)} < 0.05\} > 10$, where the cutoff value of 10 was calculated as two standard deviations above the expected value of $\sum_{m=1}^{100} 1\{p_{jk}^{(m)} < 0.05\}$ for a non DE gene. Specifically, the number of times a non DE gene is selected follows a Binomial(100,0.05), with a mean 5 and variance 4.75. We refer to these two approaches as bMIND and bMIND+DECALS, respectively. We focus on gene set enrichment analysis, which aims to capture coordinated expression changes of groups of genes instead of individual genes. To identify a set of DE genes for this purpose, a less stringent p-value cutoff, such as the threshold of 0.05 we used, is often employed (Labonté et al. 2017; Tian et al. 2020) and sometimes even all genes are included (Mootha et al. 2003).

To compare the DE gene sets identified from the two approaches for each of the five cell types, we performed enrichment analysis using QIAGEN Ingenuity Pathway Analysis (IPA, QIAGEN Inc., *https://digitalinsights.qiagen.com/IPA*). IPA identifies pathways enriched with DE genes by testing the association between the input genes and canonical pathways by first calculating the ratio of the number of genes in the input gene set that map to each pathway, and then using a Fisher's exact test to assess the statistical significance for the association between the input gene sets and canonical pathways (Krämer et al. 2013). We hypothesized that as bMIND+DECALS considered uncertainties in the cell-type proportion estimates, the inferred DE gene sets should be more enriched with biological signals as reflected from the IPA analysis. Because a larger gene set is likely more enriched for biological signals, when the gene sets inferred from bMIND and bMIND+DECALS differed in size, we only kept the top significant genes from the method with the larger gene set so that the resulting two gene sets had the same size in the enrichment analysis.

Enriched biological findings from bMIND+DECALS. Table 1 shows that the bMIND+DECALS approach implied a larger number of significant IPA canonical pathways than the bMIND approach in most cell types (see significant IPA canonical pathways in Section A9 in the supplementary materials). This result suggests that the DE gene sets identified from the bMIND+DECALS procedure can potentially offer more biological insights than those from the bMIND procedure. Moreover, a further investigation shows that bMIND+DECALS might better identify canonical pathways related to Alzheimer's disease. For instance, in oligodendrocyte (Oli), the Sumoylation pathway was only identified in bMIND+DECALS (Benjamini–Hochberg adjusted [BH] p-value = 3.24×10^{-7}). This pathway was previously reported to regulate amyloid precursor proteins, which are central to Alzheimer's disease (Li et al. 2003; Martin et al. 2007; Anderson et al. 2017). In astrocytes (Ast), the top three pathways identified in bMIND+DECALS were EIF2 Signaling (BH p-value = 1.16×10^{-25}), mTOR Signaling (BH p-value = 1.07×10^{-8}) and Regulation of eIF4 and p70S6K Signaling (BH p-value = 2.40×10^{-8}). These three pathways were previously reported to have associations with the development of Alzheimer's disease through meta-analysis (Yussof et al. 2020). Specifically, mTOR Signaling, which was already shown to be associated with Alzheimer's disease (Congdon and Sigurdsson 2018; Butterfield and Halliwell 2019), was not identified by bMIND. In microglia (Mic), Cholesterol Biosynthesis I (BH p-value = 2.63×10^{-5}), Cholesterol Biosynthesis II (BH p-value = 2.63×10^{-5}), Cholesterol Biosynthesis III (BH p-value = 2.63×10^{-5}) and Putrescine Degradation III (BH p-value = 7.08×10^{-5}) pathways were only identified by bMIND+DECALS. These pathways were reported to be related to amyloid-β peptide (Reitz, Brayne, and Mayeux 2011; Chun et al. 2020),

Table 1. Numbers of pathways selected in the IPA enrichment analysis.

	Neu	Oli	Ast	Mic	End
bMIND	0	0	6	19	0
bMIND+DECALS	1	7	5	55	0

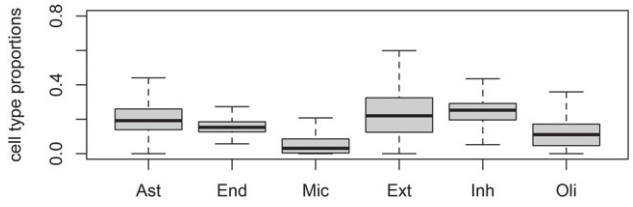

Figure 8. Estimated cell-type proportions across all samples in the GTEx data.

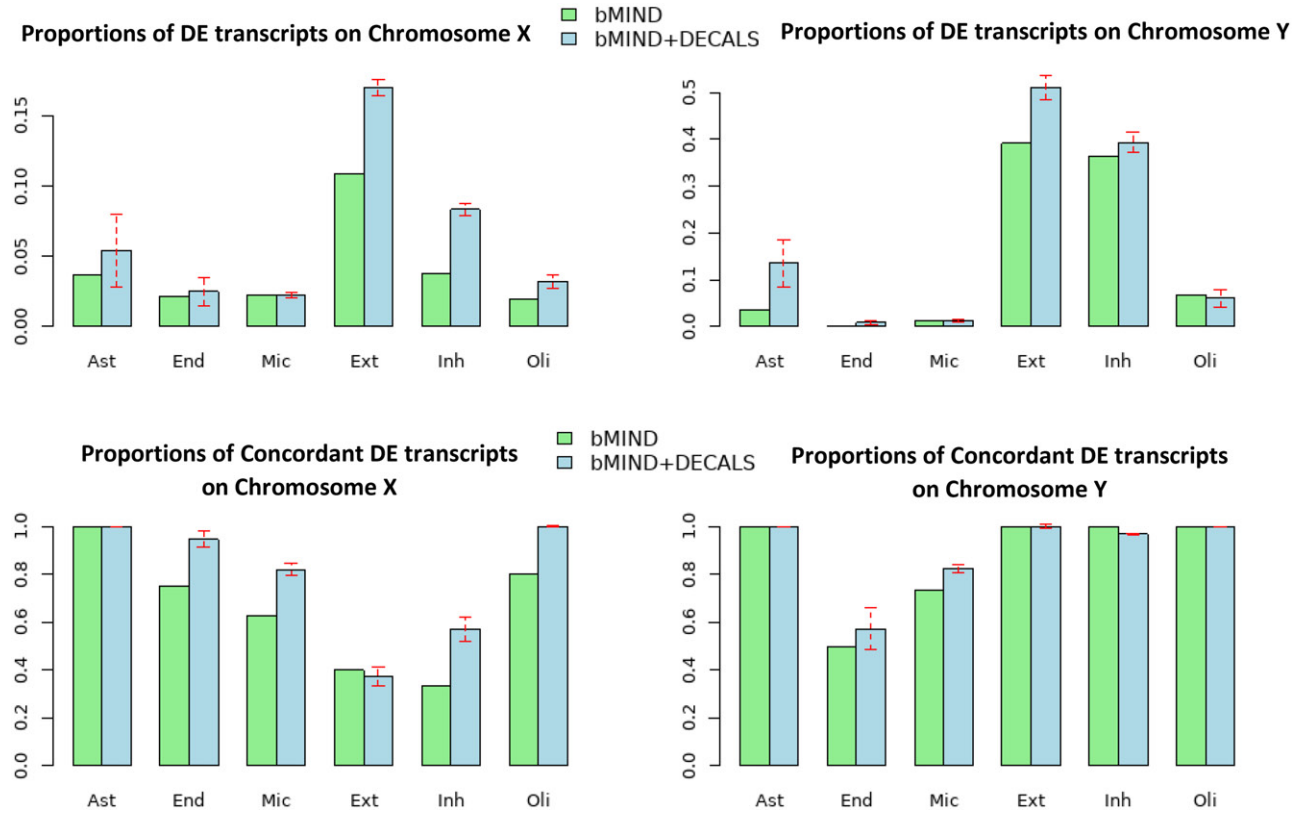

Figure 9. Identified CTS DE transcripts using the GTEx data with proportions of CTS DE transcripts on Chromosomes X and Y (top panel) and proportions of concordant CTS DE transcripts on Chromosomes X and Y, respectively (bottom panel). The red dashed lines mark mean $\pm$ 2$\times$standard errors.

which is important in Alzheimer's disease. These results further demonstrate the benefit of considering uncertainties in cell-type proportion estimates in identifying CTS DE genes.

4.3. GTEx Data

The Genotype-Tissue Expression (GTEx) project (Consortium 2020) is a major effort to collect gene expression data from postmortem donor samples at a number of non-diseased tissue sites. Our analysis focused on identifying CTS DE transcripts between males and female in brain tissues, as considered in Wang, Roeder, and Devlin (2021). We focused on $n = 1671$ brain samples from the GTEx project (*https://www.ncbi.nlm.nih.gov/gap/*), as considered in Wang, Roeder, and Devlin (2021). Among them, 509 were collected from female group and 1162 were collected from male group. The signature matrix was derived from Darmanis et al. (2015) with six cell types, including astrocyte (Ast), endothelial (End), microglia (Mic), excitatory (Ext) neuron, inhibitory (Inh) neuron and oligodendrocyte (Oli), and the same set of $p = 754$ signature genes as in Wang, Roeder, and Devlin (2021). Given the bulk sample expressions and the signature gene matrix in GTEx data, the estimated cell-type proportions from `DECALS` are shown in Figure 8, consistent with those reported in Wang, Roeder, and Devlin (2021). More discussions about CTS proportions can be found in Section A3.2 of the supplement.

More biologically interpretable findings from `bMIND+DECALS`. There are a total of 54,271 transcripts in the GTEx dataset and we consider all of them in our analysis. Following the procedure in Section 4.2, we identified CTS DE transcript sets using `bMIND` and `bMIND+DECALS`, respectively. We mapped these DE transcripts to all chromosomes including sex chromosomes X and Y. For a method that can better detect DE transcripts between males and females, we would expect larger proportions of the identified DE transcripts on the sex chromosomes. For a transcript set $\mathcal{A}$, denote $\mathcal{A}^X$ and $\mathcal{A}^Y$ as the subsets of $\mathcal{A}$ that are mapped to chromosomes X and Y, respectively. We calculate the proportions of DE transcripts that are mapped to the sex chromosomes as $|\mathcal{A}^X|/|\mathcal{A}|$ and $|\mathcal{A}^Y|/|\mathcal{A}|$, where $|\cdot|$ denotes the cardinality of a set. The top panel of Figure 9 shows that `bMIND+DECALS` has higher proportions of DE transcripts mapped to the sex chromosomes in most cell types than `bMIND`.

Next, we compare the concordance of DE transcripts on sex chromosomes. Specifically, it is expected that females will more likely have higher expression levels for DE transcripts on the X chromosome and males will more likely have higher expression levels for DE transcripts on the Y chromosome. Correspondingly, if a DE transcript on the X (Y) chromosome is over-expressed in females (males), we referred to it as a *concordant* DE transcript. For the set of DE transcripts $\mathcal{A}^X$ mapped to the chromosome X, we denote $\mathcal{A}^{XF}$ and $\mathcal{A}^{XM}$ as the subsets of $\mathcal{A}^X$ that are over-expressed in females and males, respectively. Similarly, for the set of DE transcripts $\mathcal{A}^Y$ mapped to the Y chromosome, we can define subsets $\mathcal{A}^{YF}$ and $\mathcal{A}^{YM}$. We calculate the proportions of concordant DE transcripts as $|\mathcal{A}^{XF}|/|\mathcal{A}^X|$ and $|\mathcal{A}^{YM}|/|\mathcal{A}^Y|$, respectively. The bottom panel of Figure 9 shows that the majority of CTS DE transcripts identified by

both methods are concordant. Moreover, bMIND+DECALS has higher proportions of concordant DE transcripts in most cell types, further suggesting considering the uncertainties in the estimated cell-type proportions can improve the detection of CTS DE transcripts.

5. Discussion

We have proposed a decorrelated constrained least squares (DECALS) framework that estimates cell-type proportions as well as their sampling distributions under a flexible statistical deconvolution framework that allows a general and subject-specific covariance of bulk gene expressions. We demonstrate through the analyses of bulk gene expression data from post mortem brain samples that considering the uncertainties in the estimated cell type proportions can lead to more enriched and interpretable biological findings in downstream CTS analysis. Our proposed method DECALS is flexible, easy to compute and can be combined with most CTS analysis methods using bulk samples, such as CTS gene expression and co-expression estimation, CTS DE gene and eQTL identification, to improve the accuracy and interpretability of the results.

Although our procedure estimates the covariance $\boldsymbol{\Sigma}_i$ for each individual, it in fact borrows information across individuals when making such estimates. Specifically, based on our model, $\boldsymbol{\Sigma}_i$ is calculated as the weighted sum of cell-type-specific covariances $\boldsymbol{\Sigma}^{(k)}$'s, that is, $\boldsymbol{\Sigma}_i = \sum_k \pi_{ik}^2 \boldsymbol{\Sigma}^{(k)}$. When estimating $\boldsymbol{\Sigma}^{(k)}$, the covariance in cell type k, our procedure uses the bulk expressions from all subjects in the moment-based estimation. That is, the cell-type-specific covariances are estimated by pooling information across all subjects, which are then used to estimate $\boldsymbol{\Sigma}_i$'s.

In Section A3.8, we discuss the selection of signature genes in real data analysis and evaluate the sensitivity to signature gene sets using simulation studies. Next, our approach assumes that the mean expression levels of signature genes in $\boldsymbol{W}$ are given. In cell type deconvolution analysis, $\boldsymbol{W}$ is usually gathered from pure cell types (Newman et al. 2015; Li et al. 2016) or single cell RNA-sequencing data (Wang et al. 2019; Newman et al. 2019; Jew et al. 2020). Our empirical investigations showed that DECALS is not sensitive to errors in $\boldsymbol{W}$ (see Section 3.2). As a future direction, it is possible to further extend our framework to accommodate a noisy $\boldsymbol{W}$ by formulating (2) as a measurement error model, similar to that in Xie and Wang (2022). The errors in $\boldsymbol{W}$ can possibly be quantified via modeling the scRNA-seq data. We leave the full investigation of this topic as future research.

Supplementary Materials

The online supplemental materials include proofs of theorems, additional simulation results, and details of the real data analysis.

Acknowledgments

We thank the ROSMAP team for their permission, requested at *https://www.radc.rush.edu*, to access the bulk RNA-seq and single nueclues RNA-seq data in the project.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

The ROSMAP project is supported by the following grants: P30AG72975, P30AG010161 (ADCC), R01AG015819 (RISK), R01AG017917 (MAP), U01AG46152 (AMP-AD Pipeline I) and U01AG61356 (AMP-AD Pipeline II). The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Zhang was supported by NSF grants DMS 2210469 and 2329296. Zhao was supported in part by NIH grants R01 GM134005, R56 AG074015, and U024 HG012108.

ORCID

Chang Su http://orcid.org/0000-0002-8704-1512

References

Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z., and Clark, H. F. (2009), "Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus," *PloS One*, 4, e6098. [2521,2523]

Anderson, D. B., Zanella, C. A., Henley, J. M., and Cimarosti, H. (2017), "Sumoylation: Implications for Neurodegenerative Diseases," in *SUMO Regulation of Cellular Processes*, ed. V. G. Wilson, pp. 261–281, Cham: Springer . [2529]

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011), "Network Medicine: A Network-based Approach to Human Disease," *Nature Reviews Genetics*, 12, 56–68. [2521]

Bennett, D. A., Buchman, A. S., Boyle, P. A., Barnes, L. L., Wilson, R. S., and Schneider, J. A. (2018), "Religious Orders Study and Rush Memory and Aging Project," *Journal of Alzheimer's Disease*, 64, S161–S189. [2522]

Butterfield, D. A., and Halliwell, B. (2019), "Oxidative Stress, Dysfunctional Glucose Metabolism and Alzheimer Disease," *Nature Reviews Neuroscience*, 20, 148–160. [2529]

Chun, H., Im, H., Kang, Y. J., Kim, Y., Shin, J. H., Won, W., Lim, J., Ju, Y., Park, Y. M., Kim, S., et al. (2020), "Severe Reactive Astrocytes Precipitate Pathological Hallmarks of Alzheimer's Disease via H2O2- Production," *Nature Neuroscience*, 23, 1555–1566. [2529]

Congdon, E. E., and Sigurdsson, E. M. (2018), "Tau-Targeting Therapies for Alzheimer Disease," *Nature Reviews Neurology*, 14, 399–415. [2529]

Consortium, G. (2020), "The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues," *Science*, 369, 1318–1330. [2530]

Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., Hayden Gephart, M. G., Barres, B. A., and Quake, S. R. (2015), "A Survey of Human Brain Transcriptome Diversity at the Single Cell Level," *Proceedings of the National Academy of Sciences*, 112, 7285–7290. [2521,2530]

De Strooper, B., and Karran, E. (2016), "The Cellular Phase of Alzheimer's Disease," *Cell*, 164, 603–615. [2522]

Erdmann-Pham, D. D., Fischer, J., Hong, J., and Song, Y. S. (2021), "Likelihood-based Deconvolution of Bulk Gene Expression Data Using Single-Cell References," *Genome Research*, 31, 1794–1806. [2522,2525]

Fujita, M., Gao, Z., Zeng, L., McCabe, C., White, C. C., Ng, B., Green, G. S., Rozenblatt-Rosen, O., Phillips, D., Amir-Zilberstein, L., et al. (2022), "Cell-Subtype Specific Effects of Genetic Variation in the Aging and Alzheimer Cortex," *bioRxiv*, 2022–11. [2524]

Goldfarb, D., and Idnani, A. (1982), "Dual and Primal-Dual Methods for Solving Strictly Convex Quadratic Programs," in *Numerical Analysis*, ed. J. P. Hennart, pp. 226–239, Berlin: Springer. [2523]

——— (1983), "A Numerically Stable Dual Method for Solving Strictly Convex Quadratic Programs," *Mathematical Programming*, 27, 1–33. [2523]

Greene, W. H. (2003), *Econometric Analysis*, Delhi: Pearson Education India. [2525]

Hekselman, I., and Yeger-Lotem, E. (2020), "Mechanisms of Tissue and Cell-Type Specificity in Heritable Traits and Diseases," *Nature Reviews Genetics*, 21, 137–150. [2521]

Jaakkola, M. K., and Elo, L. L. (2022), "Estimating Cell Type-Specific Differential Expression Using Deconvolution," *Briefings in Bioinformatics*, 23, bbab433. [2522]

Jew, B., Alvarez, M., Rahmani, E., Miao, Z., Ko, A., Garske, K. M., Sul, J. H., Pietiläinen, K. H., Pajukanta, P., and Halperin, E. (2020), "Accurate Estimation of Cell Composition in Bulk Expression through Robust Integration of Single-Cell Information," *Nature Communications*, 11, 1–11. [2521,2523,2526,2531]

Jin, C., Chen, M., Lin, D.-Y., and Sun, W. (2021), "Cell-Type-Aware Analysis of RNA-seq Data," *Nature Computational Science*, 1, 253–261. [2521,2522,2528]

Krämer, A., Green, J., Pollard, Jack, J., and Tugendreich, S. (2013), "Causal Analysis Approaches in Ingenuity Pathway Analysis," *Bioinformatics*, 30, 523–530. [2529]

Labonté, B., Engmann, O., Purushothaman, I., Menard, C., Wang, J., Tan, C., Scarpa, J. R., Moy, G., Loh, Y.-H. E., Cahill, M., et al. (2017), "Sex-Specific Transcriptional Signatures in Human Depression," *Nature Medicine*, 23, 1102–1111. [2529]

Li, B., Severson, E., Pignon, J.-C., Zhao, H., Li, T., Novak, J., Jiang, P., Shen, H., Aster, J. C., Rodig, S., et al. (2016), "Comprehensive Analyses of Tumor Immunity: Implications for Cancer Immunotherapy," *Genome Biology*, 17, 1–16. [2523,2531]

Li, H., Zhu, B., Xu, Z., Adams, T., Kaminski, N., and Zhao, H. (2021), "A Markov Random Field Model for Network-based Differential Expression Analysis of Single-Cell RNA-Seq Data," *BMC Bioinformatics*, 22, 524. [2521]

Li, Y., Wang, H., Wang, S., Quon, D., Liu, Y.-W., and Cordell, B. (2003), "Positive and Negative Regulation of APP Amyloidogenesis by Sumoylation," *Proceedings of the National Academy of Sciences*, 100, 259–264. [2529]

Little, P., Zhabotynsky, V., Li, Y., Lin, D., and Sun, W. (2022), "Cell Type-Specific Expression Quantitative Trait Loci," *bioRxiv*. [2521,2528]

Martin, S., Wilkinson, K. A., Nishimune, A., and Henley, J. M. (2007), "Emerging Extranuclear Roles of Protein SUMOylation in Neuronal Function and Dysfunction," *Nature Reviews Neuroscience*, 8, 948–959. [2529]

Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J. Z., Menon, M., He, L., Abdurrob, F., Jiang, X., et al. (2019), "Single-Cell Transcriptomic Analysis of Alzheimerâ's Disease," *Nature*, 570, 332–337. [2522,2528,2529]

Montembeault, M., Rouleau, I., Provost, J.-S., and Brambati, S. M. (2016), "Altered Gray Matter Structural Covariance Networks in Early Stages of Alzheimer's Disease," *Cerebral cortex*, 26, 2650–2662. [2522]

Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstr*r*ale, M., Laurila, E., et al. (2003), "PGC-1α-Responsive Genes Involved in Oxidative Phosphorylation are Coordinately Downregulated in Human Diabetes," *Nature Genetics*, 34, 267–273. [2529]

Mostafavi, S., Gaiteri, C., Sullivan, S. E., White, C. C., Tasaki, S., Xu, J., Taga, M., Klein, H.-U., Patrick, E., Komashko, V., et al. (2018), "A Molecular Network of the Aging Human Brain Provides Insights into the Pathology and Cognitive Decline of Alzheimer's Disease," *Nature Neuroscience*, 21, 811–819. [2521,2522]

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M., and Alizadeh, A. A. (2015), "Robust Enumeration of Cell Subsets from Tissue Expression Profiles," *Nature Methods*, 12, 453–457. [2521,2523,2531]

Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., Khodadoust, M. S., Esfahani, M. S., Luca, B. A., Steiner, D., et al. (2019), "Determining Cell Type Abundance and Expression from Bulk Tissues with Digital Cytometry," *Nature Biotechnology*, 37, 773–782. [2522,2523,2528,2531]

Patel, D., Zhang, X., Farrell, J. J., Chung, J., Stein, T. D., Lunetta, K. L., and Farrer, L. A. (2021), "Cell-Type-Specific Expression Quantitative Trait Loci Associated with Alzheimer Disease in Blood and Brain Tissue," *Translational Psychiatry*, 11, 1–17. [2521,2528]

Patrick, E., Taga, M., Ergun, A., Ng, B., Casazza, W., Cimpean, M., Yung, C., Schneider, J. A., Bennett, D. A., Gaiteri, C., et al. (2020), "Deconvolving the Contributions of Cell-Type Heterogeneity on Cortical Gene Expression," *PLoS Computational Biology*, 16, e1008120. [2529]

Reitz, C., Brayne, C., and Mayeux, R. (2011), "Epidemiology of Alzheimer Disease," *Nature Reviews Neurology*, 7, 137–152. [2529]

Rothman, A. J., Levina, E., and Zhu, J. (2009), "Generalized Thresholding of Large Covariance Matrices," *Journal of the American Statistical Association*, 104, 177–186. [2525]

Salat, D. H., Kaye, J. A., and Janowsky, J. S. (2001), "Selective Preservation and Degeneration Within the Prefrontal Cortex in Aging and Alzheimer Disease," *Archives of Neurology*, 58, 1403–1408. [2522]

Sims, R., Hill, M., and Williams, J. (2020), "The Multiplex Model of the Genetics of Alzheimer's Disease," *Nature Neuroscience*, 23, 311–322. [2522]

Su, C., Zhang, J., and Zhao, H. (2021), "CSNet: Estimating Cell-Type-Specific Gene Co-Expression Networks from Bulk Gene Expression Data," bioRxiv. [2521,2524,2528]

Tang, D., Park, S., and Zhao, H. (2020), "NITUMID: Nonnegative Matrix Factorization-based Immune-TUmor MIcroenvironment Deconvolution," *Bioinformatics*, 36, 1344–1350. [2521,2523]

——— (2022), "SCADIE: Simultaneous Estimation of Cell Type Proportions and Cell Type-Specific Gene Expressions Using SCAD-based Iterative Estimating Procedure," *Genome Biology*, 23, 1–23. [2521,2522]

Tian, J., Wang, J., and Roeder, K. (2021), "ESCO: Single Cell Expression Simulation Incorporating Gene Co-expression," *Bioinformatics*, 37, 2374–2381. [2527]

Tian, W., Zhang, N., Jin, R., Feng, Y., Wang, S., Gao, S., Gao, R., Wu, G., Tian, D., Tan, W., et al. (2020), "Immune Suppression in the Early Stage of COVID-19 Disease," *Nature Communications*, 11, 5859. [2529]

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012), "Differential Gene and Transcript Expression Analysis of RNA-seq Experiments with TopHat and Cufflinks," *Nature Protocols*, 7, 562–578. [2521]

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010), "Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching During Cell Differentiation," *Nature Biotechnology*, 28, 511–515. [2522,2523,2527]

Wang, J., Roeder, K., and Devlin, B. (2021), "Bayesian Estimation of Cell Type-Specific Gene Expression with Prior Derived from Single-Cell Data," *Genome Research*, 31,1807–1818. [2521,2522,2528,2530]

Wang, X., Park, J., Susztak, K., Zhang, N. R., and Li, M. (2019), "Bulk Tissue Cell Type Deconvolution with Multi-Subject Single-Cell Expression Reference," *Nature Communications*, 10, 1–9. [2521,2523,2531]

Winblad, B., Amouyel, P., Andrieu, S., Ballard, C., Brayne, C., Brodaty, H., Cedazo-Minguez, A., Dubois, B., Edvardsson, D., Feldman, H., et al. (2016), "Defeating Alzheimer's Disease and Other Dementias: A Priority for European Science and Society," *The Lancet Neurology*, 15, 455–532. [2522]

Xie, D., and Wang, J. (2022), "Robust Statistical Inference for Cell Type Deconvolution," arXiv preprint arXiv:2202.06420. [2522,2525,2526,2531]

Yang, T., Alessandri-Haber, N., Fury, W., Schaner, M., Breese, R., LaCroix-Fralish, M., Kim, J., Adler, C., Macdonald, L. E., Atwal, G. S., et al. (2021), "AdRoit is an Accurate and Robust Method to Infer Complex Transcriptome Composition," *Communications Biology*, 4, 1–14. [2521,2523]

Yussof, A., Yoon, P., Krkljes, C., Schweinberg, S., Cottrell, J., Chu, T., and Chang, S. L. (2020), "A Meta-Analysis of the Effect of Binge Drinking on the Oral Microbiome and its Relation to Alzheimer's Disease," *Scientific Reports*, 10, 19872. [2529]

Zhang, B., Gaiteri, C., Bodea, L.-G., Wang, Z., McElwee, J., Podtelezhnikov, A. A., Zhang, C., Xie, T., Tran, L., Dobrin, R., et al. (2013), "Integrated Systems Approach Identifies Genetic Nodes and Networks in Late-Onset Alzheimer's Disease," *Cell*, 153, 707–720. [2521]

Zhang, B., and Horvath, S. (2005), "A General Framework for Weighted Gene Co-expression Network Analysis," *Statistical Applications in Genetics and Molecular Biology*, 4, 17. [2521,2525]

Zhang, J., and Li, Y. (2022), "High-Dimensional Gaussian Graphical Regression Models with Covariates," *Journal of the American Statistical Association*, 118, 2088–2100. [2521]

Zhu, B., Li, H., Zhang, L., Chandra, S. S., and Zhao, H. (2022), "A Markov Random Field Model-based Approach for Differentially Expressed Gene Detection from Single-Cell RNA-Seq Data," *Briefings in Bioinformatics*, 23, bbac166. [2521]