JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 119, NO. 546, 811–824: Applications and Case Studies
https://doi.org/10.1080/01621459.2023.2297467

Estimating Cell-Type-Specific Gene Co-Expression Networks from Bulk Gene Expression Data with an Application to Alzheimer's Disease

Chang Su[a,b], Jingfei Zhang[c], and Hongyu Zhao[b]

[a]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA; [b]Department of Biostatistics, Yale University, New Haven, CT; [c]Information Systems and Operations Management, Emory University, Atlanta, GA

ABSTRACT

Inferring and characterizing gene co-expression networks has led to important insights on the molecular mechanisms of complex diseases. Most co-expression analyses to date have been performed on gene expression data collected from bulk tissues with different cell type compositions across samples. As a result, the co-expression estimates only offer an aggregated view of the underlying gene regulations and can be confounded by heterogeneity in cell type compositions, failing to reveal gene coordination that may be distinct across different cell types. In this article, we introduce a flexible framework for estimating cell-type-specific gene co-expression networks from bulk sample data, without making specific assumptions on the distributions of gene expression profiles in different cell types. We develop a novel sparse least squares estimator, referred to as CSNet, that is efficient to implement and has good theoretical properties. Using CSNet, we analyzed the bulk gene expression data from a cohort study on Alzheimer's disease and identified previously unknown cell-type-specific co-expressions among Alzheimer's disease risk genes, suggesting cell-type-specific disease mechanisms. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received January 2022
Accepted December 2023

KEYWORDS
Bulk RNA-seq;
Cell-type-specific analysis;
Deconvolution; Gene co-expression networks;
Sparse covariance estimation

1. Introduction

Alzheimer's disease is a neurodegenerative disorder that causes progressive and irreversible loss of neurons in the brain (Winblad et al. 2016). It is estimated to affect 5.8 million people in the United States and has become the fifth leading cause of death among Americans over 64 years old (Alzheimer's Association 2019). Genetic factors are known to play an important role in Alzheimer's disease, with an estimated heritability of 58%–79% for late-onset Alzheimer's disease, and large scale genome wide association studies (GWAS) have implicated dozens of regions of the human genome for their relevance for Alzheimer's disease (Sims, Hill, and Williams 2020). To understand the mechanisms of these disease associated risk genes and the pathogenesis of Alzheimer's disease, gene co-expression networks have been widely employed (Zhang et al. 2013; Wang et al. 2016; Mostafavi et al. 2018; Meng and Mei 2019; Wan et al. 2020; Wang et al. 2021a).

Gene co-expression networks characterize correlations of gene expression levels across biological samples and co-expressed genes may be regulated by the same transcription factors, functionally related, or involved in the same pathways (Gaiteri et al. 2014). Gene co-expression networks have been extensively used to identify functional modules of genes and pathways, which were further associated with disease phenotypes (Wang et al. 2016; Mostafavi et al. 2018; Meng and Mei 2019; Wang et al. 2021a). For example, a gene co-expression analysis in Zhang et al. (2013) identified TYROBP as a key regulator in an immune related module upregulated in late-onset Alzheimer's disease, which was recapitulated in vivo in mice. In this article, the links in a co-expression network characterize associations of gene expression levels across samples and do not necessarily imply gene regulatory relationships.

While the literature on gene co-expression networks is rapidly growing, most co-expression analyses to date have been performed on data collected from bulk tissue samples that aggregated the expression profiles from different cell types. As a result, the estimated co-expression networks only offer an aggregated view of the underlying gene regulations, while gene regulations may differ considerably in different cell types (Heintzman et al. 2009), and the co-expression estimates can be dominated by signals from the more abundant cell types. Moreover, as different bulk samples may have different cell type compositions, the observed co-expressions may be confounded with cell type proportions. For example, consider two genes that are both highly expressed in one cell type but are not co-expressed at the cell type level. Bulk sample data may show that these two genes co-express as their expression levels co-vary with the proportion of this specific cell type in the bulk sample. Meanwhile, increasing evidence suggests cell-type-specific pathogenesis of Alzheimer's disease (De Strooper and Karran 2016). For example, neuroinflammation represents a key causal pathway in Alzheimer's disease and involves primarily glial cells in the brain including microglia and astrocytes (Heneka et al. 2015); myelination is also implicated in the disease, which is

CONTACT Jingfei Zhang emma.zhang@emory.edu Information Systems and Operations Management, Emory University, Atlanta, GA; Hongyu Zhao hongyu.zhao@yale.edu Department of Biostatistics, Yale University, New Haven, CT.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

mainly contributed by oligodendrocytes (Cai and Xiao 2016). To gain a more accurate and comprehensive view of the biological processes underlying Alzheimer's disease, a better approach is to estimate cell-type-specific co-expression networks.

Cell-type-specific co-expressions can possibly be estimated from single cell RNA sequencing (RNA-seq) data (Hwang, Lee, and Bang 2018) that measure expression profiles in single cells. However, these data are much more limited in the number of biological samples analyzed (Stower 2019), have high noises and sparsity due to low coverage and biological noises (Kiselev, Andrews, and Hemberg 2019), and may be biased due to cell isolation and sequencing protocols (Denisenko et al. 2020). For example, a recent single cell study on Alzheimer's disease has profiled 18 samples and conducted a gene co-expression analysis after aggregating the sparse expression data over single cells (Morabito et al. 2021). The co-expression analysis findings were only made for oligodendrocytes and did not cover astrocytes and microglia which are arguably more closely related to the Alzheimer's disease mechanisms, potentially due to their lower proportions in the brain and the high sparsity of single cell data. Real data results in Section 4 show that cell-type-specific co-expressions estimated from single cell RNA-seq data are noisy and co-expressions are often only seen in the highly abundant cell types. Instead of resorting to single cell data for estimating cell-type-specific co-expressions in Alzheimer's disease, we consider the use of bulk sample data in this article.

1.1. The ROSMAP Study on Alzheimer's Disease

Our work focuses on the bulk RNA-seq data from the Religious Orders Study and Rush Memory and Aging Project (Bennett et al. ROSMAP; 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 grey matter of dorsolateral prefrontal cortex, a brain region heavily implicated in Alzheimer's disease pathology. Single cell data from the same study found eight common brain cell types in these tissue samples (Mathys et al. 2019) including excitatory neuron, astrocyte, oligodendrocyte, microglia, inhibitory neuron, endothelial cell, pericyte, oligodendrocyte precursor cell, where the average estimated proportions for the first four major cell types in the bulk RNA-seq samples are 0.50, 0.20, 0.19, 0.08, respectively, adding together to 97% (see Section 4).

While the large sample size of the ROSMAP study facilitates a better estimation of gene co-expressions in the brain (Mostafavi et al. 2018), co-expressions directly estimated via correlations of the aggregated bulk samples are subject to the issues discussed before, including confounding by cell type proportions and estimates dominated by signals from the more abundant cell types. In particular, these estimates may fail to reveal the biological functions and pathways in different brain cell types that underlie the pathogenesis of Alzheimer's disease. This motivates our study that aims to estimate cell-type-specific co-expression networks from bulk RNA-seq data, by leveraging the estimated cell type proportions in bulk samples (Newman et al. 2019), and to examine the estimated networks in cell types relevant to Alzheimer's disease for a better understanding of disease mechanisms.

1.2. Existing Methods and Our Approach

There is a recently growing literature on decomposing bulk gene expression profiles into cell-type-specific profiles (Cobos et al. 2020). Using the bulk data, various methods are available to infer mean expression levels in each cell type (Newman et al. 2019) and to infer cell type proportions (Abbas et al. 2009; Wang et al. 2019; Newman et al. 2019; Tang, Park, and Zhao 2020; Jew et al. 2020; Yang et al. 2021). More recently, methods have been proposed to infer cell-type-specific expressions in each sample, such as CIBERSORTx (Newman et al. 2019), bMIND (Wang, Roeder, and Devlin 2021), and ENIGMA (Wang et al. 2021b). For each cell type, these methods offer an indirect way to estimate the co-expressions, by calculating the correlations of estimated expression profiles across samples. However, we show in Sections 3 and 4 that these methods rely on either restrictive assumptions, or high-quality external information that is not readily available in practice.

In our work, we consider a different statistical approach and propose a flexible method to estimate **C**ell-type-**S**pecific gene co-expression **Net**works using bulk gene expression data, and call this method `CSNet`. Specifically, we formulate the problem as estimating the means and covariances of unknown densities from different cell types using data (i.e., bulk samples) generated from a convolution of these densities with varying compositions. Our method `CSNet` does not make specific assumptions on the distributions of expression levels from different cell types, and it overcomes the computational challenge in estimating the covariances in a convolution of densities, especially when the number of genes is large, through a novel least squares approach that is efficient to implement and has good theoretical properties. We further propose a sparse estimator with SCAD penalty in the high dimension regime where the number of genes p can far exceed the sample size n.

Using `CSNet`, we analyzed the bulk RNA-seq data from the ROSMAP study and estimated gene co-expression networks for eight common cell types in the brain, including four major cell types, excitatory neuron, oligodendrocyte, astrocyte and microglia, on genes with known genetic risk for Alzheimer's disease, where modules of risk genes that uniquely co-express in astrocytes and microglia were uncovered. In contrast, the estimator based on cell-type-specific expression levels imputed by Wang, Roeder, and Devlin (2021) generated co-expression estimates that were similar in all cell types and did not identify any cell-type-specific modules (see Figure S10(a)). Both astrocyte and microglia are cell types that are less abundant (less than 20%), and the co-expressions estimated from single cell RNA-seq data showed no correlations in these two cell types (see Figure S10(b)). We have also considered gene sets that function primarily in specific cell types to validate `CSNet` and conducted several sensitivity analyses to further validate our results.

The rest of the article is organized as follows. Section 2 introduces the data problem and discusses estimating cell-type-specific co-expressions from bulk samples using `CSNet`. Section 3 reports the simulation results, and Section 4 conducts an analysis of cell-type-specific gene co-expression networks on gene sets with known cell-type-specific functions and on an Alzheimer's disease risk gene set using bulk RNA-seq data

from the ROSMAP study. Section 5 concludes the article with discussions.

2. Model and Estimation

2.1. Problem Formulation

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

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

where π_{ik} and $\boldsymbol{x}_i^{(k)}$ represent the proportion and expression profile of the kth cell type in the ith sample, respectively. Let $\boldsymbol{x}_i^{(k)}$ be independent from a multivariate distribution with mean $\boldsymbol{\mu}^{(k)} \in \mathbb{R}^p$ and covariance $\Sigma^{(k)} \in \mathbb{R}^{p\times p}$, where $\boldsymbol{\mu}^{(k)}$ and $\Sigma^{(k)}$ characterize the cell-type-specific mean gene expression and co-expression, respectively. Correspondingly, we can write

$$\begin{aligned} \mathbb{E}(\boldsymbol{x}_i) &= \sum_{k=1}^{K} \pi_{ik} \boldsymbol{\mu}^{(k)}, \\ \mathrm{Cov}(\boldsymbol{x}_i) &= \sum_{k=1}^{K} \pi_{ik}^2 \Sigma^{(k)} + \sum_{k\neq k'}^{K} \pi_{ik}\pi_{ik'} \Sigma^{(k,k')}, \end{aligned} \tag{2}$$

where $\Sigma^{(k,k')} = \mathrm{cov}(\boldsymbol{x}_i^{(k)}, \boldsymbol{x}_i^{(k')})$. As the gene regulation mechanisms in functionally distinct cell types are different (Heintzman et al. 2009), it is expected that the within-cell-type covariance $\sum_{k=1}^{K} \pi_{ik}^2 \Sigma^{(k)}$ is much larger than the across-cell-type covariance $\sum_{k\neq k'}^{K} \pi_{ik}\pi_{ik'} \Sigma^{(k,k')}$. This assumption is supported by our real data analysis using a large single nucleus RNA-seq data from brain tissues in Fujita et al. (2022) with 436 individuals (see Section S2). Hence, to reduce model complexity, we assume $\sum_{k\neq k'}^{K} \pi_{ik}\pi_{ik'} \Sigma^{(k,k')} = 0$ and consider

$$\mathbb{E}(\boldsymbol{x}_i) = \sum_{k=1}^{K} \pi_{ik} \boldsymbol{\mu}^{(k)}, \qquad \mathrm{Cov}(\boldsymbol{x}_i) = \sum_{k=1}^{K} \pi_{ik}^2 \Sigma^{(k)}. \tag{3}$$

Also using the dataset in Fujita et al. (2022), we demonstrate that our method is not sensitive to the misspecification of assuming the across-cell-type covariance $\sum_{k\neq k'}^{K} \pi_{ik}\pi_{ik'} \Sigma^{(k,k')} = 0$. Details of this analysis are included in Section S2. As $\Sigma^{(k)}$ does not relate directly to the strength of gene co-expressions, due to heterogeneity in variances, we further consider correlation matrices in our analysis denoted as $\boldsymbol{R}^{(k)} = \boldsymbol{D}_k^{-1/2} \Sigma^{(k)} \boldsymbol{D}_k^{-1/2}$, where $\boldsymbol{D}_k$ is a $p \times p$ diagonal matrix with the same diagonal as $\Sigma^{(k)}$.

Denote $[m] = \{1, 2, \ldots, m\}$ for a positive integer m. We shall assume the cell type proportions π_{ik}'s in (3) are given in the ensuing development, and later demonstrate in experiments that our method is not sensitive to biases and errors in π_{ik}'s; see results in Sections 3.2, A5.1 and discussions in Section 5. To infer π_{ik}'s from bulk samples, many methods have been developed that use cell type marker genes (i.e., genes that are only highly expressed in one cell type of interest) with expression profiles gathered from pure cell types (Newman et al. 2015; Li et al. 2016) or single cell RNA-seq data (Wang et al. 2019; Newman et al. 2019; Jew et al. 2020; Dong et al. 2021; Yang et al. 2021; Chu et al. 2022). In these methods, the proportions π_{ik}'s are estimated by, for example, nonnegative least squares (Wang et al. 2019) or support vector regression (Newman et al. 2019). Given the bulk samples $\{\boldsymbol{x}_i\}_{i\in[n]}$ and cell type proportions $\{\pi_{ik}\}_{i\in[n],k\in[K]}$, our goal is to estimate the cell-type-specific correlations $\{\boldsymbol{R}^{(k)}\}_{k\in[K]}$.

2.2. Estimating $R^{(k)}$ with Large p

It is easily seen from (1) and (3) that each bulk sample $\boldsymbol{x}_i$ is from a convolution of K distributions. In this case, estimating $\{\Sigma^{(k)}\}_{k\in[K]}$ from $\{\boldsymbol{x}_i\}_{i\in[n]}$ is very challenging. For example, even in the simple Gaussian case, the loglikelihood function is, up to a constant,

$$\sum_{i=1}^{n} \log \left| \sum_{k=1}^{K} \pi_{ik}^2 \Sigma^{(k)} \right| - \sum_{i=1}^{n} \mathrm{tr} \left\{ \left(\sum_{k=1}^{K} \pi_{ik}^2 \Sigma^{(k)} \right)^{-1} \boldsymbol{z}_i \boldsymbol{z}_i^\top \right\},$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix and $\boldsymbol{z}_i = \boldsymbol{x}_i - \mathbb{E}(\boldsymbol{x}_i)$. This loglikelihood is not convex or biconvex with respect to $\{\Sigma^{(k)}\}_{k\in[K]}$, and cannot be directly optimized using existing iterative algorithmic solutions such as EM and coordinate descent. To our knowledge, there are no existing methods that can effectively estimate the covariances in a convolution of densities.

To tackle this challenge, we propose a novel moment-based approach that is efficient to implement and flexible, in that it does not assume the distributions from the K cell types to be known or of the same type. The proposed approach, named `CSNet`, first estimates $\boldsymbol{R}^{(k)}$ efficiently in an element-wise fashion and then applies a thresholding step, in the case of a large p, to give a sparse estimator. Next, we introduce the `CSNet` estimator.

Letting $\boldsymbol{\mu}^{(k)} = (\mu_1^{(k)}, \ldots, \mu_p^{(k)})$ and $\Sigma^{(k)} = (\sigma_{jj'}^{(k)})_{p\times p}$, (1) and (3) together imply

$$x_{ij} = \sum_{k=1}^{K} \pi_{ik} \mu_j^{(k)} + z_{ij}, \quad j \in [p], \tag{4}$$

$$z_{ij} z_{ij'} = \sum_{k=1}^{K} \pi_{ik}^2 \sigma_{jj'}^{(k)} + \epsilon_{ijj'}, \quad j, j' \in [p], \tag{5}$$

where $\mathbb{E}(z_{ij}) = 0$ and $\mathbb{E}(\epsilon_{ijj'}) = 0$. This formulation facilitates an efficient least squares estimation procedure to be detailed in the next paragraph. Note that (4)–(5) hold generally without parametric assumptions on the distributions from the K cell types.

Denote $\boldsymbol{x}^j = (x_{1j}, \ldots, x_{nj})$ and $\mathbf{P}_1 = (\pi_{ik})_{n\times K}$. Equation (4) entails estimation of the cell-type-specific mean $\boldsymbol{\mu}^{(k)}$ via

$$\hat{\mu}_j^{(k)} = \left[\left(\mathbf{P}_1^\top \mathbf{P}_1 \right)^{-1} \mathbf{P}_1^\top \boldsymbol{x}^j \right]_k, \quad j \in [p],\ k \in [K], \tag{6}$$

where $[\boldsymbol{x}]_k$ is the kth entry in $\boldsymbol{x} \in \mathbb{R}^K$. Let $\hat{z}_{ij} = x_{ij} - \sum_{k=1}^{K} \pi_{ik} \hat{\mu}_j^{(k)}$ and $\hat{\boldsymbol{z}}_j = (\hat{z}_{1j}, \ldots, \hat{z}_{nj}) \in \mathbb{R}^n$. Denoting $\mathbf{P}_2 = (\pi_{ik}^2)_{n\times K}$, (5) entails estimation of the cell-type-specific covariance $\sigma_{jj'}^{(k)}$ via

$$\hat{\sigma}_{jj'}^{(k)} = \left[\left(\mathbf{P}_2^\top \mathbf{P}_2 \right)^{-1} \mathbf{P}_2^\top (\hat{\boldsymbol{z}}_j \circ \hat{\boldsymbol{z}}_{j'}) \right]_k, \quad j, j' \in [p],\ k \in [K], \tag{7}$$

where $\circ$ denotes element-wise product. Then, the cell-type-specific correlation $\hat{R}^{(k)}_{jj'}$ is estimated as

$$\hat{R}^{(k)}_{jj'} = \hat{\sigma}^{(k)}_{jj'} / \sqrt{\hat{\sigma}^{(k)}_{jj} \hat{\sigma}^{(k)}_{j'j'}}, \quad j, j' \in [p],\ k \in [K]. \tag{8}$$

In Section S6.2, we show that element-wisely $\hat{R}^{(k)}_{jj'}$ is consistent with a $\sqrt{n}$-convergence rate under certain regularity conditions. In our implementation, as cell-type-specific means $\mu^{(k)}_j$'s and variances $\sigma^{(k)}_{jj}$'s are positive parameters, we adopt non-negative least squares (NNLS) to improve their estimation accuracy in finite samples. To reduce the variability in the estimates of cell-type-specific covariances $\sigma^{(k)}_{jj'}$'s, we further consider a weighted least squares approach with carefully chosen weights (see details in Section S3.1).

Though each entry in the correlation matrix estimated in (8) has a $\sqrt{n}$-convergence rate, the accumulated errors across $O(p^2)$ entries in $\hat{\boldsymbol{R}}^{(k)}$ can be excessive, especially as the number of genes p often far exceeds the sample size n in co-expression network analysis, which can negatively impact downstream analyses such as ranking, principal component analysis or clustering. This same challenge also arises in estimating large sample covariance (Bickel and Levina 2008a,b; Rothman et al. 2008; Rothman, Levina, and Zhu 2009) and correlation matrices (El Karoui 2008; Jiang 2013). To facilitate estimability and interpretability, we assume that $\Sigma^{(k)}$ (or equivalently $\boldsymbol{R}^{(k)}$) is approximately sparse for all k; see the definition of approximate sparsity in Assumption 2 (Section S1). Sparsity is plausible in our data problem, as gene co-expressions are expected to be sparse when p is large (Zhang and Horvath 2005).

Based on $\hat{\boldsymbol{R}}^{(k)}$, the proposed `CSNet` estimator computes sparse cell-type-specific correlation estimates via thresholding. Specifically, `CSNet` applies an element-wise SCAD (Fan and Li 2001) thresholding operator to $\hat{\boldsymbol{R}}^{(k)}$, written as $\mathcal{T}_{\lambda_k}(\hat{\boldsymbol{R}}^{(k)})$, with the (j,j')th thresholded entry calculated as

$$\left[\mathcal{T}_{\lambda_k}\left(\hat{\boldsymbol{R}}^{(k)}\right)\right]_{jj'} = \operatorname{sign}\left(\hat{R}^{(k)}_{jj'}\right) \times \begin{cases} \left(\left|\hat{R}^{(k)}_{jj'}\right| - \lambda_k\right)_+ & |\hat{R}^{(k)}_{jj'}| \le 2\lambda_k \\ \lambda_k + \dfrac{a-1}{a-2}\left(\left|\hat{R}^{(k)}_{jj'}\right| - 2\lambda_k\right) & \left|\hat{R}^{(k)}_{jj'}\right| \in (2\lambda_k, a\lambda_k) \\ \left|\hat{R}^{(k)}_{jj'}\right| & \text{otherwise,} \end{cases} \tag{9}$$

where λ_k is a tuning parameter and we discuss its selection in Section 2.3. As has been well established in the sparse covariance estimation literature (e.g., Bickel and Levina 2008a,b; Rothman et al. 2008; Rothman, Levina, and Zhu 2009), the thresholding procedure is easy to implement and enjoys good theoretical properties. Moreover, the SCAD thresholding has been found to give better numerical performances when compared to soft or hard thresholding (Rothman, Levina, and Zhu 2009). We set $a = 3.7$ in our experiments as recommended by Fan and Li (2001). The thresholds λ_k's may differ across different cell types to accommodate varying sparsity among different cell types. We note that the λ_k's can be selected separately (without a joint tuning) in our tuning procedure (see Section 2.3), which is an attractive computational property of our procedure. In Section S1, we show the convergence rates of `CSNet`, that is, $\mathcal{T}_{\lambda_k}(\hat{\boldsymbol{R}}^{(k)})$, in spectral and Frobenius norms, and establish its selection consistency.

As with all thresholding approaches, $\mathcal{T}_{\lambda_k}(\hat{\boldsymbol{R}}^{(k)})$ is not guaranteed to be positive definite; see more discussions in Section 5 on considerations for positive definiteness. To ensure the finite sample validity of correlation estimates, we threshold the correlation estimates to be within $[-1,1]$ in our experiments.

2.3. Time Complexity and Parameter Tuning

We first discuss the time complexity of solving (7) for all entries in covariance matrices. Though (7) is computed element-wisely, the matrix $\left(\mathbf{P}_2^\top \mathbf{P}_2\right)^{-1} \mathbf{P}_2^\top \in \mathbb{R}^{K \times n}$ is common and only needs to be calculated once. Hence, entries in $\Sigma_1, \ldots, \Sigma_K$ can be estimated efficiently via $\left(\mathbf{P}_2^\top \mathbf{P}_2\right)^{-1} \mathbf{P}_2^\top \boldsymbol{U}$, where $\boldsymbol{U}$ is an $n \times p^2$ matrix with the jj'th column set to $\hat{\boldsymbol{z}}_j \circ \hat{\boldsymbol{z}}_{j'}$. Correspondingly, the time complexity of estimating $\Sigma_1, \ldots, \Sigma_K$ is $O(Knp^2)$, while that of a naive sample covariance estimation is $O(np^2)$. As the number of cell types K is usually small, the two computing times are comparable.

Next, we discuss parameter tuning. In our procedure, the tuning parameters λ_k's are selected using cross-validation or, if available, an independent validation dataset. Here, we introduce the cross-validation procedure and note that selection with a validation dataset can be carried out similarly. We randomly split the data into two equal-sized pieces and estimate for each piece, the cell-type-specific correlation matrices as in (8). We denote the estimated correlation matrices from these two data splits as $\hat{\boldsymbol{R}}^{(k)}_1$ and $\hat{\boldsymbol{R}}^{(k)}_2$, $k \in [K]$, respectively. For each k, the tuning parameter λ_k is selected by minimizing $\|\mathcal{T}_{\lambda_k}(\hat{\boldsymbol{R}}^{(k)}_1) - \hat{\boldsymbol{R}}^{(k)}_2\|_F^2 + \|\mathcal{T}_{\lambda_k}(\hat{\boldsymbol{R}}^{(k)}_2) - \hat{\boldsymbol{R}}^{(k)}_1\|_F^2$ among a set of working values, where $\|\cdot\|_F$ denotes the Frobenius norm. We consider two equal sized data splits as sufficient samples are needed to estimate $\hat{\boldsymbol{R}}^{(k)}_1$ and $\hat{\boldsymbol{R}}^{(k)}_2$ well. This procedure is similar to what was proposed in Bickel and Levina (2008a), where the theoretical justification was provided, and it is found to give a good performance in our numerical experiments. In practice, this procedure can be overly conservative for less abundant cell types, as the estimates $\hat{\boldsymbol{R}}^{(k)}_1$ and $\hat{\boldsymbol{R}}^{(k)}_2$ can be very noisy (see Figure S11). To mitigate this issue, we propose to further consider a one standard error rule for selecting the tuning parameters in less abundant cell types (see Section S3.2). An attractive feature of our proposed tuning procedure is that λ_k's are selected separately without the need of a joint tuning, further reducing the computational cost.

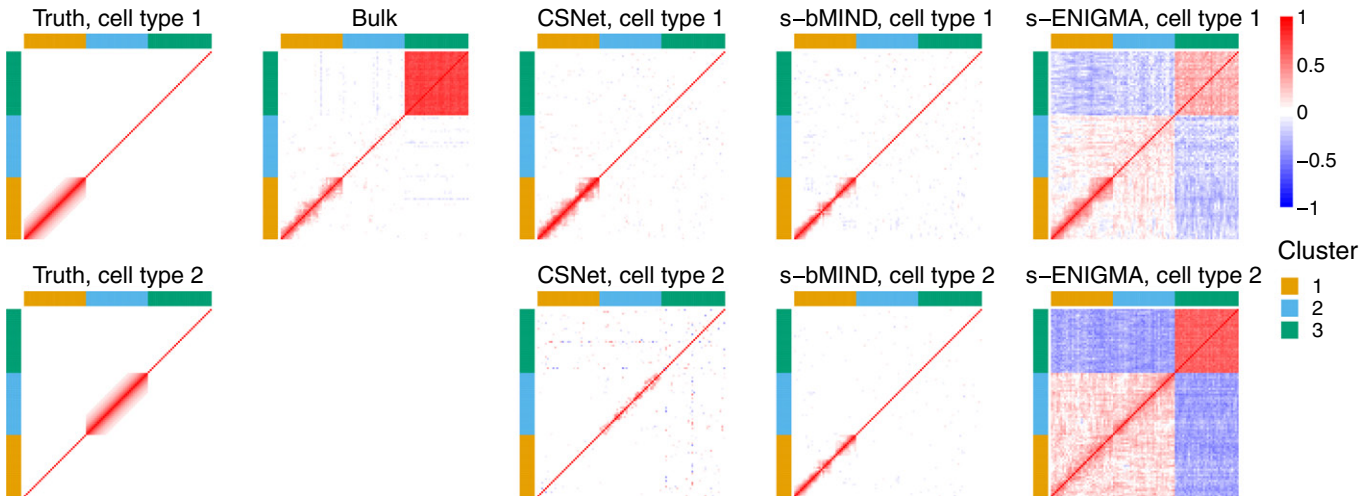

Figure 1. Correlation matrix estimate in one data replicate with $p = 100$ and $n = 150$, under a negative binomial distribution with a truncated AR(1) correlation structure in V_1 or V_2 (see Section 3). From left to right: true correlation matrices, estimates from `Bulk`, `CSNet`, `s-bMIND`, and `s-ENIGMA`. Note the `Bulk` estimate is not cell-type-specific.

3. Simulation Studies

We first investigate the finite sample performance of `CSNet` and compare it with three existing methods, and then conduct a sensitivity analysis to examine the performance of `CSNet` when the provided cell type proportions have biased noises.

3.1. Comparing `CSNet` with Other Methods

We consider two classes of estimators: dense correlation estimators without thresholding and sparse correlation estimators that implement SCAD thresholding. For dense correlation estimators, we consider the estimator calculated using bulk samples, referred to as `d-Bulk`, the cell-type-specific estimator in (8) without sparsity, referred to as `d-CSNet`, estimators that compute the sample correlations of the cell-type-specific expressions estimated by Wang, Roeder, and Devlin (2021) and Wang et al. (2021b), referred to as `bMIND` and `ENIGMA`, respectively. We apply SCAD thresholding to the above four dense estimators and obtain sparse correlation estimators referred to as `Bulk`, `CSNet`, `s-bMIND`, and `s-ENIGMA`, respectively. For these sparse estimators, tuning parameters were selected using the same cross-validation procedure to facilitate a fair comparison. The method `bMIND` considers a Bayesian mixed effects model that constructs priors using single cell data and estimates cell-type-specific expressions in each sample with posterior means. In our experiment, it was evaluated with non-informative priors to be comparable with the other methods, which do not depend on prior knowledge. We also consider informative priors for `bMIND` in Tables S2 and S5 in the supplement and the results remain similar. The method `ENIGMA` (Wang et al. 2021b) is an optimization-based method that estimates cell-type-specific expressions for each sample by minimizing the difference between estimated and observed bulk data and it was evaluated under the default parameter setting (see more details in Section S4). We chose to not compare with CIBERSORTx high-resolution expression purification (Newman et al. 2019) as it is applicable only when there are both case and control samples in the data. All methods under comparison were supplied with true cell type proportions. Sensitivity analysis with biased noises in cell type proportions is presented in Section 3.2.

We simulate n bulk samples of dimension p following $\boldsymbol{x}_i = \sum_{k=1}^{K} \pi_{ik} \boldsymbol{x}_i^{(k)}$ in (1) with $K = 2$ cell types and π_{i1}'s i.i.d from Beta(2, 1). Correspondingly, cell type 1 ($m = 2/3$) is on average twice as abundant as cell type 2 ($m = 1/3$), where m denotes the average cell type proportions. We simulate $\boldsymbol{x}_i^{(k)}$'s from multivariate negative binomial distributions, resembling read counts from bulk RNA-seq data (Love, Huber, and Anders 2014), with mean $\boldsymbol{\mu}^{(k)}$'s and covariance matrices $\Sigma^{(k)}$'s specified as follows. The p genes are divided into three equal-sized sets, denoted as V_1, V_2, and V_3; genes in V_1 and V_2 are set to co-express in cell types 1 and 2, respectively, while all other correlations are set to zero (see the left panel in Figure 1). For the co-expressed genes in V_1 or V_2, two types of structures are considered, including a truncated AR(1) structure with $\rho_{jj'} = 0.8^{|j-j'|}$ for $|j - j'| \le 10$, and a correlation structure estimated from real single cell RNA-seq data for $j \neq j' \in V_1$ or V_2 (see details in Section S3.3). We set $\log \sigma_{jj}^{(1)} = 8$ for all j, $\log \sigma_{jj}^{(2)} = 8$ for $j \in V_1, V_2$ and $\log \sigma_{jj}^{(2)} = 9$ for $j \in V_3$ with sequencing depth set to $S = 6 \times 10^7$ to mimic highly-expressed protein-coding genes in real sequencing data. The mean $\boldsymbol{\mu}^{(k)}$ is set to be a function of $\Sigma^{(k)}$ (see details in Section S3.4), consistent with the observation in real data that higher expression levels are often associated with larger variances. To simulate correlated negative binomial random variables as specified, we combine a marginal negative binomial model (Section S3.4) and a copula-based approach that can simulate multivariate count data following a pre-specified correlation matrix (Tian, Wang, and Roeder 2021; Sun et al. 2021). The tuning parameters for `CSNet` and `Bulk` are selected following Section 2.3 and the suggested procedure in Rothman, Levina, and Zhu (2009), respectively, both using a validation dataset with the same size as the observed samples. We consider network sizes $p = 100, 200$ and sample sizes $n = 150,600$.

To evaluate the estimation accuracy, we report the estimation errors in the Frobenius norm $\|\hat{\boldsymbol{R}}^{(k)} - \boldsymbol{R}^{(k)}\|_F$ and operator norm $\|\hat{\boldsymbol{R}}^{(k)} - \boldsymbol{R}^{(k)}\|$, where $\hat{\boldsymbol{R}}^{(k)}$, with a slight overuse of notation,

Table 1. Evaluation criteria of sparse estimators with varying sample size n and network size p under a negative binomial distribution with a truncated AR(1) correlation structure for genes in V_1 or V_2 (see Section 3).

			Cell type 1 ($m = 2/3$)				Cell type 2 ($m = 1/3$)			
n	p	Method	F-norm	Op.-norm	FPR	TPR	F-norm	Op.-norm	FPR	TPR
150	100	Bulk	28.11	27.29	0.26	0.79	30.16	27.29	0.28	0.39
			(0.14)	(0.12)	(0.03)	(0.06)	(0.15)	(0.12)	(0.03)	(0.06)
		CSNet	**4.34**	**2.36**	**0.11**	0.75	**7.56**	**4.53**	**0.08**	0.51
			(0.34)	**(0.39)**	**(0.01)**	(0.06)	**(0.41)**	**(0.44)**	**(0.01)**	(0.07)
		s-bMIND	5.97	4.13	0.14	0.59	11.76	6.60	0.09	0.08
			(0.55)	(0.41)	(0.03)	(0.06)	(0.58)	(0.04)	(0.05)	(0.06)
		s-ENIGMA	25.07	23.51	1.00	**1.00**	31.59	29.81	1.00	**1.00**
			(0.76)	(0.81)	(0.00)	**(0.00)**	(0.74)	(0.78)	(0.00)	**(0.00)**
	200	Bulk	54.74	53.79	0.20	0.73	56.89	53.79	0.22	0.32
			(0.19)	(0.18)	(0.02)	(0.05)	(0.18)	(0.18)	(0.02)	(0.04)
		CSNet	**7.07**	**3.10**	**0.07**	0.69	**11.84**	**5.37**	**0.05**	0.44
			(0.35)	**(0.30)**	**(0.01)**	(0.05)	**(0.34)**	**(0.26)**	**(0.01)**	(0.04)
		s-bMIND	8.48	4.36	0.08	0.60	17.17	6.97	**0.05**	0.05
			(0.55)	(0.34)	(0.01)	(0.05)	(0.53)	(0.02)	**(0.02)**	(0.03)
		s-ENIGMA	48.97	46.45	1.00	**1.00**	62.01	59.22	1.00	**1.00**
			(0.93)	(0.99)	(0.00)	**(0.00)**	(0.98)	(1.05)	(0.00)	**(0.00)**
600	100	Bulk	28.05	27.37	0.30	0.95	30.05	27.37	0.31	0.67
			(0.07)	(0.06)	(0.04)	(0.03)	(0.08)	(0.06)	(0.04)	(0.05)
		CSNet	**2.08**	**0.99**	**0.10**	0.93	**3.69**	**1.88**	**0.10**	0.80
			(0.20)	**(0.20)**	**(0.02)**	(0.03)	**(0.27)**	**(0.31)**	**(0.01)**	(0.05)
		s-bMIND	4.49	3.37	0.35	0.79	10.18	5.87	0.21	0.50
			(0.15)	(0.20)	(0.03)	(0.03)	(0.34)	(0.25)	(0.05)	(0.09)
		s-ENIGMA	18.77	17.68	1.00	**1.00**	25.26	23.89	1.00	**1.00**
			(0.34)	(0.36)	(0.00)	**(0.00)**	(0.35)	(0.36)	(0.00)	**(0.00)**
	200	Bulk	54.64	53.93	0.23	0.94	56.88	53.93	0.24	0.61
			(0.09)	(0.08)	(0.03)	(0.02)	(0.10)	(0.08)	(0.03)	(0.04)
		CSNet	**3.41**	**1.30**	**0.07**	0.91	**5.89**	**2.40**	**0.07**	0.76
			(0.21)	**(0.19)**	**(0.01)**	(0.03)	**(0.29)**	**(0.25)**	**(0.01)**	(0.04)
		s-bMIND	5.07	2.76	0.15	0.83	14.98	6.17	0.11	0.49
			(0.20)	(0.15)	(0.01)	(0.03)	(0.41)	(0.20)	(0.02)	(0.06)
		s-ENIGMA	36.25	34.92	1.00	**1.00**	49.17	47.33	1.00	**1.00**
			(0.51)	(0.53)	(0.00)	**(0.00)**	(0.48)	(0.49)	(0.00)	**(0.00)**

NOTE: The four sparse estimators under comparison are Bulk, CSNet, s-bMIND, and s-ENIGMA. We use F-norm to denote the Frobenius norm and Op.-norm to denote the operator norm. Marked in boldface are those achieving the best evaluation criteria in each setting.

Table 2. Evaluation criteria of dense estimators under the same setting as Table 1.

n		150				600			
Cell type		1 (m = 2/3)		2 (m = 1/3)		1 (m = 2/3)		2 (m = 1/3)	
p	Method	F norm	Op. norm	F norm	Op. norm	F norm	Op. norm	F norm	Op. norm
100	d-Bulk	29.02	27.77	30.78	27.78	28.30	27.49	30.10	27.49
		(0.18)	(0.18)	(0.19)	(0.19)	(0.07)	(0.07)	(0.08)	(0.07)
	d-CSNet	11.80	**4.15**	20.01	7.34	**5.80**	**1.85**	**9.52**	**3.08**
		(0.23)	**(0.39)**	(0.58)	(0.63)	**(0.10)**	**(0.16)**	**(0.19)**	**(0.29)**
	bMIND	**10.80**	4.89	**15.57**	**6.62**	7.08	4.60	10.93	5.58
		(0.24)	(0.25)	**(0.36)**	**(0.39)**	(0.08)	(0.12)	(0.31)	(0.20)
	ENIGMA	25.07	23.51	31.59	29.81	18.77	17.68	25.26	23.89
		(0.76)	(0.81)	(0.74)	(0.78)	(0.34)	(0.36)	(0.35)	(0.36)
200	d-Bulk	56.76	54.80	58.67	54.81	55.18	54.17	57.16	54.18
		(0.24)	(0.25)	(0.23)	(0.25)	(0.09)	(0.09)	(0.10)	(0.09)
	d-CSNet	23.66	6.95	40.22	12.96	11.63	**2.97**	19.17	**5.15**
		(0.27)	(0.43)	(0.74)	(0.95)	(0.12)	**(0.21)**	(0.20)	**(0.33)**
	bMIND	**19.84**	**5.81**	**26.15**	**8.78**	**11.43**	4.97	**17.27**	5.73
		(0.22)	**(0.31)**	**(0.35)**	**(0.63)**	**(0.11)**	(0.14)	**(0.34)**	(0.18)
	ENIGMA	48.97	46.45	62.01	59.22	36.25	34.92	49.17	47.33
		(0.93)	(0.99)	(0.98)	(1.05)	(0.51)	(0.53)	(0.48)	(0.49)

NOTE: The four dense estimators under comparison are d-Bulk, d-CSNet, bMIND, and ENIGMA.

denotes the estimate of $\boldsymbol{R}^{(k)}$ obtained by various methods. For sparse estimators, we also report the true positive rate (TPR) and false positive rate (FPR), which evaluate the selection accuracy of nonzero entries in $\boldsymbol{R}^{(k)}$'s. Tables 1 and 2 report the average criteria under the truncated AR(1) model for sparse and dense estimators, with standard deviations in the parentheses, over 200 data replications. Table 1 shows that CSNet achieved the best performance in terms of both estimation accuracy and selection accuracy. CSNet also performed the closest to an oracle estimator that benchmarks the performance if true cell-type-specific expressions within each bulk sample were available (Table S1). In the supplement, we demonstrate in Tables S2 that

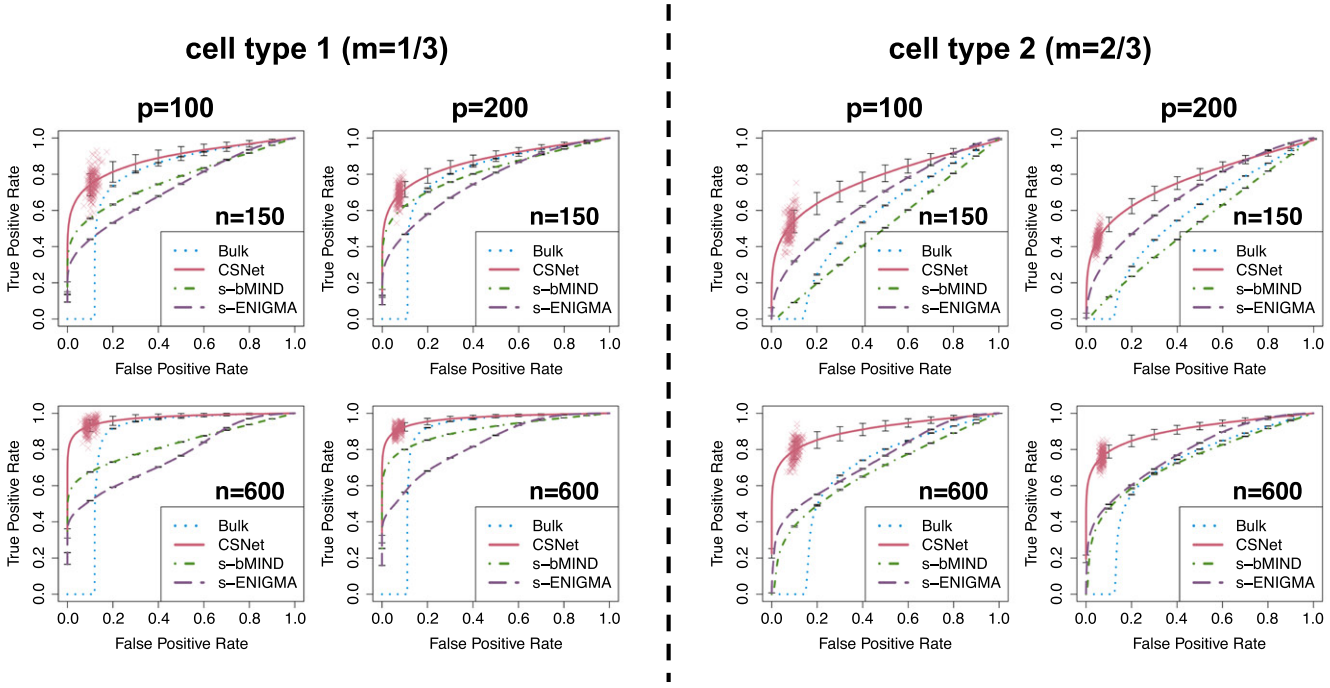

Figure 2. ROC curves with a varying sample size n and network size p under a negative binomial distribution with a truncated AR(1) correlation structure in V_1 or V_2 (see Section 3). The four methods under comparison are `Bulk`, `CSNet`, `s-bMIND`, and `s-ENIGMA`. The thresholds selected by `CSNet` are marked using red crosses on the curves.

even with informative priors derived from simulated cell-type-specific data, `CSNet` still performed better than `s-bMIND`. Table 2 further shows that `d-CSNet` yielded satisfactory performance among dense estimators. The errors of `bMIND` can be smaller than `d-CSNet` in some cases. This is because the true signal is highly sparse and `bMIND` tended to give a biased but less variable estimate (see Figure S5). In Table S3, we show that `d-CSNet` outperformed `bMIND` when denser signals were considered. We also demonstrate in Table S4 that when the true co-expressions were set by a correlation structure estimated from real data, `CSNet` again achieved the best performance among all methods.

To better visualize the estimates, Figure 1 plots heat maps of the true cell-type-specific co-expression matrices and estimates from `Bulk`, `CSNet`, `s-bMIND`, and `s-ENIGMA`. The estimates from `d-CSNet` show similar patterns as `CSNet` plus some additional noises (see Figure S5). From Figure 1, it is clearly seen that `Bulk`, `s-bMIND`, and `s-ENIGMA` give a less accurate view of the true co-expressions. Specifically, `Bulk` estimates high co-expressions in V_3 while genes in V_3 are not co-coexpressed in either cell type. True co-expression patterns in V_2 from cell type 2 are also notably attenuated in `Bulk`. Moreover, `s-bMIND` does not perform well for cell type 2, the less abundant cell type. It is seen that the true co-expressions in V_2 are not identified while co-expressions specific to cell type 1 are incorrectly inferred. `s-ENIGMA` also estimate high co-expressions in V_3 and generally give correlations with large magnitude, potentially caused by the penalty used in ENIGMA, where cell-type-specific expressions are encouraged to be similar to the reference cell-type-specific expressions. In comparison, `CSNet` was able to identify the true co-expression patterns in both cell types.

To evaluate the threshold selection procedure in Section 2.3, we plot the ROC curves that plot the TPR against the FPR across a fine grid of thresholding parameters for `Bulk`, `CSNet`, `s-bMIND`, and `s-ENIGMA`. The thresholds selected by our proposed procedure in Section 2.3 are marked on the curves for `CSNet`. The ROC curves in Figure 2 show that `CSNet` achieves the best performance and the selected thresholds generally strike a reasonable balance between TPR and FPR. As shown in Tables 1, 2 and Figures 1, 2, the improvement of `CSNet` over others is the most notable for the less abundant cell type, and this demonstrates the efficacy of our proposed method for cell types whose signals are attenuated in bulk samples.

We also consider a setting where there are $K = 4$ cell types (see Section S3.5). To mimic microglia analyzed in Section 4, a less abundant cell type with important roles in Alzheimer's disease (Tansey, Cameron, and Hill 2018), we simulate a cell type with an average proportion of 10%. Table S6 shows that `CSNet` performed better than the other methods for this cell type. We also consider a setting with $K = 10$ cell types that includes 6 rare cell types whose average proportions add up to 10% (see Section S3.5). This setting resembles our real data analysis in Section 4, as the brain tissue studied in our real data analysis have eight common brain cell types and the four least abundant cell types add up to less than 10%. Table S7 shows that the estimation and selection accuracy of `CSNet` on the four major cell types remained similar with the addition of six cell types when compared to Table S6.

3.2. Sensitivity Analysis of `CSNet`

In this section, we conduct a sensitivity analysis to examine the performance of our method when the cell type proportions π_{ik}'s

Table 3. Sensitivity analysis of CSNet with $n = 600$ and $p = 100$, under the same setting as Table S4.

				Cell type 1 ($m = 2/3$)				Cell type 2 ($m = 1/3$)			
κ	b	cor	r-RMSE	F-norm	Op.-norm	FPR	TPR	F-norm	Op.-norm	FPR	TPR
0.0	−0.2	1.00	0.26	1.32	0.57	0.05	0.42	2.56	1.54	0.07	0.43
				(0.08)	(0.10)	(0.01)	(0.02)	(0.19)	(0.22)	(0.01)	(0.02)
	−0.1	1.00	0.13	1.22	0.56	0.04	0.43	2.50	1.43	0.06	0.43
				(0.09)	(0.10)	(0.01)	(0.02)	(0.19)	(0.22)	(0.01)	(0.02)
	0.0	1.00	0.00	1.16	0.58	0.04	0.43	2.48	1.33	0.06	0.43
				(0.09)	(0.12)	(0.01)	(0.02)	(0.18)	(0.23)	(0.01)	(0.02)
	0.1	1.00	0.13	1.16	0.59	0.05	0.43	2.53	1.23	0.06	0.43
				(0.09)	(0.12)	(0.01)	(0.02)	(0.17)	(0.22)	(0.01)	(0.02)
	0.2	0.99	0.24	3.72	2.80	0.27	0.54	2.71	1.17	0.07	0.43
				(0.31)	(0.32)	(0.02)	(0.04)	(0.16)	(0.21)	(0.01)	(0.03)
0.6	−0.2	0.96	0.30	10.63	10.07	0.31	0.55	10.26	9.26	0.35	0.56
				(0.33)	(0.35)	(0.01)	(0.03)	(0.46)	(0.50)	(0.02)	(0.03)
	−0.1	0.96	0.19	10.48	9.98	0.30	0.55	10.32	9.23	0.35	0.56
				(0.29)	(0.31)	(0.02)	(0.04)	(0.49)	(0.53)	(0.01)	(0.03)
	0.0	0.96	0.14	10.16	9.69	0.32	0.56	10.39	9.18	0.34	0.56
				(0.27)	(0.29)	(0.02)	(0.04)	(0.54)	(0.58)	(0.01)	(0.03)
	0.1	0.96	0.18	9.96	9.55	0.29	0.55	10.02	8.58	0.34	0.56
				(0.26)	(0.27)	(0.01)	(0.03)	(0.61)	(0.65)	(0.02)	(0.03)
	0.2	0.95	0.26	11.25	10.92	0.25	0.52	8.45	6.57	0.32	0.55
				(0.22)	(0.22)	(0.02)	(0.04)	(0.58)	(0.64)	(0.01)	(0.03)
0.9	−0.2	0.83	0.40	20.85	20.47	0.17	0.44	26.50	25.88	0.20	0.42
				(0.29)	(0.28)	(0.01)	(0.03)	(0.30)	(0.30)	(0.02)	(0.03)
	−0.1	0.83	0.33	21.26	20.89	0.16	0.43	27.39	26.78	0.19	0.42
				(0.27)	(0.27)	(0.01)	(0.02)	(0.34)	(0.34)	(0.02)	(0.03)
	0.0	0.83	0.29	21.34	20.98	0.16	0.43	28.11	27.45	0.19	0.42
				(0.25)	(0.25)	(0.01)	(0.03)	(0.36)	(0.34)	(0.01)	(0.03)
	0.1	0.83	0.30	21.35	20.99	0.16	0.43	28.39	27.57	0.19	0.42
				(0.23)	(0.22)	(0.01)	(0.02)	(0.34)	(0.34)	(0.01)	(0.03)
	0.2	0.83	0.34	21.51	21.15	0.16	0.44	28.06	27.10	0.22	0.44
				(0.20)	(0.20)	(0.01)	(0.03)	(0.36)	(0.38)	(0.01)	(0.03)

NOTE: Noisy cell type proportions were set to $\pi_{i1} + N(b \times \bar{\pi}_1, \kappa \times 0.04)$, and then thresholded to be within [0, 1]. We use cor to denote the Pearson correlation $\text{cor}(\pi_{i1}, \tilde{\pi}_{i1})$ and r-RMSE to denote the relative root mean squared error $\sqrt{\sum_{i=1}^{n}(\pi_{i1} - \tilde{\pi}_{i1})^2/n}/\bar{\pi}_1$.

used in CSNet are measured with biased noises. We consider $n = 600, p = 100$ and the same simulation setting as in Table S4 where the co-expression structures are those estimated from real data. Let $\{\pi_{i1}\}_{i=1}^n$ and $\{\tilde{\pi}_{i1}\}_{i=1}^n$ denote the true and noisy cell type proportions for cell type 1, respectively. We let

$$\tilde{\pi}_{i1} = \pi_{i1} + e_i, \quad e_i \sim N(b \times \bar{\pi}_1, \kappa \times 0.04),$$

where $\bar{\pi}_1 = \sum_{i=1}^n \pi_{i1}/n$ and b controls the magnitude of relative bias (e.g., $b = 0.2$ indicates that cell type proportions are on average over-estimated by 20% in cell type 1). We consider $b \in \{-0.2, -0.1, 0, 0.1, 0.2\}$ and $\kappa \in \{0, 0.6, 0.9\}$. We threshold $\tilde{\pi}_{i1}$ to be between $[0, 1]$ and set $\tilde{\pi}_{i2} = 1 - \tilde{\pi}_{i1}$. To quantify the noise level, we adopt two metrics: Pearson correlation $\text{cor}(\pi_{i1}, \tilde{\pi}_{i1})$ and relative root mean squared error (r-RMSE) $\sqrt{\frac{1}{n}\sum_{i=1}^n(\pi_{i1} - \tilde{\pi}_{i1})^2}/\bar{\pi}_1$. We present these metrics for cell type 1 as there are two cell types in the experiment. The noise levels in this sensitive analysis (see Table 3) are similar to those achieved by recently developed methods for estimating cell type proportions from bulk RNA-seq data. For example, Newman et al. (2019) and Chu et al. (2022) showed that the correlations between their estimated proportions and ground truth proportions, measured by technologies such as flow cytometry, were generally around 0.9.

Table 3 presents the sensitivity analysis results. It is seen that at any fixed value of κ (i.e., noise variance), the estimation and selection errors are similar across different values of b (i.e., noise bias), suggesting that CSNet is not overly sensitive to biases. When $\kappa = 0$ (only bias, no random errors), the evaluation metrics are comparable to the case with accurate cell type proportions (i.e., $\kappa = b = 0$). When $\kappa = 0.6$, the TPRs remain high in both cell types with a reasonable false positive control, though the estimation errors increase. When $\kappa = 0.9$, the correlations can be as low as 0.83 and r-RMSE can be as large as 0.40, suggesting a considerable amount of noises in cell type proportions. Though the estimation errors increase, the FPRs and TPRs remain reasonably satisfactory.

4. Cell-Type-Specific Co-expressions of Different Gene Sets for an Alzheimer's Disease Cohort

We focus on estimating cell-type-specific co-expressions using bulk RNA-seq data from the Religious Orders Study and Rush Memory and Aging Project (ROSMAP) study on Alzheimer's disease (Bennett et al. 2018). Gene expressions were profiled by bulk RNA-seq for $n = 541$ postmortem brain samples, expression unit FPKM (Trapnell et al. 2010) was used to quantify gene expressions, and no notable batch effects were observed from these samples (see Figure S6). The cell type proportions for $K = 8$ cell types were estimated using CIBERSORTx (Newman et al. 2019) with the signature matrix built from the single nucleus RNA-seq data (Mathys et al. 2019) collected on the same brain region in a subset of 48 samples.

We applied CSNet as defined in (9), with the tuning parameters selected using the cross-validation procedure discussed in Section 2.3, to estimate cell-type-specific co-expressions for

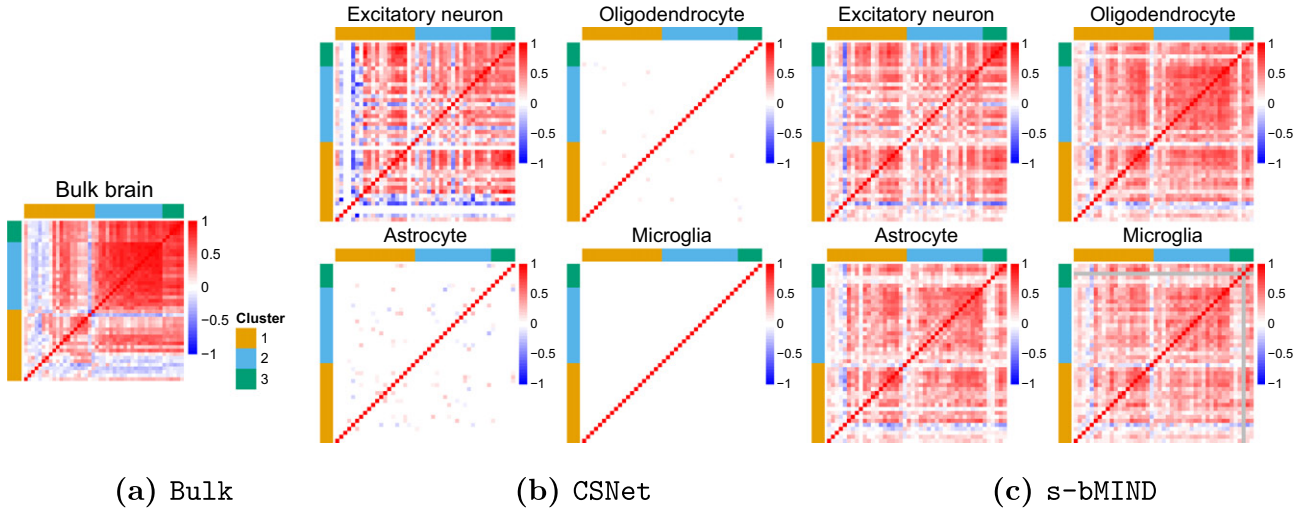

Figure 3. Co-expressions of *excitatory synapse* genes in different cell types. From left to right: (a) sample correlation matrix estimated from bulk RNA-seq data, (b) `CSNet` estimates and (c) `s-bMIND` estimates. For `s-bMIND`, genes with constant expression estimates across samples are marked in gray.

all $K = 8$ cell types. Additionally, we compared `CSNet` to two alternative approaches, one through `s-bMIND`, the best performing alternative in Section 3, and one using single cell data (Mathys et al. 2019), respectively. For `s-bMIND` estimates, we followed Wang, Roeder, and Devlin (2021) to infer priors from single cell data and supplied these priors to estimate cell-type-specific expressions in each sample; see details in Section S4; correlation estimates were then computed using the estimated cell-type-specific expressions across different samples. To estimate cell-type-specific co-expressions from single cell data, we first calculated cell-type-specific expressions in each sample. Specifically, the expression profile of gene j in cell type k for sample i was calculated by first summing over the UMI counts of gene j from all cells of cell type k in sample i, and then normalized by the total number of UMI counts in cell type k from sample i. These cell-type-specific expressions calculated for different samples were then used to estimate the co-expression (i.e., correlation matrix) in each cell type, and the correlation matrices were further thresholded following the procedure in Rothman, Levina, and Zhu (2009) with the SCAD penalty. For all methods, we visualized the estimated co-expressions using heat maps, with genes ordered into clusters (or modules) identified by WGCNA (Langfelder and Horvath 2008), a gene clustering method, applied to bulk samples. In the ensuing analysis, we focus on the four most abundant cell types: excitatory neuron (Ex), oligodendrocyte (Oli), astrocyte (Ast), and microglia (Mic). The average proportions for these four cell types are 0.50, 0.19, 0.20, 0.08, respectively.

4.1. Gene Sets with Known Cell-type-specific Functions

The gene co-expressions estimated from different methods were compared on a few sets of genes. We first considered three sets of genes obtained from Gene Ontology (GO) (Ashburner et al. 2000; Consortium 2021) including the *excitatory synapse* genes (GO:0060076, $p = 46$), *myelin sheath* genes (GO:0043209, $p = 42$) and *astrocyte differentiation* genes (GO:0048708, $p = 72$), primarily functioning in excitatory neurons, oligodendrocytes and astrocytes, respectively. Specifically, the *excitatory synapse* gene set contains genes whose products function mainly in excitatory synapses, and the *myelin sheath* gene set has genes related to myelin sheath, which is supplied by oligodendrocytes to the central nervous system; the *astrocyte differentiation* gene set contains genes involved in the differentiation process of an astrocyte. These gene sets, according to their GO definitions, are expected to express and/or co-express primarily in the cell types that are relevant to their functions. In our analysis of these three gene sets, we focused on genes expressed in more than 25% of the ROSMAP bulk samples, resulting in sets of sizes $p = 45, 41$ and 68, respectively.

Figure 3 shows the co-expression estimates from `Bulk`, `CSNet`, and `s-bMIND` for the *excitatory synapse* gene set. It is seen that `CSNet` identified co-expressions specific to excitatory neurons, while `s-bMIND` suggested similar co-expression patterns in all four cell types. We also estimated cell-type-specific co-expressions for the *myelin sheath* and *astrocyte differentiation* gene sets, shown in Figures S7 and S8, respectively. These plots show that `CSNet` identified co-expressions specific to oligodendrocytes and astrocytes, respectively, while `s-bMIND` again estimated similar co-expressions across four cell types. We also conducted an analysis that simultaneously considers all three GO gene sets and found that `CSNet` can distinguish the three gene sets in the corresponding cell types (Figure S9). In comparison, the bulk estimates cannot distinguish them and the co-expressions of *myelin sheath* and *astrocyte differentiation* gene sets are much weaker (Figure S9), suggesting that bulk estimates may miss co-expressions from these less abundant cell types. Finally, Figure 4 shows that the estimates based on single cell data are noisy and do not show any cell-type-specific co-expression patterns. Also in Figure 4, for all three gene sets, the strongest co-expressions are always observed in excitatory neurons, likely driven by the fact that it is the most abundant cell type (Mathys et al. 2019).

4.2. Alzheimer's Disease Risk Gene Set

Next, we focused on Alzheimer's disease risk genes from GWAS (see gene names in Table 4), which capture around 50% of the

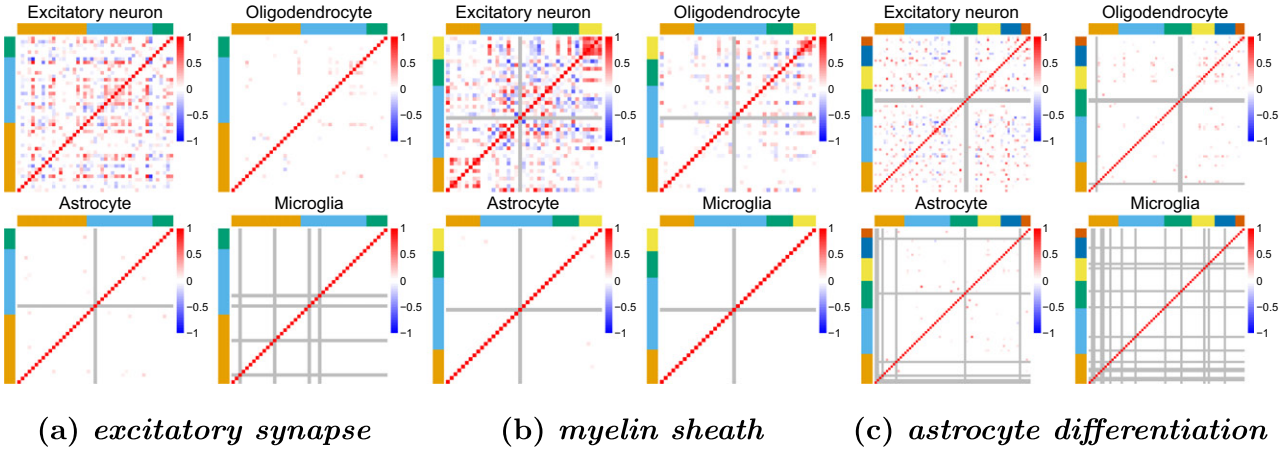

Figure 4. Co-expressions of *excitatory synapse*, *myelin sheath*, and *astrocyte differentiation* genes in different cell types, estimated using the ROSMAP single cell data (Mathys et al. 2019). Marked in gray are genes with no variation or no expression within a given cell type.

Table 4. The list of Alzheimer's disease risk genes.

Cluster	Genes
1	MEF2C, PILRA, MS4A8, EPHA1, IL34, CELF1, KAT8, KANSL1, HS3ST1, ACE, PTK2B, MAPT, NDUFAF7, TSPOAP1-AS1, SLC24A4, SORL1, PPARGC1A, C17orf107, APP, TPBG, CNTNAP2
2	IGHG3, HLA-DRB1, TRIP4, CASS4, MS4A14, PLCG2, INPP5D, SCIMP, MS4A7, CD33, TREM2, MS4A4A, MS4A6A, ABI3, SPI1
3	PRKD3, TP53INP1, CD2AP, ZNF423, CLU, SPPL2A, FERMT2, IQCK, RORA, ECHDC3
4	APOE, MS4A4E, ZCWPW1, ABCA7, BIN1, SHARPIN
5	HESX1, APH1B, ADAM10, PICALM, ZNF655
6	CR1, HBEGF, ADAMTS1, ADAMTS4

NOTE: Gene names are displayed by cluster and in the same order as they appear in Figure 5. Our analysis included only the sequenced genes with an FPKM greater than 0.1 in at least 50 ROSMAP samples.

heritability in late-onset AD (Sims, Hill, and Williams 2020). Our analysis focused on 61 genes with an FPKM greater than 0.1 in at least 50 ROSMAP samples. There is a growing literature on the molecular mechanisms and related cell types for these risk genes. Besides the well studied pathways of amyloid-β and tau processing, several other pathways have also been implicated (Pimenova, Raj, and Goate 2018; Sims, Hill, and Williams 2020), among which neuroinflammation was recently highlighted as one of the most important causal pathways in Alzheimer's disease (Heneka et al. 2015). Both microglia and astrocyte are the key cell types involved in such immune responses, and microglia, the innate immune cells in central nervous system, was prioritized as the cell type most enriched for GWAS associations (Skene and Grant 2016; Tansey, Cameron, and Hill 2018). Our analysis aims to use `CSNet` to explore the cell-type-specific co-expression patterns among these Alzheimer's disease risk genes.

Figure 5 shows the estimates from `Bulk`, `d-CSNet`, and `CSNet`, respectively. We applied the one standard error rule when selecting the tuning parameters for less abundant cell types (Section S3.2). The `s-bMIND` estimates are again similar across cell types, and are relegated to Figure S10(a) in the supplement. In Figure 5, some within cluster co-expressions from bulk samples are no longer seen in the cell-type-specific estimates, likely due to the confounding effect of cell type proportions. The `d-CSNet` and `CSNet` estimates in Figure 5(b) and (c) show that genes in Cluster 4 (colored in yellow) were co-expressed in astrocytes. This gene cluster includes APOE, a major Alzheimer's disease risk gene known to be highly expressed in astrocytes (Yamazaki et al. 2019). APOE protein is primarily produced in astrocytes, which then interacts with amyloid-β, which is involved in a central pathway of Alzheimer's disease (Yamazaki et al. 2019). Besides, both APOE and ABCA7 contribute to lipid metabolism and phagocytosis (Pimenova, Raj, and Goate 2018), consistent with their high co-expressions found in Cluster 4. The `CSNet` estimates for Cluster 4 further highlight their connections with several other Alzheimer's disease risk genes in astrocytes. Additionally, the `d-CSNet` and `CSNet` estimates in Figure 5(b) and (c) suggest that genes in Cluster 2 (colored in blue) were co-expressed in microglia, a finding supported by existing literature on Alzheimer's disease. First, 9 out of 15 genes in Cluster 2 are known to be involved in neuroinflammation and Alzheimer's disease mechanisms via microglia. Among them, the coding variants in PLCG2, TREM2, ABI3 implicate innate immunity in Alzheimer's disease as mediated by microglia (Sims et al. 2017); CD33 inhibits the uptake of amyloid-β in microglia (Griciuc et al. 2013); MS4A gene cluster is a key modulator of TREM2 in microglia (Deming et al. 2019) and SPI1 is a central regulator of microglia expression and Alzheimer's disease risk (Kosoy et al. 2021). In addition, 9 genes are known to express uniquely in microglia, including HLA-DRB1, PLCG2, CD33, TREM2, ABI3 and the MS4A gene cluster (Sims et al. 2017; Pimenova, Raj, and Goate 2018). The `d-CSNet` and `CSNet` estimates were able to identify cell-type-specific co-expression patterns of these genes, while single cell data based estimates could not (see Figure S10(b)), which possibly offered new insights into regulations of Alzheimer's disease risk genes. The estimated co-expressions in gene Clusters 2 and 4 reveal previously unknown cell-type-specific co-expressions among Alzheimer's disease risk genes, and may suggest cell-type-specific disease mechanisms.

Finally, the sensitivity analysis in Section S5.1 shows that `CSNet` remained robust as a reasonable amount of noise was added to the cell type proportions. We have also conducted a negative control experiment where cell type proportion vectors for different samples were randomly permuted. Figure S14 shows that the resulting estimates in excitatory neurons, the most abundant cell type, always resemble the bulk co-expression

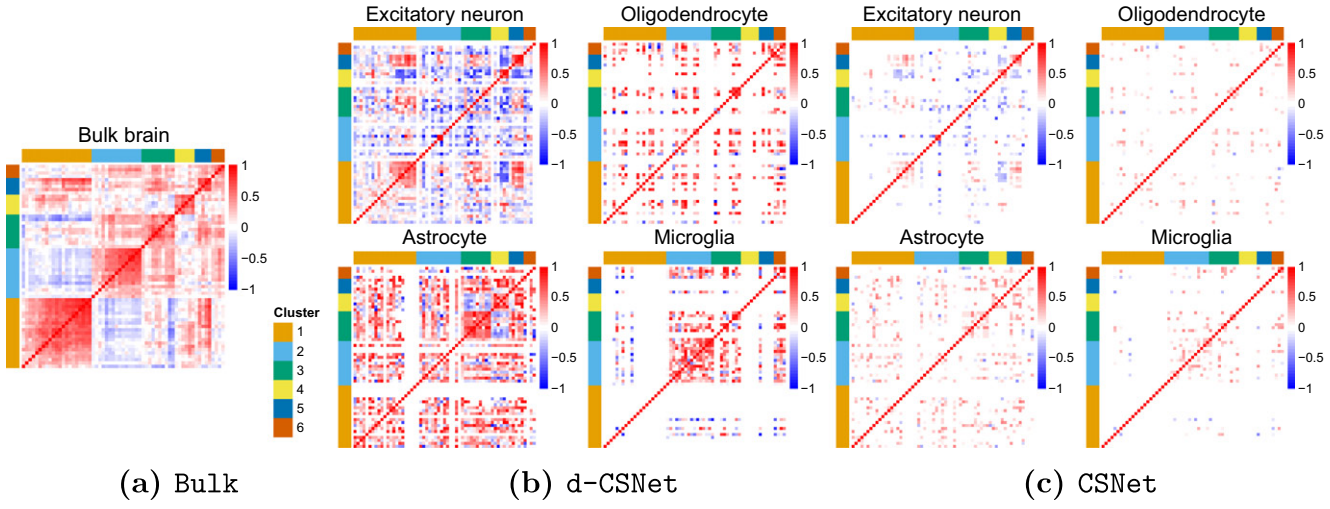

Figure 5. Co-expression networks of Alzheimer's disease risk genes inferred from the ROSMAP data. From left to right: sample correlation matrix estimated from bulk RNA-seq data, `d-CSNet` estimates, and `CSNet` estimates.

estimate, while the previously uncovered cell-type-specific co-expression patterns are no longer seen.

5. Discussion

Our model (1) is designed for gene expression data measured by the RNA-seq protocol, where sequencing read counts capture the expression levels for all cells in a tissue sample. We caution that the same model may not be applicable to microarray data, where expression levels have been transformed for normalization (Zhong and Liu 2012).

We have assumed that the cell type proportions π_{ik}'s are given in our analysis. In practice, they are estimated with existing methods such as Newman et al. (2019) and Chu et al. (2022), where it is shown that the correlations between their estimated proportions and ground truth proportions were generally around 0.9. Our empirical investigations showed that `CSNet` is not overly sensitive to errors in π_{ik}'s (see sensitivity analysis in Sections 3.2 and S5.1). It is possible to further extend our framework to accommodate noisy π_{ik}'s. In this case, we may further consider the $\sigma_{jj'}^{(k)}$'s to be estimated from $z_{ij}z_{ij'} = \sum_{k=1}^{K} \hat{\pi}_{ik}^2 \sigma_{jj'}^{(k)} + \xi_{ijj'}, j, j' \in [p]$, where $\xi_{ijj'} = \epsilon_{ijj'} + \sum_{k=1}^{K}(\pi_{ik}^2 - \hat{\pi}_{ik}^2)\sigma_{jj'}^{(k)}$. Hence, if the error $|\pi_{ik}^2 - \hat{\pi}_{ik}^2|$ is small, $\sigma_{jj'}^{(k)}$ should still be well estimated. We leave the full investigation of this topic as future research.

Our proposed work focuses on the estimation of correlation matrices $\boldsymbol{R}^{(k)}$'s that characterize marginal associations amongst genes. It may also be of interest to estimate the precision matrix defined as $\Omega^{(k)} = \Sigma^{(k)^{-1}}$. Under the Gaussian assumption, a zero (nonzero) entry in $\Omega^{(k)}$ indicates independence (dependence) between two genes conditional on all other genes. Without making the Gaussian assumption, the nonzero entries in $\Omega^{(k)}$ characterize partial correlations instead of conditional dependence. Under our framework, one feasible approach to estimate $\Omega^{(k)}$'s is to consider

$$\min_{\Omega^{(k)}} \|\hat{\Sigma}^{(k)}\Omega^{(k)} - \boldsymbol{I}\|_F^2 + \mathcal{P}_{\lambda_k}(\Omega^{(k)}) \tag{10}$$

where $\hat{\Sigma}^{(k)}$ is estimated from the first step of `CSNet` and $\mathcal{P}_{\lambda_k}(\cdot)$ is a sparse penalty function. This type of problems can be solved, for example, using the CLIME method (Cai, Liu, and Luo 2011). The formulation in (10) may incur a high computational cost, and the results can be sensitive to the selection of gene sets. We leave a full investigation of (10) to future work. In our current work, we focus on a co-expression network due to its computational efficiency, model flexibility and wide adoption in biomedical research (Langfelder and Horvath 2008).

In terms of estimation, instead of the two-step procedure considered in `CSNet`, an alternative approach is to integrate these two steps and consider

$$\sum_{i=1}^{n}\sum_{j,j'}\left(z_{ij}z_{ij'} - \sum_{k}\pi_{ik}^2\sigma_{jj'}^{(k)}\right)^2 + \sum_{k=1}^{K}\sum_{j\neq j'}\mathcal{P}_{\lambda_k}\left(\sigma_{jj'}^{(k)}/\sqrt{\sigma_{jj}^{(k)}\sigma_{j'j'}^{(k)}}\right), \tag{11}$$

where $\mathcal{P}_\lambda(x)$ denotes a penalty function, such as lasso or SCAD, with a regularization parameter λ. Estimating $\sigma_{jj'}^{(k)}$'s from (11) can be computationally costly, as there are $O(Kp^2)$ parameters to optimize together and it is necessary to tune the regularization parameters $\lambda_1, \ldots, \lambda_K$ jointly. In comparison, the two-step procedure in `CSNet` allows for a separate tuning of λ_k, and given the tuning parameters, the computational complexity of `CSNet` is comparable to that of naive covariance estimators for i.i.d. samples. While the two-step procedure in `CSNet` is not guaranteed to optimize (11), it is highly computationally efficient, which is appealing for practitioners. We also show in Theorem S1.1 that the error rate of `CSNet` is comparable to that of estimating sparse covariance matrices directly from i.i.d samples (Rothman et al. 2008). For finite sample cases, it may be desirable to ensure the positive definiteness of the final estimator. One strategy is to solve a constrained optimization problem, subject to positive definiteness, to find the nearest correlation matrix in Frobenius norm. This can be carried out efficiently using existing solvers (e.g., Qi and Sun 2006; Sun and Vandenberghe 2015).

Some recent methods have been developed to estimate cell-specific co-expression networks using single cell data, including Dai et al. (2019) and Wang, Choi, and Roeder (2021). It is important to note that cell-specific and cell-type-specific networks do not necessarily capture the same type of correlation of expression levels. For example, if two genes have the same expression level for all cells in a tissue sample and this expression level varies from sample to sample, then these two genes are not correlated in the cell-specific networks but are highly correlated in the network estimated using bulk samples, and this correlation may be due to regulations (e.g., genetic variants) at the sample level. Moreover, cell-specific network estimates may be subject to noises and may require averaging to facilitate interpretation (Wang, Choi, and Roeder 2021).

Recently, new computational tools have been developed to more accurately estimate the proportions of cell subtypes in bulk tissues, such as Chu et al. (2022) and Huang et al. (2023). Our proposed method can also be combined with these estimates to understand co-expressions in cell subtypes and to address potential confounding due to cell subtypes.

Finally, with an increasing number of studies collecting single cell data, we may also obtain more accurate co-expression estimates. For example, if a more accurate or refined reference signature matrix is available from the new single cell studies, one can generate updated cell type proportion estimates for bulk data (Newman et al. 2019) and use them in our procedure. As cells can be annotated into cell types, statistical methods have also been developed to directly infer cell-type-specific co-expression networks from single cell data, such as Su et al. (2022) and Lu and Keles (2023). Moreover, an integrated analysis of bulk and single cell data may also improve the co-expression estimates. For example, bMIND and ENIGMA have explored ways to extract information from single cell data to help with the estimation. However, platform differences and batch effects are prominent in integration, and have not been addressed well in these methods. We plan to explore along this direction in our future research.

Supplementary Materials

The supplementary materials comprise theoretical results, supplementary methods, figures, and additional numerical results. The codes for reproducing the results in this article are available at *https://github.com/ChangSuBiostats/CSNet_analysis*. An R package that implements `CSNet` is provided at *https://github.com/ChangSuBiostats/CSNet*.

Acknowledgments

We thank the ROSMAP project for their permission, requested at *https://www.radc.rush.edu*, to access the bulk and single nucleus RNA-seq data in the project. We are grateful to the Editor, the AE and three anonymous referees for their insightful comments that have substantially improved the quality, the presentation, and the reproducibility of the manuscript. We also thank Dr. Jiawei Wang at Yale University for helpful discussions on real data analysis.

Disclosure Statement

The authors report there are no competing interests to declare.

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). Su and Zhao were supported in part by NIH grants R01 GM134005 and R56 AG074015. Zhang was supported by NSF grant DMS 2210469 and DMS 2329296.

ORCID

Chang Su https://orcid.org/0000-0002-8704-1512
Jingfei Zhang https://orcid.org/0000-0001-9700-1103
Hongyu Zhao https://orcid.org/0000-0003-1195-9607

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. [812]

Alzheimer's Association. (2019), "2019 Alzheimer's Disease Facts and Figures," *Alzheimer's & Dementia*, 15, 321–387. [811]

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000), "Gene Ontology: Tool for the Unification of Biology," *Nature Genetics*, 25, 25–29. [819]

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. [812,818]

Bickel, P. J., and Levina, E. (2008a), "Covariance Regularization by Thresholding," *The Annals of Statistics*, 36, 2577–2604. [814]

——— (2008b), "Regularized Estimation of Large Covariance Matrices," *The Annals of Statistics*, 36, 199–227. [814]

Cai, T., Liu, W., and Luo, X. (2011), "A constrained l1 minimization approach to sparse precision matrix estimation," *Journal of the American Statistical Association*, 106, 594–607. [821]

Cai, Z., and Xiao, M. (2016), "Oligodendrocytes and Alzheimer's Disease," *International Journal of Neuroscience*, 126, 97–104. [812]

Chu, T., Wang, Z., Pe'er, D., and Danko, C. G. (2022), "Cell Type and Gene Expression Deconvolution with BayesPrism Enables Bayesian Integrative Analysis Across Bulk and Single-Cell RNA Sequencing in Oncology," *Nature Cancer*, 3, 505–517. [813,818,821,822]

Cobos, F. A., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P., and De Preter, K. (2020), "Benchmarking of Cell Type Deconvolution Pipelines for Transcriptomics Data," *Nature Communications*, 11, 1–14. [812]

Consortium, T. G. O. (2021), "The Gene Ontology Resource: Enriching a GOld Mine," *Nucleic Acids Research*, 49, D325–D334. [819]

Dai, H., Li, L., Zeng, T., and Chen, L. (2019), "Cell-Specific Network Constructed by Single-Cell RNA Sequencing Data," *Nucleic Acids Research*, 47, e62. [822]

De Strooper, B., and Karran, E. (2016), "The Cellular Phase of Alzheimer's Disease," *Cell*, 164, 603–615. [811]

Deming, Y., Filipello, F., Cignarella, F., Cantoni, C., Hsu, S., Mikesell, R., Li, Z., Del-Aguila, J. L., Dube, U., Farias, F. G., et al. (2019), "The MS4A Gene Cluster is a Key Modulator of Soluble TREM2 and Alzheimer's Disease Risk," *Science Translational Medicine*, 11, eaau2291. [820]

Denisenko, E., Guo, B. B., Jones, M., Hou, R., De Kock, L., Lassmann, T., Poppe, D., Clément, O., Simmons, R. K., Lister, R., et al. (2020), "Systematic Assessment of Tissue Dissociation and Storage Biases in Single-Cell and Single-Nucleus RNA-Seq Workflows," *Genome Biology*, 21, 1–25. [812]

Dong, M., Thennavan, A., Urrutia, E., Li, Y., Perou, C. M., Zou, F., and Jiang, Y. (2021), "SCDC: Bulk Gene Expression Deconvolution by Multiple Single-Cell RNA Sequencing References," *Briefings in Bioinformatics*, 22, 416–427. [813]

El Karoui, N. (2008), "Operator Norm Consistent Estimation of Large-Dimensional Sparse Covariance Matrices," *The Annals of Statistics*, 36, 2717–2756. [814]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American statistical Association*, 96, 1348–1360. [814]

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. [813]

Gaiteri, C., Ding, Y., French, B., Tseng, G. C., and Sibille, E. (2014), "Beyond Modules and Hubs: The Potential of Gene Coexpression Networks for Investigating Molecular Mechanisms of Complex Brain Disorders," *Genes, Brain and Behavior*, 13, 13–24. [811]

Griciuc, A., Serrano-Pozo, A., Parrado, A. R., Lesinski, A. N., Asselin, C. N., Mullin, K., Hooli, B., Choi, S. H., Hyman, B. T., and Tanzi, R. E. (2013), "Alzheimer's Disease Risk Gene CD33 Inhibits Microglial Uptake of Amyloid Beta," *Neuron*, 78, 631–643. [820]

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., et al. (2009), "Histone Modifications at Human Enhancers Reflect Global Cell-Type-Specific Gene Expression," *Nature*, 459, 108–112. [811,813]

Heneka, M. T., Carson, M. J., El Khoury, J., Landreth, G. E., Brosseron, F., Feinstein, D. L., Jacobs, A. H., Wyss-Coray, T., Vitorica, J., Ransohoff, R. M., et al. (2015), "Neuroinflammation in Alzheimer's Disease," *The Lancet Neurology*, 14, 388–405. [811,820]

Huang, P., Cai, M., Lu, X., McKennan, C., and Wang, J. (2023), "Accurate Estimation of Rare Cell Type Fractions from Tissue Omics Data via Hierarchical Deconvolution," bioRxiv, 2023–03. [822]

Hwang, B., Lee, J. H., and Bang, D. (2018), "Single-Cell RNA Sequencing Technologies and Bioinformatics Pipelines," *Experimental & Molecular Medicine*, 50, 1–14. [812]

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. [812,813]

Jiang, B. (2013), "Covariance Selection by Thresholding the Sample Correlation Matrix," *Statistics & Probability Letters*, 83, 2492–2498. [814]

Kiselev, V. Y., Andrews, T. S., and Hemberg, M. (2019), "Challenges in Unsupervised Clustering of Single-Cell RNA-Seq Data," *Nature Reviews Genetics*, 20, 273–282. [812]

Kosoy, R., Fullard, J., Zeng, B., Bendl, J., Dong, P., Rahman, S., Kleopoulos, S., Shao, Z., Humphrey, J., de Paiva Lopes, K., et al. (2021), "Genetics of the Human Microglia Regulome Refines Alzheimer's Disease Risk Loci," medRxiv. [820]

Langfelder, P., and Horvath, S. (2008), "WGCNA: An R Package for Weighted Correlation Network Analysis," *BMC Bioinformatics*, 9, 1–13. [819,821]

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. [813]

Love, M. I., Huber, W., and Anders, S. (2014), "Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2," *Genome Biology*, 15, 1–21. [815]

Lu, S., and Keles, S. (2023), "Debiased Personalized Gene Coexpression Networks for Population-Scale scRNA-seq Data," *Genome Research*, gr–277363. [822]

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 Alzheimer's Disease," *Nature*, 570, 332–337. [812,818,819,820]

Meng, G., and Mei, H. (2019), "Transcriptional Dysregulation Study Reveals a Core Network Involving the Progression of Alzheimer's Disease," *Frontiers in Aging Neuroscience*, 11, 101. [811]

Morabito, S., Miyoshi, E., Michael, N., Shahin, S., Martini, A. C., Head, E., Silva, J., Leavy, K., Perez-Rosendahl, M., and Swarup, V. (2021), "Single-Nucleus Chromatin Accessibility and Transcriptomic Characterization of Alzheimer's Disease," *Nature Genetics*, 53, 1143–1155. [812]

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. [811,812]

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. [813]

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. [812,813,815,818,821,822]

Pimenova, A. A., Raj, T., and Goate, A. M. (2018), "Untangling Genetic Risk for Alzheimer's Disease," *Biological Psychiatry*, 83, 300–310. [820]

Qi, H., and Sun, D. (2006), "A Quadratically Convergent Newton Method for Computing the Nearest Correlation Matrix," *SIAM Journal on Matrix Analysis and Applications*, 28, 360–385. [821]

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008), "Sparse Permutation Invariant Covariance Estimation," *Electronic Journal of Statistics*, 2, 494–515. [814,821]

Rothman, A. J., Levina, E., and Zhu, J. (2009), "Generalized Thresholding of Large Covariance Matrices," *Journal of the American Statistical Association*, 104, 177–186. [814,815,819]

Sims, R., Hill, M., and Williams, J. (2020), "The Multiplex Model of the Genetics of Alzheimer's Disease," *Nature Neuroscience*, 23, 311–322. [811,820]

Sims, R., Van Der Lee, S. J., Naj, A. C., Bellenguez, C., Badarinarayan, N., Jakobsdottir, J., Kunkle, B. W., Boland, A., Raybould, R., Bis, J. C., et al. (2017), "Rare Coding Variants in PLCG2, ABI3, and TREM2 Implicate Microglial-Mediated Innate Immunity in Alzheimer's Disease," *Nature Genetics*, 49, 1373–1384. [820]

Skene, N. G., and Grant, S. G. (2016), "Identification of Vulnerable Cell Types in Major Brain Disorders Using Single Cell Transcriptomes and Expression Weighted Cell Type Enrichment," *Frontiers in Neuroscience*, 10, 16. [820]

Stower, H. (2019), "Single-Cell Insights into Neurology," *Nature Medicine*, 25, 1799–1799. [812]

Su, C., Xu, Z., Shan, X., Cai, B., Zhao, H., and Zhang, J. (2022), "Cell-Type-Specific Co-Expression Inference from Single Cell RNA-Sequencing Data," *Nature Communications*, 14, 4846. [822]

Sun, T., Song, D., Li, W. V., and Li, J. J. (2021), "scDesign2: A Transparent Simulator that Generates High-Fidelity Single-Cell Gene Expression Count Data with Gene Correlations Captured," *Genome Biology*, 22, 1–37. [815]

Sun, Y., and Vandenberghe, L. (2015), "Decomposition Methods for Sparse Matrix Nearness Problems," *SIAM Journal on Matrix Analysis and Applications*, 36, 1691–1717. [821]

Tang, D., Park, S., and Zhao, H. (2020), "NITUMID: Nonnegative Matrix Factorization-based Immune-TUmor MIcroenvironment Deconvolution," *Bioinformatics*, 36, 1344–1350. [812]

Tansey, K. E., Cameron, D., and Hill, M. J. (2018), "Genetic Risk for Alzheimer's Disease is Concentrated in Specific Macrophage and Microglial Transcriptional Networks," *Genome Medicine*, 10, 1–10. [817,820]

Tian, J., Wang, J., and Roeder, K. (2021), "ESCO: Single Cell Expression Simulation Incorporating Gene Co-Expression," *Bioinformatics*, 37, 2374–2381. [815]

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. [818]

Wan, Y.-W., Al-Ouran, R., Mangleburg, C. G., Perumal, T. M., Lee, T. V., Allison, K., Swarup, V., Funk, C. C., Gaiteri, C., Allen, M., et al. (2020), "Meta-Analysis of the Alzheimer's Disease Human Brain Transcriptome and Functional Dissection in Mouse Models," *Cell Reports*, 32, 107908. [811]

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*, gr–268722. [812,815,819]

Wang, M., Li, A., Sekiya, M., Beckmann, N. D., Quan, X., Schrode, N., Fernando, M. B., Yu, A., Zhu, L., Cao, J., et al. (2021a), "Transformative

Network Modeling of Multi-Omics Data Reveals Detailed Circuits, Key Regulators, and Potential Therapeutics for Alzheimer's Disease," *Neuron*, 109, 257–272. [811]

Wang, M., Roussos, P., McKenzie, A., Zhou, X., Kajiwara, Y., Brennand, K. J., De Luca, G. C., Crary, J. F., Casaccia, P., Buxbaum, J. D., et al. (2016), "Integrative Network Analysis of Nineteen Brain Regions Identifies Molecular Signatures and Networks Underlying Selective Regional Vulnerability to Alzheimer's Disease," *Genome Medicine*, 8, 1–21. [811]

Wang, W., Yao, J., Wang, Y., Zhang, C., Tao, W., Zou, J., and Ni, T. (2021b), "Improved Estimation of Cell Type-Specific Gene Expression through Deconvolution of Bulk Tissues with Matrix Completion," bioRxiv, 2021–06. [812,815]

Wang, X., Choi, D., and Roeder, K. (2021), "Constructing Local Cell-Specific Networks from Single-Cell Data," *Proceedings of the National Academy of Sciences*, 118, e2113178118. [822]

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. [812,813]

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. [811]

Yamazaki, Y., Zhao, N., Caulfield, T. R., Liu, C.-C., and Bu, G. (2019), "Apolipoprotein E and Alzheimer Disease: Pathobiology and Targeting Strategies," *Nature Reviews Neurology*, 15, 501–518. [820]

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. [812,813]

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. [811]

Zhang, B., and Horvath, S. (2005), "A General Framework for Weighted Gene Co-expression Network Analysis," *Statistical Applications in Genetics and Molecular Biology*, 4. [814]

Zhong, Y., and Liu, Z. (2012), "Gene Expression Deconvolution in Linear Space," *Nature Methods*, 9, 8–9. [821]