Estimating cell-type-specific gene co-expression networks from bulk gene expression data with an application to Alzheimer's disease

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

[a] Department of Biostatistics, Yale University

[b] Department of Management Science, University of Miami

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 aggregate 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 paper, we describe 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 pathology for Alzheimer's disease.

Keywords: bulk RNA-seq; cell-type-specific analysis; deconvolution; gene co-expression networks; sparse covariance estimation.

1 Introduction

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). In the past decades, 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., 2021b). 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 and is now a new therapeutic target for the disease.

While the literature on gene co-expression networks is readily growing, most co-expression analyses to date have been performed on data collected from bulk tissue samples that aggregate 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-expressed as their expression levels co-vary with the proportion of this specific cell type in the bulk sample. To avoid such confounded results and to gain a more accurate and comprehensive view of the underlying biological processes, 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 et al., 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 due to low coverage and biological noises (Kiselev et al., 2019), and may be biased due to cell isolation and sequencing protocols (Denisenko et al., 2020). 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 cell-type-specific co-expression analysis, we consider the use of bulk sample data for such analysis in this paper. With the readily available rich collections of bulk gene expression data, there is a great need to develop methods for estimating cell-type-specific co-expressions from bulk samples. There is a recent 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), or to infer cell type proportions (Abbas et al., 2009; Wang et al., 2019; Newman et al., 2019; Tang et al., 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) and bMIND (Wang et al., 2021a). 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.

Our real data application focuses on the Alzheimer's disease, 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 et al., 2020). To understand the mechanisms of these disease associated risk genes and the pathology 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., 2021b). However, most co-expression analyses focus on correlations between bulk samples, that may be confounded with cell type compositions and only offer an aggregated view of the biological processes in different cell types. Recently, more evidence suggests cell-type-specific pathology of Alzheimer's disease. 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 mainly contributed by oligoden-

drocytes (Cai and Xiao, 2016). While such evidence is rapidly increasing, there have been very few cell-type-specific co-expression analyses for Alzheimer's disease. In our analysis, we focused on the bulk RNA-seq data 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. Using `CSNet`, we estimated gene co-expression networks for four major cell types in the brain, including excitatory neurons, oligodendrocytes, astrocytes 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. Both astrocytes 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 co-expressions in these two cell types [see Figure S6(b)]. We have also considered gene sets that function primarily in specific cell types to validate `CSNet` and added several sensitivity analyses to further validate our results.

The rest of the paper is organized as follows. Section 2 introduces the 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 an Alzheimer's disease risk gene set using bulk RNA-seq data from the ROSMAP study. Section 5 investigates theoretical properties of the proposed estimator. Section 6 concludes the paper with a brief discussion.

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 $\boldsymbol{\Sigma}^{(k)} \in \mathbb{R}^{p \times p}$, where $\boldsymbol{\mu}^{(k)}$ and $\boldsymbol{\Sigma}^{(k)}$ characterize, respectively, the cell-type-specific mean gene expression and co-expression across samples. As the gene regulation mechanisms in functionally distinct cell types are different (Heintzman et al., 2009), we further assume that $\boldsymbol{x}_i^{(1)}, \ldots, \boldsymbol{x}_i^{(K)}$ are independent (see discussions in Section 6). Correspondingly, we can write

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

As $\boldsymbol{\Sigma}^{(k)}$ 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 $\boldsymbol{R}^{(k)} = \boldsymbol{D}_k^{-1/2} \boldsymbol{\Sigma}^{(k)} \boldsymbol{D}_k^{-1/2}$, where $\boldsymbol{D}_k$ is a $p \times p$ diagonal matrix with the same diagonal as $\boldsymbol{\Sigma}^{(k)}$.

Denote $[m] = \{1, 2, \ldots, m\}$ for a positive integer m. We shall assume the cell-type proportions π_{ik}'s in (2) are given in the ensuing development, and later demonstrate in experiments that our method is not sensitive to uncertainties in π_{ik}'s; see discussions in Section 6. To infer π_{ik}'s from bulk samples, many methods have been developed that utilize cell type marker genes (i.e., genes that are only highly expressed in one cell type of interest) with expressions 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). 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 $\boldsymbol{R}^{(k)}$ with large p

It is easily seen from (1) and (2) that each bulk sample $\boldsymbol{x}_i$ is from a convolution of K distributions. In this case, estimating $\{\boldsymbol{\Sigma}^{(k)}\}_{k\in[K]}$ 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,

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

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 $\{\boldsymbol{\Sigma}^{(k)}\}_{k\in[K]}$, 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, especially when $p > n$ as in our problem.

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 $\boldsymbol{\Sigma}^{(k)} = (\sigma_{jj'}^{(k)})_{p\times p}$, (1) and (2) together imply

$$x_{ij} = \sum_{k=1}^{K}\pi_{ik}\mu_j^{(k)} + z_{ij}, \quad j\in[p], \tag{3}$$

$$z_{ij}z_{ij'} = \sum_{k=1}^{K}\pi_{ik}^2\sigma_{jj'}^{(k)} + \epsilon_{ijj'}, \quad j,j'\in[p], \tag{4}$$

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 (3)-(4) hold generally without parametric assumptions on the distributions from the K cell types.

Denote $\boldsymbol{y}_j = (x_{1j}, \ldots, x_{nj})$ and $\boldsymbol{D} = (\pi_{ik})_{n\times K}$. Equation (3) entails estimation of the cell-type-specific mean $\boldsymbol{\mu}^{(k)}$ via

$$\hat{\mu}_j^{(k)} = \left[\left(\boldsymbol{D}^\top \boldsymbol{D}\right)^{-1} \boldsymbol{D}^\top \boldsymbol{y}_j\right]_k, \quad j \in [p],\ k \in [K], \tag{5}$$

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 $\boldsymbol{H} = (\pi_{ik}^2)_{n\times K}$, equation (4) entails estimation of the cell-type-specific covariance $\sigma_{jj'}^{(k)}$ via

$$\hat{\sigma}_{jj'}^{(k)} = \left[\left(\boldsymbol{H}^\top \boldsymbol{H}\right)^{-1} \boldsymbol{H}^\top (\hat{\boldsymbol{z}}_j \circ \hat{\boldsymbol{z}}_{j'})\right]_k, \quad j, j' \in [p],\ k \in [K], \tag{6}$$

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

$$\hat{R}_{jj'}^{(k)} = \hat{\sigma}_{jj'}^{(k)} / \sqrt{\hat{\sigma}_{jj}^{(k)} \hat{\sigma}_{j'j'}^{(k)}}, \quad j, j' \in [p],\ k \in [K]. \tag{7}$$

In Section A3.2, we show that element-wisely $\hat{R}_{jj'}^{(k)}$ is consistent with a $\sqrt{n}$-convergence rate under certain regularity conditions.

Though each entry in the correlation matrix estimated in (7) 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, 2009) and correlation matrices (El Karoui, 2008; Jiang, 2013). To facilitate estimability and interpretability, we assume that $\boldsymbol{\Sigma}^{(k)}$ (or equivalently $\boldsymbol{R}^{(k)}$) is approximately sparse for all k; see the definition of approximate sparsity in Assumption 2. 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

$$\left[\mathcal{T}_{\lambda_k}\left(\hat{\boldsymbol{R}}^{(k)}\right)\right]_{jj'} = \text{sign}\left(\hat{R}^{(k)}_{jj'}\right) \times \begin{cases} \left(\left|\hat{R}^{(k)}_{jj'}\right| - \lambda_k\right)_+ & |\hat{R}^{(k)}_{jj'}| \leq 2\lambda_k \\ \lambda_k + \dfrac{a-1}{a-2}\left(\left|\hat{R}^{(k)}_{jj'}\right| - 2\lambda_k\right) & \left|\hat{R}^{(k)}_{jj'}\right| \in (2\lambda_k, a\lambda_k) \\ \left|\hat{R}^{(k)}_{jj'}\right| & \text{otherwise,} \end{cases} \tag{8}$$

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, 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 et al., 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. However, 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 5, we show the convergence rates of `CSNet`, i.e., $\mathcal{T}_{\lambda_k}(\hat{\boldsymbol{R}}^{(k)})$, 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, though it is asymptotically positive definite as ensured by Theorems 5.1-5.2; see more discussions in Section 6 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 (6) for all entries in covariance matrices. Though (6) is computed element-wisely, the matrix $\left(\boldsymbol{H}^\top \boldsymbol{H}\right)^{-1} \boldsymbol{H}^\top \in \mathbb{R}^{K\times n}$ is common and only needs to be calculated once. Hence, entries in $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K$ can be estimated efficiently via $\left(\boldsymbol{H}^\top \boldsymbol{H}\right)^{-1} \boldsymbol{H}^\top \boldsymbol{Y}$, where $\boldsymbol{Y}$ 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 $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K$ is $O(Kn^2p^2)$, while that of a naive sample covariance estimation is $O(np^2)$. As the number of cell types K is usually small, the proposed element-wise estimation is computationally feasible when n, the number of bulk RNA-seq samples, is moderate.

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 data set. Here, we introduce the cross validation procedure and note that selection with a validation data set 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 (7). We denote the estimated correlation matrices from these two data splits as $\hat{\boldsymbol{R}}_1^{(k)}$ and $\hat{\boldsymbol{R}}_2^{(k)}$, $k \in [K]$, respectively. For each k, the tuning parameter λ_k is selected by minimizing $\|\mathcal{T}_{\lambda_k}(\hat{\boldsymbol{R}}_1^{(k)}) - \hat{\boldsymbol{R}}_2^{(k)}\|_F^2 + \|\mathcal{T}_{\lambda_k}(\hat{\boldsymbol{R}}_2^{(k)}) - \hat{\boldsymbol{R}}_1^{(k)}\|_F^2$ among a set of working values, where $\|\cdot\|_F$ denotes Frobenius norm. We consider two equal sized data splits as sufficient samples are needed to estimate $\hat{\boldsymbol{R}}_1^{(k)}$ and $\hat{\boldsymbol{R}}_2^{(k)}$ 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. 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.

3 Simulation Studies

We first investigate the finite sample performance of `CSNet` and compare it with two existing methods, and then conduct a sensitive analysis to examine the performance of `CSNet` when provided cell type proportions are inaccurate. A notable contribution from our simulation studies is a new and efficient method for simulating from multivariate negative binomial distributions with large and structured covariances, presented in Algorithm 1 in the supplement, which can be of independent interest.

3.1 Comparing `CSNet` with other methods

In this section, we consider [1] the proposed `CSNet` estimator calculated as in (8), [2] the cell-type-specific estimator in (7) without sparsity, referred to as `d-CSNet`, [3] the sparse estimator with SCAD thresholding in Rothman et al. (2009) for bulk samples (i.e., not cell-type-specific), referred to as `Bulk`, and [4] the estimator computed from the sample level cell-type-specific expressions estimated by `bMIND` (Wang et al., 2021a). 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 experiments, `bMIND` is 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 S1 and S3 in the supplement and the results remain similar. 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. Both `CSNet` and `d-CSNet` are computed using the procedure in Section 2.2 with nonnegative least squares to estimate the unknown $\boldsymbol{\mu}^{(k)}$'s.

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

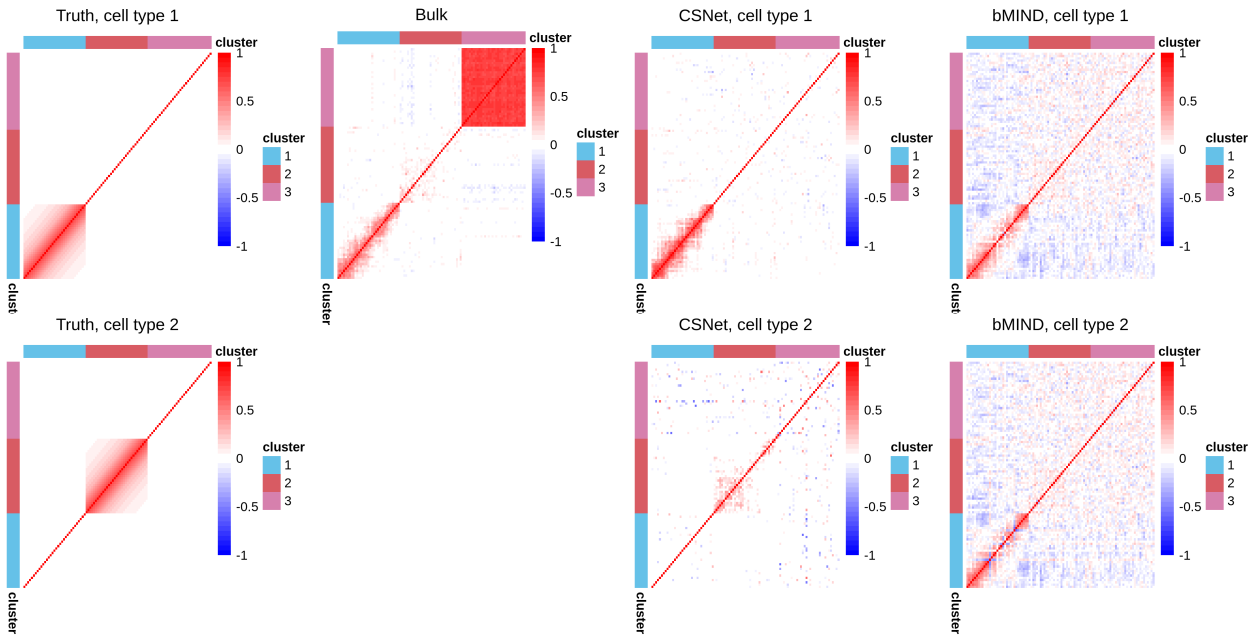


Figure 1: Correlation matrices in one data replicate with $p = 100$ and $n = 150$, under a negative binomial distribution with an AR(1)-type correlation structure in V_1 or V_2 (See Section 3). From left to right: true correlation matrices, estimates from `Bulk`, `CSNet` and `bMIND`. Note the `Bulk`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

$K = 2$ 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 et al., 2014), with mean $\boldsymbol{\mu}^{(k)}$'s and covariance matrices $\boldsymbol{\Sigma}^{(k)}$'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 an MA(1) structure with $\rho_{jj'} = 0.39 \times \mathbf{1}_{|j-j'|=1}$ and an AR(1)-type structure with $\rho_{jj'} = 0.70 \times 0.9^{|j-j'-1|}$, for $j \neq j' \in V_1$ or V_2. We set $\log \sigma_{jj}^{(1)} = 8$ for all j, $\log \sigma_{jj}^{(2)} = 8$ for $j \in V_1, V_2$ and $\log \sigma_{jj}^{(2)} = 7$ for $j \in V_3$ with sequencing depth set to $S = 6 \times 10^7$ in Algorithm 1 to mimic highly-expressed protein-coding genes in real sequencing data. The mean $\boldsymbol{\mu}^{(k)}$ is set to be a function of $\boldsymbol{\Sigma}^{(k)}$ (see details in Algorithm1), consistent with the observation in real data that higher expression

levels are often associated with larger variances. The tuning parameters for `CSNet` and `Bulk` are selected following Section 2.3 and the suggested procedure in Rothman et al. (2009), respectively, both using a validation data set with 150 independent 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 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, denotes the estimate of $\boldsymbol{R}^{(k)}$ obtained by various methods. Also reported are the true positive rate (TPR) and false positive rate (FPR), which evaluate the selection accuracy of nonzero entries in $\boldsymbol{R}^{(k)}$'s. The TPR and FPR are only reported for sparse estimators `Bulk` and `CSNet`. Table 1 reports the average criteria under the MA(1) model for four methods, with standard deviations in the parentheses, over 200 data replications. The results for AR(1)-type are similar and relegated to Table S2 in the supplement due to space limitation. It shows that `CSNet` achieves the best performance in terms of both estimation accuracy and selection accuracy. In the supplement, we demonstrate in Tables S1 and S3 that even with informative priors derived from simulated cell-type-specific data, `CSNet` still performs better than `bMIND`.

To better visualize the estimates, Figure 1 plots heat maps of the true cell-type-specific co-expression matrices and estimates from `Bulk`, `CSNet` and `bMIND`. The estimates from `d-CSNet` show similar patterns as `CSNet` plus some additional noises, and are omitted due to space limitation. From Figure 1, it is clearly seen that both `Bulk` and `bMIND` 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, `bMIND` does not perform well for cell type 2, the less abundent 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. In

			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`	25.13 (0.17)	24.64 (0.17)	0.16 (0.02)	**0.97 (0.02)**	25.24 (0.17)	24.64 (0.17)	0.16 (0.02)	**0.69 (0.03)**
		`d-CSNet`	11.91 (0.16)	3.82 (0.21)	- -	- -	25.94 (0.79)	9.48 (0.68)	- -	- -
		`CSNet`	**2.35 (0.11)**	**0.63 (0.04)**	**0.03 (0.01)**	0.96 (0.02)	**3.11 (0.05)**	**0.77 (0.01)**	**0.00 (0.00)**	0.61 (0.01)
		`bMIND`	9.40 (0.13)	2.79 (0.15)	- -	- -	10.05 (0.15)	2.97 (0.16)	- -	- -
	200	`Bulk`	49.06 (0.20)	48.54 (0.20)	0.14 (0.01)	**0.96 (0.02)**	49.17 (0.20)	48.54 (0.20)	0.14 (0.01)	**0.67 (0.02)**
		`d-CSNet`	23.90 (0.16)	6.59 (0.30)	- -	- -	51.64 (1.21)	17.17 (0.97)	- -	- -
		`CSNet`	**3.58 (0.10)**	**0.68 (0.02)**	**0.02 (0.00)**	0.93 (0.02)	**4.45 (0.03)**	**0.77 (0.00)**	**0.00 (0.00)**	0.60 (0.00)
		`bMIND`	18.76 (0.15)	4.70 (0.19)	- -	- -	19.42 (0.17)	4.84 (0.20)	- -	- -
600	100	`Bulk`	25.17 (0.08)	24.77 (0.08)	0.16 (0.02)	**1.00 (0.00)**	25.41 (0.08)	24.77 (0.08)	0.16 (0.02)	0.85 (0.05)
		`d-CSNet`	5.87 (0.06)	1.53 (0.07)	- -	- -	12.13 (0.22)	3.46 (0.20)	- -	- -
		`CSNet`	**0.91 (0.08)**	**0.27 (0.04)**	**0.04 (0.01)**	**1.00 (0.00)**	**2.56 (0.13)**	**0.65 (0.05)**	**0.03 (0.01)**	**0.94 (0.04)**
		`bMIND`	5.02 (0.05)	1.24 (0.06)	- -	- -	5.88 (0.09)	1.63 (0.13)	- -	- -
	200	`Bulk`	49.18 (0.12)	48.77 (0.12)	0.14 (0.01)	**1.00 (0.00)**	49.43 (0.12)	48.77 (0.12)	0.14 (0.01)	0.82 (0.04)
		`d-CSNet`	11.79 (0.07)	2.46 (0.09)	- -	- -	24.23 (0.29)	5.90 (0.29)	- -	- -
		`CSNet`	**1.47 (0.08)**	**0.32 (0.03)**	**0.02 (0.01)**	**1.00 (0.00)**	**3.89 (0.14)**	**0.71 (0.07)**	**0.01 (0.01)**	**0.91 (0.05)**
		`bMIND`	9.88 (0.06)	1.96 (0.06)	- -	- -	11.04 (0.11)	3.02 (0.22)	- -	- -

Table 1: Evaluation criteria with varying sample size n and network size p under a negative binomial distribution with an MA(1) correlation structure in V_1 or V_2 (See Section 3). The four methods under comparison are `Bulk`, `d-CSNet`, `CSNet` and `bMIND`. 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.

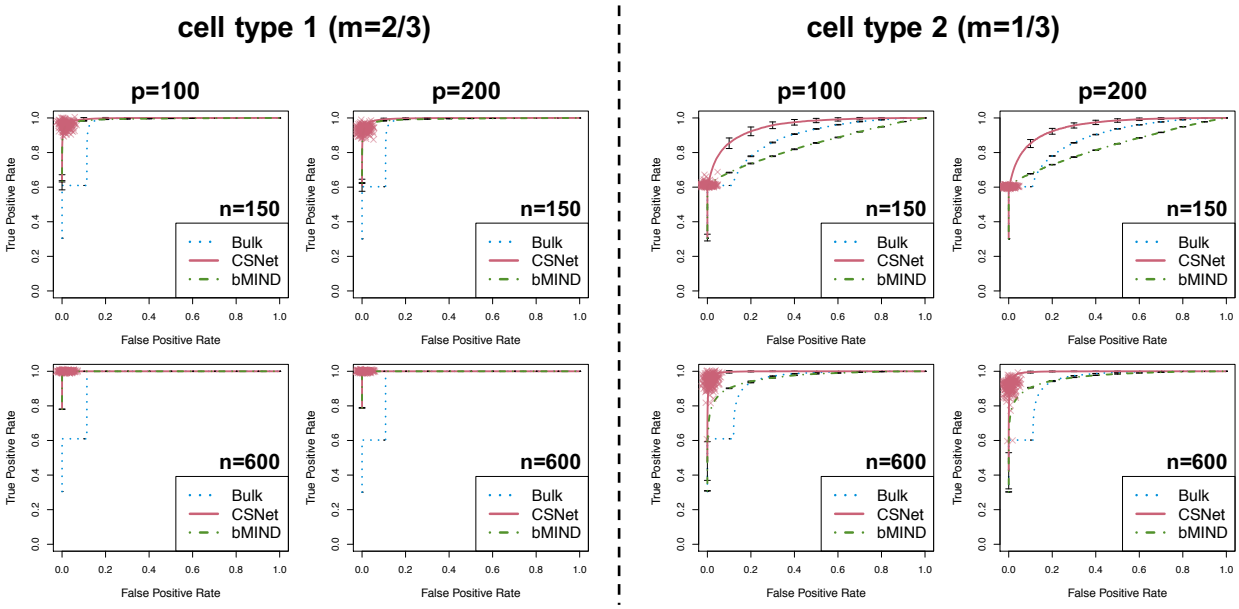


Figure 2: ROC curves with varying sample size n and network size p under a negative binomial distribution with an MA(1) correlation structure in V_1 or V_2 (See section 3). The three methods under comparison are `Bulk`, `CSNet` and `bMIND`. The thresholds selected by our proposed procedure in Section 2.3 are marked on the curves (in red crosses).

comparison, `CSNet` was able to identify the true co-expression patterns in both cell types.

Finally, 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` and `bMIND`. Though `bMIND` is not sparse, we applied the SCAD thresholding operator to the `bMIND` estimates to examine its performance. The thresholds selected by our proposed procedure in Section 2.3 are marked on the curves for `CSNet`. The ROC curves in Figures 2 show that `CSNet` achieves the best performance and the selected thresholds generally strike a reasonable balance between TPR and FPR. As shown in Table 1 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.

	$m = 2/3$				$m = 1/3$			
	F-norm	Op.-norm	FPR	TPR	F-norm	Op.-norm	FPR	TPR
$\kappa = 0$	2.35 (0.11)	0.63 (0.04)	0.03 (0.01)	0.96 (0.02)	3.11 (0.05)	0.77 (0.01)	0.00 (0.00)	0.61 (0.01)
$\kappa = 0.25$	2.38 (0.11)	0.63 (0.04)	0.03 (0.01)	0.95 (0.02)	3.11 (0.05)	0.77 (0.01)	0.00 (0.00)	0.61 (0.01)
$\kappa = 0.5$	2.72 (0.10)	0.68 (0.03)	0.03 (0.01)	0.89 (0.04)	3.11 (0.04)	0.77 (0.00)	0.00 (0.00)	0.61 (0.01)
$\kappa = 1$	3.09 (0.01)	0.77 (0.00)	0.00 (0.00)	0.62 (0.01)	3.10 (0.02)	0.77 (0.00)	0.00 (0.01)	0.61 (0.00)

Table 2: Sensitivity analysis of `CSNet` with $n = 150$ and $p = 100$, under the same setting as Table 1. Noisy cell type proportions were set to $\pi_{i1} + \kappa \times N(0, 0.04)$, and then thresholded to be within $[0, 1]$.

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 used in `CSNet` are inaccurate. We consider the same simulation setting as in Table 1 with $n = 150$, $p = 100$ and $K = 2$. For $i \in [n]$, let $\tilde{\pi}_{i1} = \pi_{i1} + e_i$ (constrained to be within $[0, 1]$), where e_i's are independent from $\kappa \times N(0, 0.04)$ and κ is set to $\{0, 0.25, 0.5, 1\}$. When $\kappa = 0$, $\tilde{\pi}_{i1} = \pi_{i1}$ and the cell type proportions are accurate with no errors. The estimation is then carried out following the procedure in Section 2.2 with inaccurate cell type proportions $\tilde{\pi}_{ik}$'s.

Table 2 reports the evaluation criteria for estimating cell-type-specific correlation matrices $\boldsymbol{R}^{(k)}$'s with `CSNet` under various noise levels. It is seen that under this inaccurate cell type proportion setting, our method still performs reasonably well.

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) (Bennett et al.,

2018), a clinical-pathologic cohort study of Alzheimer's disease. In the ROSMAP study, post-mortem brain samples from $n = 541$ subjects were collected from the grey matter of dorsolateral prefrontal cortex (DLPFC), a brain region heavily implicated in Alzheimer's disease pathology. 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 S3). 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.

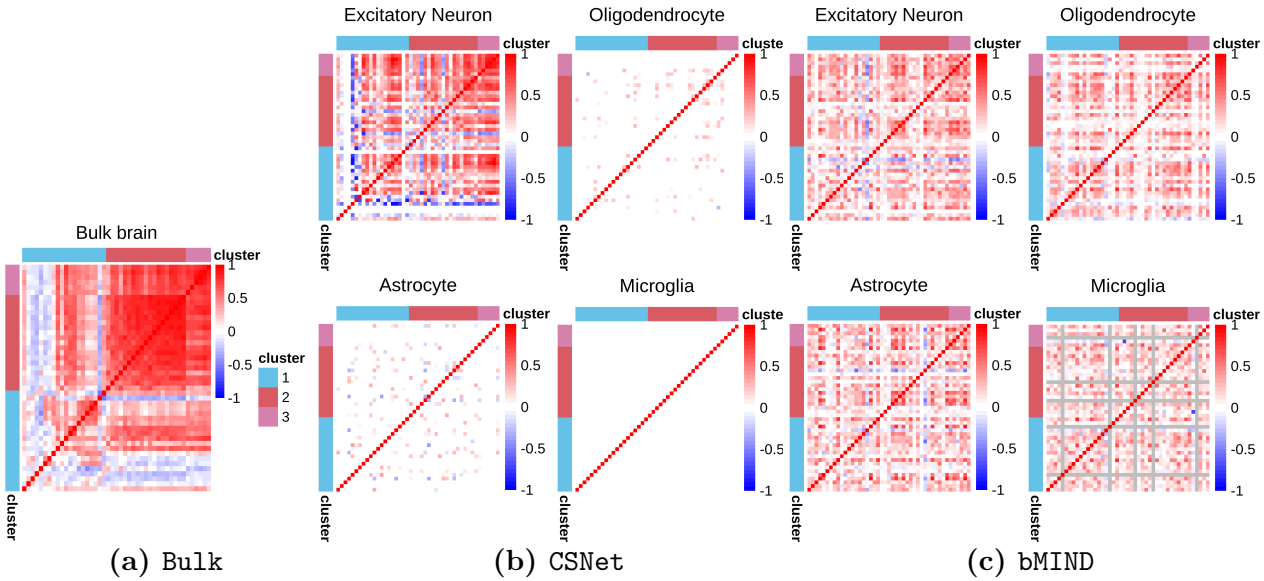
In the ensuing analysis, we focus on four most abundant cell types: excitatory neurons (Ex), oligodendrocytes (Oli), astrocytes (Ast) and microglia (Mic). The average proportions for these four cell types are 0.48, 0.21, 0.18 and 0.08, respectively. We applied `CSNet` as defined in (8), with the tuning parameter selected using the cross validation procedure discussed in Section 2.3. Additionally, we compared `CSNet` to two alternative approaches, one through `bMIND`, the best performing alternative in Section 3, and one using single cell data (Mathys et al., 2019), respectively. For `bMIND` estimates, we followed Wang et al. (2021a) to infer priors from single cell data and supplied these priors to estimate cell-type-specific expressions in each sample; see details in Section A4; 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 et al. (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.

4.1 Gene sets with known cell-type-specific functions

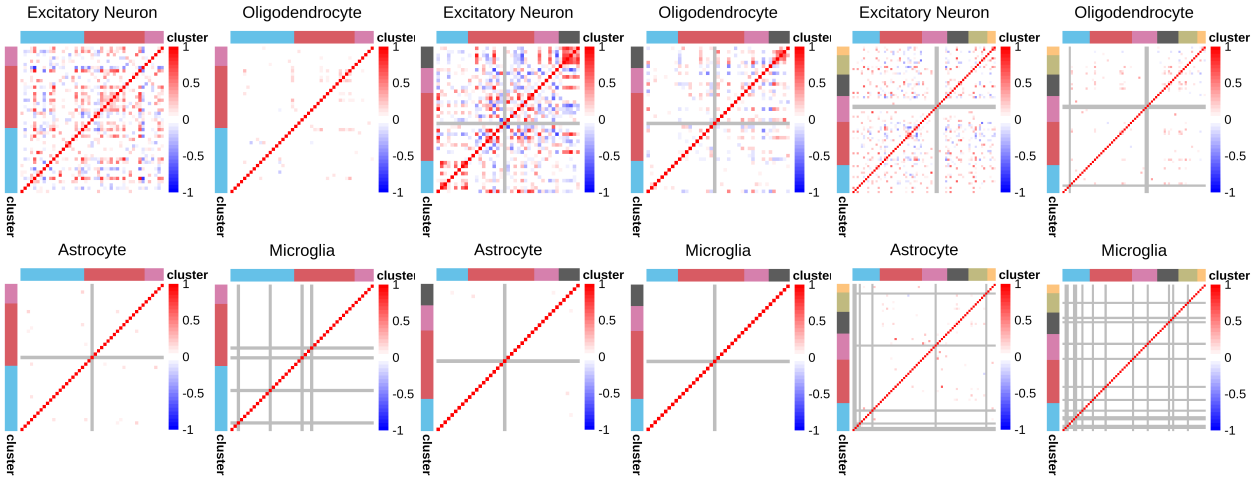
The gene co-expressions estimated from different methods were compared on a few sets of genes. We first considered three sets of genes obtained from Gene Ontology (GO) (Ashburner et al., 2000; Consortium, 2021) including the *excitatory synapse* genes (GO:0060076, $p = 46$), *myelin sheath* genes (GO:0043209, $p = 42$) 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 `bMIND` for the *excitatory synapse* gene set. It is seen that `CSNet` identified co-expressions specific to excitatory neurons, while `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 S4 and S5, respectively. These plots show that `CSNet` identified co-expressions specific to oligodendrocyte and astrocyte, respectively, while `bMIND` again estimated similar co-expressions across four cell types. Finally, Figure 4 shows that the esti-



(a) `Bulk` **(b)** `CSNet` **(c)** `bMIND`

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) `bMIND` estimates. For `bMIND`, genes with constant expression estimates across samples are marked in gray.



(a) ***excitatory synapse*** **(b)** ***myelin sheath*** **(c)** ***astrocyte differentiation***

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 expressions within a given cell type.

mates 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).

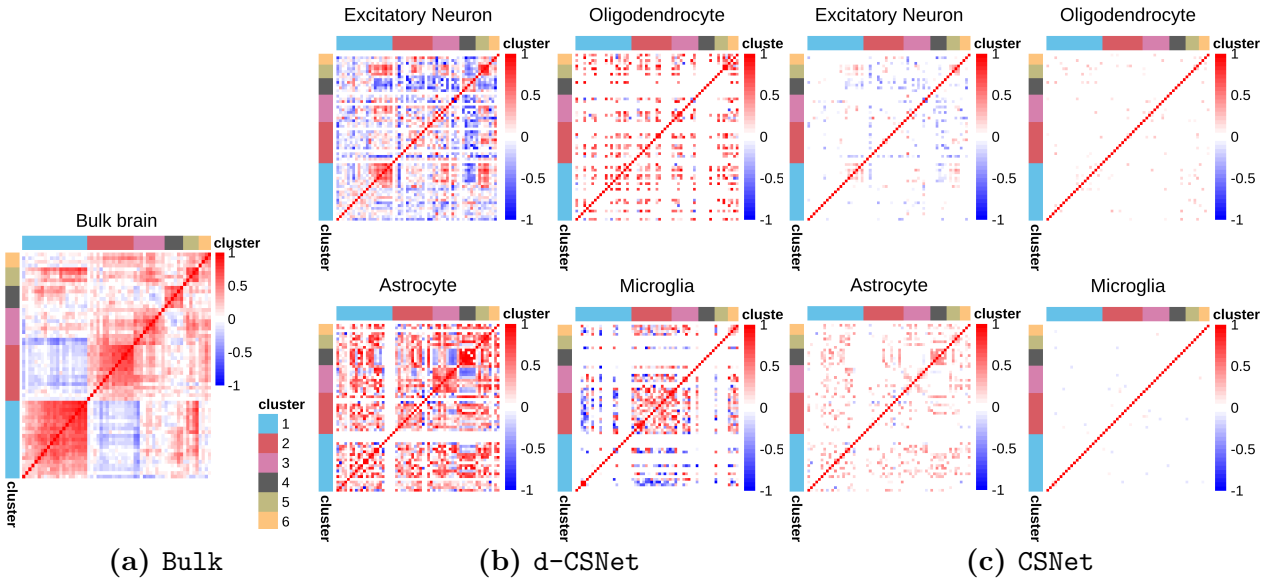
Cluster	Genes
1	MEF2C, PILRA, MS4A8, EPHA1, IL34, CELF1, KAT8, KANSL1, HS3ST1, ACE, PTK2B, MAPT, NDUFAF7, TSPOAP1-AS1, SLC24A4, SORL1, PPARGC1A, C17orf107, APP, TPBG, CNTNAP2
2	IGHG3, HLA-DRB1, TRIP4, CASS4, MS4A14, PLCG2, INPP5D, SCIMP, MS4A7, CD33, TREM2, MS4A4A, MS4A6A, ABI3, SPI1
3	PRKD3, TP53INP1, 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

Table 3: The list of Alzheimer's disease risk genes. 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.

4.2 Alzheimer's disease risk gene set

Next, we focused on Alzheimer's disease risk genes from GWAS (see gene names in Table 3), which capture around 50% of the heritability in late-onset AD (Sims et al., 2020). Our analysis focused on 61 genes with a 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 et al., 2018; Sims et al., 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, were prioritized as the cell type most enriched for GWAS associations (Skene and Grant, 2016; Tansey et al., 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. The `bMIND` estimates are again similar across cell types, and are relegated to Figure S6(a) in the supplement. In Figure 5, some within cluster co-expressions from bulk samples are no longer seen in



(a) `Bulk` (b) `d-CSNet` (c) `CSNet`

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 [panel (a)], d-CSNet estimates [panel (b)], and CSNet estimates [panel (c)].

the cell-type-specific estimates, likely due to the confounding effect of cell type proportions. The `CSNet` estimates in Figure 5(c) show that Cluster 4 (colored in black) were highly 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 et al., 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` estimates in Figure 5(b) suggest Cluster 2 (colored in red) were co-expressed in microglia, though the signals are relatively weak and the `CSNet` estimator for microglia is nearly diagonal. Nevertheless, the co-expression in Cluster 2 is likely microglia specific supported by several existing findings in Alzheimer's disease. Firstly, 9 out of 15 genes in Cluster 2 are known to be involved in neuroinflammation and Alzheimer's disease

pathology 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 et al., 2018). The `d-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 S6(b)], and possibly offer new insights into regulations of Alzheimer's disease risk genes. The estimated co-expressions in gene Clusters 1 and 2 reveal previously unknown cell-type-specific co-expressions among Alzheimer's disease risk genes, and may suggest cell-type-specific disease pathology for AD.

Finally, the sensitivity analysis in Section A2.1 shows that `CSNet` remains robust as a reasonable amount of noise is added to the cell type proportions. We have also added a negative control experiment where cell type proportion vectors for different samples were randomly permuted. Figure S8 shows that the resulting estimates in excitatory neurons, the most abundent cell type, always resemble the bulk co-expression estimate, while the previously uncovered cell-type-specific co-expression patterns are no longer seen.

5 Theoretical Properties

In this section, we establish the non-asymptotic convergence rate of `CSNet`, the sparse cell-type-specific correlation estimates, and also establish variable selection consistency, ensuring that we correctly identify edges in the cell-type-specific co-expression networks with probability tending to 1. Our theoretical analysis is challenging with a few unique aspects. First,

we assume the expression profile from each cell type follows (marginally) a sub-exponential distribution to accommodate the commonly used negative binomial distributions in modeling read counts from RNA-seq data (Robinson et al., 2010). In this case, each element $z_{ij}z_{ij'}$ in the response vector in (8) is the product of two sub-exponential random variables, which requires a new concentration result (see Lemma A3.2 and its proof in Section A3.4). Second, our procedure considers correlation estimates, which are normalized with estimated variances. Thus, a more delicate analysis is needed to find the non-asymptotic convergence rate of $\hat{R}_{jj'}^{(k)}$ to $R_{jj'}^{(k)}$. Third, as $\boldsymbol{\mu}^{(k)}$'s are unknown, the response $\hat{\boldsymbol{z}}_j \circ \hat{\boldsymbol{z}}_{j'}$ in (8) inherits errors from estimating the mean parameter $\boldsymbol{\mu}^{(k)}$'s. Next, we state a few regularity conditions.

Assumption 1. *Let $x_{ij}^{(k)}$ follow a sub-exponential distribution, $i \in [n]$, $j \in [p]$, $k \in [K]$ and $\max_{j,k} \|x_{ij}^{(k)}\|_{\psi_1} \leq \gamma_0$ for some positive constant γ_0, where $\|X\|_{\psi_1} = \sup_{p\geq 1}\{p^{-1}\left(\mathbb{E}|X|^p\right)^{1/p}\}$.*

Assumption 1 imposes a marginal distribution assumption on $\boldsymbol{x}_i^{(k)}$, the expression profile from the kth cell type in sample i, in that each element $x_{ij}^{(k)}$ follows a sub-exponential distribution with a bounded sub-exponential norm. The sub-exponential assumption is more relaxed than the Gaussian or sub-Gaussian assumption commonly imposed (Bickel and Levina, 2008a,b; Rothman et al., 2008, 2009) and includes distributions such as negative binomial, which is often used to model read counts from RNA-seq data (Robinson et al., 2010).

Assumption 2. *Let $\boldsymbol{R}^{(k)} \in \mathcal{U}(q, s_p)$ for $0 \leq q < 1$, where $\mathcal{U}(q, s_p)$ is a class of approximately sparse matrices with sparsity parameter s_p (can be a function of p), defined as*

$$\mathcal{U}(q, s_p) = \left\{ \boldsymbol{R} : \boldsymbol{R} \succ 0,\ \max_{j'} \sum_{j=1}^{p} |R_{jj'}|^q \leq s_p \right\}. \tag{9}$$

This assumption stipulates that the cell-type-specific correlation matrices are approximately sparse. The definition of approximate sparsity in (9) is commonly employed in large covariance matrix estimation (e.g., Bickel and Levina, 2008a; Rothman et al., 2008). When $q = 0$,

Assumption 2 imposes that the number of non-zero entries in each column of $\boldsymbol{R}^{(k)}$ is less than s_p, $k \in [K]$. In Assumption 2, the sparse parameters q and s_p may vary across k, i.e., $\boldsymbol{R}^{(k)} \in \mathcal{U}(q^{(k)}, s_p^{(k)})$, though we dropped the superscript k to simplify notation, which can be viewed as taking $q = \max_k q^{(k)}$ and $s_p = \max_k s_p^{(k)}$.

Define two matrices $\boldsymbol{\Psi}_{K\times K}$ and $\boldsymbol{\Phi}_{K\times K}$, where $\boldsymbol{\Psi}_{kl} = \sum_{i=1}^n \pi_{ik}\pi_{il}/n$ and $\boldsymbol{\Phi}_{kl} = \sum_{i=1}^n \pi_{ik}^2\pi_{il}^2/n$, for $k, l \in [K]$. It then holds that $\boldsymbol{\Psi}_{K\times K} = \boldsymbol{D}^\top \boldsymbol{D}/n$ and $\boldsymbol{\Phi}_{K\times K} = \boldsymbol{H}^\top \boldsymbol{H}/n$, where $\boldsymbol{D}$ and $\boldsymbol{H}$ are as defined in (5) and (6), respectively.

Assumption 3. *There exist constants $\phi_1', \phi_1, \phi_2 > 0$ such that the eigenvalues of $\boldsymbol{\Psi}_{K\times K}$ are lower-bounded by ϕ_1', the eigenvalues of $\boldsymbol{\Phi}_{K\times K}$ are lower-bounded by ϕ_1 and $\min_k \boldsymbol{\Phi}_{kk} \geq \phi_2$.*

Assumption 3 places two regularity conditions on the cell type proportion matrices $\boldsymbol{D}, \boldsymbol{H} \in \mathbb{R}^{n\times K}$, which can be viewed as the fixed design matrices in (3) and (4). The bounded eigenvalue conditions on $\boldsymbol{\Psi}_{K\times K}$ and $\boldsymbol{\Phi}_{K\times K}$ are not restrictive as the number of cell types K is treated as fixed. The condition $\min_k \boldsymbol{\Phi}_{kk} \geq \phi_2$ stipulates that the average proportion for each cell type (over all bulk samples) should be bounded away from zero, which is a mild condition. In fact, if the cell type proportions are from an underlying multinomial or Dirichlet distribution with fixed parameters, all conditions in Assumption 3 are expected to hold; see proof in Section A3.6.

Theorem 5.1. *Let $\mathcal{T}_\lambda(\hat{\boldsymbol{R}}^{(k)})$ be the estimator as defined in (8). Under Assumptions 1-3, $\log p = o(n^{1/3})$ and $\lambda = M\sqrt{\log p/n}$ for a sufficiently large constant $M > 0$, we have, for $k \in [K]$,*

$$\left\| \mathcal{T}_\lambda\left(\hat{\boldsymbol{R}}^{(k)}\right) - \boldsymbol{R}^{(k)} \right\| \leq M s_p \left(\frac{\log p}{n}\right)^{(1-q)/2} \tag{10}$$

with probability at least $1 - C_1 \exp(-C_2 \log p)$, where C_1 and C_2 are some positive constants.

Theorem 5.1 implies that the number of genes p can far exceed the sample size n, as long as $s_p(\log p/n)^{(1-q)/2}$ tends to zero. The condition $\log p = o(n^{1/3})$ is needed as $x_{ij}^{(k)}$'s are assumed

to follow a sub-exponential distribution. When $q = 0$, the convergence rate in (10) further simplifies to $\|\mathcal{T}_\lambda(\hat{\boldsymbol{R}}^{(k)}) - \boldsymbol{R}^{(k)}\| = O_p(s_p\sqrt{\log p/n})$, which is comparable with the convergence rate derived in estimating sparse covariance matrices (e.g., Rothman et al., 2008). Based on (10) and using a similar argument as in Bickel and Levina (2008a), it can be shown that the convergence rate in Frobenius norm is $\frac{1}{p}\|\mathcal{T}_\lambda(\hat{\boldsymbol{R}}^{(k)}) - \boldsymbol{R}^{(k)}\|_F^2 = O_p(s_p(\log p/n)^{1-q/2})$.

Denote the support of a matrix $\boldsymbol{R}$ by $\Omega(\boldsymbol{R}) = \{(j, j') : R_{jj'} \neq 0\}$. We show that the support of $\boldsymbol{R}^{(k)}$ is recovered with high probability assuming a minimal signal condition.

Theorem 5.2. *Suppose Assumptions 1 and 3 hold. Let $\mathcal{T}_\lambda(\hat{\boldsymbol{R}}^{(k)})$ be the estimator as defined in (8), $\log p = o(n^{1/3})$, $\lambda = M\sqrt{\log p/n}$ for a sufficiently large constant $M > 0$. If $|R_{jj'}^{(k)}| > \rho_{\min}$ for some $\rho_{\min} > 0$ such that $\sqrt{n}(\rho_{\min} - \lambda) \to \infty$, then we have for $k \in [K]$*

$$\mathbb{P}\left\{\Omega\left[\mathcal{T}_\lambda\left(\hat{\boldsymbol{R}}^{(k)}\right)\right] = \Omega\left(\boldsymbol{R}^{(k)}\right)\right\} \to 1.$$

Theorem 5.2 holds without the approximately sparse condition in Assumption 3, though it requires a minimal signal condition on the nonzero elements in $\boldsymbol{R}^{(k)}$'s. Such a condition is commonly considered in establishing selection consistency (e.g., El Karoui, 2008) and our result allows the minimal signal to tend to zero as n increases (e.g., $\rho_{\min} = 2\lambda$). Theorem 5.2 also has an important implication in practice as it ensures that the selected set of edges after thresholding is the same as the true set of edges with probability tending to one.

6 Discussion

We conclude the article with a few remarks. First, the 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). In (1), $\boldsymbol{x}_i^{(1)}, \ldots, \boldsymbol{x}_i^{(K)}$ are assumed independent, as we focus on cell types with distinct functions and gene expression profiles across different cell types are likely not highly correlated (see Figure S7.) When this assumption does not hold, we can expand the covariance expression in (2) to include the cross product terms $\sum_{k \neq k'} \pi_{ik}\pi_{ik'}\boldsymbol{\Sigma}^{(k,k')}$, where $\boldsymbol{\Sigma}^{(k,k')} \in \mathbb{R}^{p \times p}$ is the cross-covariance $\boldsymbol{\Sigma}^{(k,k')} = \mathrm{Cov}(\boldsymbol{x}_i^{(k)}, \boldsymbol{x}_i^{(k')})$. In this case, the estimation procedure can be carried out similarly, though the total number of parameters increase from $O(Kp^2)$ to $O(K^2p^2)$.

Next, we have assumed that the cell-type proportions π_{ik}'s are given in our analysis. For example, they can be inferred using existing methods such as CIBERSORTx (Newman et al., 2019). Our empirical investigations showed that `CSNet` is not overly sensitive to errors in π_{ik}'s (see sensitive analysis in Section 3.2 and A2.1). 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. As the correlation matrices are estimated by element and then thresholded, the correlation matrix estimates are not guaranteed to be positive definite. However, as the true correlation matrices are positive definite, it follows from Theorems 5.1-5.2 that the estimated correlation matrices are asymptotically positive definite. 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., Higham, 2002).

Finally, we have focused on using bulk samples for co-expression analysis. With an increasing number of studies collecting single cell data, we may obtain more accurate co-expression estimates through integrated analysis of bulk and single cell data. For example,

Morabito et al. (2021) and bMIND 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.

Acknowledgements

We thank the ROSMAP project 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. 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).

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.

Adamczak, R. and Wolff, P. (2015), "Concentration inequalities for non-Lipschitz functions with bounded derivatives of higher order," *Probability Theory and Related Fields*, 162, 531–586.

Alzheimer's Association (2019), "2019 Alzheimer's disease facts and figures," *Alzheimer's & dementia*, 15, 321–387.

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.

Balasubramanian, K., Fan, J., and Yang, Z. (2018), "Tensor methods for additive index models under discordance and heterogeneity," *arXiv preprint arXiv:1807.06693*.

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.

Bickel, P. J. and Levina, E. (2008a), "Covariance regularization by thresholding," *The Annals of Statistics*, 36, 2577–2604.

— (2008b), "Regularized estimation of large covariance matrices," *The Annals of Statistics*, 36, 199–227.

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

Chen, Y., Lun, A. T., and Smyth, G. K. (2014), "Differential expression analysis of complex RNA-seq experiments using edgeR," *Statistical analysis of next generation sequencing data*, 51–74.

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.

Consortium, T. G. O. (2021), "The Gene Ontology resource: enriching a GOld mine," *Nucleic Acids Research*, 49, D325–D334.

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.

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.

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.

El Karoui, N. (2008), "Operator norm consistent estimation of large-dimensional sparse covariance matrices," *The Annals of Statistics*, 36, 2717–2756.

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.

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.

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.

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.

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.

Higham, N. J. (2002), "Computing the nearest correlation matrix—a problem from finance," *IMA journal of Numerical Analysis*, 22, 329–343.

Hwang, B., Lee, J. H., and Bang, D. (2018), "Single-cell RNA sequencing technologies and bioinformatics pipelines," *Experimental & molecular medicine*, 50, 1–14.

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.

Jiang, B. (2013), "Covariance selection by thresholding the sample correlation matrix," *Statistics & Probability Letters*, 83, 2492–2498.

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.

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*.

Langfelder, P. and Horvath, S. (2008), "WGCNA: an R package for weighted correlation network analysis," *BMC bioinformatics*, 9, 1–13.

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.

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.

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.

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.

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.

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.

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.

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.

Pimenova, A. A., Raj, T., and Goate, A. M. (2018), "Untangling genetic risk for Alzheimer's disease," *Biological psychiatry*, 83, 300–310.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010), "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, 26, 139–140.

Ronning, G. (1977), "A simple scheme for generating multivariate gamma distributions with non-negative covariance matrix," *Technometrics*, 19, 179–183.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008), "Sparse permutation invariant covariance estimation," *Electronic Journal of Statistics*, 2, 494–515.

Rothman, A. J., Levina, E., and Zhu, J. (2009), "Generalized thresholding of large covariance matrices," *Journal of the American Statistical Association*, 104, 177–186.

Sims, R., Hill, M., and Williams, J. (2020), "The multiplex model of the genetics of Alzheimer's disease," *Nature neuroscience*, 23, 311–322.

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.

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.

Stower, H. (2019), "Single-cell insights into neurology," *Nature medicine*, 25, 1799–1799.

Tang, D., Park, S., and Zhao, H. (2020), "NITUMID: nonnegative matrix factorization-based immune-TUmor MIcroenvironment Deconvolution," *Bioinformatics*, 36, 1344–1350.

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.

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.

Vershynin, R. (2018), *High-dimensional probability: An introduction with applications in data science*, vol. 47, Cambridge university press.

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.

Wang, J., Roeder, K., and Devlin, B. (2021a), "Bayesian estimation of cell type-specific gene expression with prior derived from single-cell data," *Genome Research*, gr–268722.

Wang, M., Li, A., Sekiya, M., Beckmann, N. D., Quan, X., Schrode, N., Fernando, M. B., Yu, A., Zhu, L., Cao, J., et al. (2021b), "Transformative network modeling of multi-omics

data reveals detailed circuits, key regulators, and potential therapeutics for Alzheimer's disease," *Neuron*, 109, 257–272.

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.

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.

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.

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.

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.

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.

Zhang, B. and Horvath, S. (2005), "A General Framework for Weighted Gene Co-Expression Network Analysis," *Statistical Applications in Genetics and Molecular Biology*, 4, 1–45.

Zhong, Y. and Liu, Z. (2012), "Gene expression deconvolution in linear space," *Nature methods*, 9, 8–9.

Supplementary Materials

Estimating cell-type-specific gene co-expression networks from bulk gene expression data with an application to Alzheimer's disease

A1 Simulating from a multivariate negative binomial distribution with a large and structured covariance

Consider a negative binomial (NB) random variable distributed as

$$X \sim \mathrm{NB}(T, \frac{1}{1+S\phi}), \tag{S1}$$

such that

$$\mathbb{E}(X) = TS\phi, \quad \mathrm{Var}(X) = TS^2\phi^2(1+\frac{1}{S\phi}). \tag{S2}$$

Our procedure simulates a Poisson-Gamma mixture that follows (S1) by simulating

- **Gamma**: $\theta \sim \mathrm{Gamma}(T, \phi)$, where $\mathbb{E}[\theta] = T\phi$ and $\mathrm{Var}[\theta] = T\phi^2$;
- **Poisson**: $X|\theta, S \sim \mathrm{Poisson}(S\theta)$.

This is a model commonly employed in differential expression analysis methods for bulk RNA-seq data (Robinson et al., 2010; Love et al., 2014), where θ models gene abundance, S denotes the total number of mRNA transcripts (i.e., sequencing depth) and $X|\theta$ models the procedure of sampling a transcript from the pool. The interpretation of parameter T will be explained in the following paragraph. In addition, we note that this model (S1) sets the variance to be a quadratic function of the mean, which is supported by observations on real bulk RNA-seq data (Chen et al., 2014).

Before we proceed, we first discuss a useful property of Gamma random variables which entails that if g_1 and g_2 are two independent samples from $\mathrm{Gamma}(T, \phi)$, then $g_1 + g_2 \sim$

Algorithm 1 Algorithm for simulating multivariate negative binomial random variables

Input: Sequencing depth S, number of genes p, variance σ^2 and correlation matrix $\boldsymbol{R}$.

Step 1: *design the sharing between Gamma random variables*

for $j < j' \in [p]$, set $T_{jj'} = \lfloor T\rho_{jj'} \rfloor$ with T set to a large positive integer, e.g. 200, draw $T_{jj'}$ samples $A_{jj'} = \{\eta_h\}_{h\in[T_{jj'}]}$ i.i.d. from Exp(1) and set $A_{j'j} = A_{jj'}$.

Step 2: *simulate Gamma random variables with correlation* $\boldsymbol{R}$

Step 2.1: set $\phi = \frac{1}{S}\sqrt{\frac{\sigma^2}{T} + \frac{1}{4}} - \frac{1}{2S}$;

Step 2.2: for $j \in [p]$, draw $T - \sum_{j\neq j'} T_{jj'}$ i.i.d. samples $A_j = \{\eta_h\}_{h\in[T-\sum_{j\neq j'} T_{jj'}]}$ from Exp(1), and get $\theta_j = \phi\left(\sum_{h\in\cup_{j\neq j'} A_{jj'}} \eta_h + \sum_{h\in A_j} \eta_h\right)$.

Step 3: *sample from Poisson-Gamma mixture to get negative binomial samples*

for $j \in [p]$, draw X_j from Poisson$(S\theta_j)$.

Output: $(X_1, \ldots, X_p)$

Gamma$(2T, \phi)$. This property facilitates an effective procedure for generating correlated Gamma random variables. For example, if g_1, g_2, and g_3 are independent Gamma(T, ϕ) random variables, then Cor$(g_1 + g_2,\ g_2 + g_3)$=0.5, with g_2 being the component that is being shared by two Gamma random variables g_1+g_2 and g_2+g_3. This idea of simulating correlated Gamma random variables has previously been explored by Ronning (1977).

Our proposed procedure that simulates correlated negative binomial random variables, detailed in Algorithm 1, is divided into the following three steps. Step 1 designs the sharing between Gamma random variables, Step 2 simulates correlated Gamma random variables and Step 3 samples from the Poisson-Gamma mixture to get desired negative binomial samples. Specifically, in Algorithm 1, Step 1 sets T to be a large constant, whose subsets will be shared across samples as specified in Step 2.2, Step 2 draws 'building blocks' of Exp(1) random variables (i.e., Gamma(1,1)) to be shared and then generates correlated Gamma random variables and Step 3 employs Poisson sampling to obtain negative binomial random variables. We note that the assumption of homogeneous variances across all genes can be relaxed by allowing ϕ to vary across samples.

The output of Algorithm 1 is a p-dimensional random vector $(X_1, \ldots, X_p)$ with Var$[X_j] =$

σ^2 for $j \in [p]$ and correlation matrix $\boldsymbol{R} \times [1/(1+1/S\phi)]$. It is seen that the correlation is biased from the specified correlation matrix $\boldsymbol{R}$ by a multiplicative factor $b = 1/(1+1/S\phi)$. This factor is close to 1 when $S\phi = \sqrt{\frac{\sigma^2}{T} + \frac{1}{4}} - \frac{1}{2}$ is large, as implied by (S2). Therefore, this bias is negligible for genes with large variances. Our own analysis of real bulk RNA-seq data found $S = 6 \times 10^7$ to be a common sequencing depth, and $\sigma^2 = \exp(8)$ to be representative of gene variances for highly-expressed protein coding genes. Under this setting, the multiplicative bias $b = 0.77$. Correspondingly, in Section 3 we simulate MA(1) structure with $\rho = 0.5 \times b \approx 0.39$ and an AR(1)-type structure with $\rho_{jj'} = 0.70 \times 0.9^{|j-j'-1|}$ for $j \neq j'$, where $0.70 \approx 0.9 \times b$. Finally, as an example, Figure S9 demonstrates that Algorithm 1 simulates samples that are faithful to the prescribed AR(1)-type correlation structures in Figure 1.

A2 More Results from Data Analysis in Section 4

A2.1 Sensitivity analysis

For the four gene sets in Section 4, we conduct a sensitivity analysis to evaluate the robustness of the `CSNet` estimates against perturbations in the cell type proportions inferred using CIBERSORTx. For each sample i and given $\boldsymbol{\pi}_i$, the inferred cell type proportions using CIBERSORTx, the perturbed cell type proportions $\tilde{\boldsymbol{\pi}}_i$ are independently simulated from Dirichlet$(\boldsymbol{\pi}_i \times \kappa)$, where κ is set to $\{10, 100, 1000\}$. A smaller κ corresponds to a larger noise level, as illustrated in Figure S1. Furthermore, we evaluate the difference between the $\hat{\boldsymbol{R}}^{(k)}$, `CSNet` estimated with $\boldsymbol{\pi}_i$, and $\tilde{\boldsymbol{R}}^{(k)}$, `CSNet` estimated with $\tilde{\boldsymbol{\pi}}_i$, by Frobenius norm measured as $\|\hat{\boldsymbol{R}}^{(k)} - \tilde{\boldsymbol{R}}^{(k)}\|_F$, TPR measured as $\sum_{j,j' \in [p]} \mathbf{1}\{\hat{\boldsymbol{R}}^{(k)}_{jj'} > 0, \tilde{\boldsymbol{R}}^{(k)}_{jj'} > 0\} / \sum_{j,j' \in [p]} \mathbf{1}\{\hat{\boldsymbol{R}}^{(k)}_{jj'} > 0\}$ and FPR measured as $\sum_{j,j' \in [p]} \mathbf{1}\{\hat{\boldsymbol{R}}^{(k)}_{jj'} = 0, \tilde{\boldsymbol{R}}^{(k)}_{jj'} > 0\} / \sum_{j,j' \in [p]} \mathbf{1}\{\hat{\boldsymbol{R}}^{(k)}_{jj'} = 0\}$. Figure S2 shows that `CSNet` estimates are robust under a reasonable amount of noise.

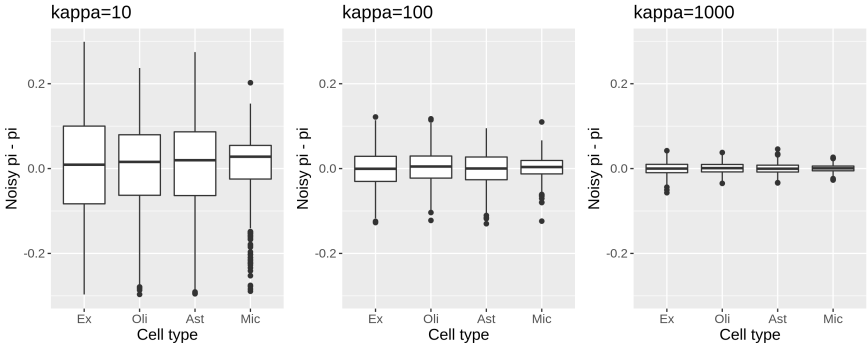


Figure S1: Boxplots of $\pi_{ik} - \tilde{\pi}_{ik}$ for four most abundant cell types: excitatory neuron (Ex), oligodendrocyte (Oli), astrocyte (Ast), microglia (Mic).

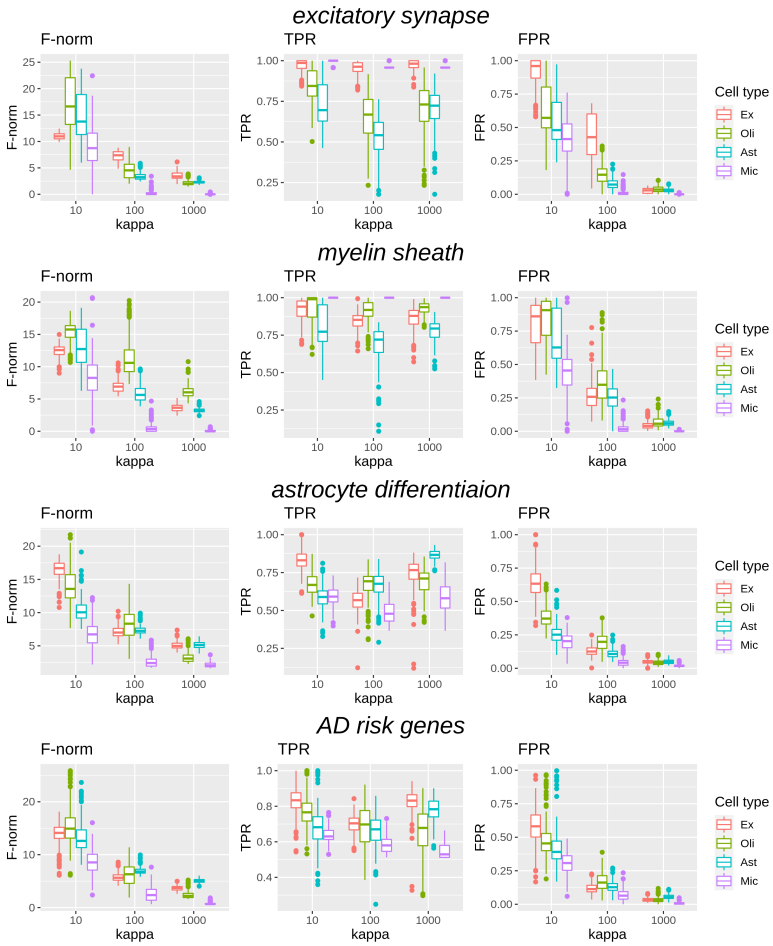


Figure S2: Evaluation criteria with varying noise level κ for four gene sets. From top to bottom: *excitatory synapse*, *myelin sheath*, *astrocyte differentiation* and Alzheimer's disease (AD) risk genes. Results are shown for four most abundant cell types: excitatory neuron (Ex), oligodendrocyte (Oli), astrocyte (Ast), microglia (Mic).

A2.2 Additional figures

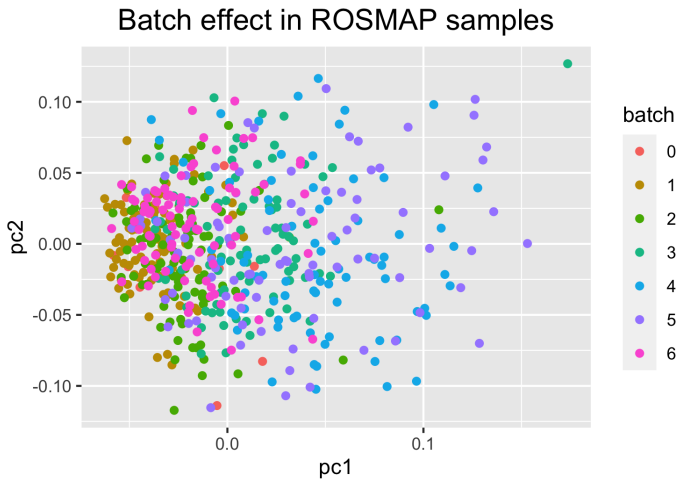


Figure S3: Top two PCs from the principal component analysis of the 541 ROSMAP samples using protein coding genes with average FPKM> 0.25, with points colored by processing batches.

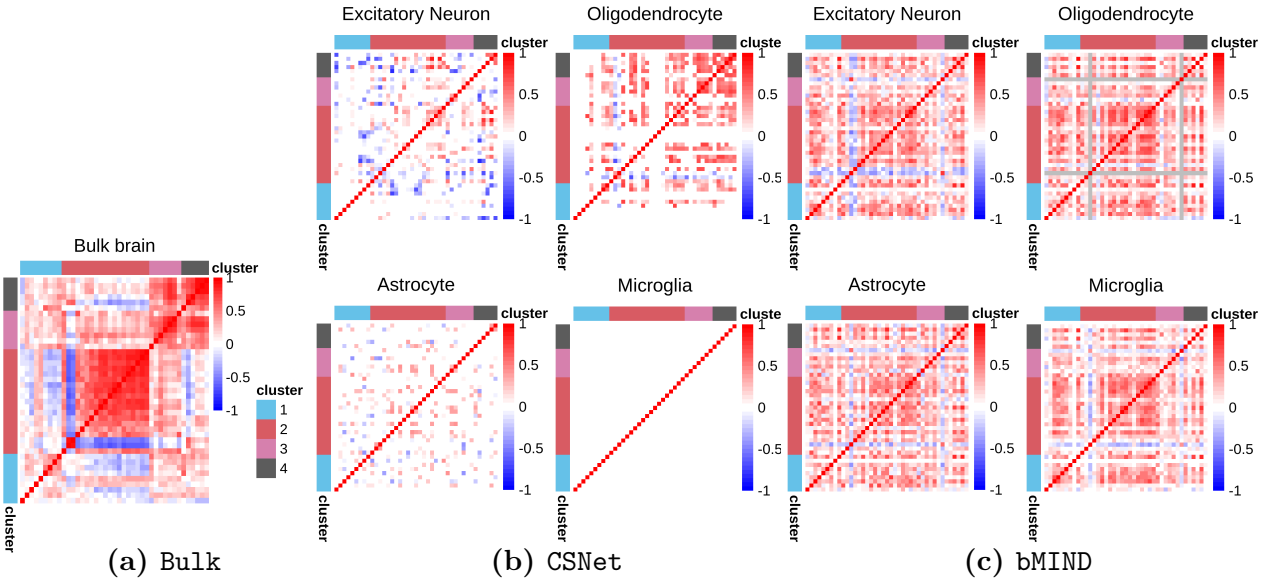
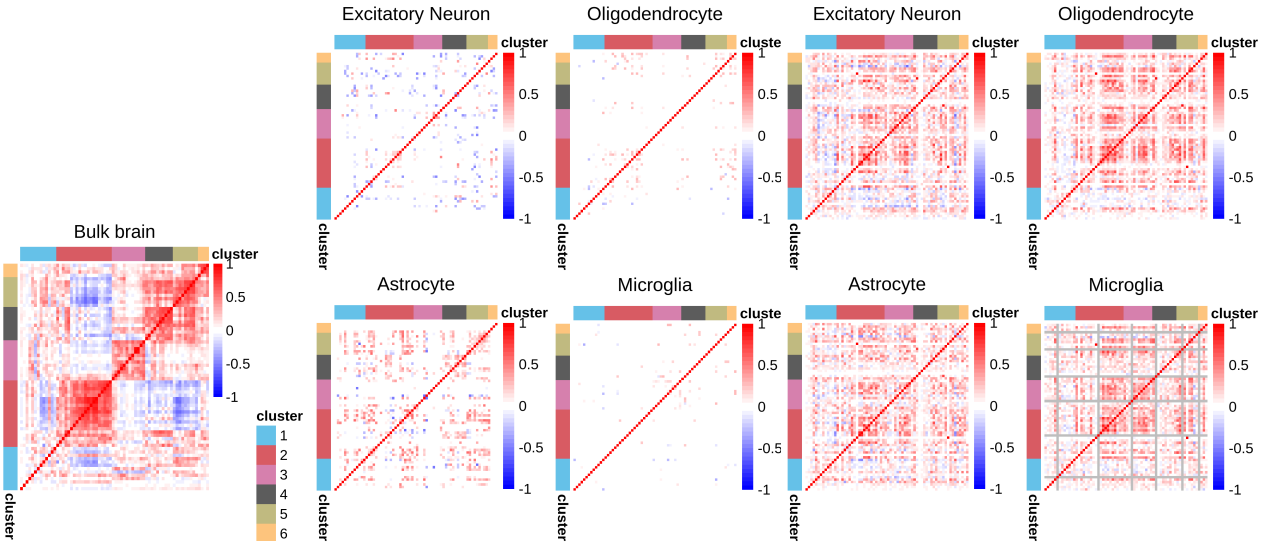
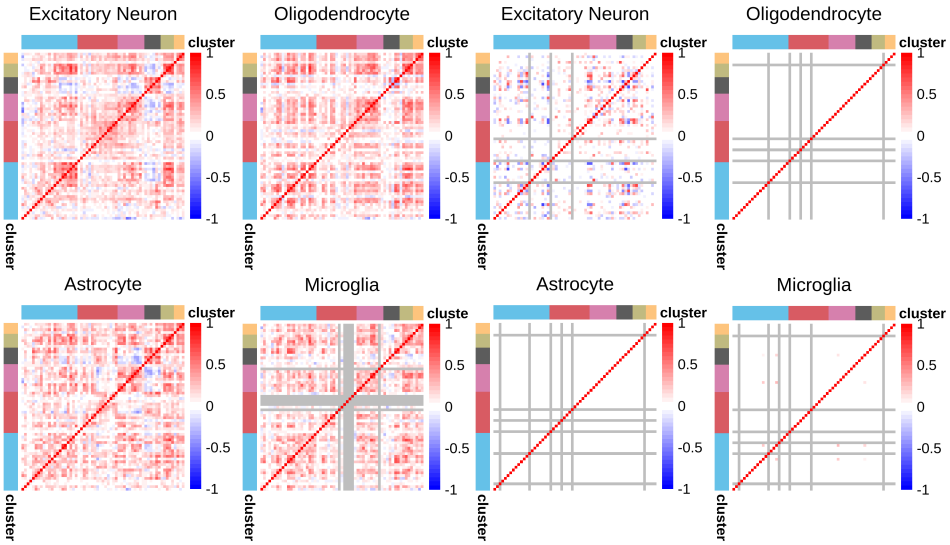


Figure S4: Co-expressions of *myelin sheath* genes in different cell types. From left to right: (a) sample correlation matrix estimated from bulk RNA-seq data, (b) `CSNet` estimates and (c) `bMIND` estimates. For `bMIND`, genes with constant expression estimates across samples are marked in gray.



(a) `Bulk` **(b)** `CSNet` **(c)** `bMIND`

Figure S5: Co-expressions of *astrocyte differentiation* genes in different cell types. From left to right: (a) sample correlation matrix estimated from bulk RNA-seq data, (b) `CSNet` estimates and (c) `bMIND` estimates. For `bMIND`, genes with constant expression estimates across samples are marked in gray.



(a) `bMIND` **(b)** `Single cell`

Figure S6: Co-expressions networks of Alzheimer's disease risk genes estimated with (a) `bMIND`, (b) single cell data from Mathys et al. (2019). Genes with constant expression estimates across samples are marked in gray.

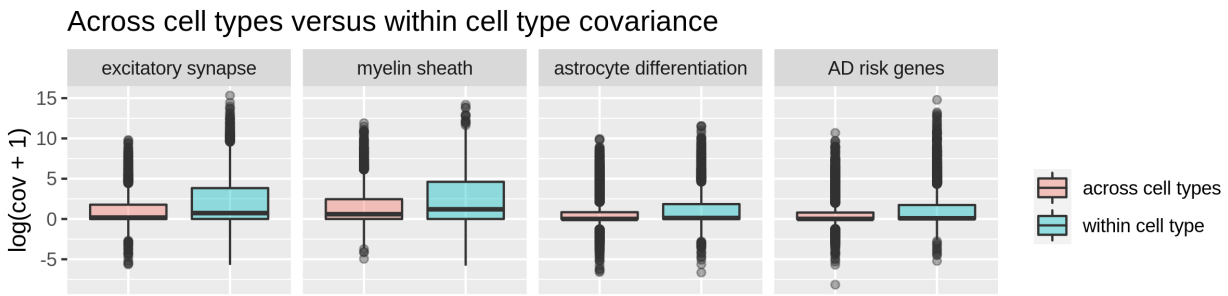
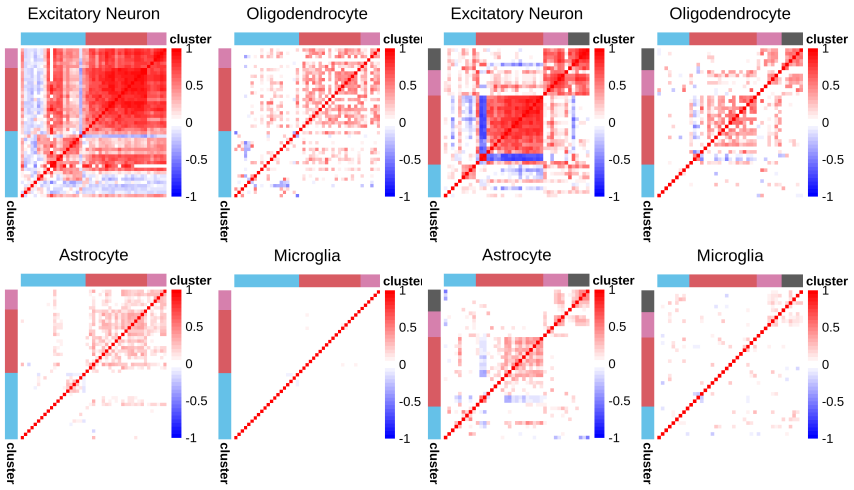
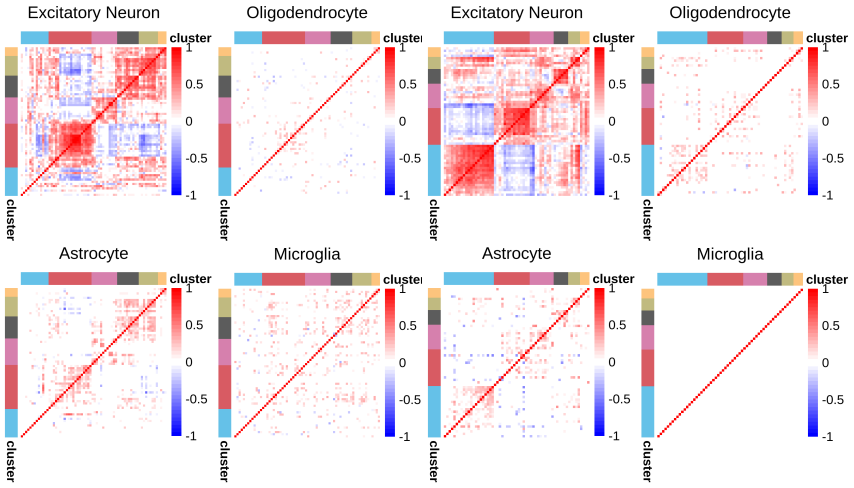


Figure S7: Cross-cell-type covariances $\mathrm{Cov}(x_{ij}^{(k)}, x_{ij'}^{(k')})$, $j, j' \in [p], k, k' \in [K], k \neq k'$ and within-cell-type covariances $\mathrm{Cov}(x_{ij}^{(k)}, x_{ij'}^{(k)})$, $j, j' \in [p], k \in [K]$ for four gene sets, evaluated with single cell data from Mathys et al. (2019) following a similar procedure as described in Section 4.



(a) *excitatory synapse* (b) *myelin sheath*



(c) *astrocyte differentiation* (d) *AD risk genes*

Figure S8: `CSNet` estimates of co-expressions in the *excitatory synapse*, *myelin sheath*, *astrocyte differentiation* and Alzheimer's disease (AD) risk genes gene sets, with proportions randomly permuted across samples.

A3 Proofs of Main Results

A3.1 Technical Lemmas

We first present a set of technical lemmas. The proofs for Lemmas A3.2 and A3.5 are presented in Sections A3.4 and A3.5, respectively.

Lemma A3.1. *[Theorem 2.8.11 in Vershynin (2018)] Let $X_1, \ldots, X_n$ be independent mean zero sub-exponential random variables with $\max_i \|X_i\|_{\psi_1} \leq \gamma_1$ for some constant $\gamma_1 > 0$. Given $\boldsymbol{a} = (a_1, \cdots, a_n) \in \mathbb{R}^n$, for every $t > 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i X_i\right| \geq t\right) \leq 2\exp\left\{-c_1 \min\left[\left(\frac{t}{\gamma_1\|\boldsymbol{a}\|_2}\right)^2, \frac{t}{\gamma_1\|\boldsymbol{a}\|_\infty}\right]\right\}, \tag{S3}$$

for some positive constant c_1.

Lemma A3.2. *Let $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_n\}$ be two sets of independent sub-exponential random variables with $\max_i \|X_i\|_{\psi_1} \leq \gamma_X$ and $\max_i \|Y_i\|_{\psi_1} \leq \gamma_Y$, respectively. Let $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$ and $\gamma_2 = \gamma_X\gamma_Y$. Then there exist some constants $d_1, d_2, c_2 > 0$ such that for $0 < t < d_1 \times \gamma_2 \times \|\boldsymbol{a}\|_\infty$,*

$$\mathbb{P}\left\{\left|\sum_{i=1}^n [a_i X_i Y_i - a_i \mathbb{E}(X_i Y_i)]\right| > t\right\} \leq 4\exp\left\{-c_2 \min\left[\left(\frac{t}{\gamma_2\|\boldsymbol{a}\|_2}\right)^2, \left(\frac{t}{\gamma_2\|\boldsymbol{a}\|_\infty}\right)\right]\right\}, \tag{S4}$$

and for any $t \geq d_2 \times \gamma_2 \times \|\boldsymbol{a}\|_\infty$,

$$\mathbb{P}\left\{\left|\sum_{i=1}^n [a_i X_i Y_i - a_i \mathbb{E}(X_i Y_i)]\right| > t\right\} \leq 4\exp\left\{-c_2 \min\left[\left(\frac{t}{\gamma_2\|\boldsymbol{a}\|_2}\right)^2, \left(\frac{t}{\gamma_2\|\boldsymbol{a}\|_\infty}\right)^{1/2}\right]\right\}. \tag{S5}$$

Lemma A3.3. *[Theorem 1.4 Adamczak and Wolff (2015)] Let $X = (X_1, \ldots, X_n)$ be a random vector with independent components, such that for all $i \in [n]$, $\|X_i\|_{\psi_2} \leq \gamma_3$, where $\|X\|_{\psi_2} = \sup_{p\geq 1}\{p^{-1/2}(\mathbb{E}|X|^p)^{1/p}\}$. Let $\mathcal{P}_d$ denote the set of partitions of $\{1, \ldots, d\}$ into*

nonempty, pairwise disjoint sets and $\|\cdot\|_{\mathcal{J}}$ *denote a tensor norm whose definition is relegated to Section A3.4. Then for every polynomial* $f\colon \mathbb{R}^n \to \mathbb{R}$ *of degree* D *and any* $t > 0$,

$$\mathbb{P}\left(|f(X) - \mathbb{E}f(X)| \geq t\right) \leq 2\exp\left(-\frac{1}{C_D}\eta_f(t)\right),$$

where $\eta_f(t) = \min_{1\leq d\leq D}\min_{\mathcal{J}\in\mathcal{P}_d}\left(\frac{t}{\gamma_3^d\|\mathbb{E}\mathcal{D}^d f(X)\|_{\mathcal{J}}}\right)^{2/|\mathcal{J}|}$, $\mathcal{D}^d f$ *denotes the* d*-th derivative of* f *and* C_D *is some positive constant depending on* D.

Lemma A3.4. *[Lemma 1 and Theorem 1 Rothman et al. (2009)] Consider the class of sparse covariance matrices* $\mathcal{U}(q, s_p)$ *as defined in Assumption 2. If* $\boldsymbol{R} \in \mathcal{U}(q, s_p)$, *then the thresholding operator* $\mathcal{T}$ *in (8) with threshold* λ *satisfies*

$$\sum_{j=1}^{p}|\mathcal{T}_\lambda(R_{jj'}) - R_{jj'}| \leq s_p\lambda^{1-q}. \tag{S6}$$

In addition, suppose we have an estimator $\hat{\boldsymbol{R}}$ *such that*

$$\mathbb{P}\left(\max_{jj'}\left|R_{jj'} - \hat{R}_{jj'}\right| \leq \lambda\right) \geq 1 - c_3\exp(-c_4\log p) \tag{S7}$$

for some $\lambda > 0$ *and positive constants* c_3 *and* c_4. *Then,* $\mathcal{T}_\lambda(\hat{\boldsymbol{R}})$ *satisfies*

$$\mathbb{P}\left(\max_{j}\sum_{j'=1}^{p}\left|\mathcal{T}_\lambda(\hat{R}_{jj'}) - R_{jj'}\right| \leq s_p\lambda^{1-q}\right) \geq 1 - c_5\exp(-c_6\log p) \tag{S8}$$

for some positive constants c_5 *and* c_6.

Lemma A3.5. *Let* $X = (X_1, \ldots, X_n)$ *be a vector of sub-exponential random variables with* $\max_i \|X_i\|_{\psi_1}^2 \leq \gamma_4$. *Let* $\mu = \sqrt{[\mathbb{E}\|X\|_2^2]/n}$. *Then, for* $t \leq \mu$,

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{n}}\|X\|_2 - \mu\right| \geq t\right) \leq 4\exp\left[-c_7\min\left(\frac{\mu^2}{\gamma_4^2}nt^2, \sqrt{\frac{\mu}{\gamma_4}nt}\right)\right]. \tag{S9}$$

A3.2 Proof of Theorem 5.1

The proof is divided into three steps. In step 1, we establish an element-wise concentration inequality for the covariance estimates $\hat{\sigma}_{jj'}^{(k)}$. In step 2, we get an element-wise concentration inequality for the correlation estimates $\hat{R}_{jj'}^{(k)}$. Step 3 puts together the previous steps and finds the appropriate thresholds that lead to the desired result.

Step 1. We first show that the element-wise covariance estimate $\hat{\sigma}_{jj'}^{(k)}$ satisfies

$$\mathbb{P}\left(\left|\hat{\sigma}_{jj'}^{(k)} - \sigma_{jj'}^{(k)}\right| > t\right) \leq c_1' \exp\left[-c_2' \min\left(nt^2, \sqrt{nt}, nt, n\sqrt{t}, \sqrt{n}t^{1/4}\right)\right], \tag{S10}$$

for $k \in [K], j, j' \in [p]$ and some positive constants c_1', c_2'.

Denote $\boldsymbol{\beta}_{jj'} = (\sigma_{jj'}^{(1)}, \ldots, \sigma_{jj'}^{(K)})$. Based on (4) and (5), we have $\hat{\boldsymbol{z}}_j \circ \hat{\boldsymbol{z}}_{j'} = \boldsymbol{H}\boldsymbol{\beta}_{jj'} + \hat{\boldsymbol{\epsilon}}_{jj'}$, where $\hat{\boldsymbol{\epsilon}}_{jj'} = \boldsymbol{\epsilon}_{jj'} + \hat{\boldsymbol{z}}_j \circ \hat{\boldsymbol{z}}_{j'} - \boldsymbol{z}_j \circ \boldsymbol{z}_{j'}$, and $\boldsymbol{z}_j$'s, $\hat{\boldsymbol{z}}_j$'s are as defined in Section 2. We consider the least squares estimates of covariances in (6), i.e. $\hat{\boldsymbol{\beta}}_{jj'} = \left(\boldsymbol{H}^\top \boldsymbol{H}\right)^{-1} \boldsymbol{H}^\top (\hat{\boldsymbol{z}}_j \circ \hat{\boldsymbol{z}}_{j'})$, which satisfies

$$\|(\hat{\boldsymbol{z}}_j \circ \hat{\boldsymbol{z}}_{j'}) - \boldsymbol{H}\hat{\boldsymbol{\beta}}_{jj'}\|_2^2 \leq \|(\hat{\boldsymbol{z}}_j \circ \hat{\boldsymbol{z}}_{j'}) - \boldsymbol{H}\boldsymbol{\beta}_{jj'}\|_2^2. \tag{S11}$$

Recall that $\boldsymbol{\epsilon}_{jj'} = \boldsymbol{z}_j \circ \boldsymbol{z}_{j'} - \boldsymbol{H}\boldsymbol{\beta}_{jj'}$. Letting $\boldsymbol{d}_{jj'} = \hat{\boldsymbol{z}}_j \circ \hat{\boldsymbol{z}}_{j'} - \boldsymbol{z}_j \circ \boldsymbol{z}_{j'}$ and $\boldsymbol{\Delta}_{jj'} = \hat{\boldsymbol{\beta}}_{jj'} - \boldsymbol{\beta}_{jj'}$, we have $(\hat{\boldsymbol{z}}_j \circ \hat{\boldsymbol{z}}_{j'}) - \boldsymbol{H}\hat{\boldsymbol{\beta}}_{jj'} = \boldsymbol{\epsilon}_{jj'} + \boldsymbol{d}_{jj'} - \boldsymbol{H}\boldsymbol{\Delta}_{jj'}$ and $(\hat{\boldsymbol{z}}_j \circ \hat{\boldsymbol{z}}_{j'}) - \boldsymbol{H}\boldsymbol{\beta}_{jj'} = \boldsymbol{\epsilon}_{jj'} + \boldsymbol{d}_{jj'}$. Then, (S11) gives that

$$\|\boldsymbol{H}\boldsymbol{\Delta}_{jj'}\|_2^2 \leq 2\langle \boldsymbol{H}\boldsymbol{\Delta}_{jj'}, \boldsymbol{\epsilon}_{jj'} + \boldsymbol{d}_{jj'}\rangle. \tag{S12}$$

By Hölder's inequality, we have

$$|\langle \boldsymbol{H}\boldsymbol{\Delta}_{jj'}, \boldsymbol{\epsilon}_{jj'} + \boldsymbol{d}_{jj'}\rangle| \leq \|\boldsymbol{H}^\top (\boldsymbol{\epsilon}_{jj'} + \boldsymbol{d}_{jj'})\|_\infty \|\boldsymbol{\Delta}_{jj'}\|_1 \leq \sqrt{K}\|\boldsymbol{H}^\top (\boldsymbol{\epsilon}_{jj'} + \boldsymbol{d}_{jj'})\|_\infty \|\boldsymbol{\Delta}_{jj'}\|_2. \tag{S13}$$

In addition, under Assumption 3, it holds that

$$\phi_1 \|\boldsymbol{\Delta}_{jj'}\|_2^2 \leq \frac{1}{n}\|\boldsymbol{H}\boldsymbol{\Delta}_{jj'}\|_2^2. \tag{S14}$$

Putting (S12)-(S14) together, we arrive at that

$$\begin{aligned}\|\boldsymbol{\Delta}_{jj'}\|_2 &\leq \frac{2\sqrt{K}}{\phi_1}\frac{1}{n}\|\boldsymbol{H}^\top(\boldsymbol{\epsilon}_{jj'}+\boldsymbol{d}_{jj'})\|_\infty \\ &\leq \frac{2\sqrt{K}}{\phi_1}\frac{1}{n}\Big(\underbrace{\|\boldsymbol{H}^\top\boldsymbol{\epsilon}_{jj'}\|_\infty}_{\text{(I)}}+\underbrace{\|\boldsymbol{H}^\top\boldsymbol{d}_{jj'}\|_\infty}_{\text{(II)}}\Big).\end{aligned} \tag{S15}$$

Next, we bound terms (I) and (II) separately. First, we focus on term (I) and bound it with Lemma A3.2. By the union sum inequality, we have

$$\mathbb{P}\left(\frac{2\sqrt{K}}{\phi_1}\frac{1}{n}\times(\text{I})\geq t\right)\leq K\mathbb{P}\left(\frac{1}{c_3'n}\max_k\left|\langle\boldsymbol{\pi}_k^2,\boldsymbol{\epsilon}_{jj'}\rangle\right|\geq t\right) \tag{S16}$$

where $c_3' = \dfrac{\phi_1}{2\sqrt{K}}$.

We note that $\epsilon_{ijj'} = z_{ij}z_{ij'} - \mathbb{E}[z_{ij}z_{ij'}]$. By Assumption 1, we have $\left\|x_{ij}^{(k)}\right\|_{\psi_1} \leq \gamma_0$ for some constant $\gamma_0 > 0$, $j \in [p]$, $k \in [K]$. It follows that $\left\|x_{ij}^{(k)} - \mathbb{E}x_{ij}^{(k)}\right\|_{\psi_1} \leq 2\gamma_0$. Therefore, there exists a constant $K_0 > 0$ such that $\|z_{ij}\|_{\psi_1} \leq 2K_0\gamma_0$ for $j \in [p]$. Then, let $X_i = z_{ij}, Y_i = z_{ij'}, a_i = \dfrac{1}{c_3'n}\pi_{ik}^2$ for $i \in [n]$ and $\gamma_X = \gamma_Y = 2K_0\gamma_0$ in Lemma A3.2, we have

$$\begin{aligned}\mathbb{P}\left(\frac{1}{c_3'n}\left|\langle\boldsymbol{\pi}_k^2,\boldsymbol{\epsilon}_{jj'}\rangle\right|\geq t\right) &\leq 4\exp\left\{-c_4'\min\left[\left(\frac{c_3'nt}{\gamma\|\boldsymbol{\pi}_k^2\|_2}\right)^2,\sqrt{\frac{c_3'nt}{\gamma\|\boldsymbol{\pi}_k^2\|_\infty}}\right]\right\} \\ &\leq 4\exp\left[-c_5'\min\left(\frac{nt^2}{\gamma^2},\sqrt{\frac{nt}{\gamma}}\right)\right]\end{aligned} \tag{S17}$$

for $k \in [K]$ and some positive constants c_4', c_5', where the last inequality is implied by $\frac{1}{n}\|\boldsymbol{\pi}_k^2\|_2^2 = O(1)$ (Assumption 3) and $\gamma = (2K_0\gamma_0)^2$. Note both cases in Lemma A3.2 are considered to give the result in (S17).

Next, we establish a bound for the second term (II) in (S15). Let $\boldsymbol{\eta}_j = (\hat{z}_{1j} - z_{1j}, \ldots, \hat{z}_{nj} -$

$z_{nj}) \in \mathbb{R}^n$ for $j \in [p]$. Then, we have $\boldsymbol{d}_{jj'} = \hat{\boldsymbol{z}}_j \hat{\boldsymbol{z}}_{j'} - \boldsymbol{z}_j \boldsymbol{z}_{j'} = \boldsymbol{\eta}_j \boldsymbol{\eta}_{j'} + \boldsymbol{z}_j \boldsymbol{\eta}_{j'} + \boldsymbol{z}_{j'} \boldsymbol{\eta}_j$, and

$$\begin{aligned}\|\boldsymbol{H}^\top \boldsymbol{d}_{jj'}\|_\infty &= \max_k |\langle \boldsymbol{\pi}_k^2, \boldsymbol{d}_{jj'} \rangle| \\ &\leq \max_k |\langle \boldsymbol{\pi}_k^2, \boldsymbol{\eta}_j \boldsymbol{\eta}_{j'} \rangle| + \max_k |\langle \boldsymbol{\pi}_k^2, \boldsymbol{z}_j \boldsymbol{\eta}_{j'} \rangle| + \max_k |\langle \boldsymbol{\pi}_k^2, \boldsymbol{\eta}_j \boldsymbol{z}_{j'} \rangle|,\end{aligned}$$

which can be bounded with the following two inequalities:

$$|\langle \boldsymbol{\pi}_k^2, \boldsymbol{\eta}_j \boldsymbol{\eta}_{j'} \rangle| \leq \frac{1}{2} \|\boldsymbol{\pi}_k^2 \circ \boldsymbol{\eta}_j\|_2^2 + \frac{1}{2} \|\boldsymbol{\eta}_{j'}\|_2^2 \leq \frac{1}{2} \|\boldsymbol{\eta}_j\|_2^2 + \frac{1}{2} \|\boldsymbol{\eta}_{j'}\|_2^2 \tag{S18}$$

$$|\langle \boldsymbol{\pi}_k^2, \boldsymbol{z}_j \boldsymbol{\eta}_{j'} \rangle| \leq \|\boldsymbol{\pi}_k^2 \circ \boldsymbol{z}_j\|_2 \|\boldsymbol{\eta}_{j'}\|_2 \leq \|\boldsymbol{z}_j\|_2 \|\boldsymbol{\eta}_{j'}\|_2. \tag{S19}$$

Next, we bound the two terms $\|\boldsymbol{\eta}_j\|_2$ and $\|\boldsymbol{z}_j\|_2$, respectively. For $\|\boldsymbol{\eta}_j\|_2$, we have $\boldsymbol{\eta}_j = \boldsymbol{D}(\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j)$, where $\hat{\boldsymbol{\mu}}_j$ is as defined in (5). By Assumption 3 and using similar arguments as in (S12)-(S14), we have

$$\|\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j\|_2 \leq \frac{2\sqrt{K}}{\phi_1'} \frac{1}{n} \left\|\boldsymbol{D}^\top \boldsymbol{z}_j\right\|_\infty, \quad \|\boldsymbol{\eta}_j\|_2^2 \leq 2\sqrt{K} \|\boldsymbol{D}^\top \boldsymbol{z}_j\|_\infty \|\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j\|_2. \tag{S20}$$

By Hölder's inequality, we can further combine the equations in (S20) and get

$$\|\boldsymbol{\eta}_j\|_2^2 \leq \frac{4K}{\phi_1'} \frac{1}{n} \left\|\boldsymbol{D}^\top \boldsymbol{z}_j\right\|_\infty^2. \tag{S21}$$

Next, we rewrite term (II) in (S15) as

$$\begin{aligned}\frac{2\sqrt{K}}{\phi_1} \frac{1}{n} \times \text{(II)} &= \frac{2\sqrt{K}}{\phi_1} \frac{1}{n} \|\boldsymbol{H}^\top \boldsymbol{d}_{jj'}\|_\infty \\ &\leq c_6' \frac{1}{n^2} \left(\left\|\boldsymbol{D}^\top \boldsymbol{z}_j\right\|_\infty^2 + \left\|\boldsymbol{D}^\top \boldsymbol{z}_{j'}\right\|_\infty^2 \right) + c_7' \frac{1}{n^{3/2}} \left(\left\|\boldsymbol{D}^\top \boldsymbol{z}_j\right\|_\infty \|\boldsymbol{z}_{j'}\|_2 + \left\|\boldsymbol{D}^\top \boldsymbol{z}_{j'}\right\|_\infty \|\boldsymbol{z}_j\|_2 \right),\end{aligned} \tag{S22}$$

where c_6' and c_7' are some positive constants.

To bound $\frac{1}{n} \left\|\boldsymbol{D}^\top \boldsymbol{z}_j\right\|_\infty = \frac{1}{n} \max |\sum_{i=1}^n \pi_{ik} z_{ij}|$, we note that z_{ij}'s are sub-exponential with $\|z_{ij}\| \leq 2K_0\gamma_0$. Therefore, let $a_i = \pi_{ik}, X_i = z_{ij}$ for $i \in [n]$ and $\gamma = (2K_0\gamma_0)^2$, Lemma A3.1

implies

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n}\pi_{ik}z_{ij}\right|\geq t\right)\leq 2\exp\left\{-c_8'\min\left[\left(\frac{nt}{\gamma^{1/2}\|\boldsymbol{\pi}_k\|_2}\right)^2,\frac{nt}{\gamma^{1/2}\|\boldsymbol{\pi}_k\|_\infty}\right]\right\}$$
$$\leq 2\exp\left[-c_9'\min\left(\frac{nt^2}{\gamma},\frac{nt}{\gamma^{1/2}}\right)\right]$$

for $k\in[K]$ and some positive constants c_8',c_9', where the last inequality is implied by Assumption 3 (i.e., $\frac{1}{n}\sum_{i=1}^n\pi_{ik}^2\geq\frac{1}{n}\sum_{i=1}^n\pi_{ik}^4>\phi_2$). Then, we have

$$\mathbb{P}\left(\frac{1}{n}\left\|\boldsymbol{D}^\top\boldsymbol{z}_j\right\|_\infty>t\right)\leq 2K\exp\left[-c_9'\min\left(\frac{nt^2}{\gamma},\frac{nt}{\gamma^{1/2}}\right)\right]. \tag{S23}$$

By Lemma A3.5, we have the following result for $\frac{1}{\sqrt{n}}\|\boldsymbol{z}_j\|_2$:

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{n}}\|\boldsymbol{z}_j\|_2-m_j\right|\geq t\right)\leq 4\exp\left[-c_{10}'\min\left(\frac{m_j^2}{\gamma^2}nt^2,\sqrt{\frac{m_j}{\gamma}nt}\right)\right], \tag{S24}$$

where $m_j=\sqrt{\frac{\mathbb{E}[\|\boldsymbol{z}_j\|_2^2]}{n}}=\sqrt{\frac{1}{n}\sum_k\pi_{ik}^2\sigma_{jj}^{(k)}}$ and c_{10}' is a positive constant.

Putting (S17) for term (I) and (S23), (S24) for term (II) together, we have

$$\begin{aligned}\mathbb{P}\left(\|\boldsymbol{\Delta}_{jj'}\|_2\geq t\right)&\leq\mathbb{P}\left[(\mathrm{I})>t/2\right]+\mathbb{P}\left[(\mathrm{II})>t/2\right]\\&\leq c_{11}'\exp\left[-c_{12}'\min\left(nt^2,\sqrt{nt},nt,n\sqrt{t},\sqrt{n}t^{1/4}\right)\right],\end{aligned} \tag{S25}$$

for some positive constants c_{11}' and c_{12}'. Finally, as $[\hat{\boldsymbol{\beta}}_{jj'}]_k=\hat{\sigma}_{jj'}^{(k)}$ and $\left|\hat{\sigma}_{jj'}^{(k)}-\sigma_{jj'}^{(k)}\right|\leq\|\boldsymbol{\Delta}_{jj'}\|_2$, setting $c_1'=c_{11}'$ and $c_2'=c_{12}'$, we achieve the desired result in (S10).

Step 2. Recall $\hat{R}_{jj'}^{(k)}=\frac{\hat{\sigma}_{jj'}^{(k)}}{\sqrt{\hat{\sigma}_{jj}^{(k)}\hat{\sigma}_{j'j'}^{(k)}}}$. In this step, we show

$$\mathbb{P}\left(\left|\hat{R}_{jj'}^{(k)}-R_{jj'}^{(k)}\right|\geq t\right)\leq c_{13}'\exp\left[-c_{14}'\min\left(nt^2,\sqrt{nt},nt,n\sqrt{t},\sqrt{n}t^{1/4}\right)\right] \tag{S26}$$

for $k\in[K],j,j'\in[p]$ and some positive constants c_{13}' and c_{14}'. The proof in this step is motivated by the techniques in Jiang (2013).

We first demonstrate that if $\mathbb{P}\left(|\hat{\sigma}_{jj'} - \sigma_{jj'}| > t\right) \leq c'_{15} \exp(-c'_{16} \times t^{\alpha})$ for $j, j' \in [p]$, some positive constants c'_{15}, α, and some positive scalar c'_{16} that may depend on n. then the corresponding correlation estimate satisfies

$$\mathbb{P}\left(\left|\hat{R}_{jj'} - R_{jj'}\right| > t\right) \leq c'_{17} \exp\left(-c'_{18} \times t^{\alpha}\right),$$

where $c'_{17} = 4c'_{15}, c'_{18} = c'_{16}[\min(\sigma_{jj}, \sigma_{j'j'})/8]^{\alpha}$. Note that the superscript (k) is dropped here and in the ensuing arguments in Step 2, while the derivation holds for all $\sigma_{jj}^{(k)}, \sigma_{j'j'}^{(k)}, \sigma_{jj'}^{(k)}$ for $k \in [K]$.

Let $\tilde{R}_{jj'} = \dfrac{\hat{\sigma}_{jj'}}{\sqrt{\sigma_{jj}\sigma_{j'j'}}}$ and $s_{jj'} = \dfrac{\sqrt{\sigma_{jj}\sigma_{j'j'}}}{\sqrt{\hat{\sigma}_{jj}\hat{\sigma}_{j'j'}}}$. For $0 < t < 2$, we have

$$\mathbb{P}\left(\left|\hat{R}_{jj'} - R_{jj'}\right| > t\right) \leq \underbrace{\mathbb{P}\left(\left|\tilde{R}_{jj'} - R_{jj'}\right| > t/2\right)}_{\text{(I)}} + \underbrace{\mathbb{P}\left(\left|\hat{R}_{jj'} - \tilde{R}_{jj'}\right| > t/2\right)}_{\text{(II)}}. \tag{S27}$$

Assuming $\sigma_{jj}, \sigma_{j'j'} > 0$, we have

$$\text{(I)} = \mathbb{P}\left(\left|\tilde{R}_{jj'} - R_{jj'}\right| > t/2\right) = \mathbb{P}(|\hat{\sigma}_{jj'} - \sigma_{jj'}| > \sqrt{\sigma_{jj}\sigma_{j'j'}} \times t/2). \tag{S28}$$

For term (II) in (S27), consider

$$a = \frac{t}{2(|R_{jj'}| + t)} \tag{S29}$$

and

$$\begin{aligned}
\text{(II)} = \mathbb{P}\left(\left|\hat{R}_{jj'} - \tilde{R}_{jj'}\right| > t/2\right) &= \mathbb{P}\left(\left|\tilde{R}_{jj'}(s_{jj'} - 1)\right| > t/2\right) \\
&\leq \mathbb{P}\left(\left|\tilde{R}_{jj'}\right| > t/2a\right) + \mathbb{P}(|s_{jj'} - 1| > a) \\
&\leq \mathbb{P}\left(\left|\tilde{R}_{jj'} - R_{jj'}\right| > t\right) + \mathbb{P}(|s_{jj'} - 1| > a).
\end{aligned} \tag{S30}$$

The term $\mathbb{P}(|\tilde{R}_{jj'} - R_{jj'}| > t)$ can again be bounded using (S28). Note that $a < 1$, the

following holds for the second term

$$\begin{aligned}
\mathbb{P}(|s_{jj'}-1|>a) &= \mathbb{P}(s_{jj'}>a+1)+\mathbb{P}(s_{jj'}<1-a) \\
&\le \mathbb{P}\left[\frac{1}{s_{jj'}^2}<\frac{1}{(a+1)^2}\right]+\mathbb{P}\left[\frac{1}{s_{jj'}^2}>\frac{1}{(1-a)^2}\right] \\
&\le \mathbb{P}\left[\hat{\sigma}_{jj}<\left(1-\frac{a}{a+1}\right)\times\sigma_{jj}\right]+\mathbb{P}\left[\hat{\sigma}_{j'j'}<\left(1-\frac{a}{a+1}\right)\times\sigma_{j'j'}\right]+ \\
&\quad \mathbb{P}\left[\hat{\sigma}_{jj}>\left(1+\frac{a}{1-a}\right)\times\sigma_{jj}\right]+\mathbb{P}\left[\hat{\sigma}_{j'j'}>\left(1+\frac{a}{1-a}\right)\times\sigma_{j'j'}\right] \\
&\le \mathbb{P}\left(|\hat{\sigma}_{jj}-\sigma_{jj}|>\frac{a}{a+1}\sigma_{jj}\right)+\mathbb{P}\left(|\hat{\sigma}_{j'j'}-\sigma_{j'j'}|>\frac{a}{a+1}\sigma_{j'j'}\right).
\end{aligned} \tag{S31}$$

Finally, by (S29) and the fact that $t\in(0,2)$, we have $\frac{a}{a+1}\ge\frac{t}{3\times2+2}=\frac{t}{8}$. Therefore, summarizing the above terms, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{R}_{jj'}-R_{jj'}\right|>t\right)\le 2\mathbb{P}\left(\left|\tilde{R}_{jj'}-R_{jj'}\right|>t/2\right)+\mathbb{P}\left(|\hat{\sigma}_{jj}-\sigma_{jj}|>\frac{t}{8}\sigma_{jj}\right)+\mathbb{P}\left(|\hat{\sigma}_{j'j'}-\sigma_{j'j'}|>\frac{t}{8}\sigma_{j'j'}\right) \\
&\le 2c_{15}'\exp\left[-c_{16}'\left(\frac{\sqrt{\sigma_{jj}\sigma_{j'j'}}}{2}\right)^{\alpha}t^{\alpha}\right]+c_{15}'\exp\left[-c_{16}'\left(\frac{\sigma_{jj}}{8}\right)^{\alpha}t^{\alpha}\right]+c_{15}'\exp\left[-c_{16}'\left(\frac{\sigma_{j'j'}}{8}\right)^{\alpha}t^{\alpha}\right] \\
&\le c_{17}'\exp\left(-c_{18}'t^{\alpha}\right),
\end{aligned}$$

where $c_{17}'=4c_{15}', c_{18}'=c_{16}'[\min(\sigma_{jj},\sigma_{j'j'})/8]^{\alpha}$. These arguments, combined with the results from Step 1, can be used to establish (S26).

Step 3. By Steps 1 and 2, we can then bound the max norm of $\hat{\boldsymbol{R}}^{(k)}-\boldsymbol{R}^{(k)}$ as

$$\mathbb{P}\left(\max_{j,j'}\left|\hat{R}_{jj'}^{(k)}-R_{jj'}^{(k)}\right|>t\right)\le p^2\times c_{13}'\exp\left[-c_{14}'\min\left(nt^2,\sqrt{n}t,nt,n\sqrt{t},\sqrt{n}t^{1/4}\right)\right], \tag{S32}$$

where c_{13}' and c_{14}' are some positive constants. Suppose $\log p=o(n^{1/3})$ and set $\lambda=\sqrt{\frac{\log p}{n}}\times M$ for some large constant M. Then, letting $t=\lambda$, we have $\min\left(nt^2,\sqrt{n}t,nt,n\sqrt{t},\sqrt{n}t^{1/4}\right)=nt^2$ and (S32) can be written as

$$\max_{j,j'}\left|\hat{R}_{jj'}^{(k)}-R_{jj'}^{(k)}\right|\le\sqrt{\frac{\log p}{n}}\times M, \tag{S33}$$

holds with probability at least $1-c'_{13}exp(-c'_{19}\log p)$, where $c'_{19}=c'_{14}M^2-2$ is a large positive constant. Combined with Lemma A3.4, we have

$$\left\|\mathcal{T}_\lambda\left(\hat{\boldsymbol{R}}^{(k)}\right)-\boldsymbol{R}^{(k)}\right\|\leq s_p\left(\frac{\log p}{n}\right)^{(1-q)/2}M \tag{S34}$$

with probability at least $1-c'_{20}\exp(-c'_{21}\log p)$ for some positive constants c'_{20} and c'_{21}.

A3.3 Proof of Theorem 5.2

As it holds that

$$\bigcup_{i,j}\left\{\mathcal{T}_\lambda\left(\hat{R}^{(k)}_{jj'}\right)\neq 0, R^{(k)}_{jj'}=0\right\}\subseteq\bigcup_{i,j}\left\{\left|\hat{R}^{(k)}_{jj'}-R^{(k)}_{jj'}\right|\geq\lambda\right\}=\mathbf{1}\left\{\max_{j,j'}\left|\hat{R}^{(k)}_{jj'}-R^{(k)}_{jj'}\right|\geq\lambda\right\},$$

we have by (S32), $\log p=o(n^{1/3})$ and $\lambda=\sqrt{\frac{\log p}{n}}\times M$ for some large constant M that

$$\text{(I)}=\mathbb{P}\left(\sum_{j,j'}\mathbf{1}\left\{\mathcal{T}_\lambda\left(\hat{R}^{(k)}_{jj'}\right)\neq 0, R^{(k)}_{jj'}=0\right\}>0\right)\leq c'_{13}\exp(-c'_{19}\log p) \tag{S35}$$

for c'_{13} and c'_{19} as defined in Section A3.2.

Given $|R^{(k)}_{jj'}|>\rho_{\min}$ and $\sqrt{n}(\rho_{\min}-\lambda)\to\infty$, we have

$$\bigcup_{j,j'}\left\{\mathcal{T}_\lambda\left(\hat{R}^{(k)}_{jj'}\right)R^{(k)}_{jj'}\leq 0\right\}\subseteq\bigcup_{j,j'}\left\{\left|\hat{R}^{(k)}_{jj'}-R^{(k)}_{jj'}\right|\geq\rho_{\min}-\lambda\right\}=\mathbf{1}\left\{\max_{j,j'}\left|\hat{R}^{(k)}_{jj'}-R^{(k)}_{jj'}\right|\geq\rho_{\min}-\lambda\right\}.$$

Combined with (S32), we obtain

$$\text{(II)}=\mathbb{P}\left(\sum_{i,j}\mathbf{1}\left\{\mathcal{T}_\lambda\left(\hat{R}^{(k)}_{jj'}\right)=0, R^{(k)}_{jj'}\neq 0\right\}>0\right)\to 0 \tag{S36}$$

as $\sqrt{n}(\rho_{\min}-\lambda)\to\infty$. Finally, as $1-\mathbb{P}\{\Omega[\mathcal{T}_\lambda(\hat{\boldsymbol{R}}^{(k)})]=\Omega(\boldsymbol{R}^{(k)})\}=\text{(I)}+\text{(II)}$, we obtain $\mathbb{P}\{\Omega[\mathcal{T}_\lambda(\hat{\boldsymbol{R}}^{(k)})]=\Omega(\boldsymbol{R}^{(k)})\}=\Omega(\boldsymbol{R}^{(k)})]\to 1$.

A3.4 Proof of Lemma A3.2

The proof is divided into two steps. In step 1, we will establish the tail bound for the weighted sum of fourth degree polynomials of sub-Gaussian random variables. Next, in step 2, we will apply the established tail bound to product of sub-exponential random variables by representing centered products of sub-exponential random variables as a fourth degree polynomial of sub-Gaussian random variables.

Step 1: Following Adamczak and Wolff (2015), we define a norm for tensors that will be used in our concentration result. Let $d \in \mathbb{N}_+$ be a positive integer. We denote by $\mathcal{P}_d$ the set of its partitions of $[d]$ into non-empty and non-intersecting disjoint sets. Moreover, let $A = (a_{\boldsymbol{i}})_{\boldsymbol{i}\in[n]^d}$ be a tensor of order-d, whose entries are of the form

$$a_{\boldsymbol{i}} = a_{i_1,i_2,\ldots,i_d}, \text{ where } \boldsymbol{i} = (i_1, i_2, \ldots, i_d).$$

Let $\mathcal{J} = \{J_1, \ldots, J_k\} \in \mathcal{P}_d$ be a fixed partition of $[d]$, where $J_j \subseteq [d]$ for each $j \in [k]$. Let $|\mathcal{J}|$ denote the cardinality of the $\mathcal{J}$, which is equal to k. We define a norm $\|\cdot\|_{\mathcal{J}}$ by

$$\|A\|_{\mathcal{J}} = \sup\left\{\sum_{\boldsymbol{i}\in[n]^d} a_{\boldsymbol{i}} \prod_{j=1}^{k} \boldsymbol{x}^{(j)}_{\boldsymbol{i}_{J_j}} : \|\boldsymbol{x}^{(j)}\|_2 \leq 1, \boldsymbol{x}^{(j)} \in \mathbb{R}^{n^{|J_j|}}, 1 \leq j \leq k\right\}, \tag{S37}$$

where we write $\boldsymbol{i}_I = (i_k)_{k\in I}$ for any $I \subseteq [d]$ and the supremum is taken over all possible k vectors $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(k)}\}$.

Let $f(u) = u^4$ and consider a polynomial of degree 4:

$$F : \mathbb{R}^n \to \mathbb{R}, F(\boldsymbol{u}) = \sum_{i=1}^{n} a_i f(u_i). \tag{S38}$$

Let $\boldsymbol{Z} = [Z_1, \ldots, Z_n]^\top$ be a vector of independent components with $\|Z_i\|_{\psi_2} \leq (\gamma')^{1/4}$, where

γ' is some positive constant. Plugging in $F(Z)$, letting $D = 4$, Lemma A3.3 gives

$$\mathbb{P}\left\{|F(\boldsymbol{Z}) - \mathbb{E}[F(\boldsymbol{Z})]| \geq t\right\} \leq 2\exp(-c_0 \times \eta_F(t)), \tag{S39}$$

and

$$\eta_F(t) = \min_{1\leq d\leq 4} \min_{\mathcal{J}\in\mathcal{P}_d} \left(\frac{t}{(\gamma')^{d/4} \times \|\mathbb{E}[\mathcal{D}^d F(\boldsymbol{Z})]\|_{\mathcal{J}}}\right)^{2/|\mathcal{J}|}. \tag{S40}$$

Following Lemma 3 of Balasubramanian et al. (2018), we have

$$\mathcal{D}^d F(Z) = \mathrm{diag}_d\left[a_1 f^{(d)}(Z_1), \ldots, a_n f^{(d)}(Z_n)\right].$$

Using (31) in Balasubramanian et al. (2018) and the fact that $|f^{(d)}(u)| \leq 4!|u|^{4-d}$ for $d = 1, \ldots, 4$, we obtain

$$\begin{aligned}\left\|\mathbb{E}\left[\mathcal{D}^d F(Z)\right]\right\|_{\mathcal{J}} &\leq \max_i \left|\mathbb{E}\left[f^{(d)}(Z_i)\right]\right| \times (\mathbf{1}\{|\mathcal{J}| = 1\}\|\boldsymbol{a}\|_2 + \mathbf{1}\{|\mathcal{J}|) \geq 2\}\|\boldsymbol{a}\|_\infty) \\ &\leq 4! \times \max_i \mathbb{E}\left(|Z_i|^{4-d}\right) \times (\mathbf{1}\{|\mathcal{J}| = 1\}\|\boldsymbol{a}\|_2 + \mathbf{1}\{|\mathcal{J}| \geq 2\}\|\boldsymbol{a}\|_\infty).\end{aligned}$$

By the definition of ψ_2-norm, $\mathbb{E}\left(|Z_i|^{4-d}\right) \leq (4-d)^{(4-d)/2}\|Z_i\|_{\psi_2}^{4-d}$. In addition, recall that $\max_i \|Z_i\|_{\psi_2} \leq \gamma'^{1/4}$, then

$$\begin{aligned}\left\|\mathbb{E}\left[\mathcal{D}^d F(Z)\right]\right\|_{\mathcal{J}} &\leq 4! \times (4-d)^{(4-d)/2} \max_i \|Z_i\|_{\psi_2}^{4-d} \times (\mathbf{1}\{|\mathcal{J}| = 1\}\|\boldsymbol{a}\|_2 + \mathbf{1}\{|\mathcal{J}| \geq 2\}\|\boldsymbol{a}\|_\infty) \\ &\leq \tilde{c}_d \times (\gamma')^{(4-d)/4} \times (\mathbf{1}\{|\mathcal{J}| = 1\}\|\boldsymbol{a}\|_2 + \mathbf{1}\{|\mathcal{J}| \geq 2\}\|\boldsymbol{a}\|_\infty),\end{aligned}$$

where $\tilde{c}_d = 4! \times (4-d)^{(4-d)/2}$ for $d = 1, \ldots, 4$. Then,

$$(\gamma')^{d/4} \times \|\mathbb{E}[D^\alpha F(Z)]\|_{\mathcal{J}} \leq \begin{cases} \tilde{c}_d \gamma' \|\boldsymbol{a}\|_2 & \text{if } |\mathcal{J}| = 1 \\ \tilde{c}_d \gamma' \|\boldsymbol{a}\|_\infty & \text{if } |\mathcal{J}| \geq 2 \end{cases} \tag{S41}$$

Combining (S40) and (S41), we have

$$\eta_F(t) \geq \min_{1\leq d\leq 4} \left\{\min\left[\left(\frac{t}{\tilde{c}_d \gamma' \|\boldsymbol{a}\|_2}\right)^2, \min_{2\leq|\mathcal{J}|\leq d}\left(\frac{t}{\tilde{c}_d \gamma' \|\boldsymbol{a}\|_\infty}\right)^{2/|\mathcal{J}|}\right]\right\}.$$

If $t < c_{\min} \times \gamma' \times \|\boldsymbol{a}\|_\infty$ for $c_{\min} = \min_{d\in\{2,3,4\}} \{\tilde{c}_d\}$, then

$$\begin{aligned}\eta_F(t) &\geq \min\left[\min_{1\leq d\leq 4}\left(\frac{t}{\tilde{c}_d\gamma'\|\boldsymbol{a}\|_2}\right)^2, \min_{2\leq d\leq 4}\frac{t}{\tilde{c}_d\gamma'\|\boldsymbol{a}\|_\infty}\right] \\ &\geq \min\left[\left(\frac{t}{c_{\max}\gamma'\|\boldsymbol{a}\|_2}\right)^2, \frac{t}{c_{\max}\gamma'\|\boldsymbol{a}\|_\infty}\right].\end{aligned} \tag{S42}$$

where $c_{\max} = \max_{d\in\{1,\ldots,4\}} \{\tilde{c}_d\}$. Plugging it in Lemma A3.3, we obtain

$$\mathbb{P}\left[\left|\sum_{i=1}^n a_i Z_i^4 - a_i\mathbb{E}(Z_i^4)\right| > t\right] \leq 2\exp\left\{-c_0 \min\left[\left(\frac{t}{c_{\max}\gamma'\|\boldsymbol{a}\|_2}\right)^2, \frac{t}{c_{\max}\gamma'\|\boldsymbol{a}\|_\infty}\right]\right\}. \tag{S43}$$

If $t \geq c_{\max} \times \gamma' \times \|\boldsymbol{a}\|_\infty$, then

$$\begin{aligned}\eta_F(t) &\geq \min\left[\min_{1\leq d\leq 4}\left(\frac{t}{\tilde{c}_l\gamma'\|\boldsymbol{a}\|_2}\right)^2, \min_{2\leq d\leq 4}\left(\frac{t}{\tilde{c}_l\gamma'\|\boldsymbol{a}\|_\infty}\right)^{2/l}\right] \\ &\geq \min\left[\left(\frac{t}{c_{\max}\gamma'\|\boldsymbol{a}\|_2}\right)^2, \left(\frac{t}{c_{\max}\gamma'\|\boldsymbol{a}\|_\infty}\right)^{1/2}\right].\end{aligned}$$

Similarly, we obtain

$$\mathbb{P}\left[\left|\sum_{i=1}^n a_i Z_i^4 - a_i\mathbb{E}(Z_i^4)\right| > t\right] \leq 2\exp\left\{-c_0 \times \min\left[\left(\frac{t}{c_{\max}\gamma'\|\boldsymbol{a}\|_2}\right)^2, \left(\frac{t}{c_{\max}\gamma'\|\boldsymbol{a}\|_\infty}\right)^{1/2}\right]\right\}. \tag{S44}$$

Step 2: Next, we show how a product of sub-exponential random variables can be represented as a fourth degree polynomial of Gaussian random variables, and then leverage the bound established in step 1 here.

Let $B_i^+ = X_iY_i\mathbf{1}\{X_iY_i \geq 0\}, B_i^- = -X_iY_i\mathbf{1}\{X_iY_i < 0\}$. It follows that $X_iY_i = B_i^+ - B_i^-$. Furthermore, define $Z_i = \eta_i|B_i^+|^{1/4}$ for $i = 1,\ldots,n$, where $\{\eta_i\}_{i=1}^n$ are independent Rademacher random variables. Then, we have

$$\mathbb{E}|Z_i|^p \leq \mathbb{E}|X_iY_i|^{p/4} \leq \sqrt{\mathbb{E}\left(|X_i|^{p/2}\right)\mathbb{E}\left(|Y_i|^{p/2}\right)} \leq \left(\|X_i\|_{\psi_1}\|Y_i\|_{\psi_1}\right)^{p/4}(p/2)^{p/2},$$

where the second inequality is given by Cauchy-Schwarz inequality and the third is given by the definition of ψ_1-norm in Assumption 1. Then, by the definition of ψ_2-norm,

$$\|Z\|_{\psi_2} = \sup_p p^{-1/2} \left(\mathbb{E}|Z|^p\right)^{1/p} \leq \frac{1}{\sqrt{2}} \left(\|X_i\|_{\psi_1} \|Y_i\|_{\psi_1}\right)^{1/4}.$$

Therefore, for $i = 1, \ldots, n$, Z_i are sub-Gaussian random variables with

$$\|Z_i\|_{\psi_2} \leq \max_i \frac{1}{\sqrt{2}} \left(\|X_i\|_{\psi_1} \|Y_i\|_{\psi_1}\right)^{1/4} = \frac{1}{\sqrt{2}} \gamma_2^{1/4}. \tag{S45}$$

Next, we start to apply the results obtained in step 1. First consider the case where $0 < t < \frac{c_{\min}}{2} \times \gamma_2 \times \|\boldsymbol{a}\|_\infty$. Then, plugging $\gamma' = \gamma_2/4$ and the tail cutoff $t/2$ into (S43), we obtain

$$\mathbb{P}\left[\left|\sum_{i=1}^n a_i Z_i^4 - a_i \mathbb{E}(Z_i^4)\right| > t/2\right] \leq 2\exp\left\{-c_0 \min\left[\left(\frac{t}{c_{\max}\gamma_2 \|\boldsymbol{a}\|_2/2}\right)^2, \frac{t}{c_{\max}\gamma_2 \|\boldsymbol{a}\|_\infty/2}\right]\right\}. \tag{S46}$$

As $Z_i^4 = |B_i^+| = B_i^+$, it follows that

$$\mathbb{P}\left[\left|\sum_{i=1}^n a_i B_i^+ - a_i \mathbb{E}(B_i^+)\right| > t/2\right] \leq 2\exp\left\{-c_0 \min\left[\left(\frac{t}{c_{\max}\gamma_2 \|\boldsymbol{a}\|_2/2}\right)^2, \frac{t}{c_{\max}\gamma_2 \|\boldsymbol{a}\|_\infty/2}\right]\right\}. \tag{S47}$$

By symmetry, similar derivations give the same bound for B_i^- as (S47). Then

$$\begin{aligned}
&\mathbb{P}\left[\left|\sum_{i=1}^n a_i X_i Y_i - a_i \mathbb{E}(X_i Y_i)\right| > t\right] \\
&\leq \mathbb{P}\left[\left|\sum_{i=1}^n a_i B_i^+ - a_i \mathbb{E}(B_i^+)\right| > t/2\right] + \mathbb{P}\left[\left|\sum_{i=1}^n a_i B_i^- - a_i \mathbb{E}(B_i^-)\right| > t/2\right] \\
&\leq 4\exp\left\{-c_0 \min\left[\left(\frac{t}{c_{\max}\gamma_2 \|\boldsymbol{a}\|_2/2}\right)^2, \frac{t}{c_{\max}\gamma_2 \|\boldsymbol{a}\|_\infty/2}\right]\right\}.
\end{aligned}$$

Finally, consider the case where $t > \frac{c_{\max}}{2} \times \gamma_2 \times \|\boldsymbol{a}\|_\infty$. Following a similar argument, we

get

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} a_i X_i Y_i - a_i \mathbb{E}(X_i Y_i)\right| > t\right] \leq 4\exp\left\{-c_0 \min\left[\left(\frac{t}{c_{\max}\gamma_2\|\boldsymbol{a}\|_2/2}\right)^2, \left(\frac{t}{c_{\max}\gamma_2\|\boldsymbol{a}\|_\infty/2}\right)^{1/2}\right]\right\}.$$

Then, setting $d_1 = c_{\min}/2$, $c_2 = 4c_0/c_{\max}^2$, $d_2 = c_{\max}/2$ gives the desired inequalities.

A3.5 Proof of Lemma A3.5

Given $\mu = \sqrt{\frac{1}{n}[\mathbb{E}\|X\|_2^2]}$, by Lemma A3.2, we have

$$\mathbb{P}\left(\left|\frac{1}{n}\|X\|_2^2 - \mu^2\right| \geq \delta\right) \leq 4\exp\left\{-C_1' \times \min\left[\left(\frac{\sqrt{n}\delta}{\gamma_4}\right)^2, \left(\frac{n\delta}{\gamma_4}\right)^{1/2}\right]\right\}, \tag{S48}$$

for any positive constant δ and some positive constant C_1'. Equivalently, let $u = \frac{\delta}{\mu^2}$, then

$$\mathbb{P}\left(\left|\frac{\|X\|_2^2}{n\mu^2} - 1\right| \geq u\right) \leq 4\exp\left\{-C_1' \times \min\left[\left(\frac{\sqrt{n}u\mu^2}{\gamma_4}\right)^2, \left(\frac{nu\mu^2}{\gamma_4}\right)^{1/2}\right]\right\}. \tag{S49}$$

As $|z-1| > u$ implies $|z^2 - 1| > \max(u, u^2)$ for any $z \geq 0$. Then, if $u \leq 1$,

$$\begin{aligned}\mathbb{P}\left(\left|\frac{\|X\|_2}{\sqrt{n\mu^2}} - 1\right| \geq u\right) &\leq \mathbb{P}\left[\left|\frac{\|X\|_2^2}{n\mu^2} - 1\right| \geq \max(u, u^2)\right] \\ &\leq 4\exp\left\{-C_1' \times \min\left[\left(\frac{\sqrt{n}u\mu^2}{\gamma_4}\right)^2, \left(\frac{nu\mu^2}{\gamma_4}\right)^{1/2}\right]\right\}\end{aligned} \tag{S50}$$

Setting $t = u\mu$ and $c_7 = C_1'$ gives (S9).

A3.6 Extended discussions on Assumptions 3

Recall

$$\boldsymbol{\Psi} = \frac{1}{n}\boldsymbol{D}^\top \boldsymbol{D} = \frac{1}{n}\begin{pmatrix} \sum_{i=1}^n \pi_{i1}^2 & \sum_{i=1}^n \pi_{i1}\pi_{i2} & \cdots & \sum_{i=1}^n \pi_{i1}\pi_{iK} \\ \sum_{i=1}^n \pi_{i1}\pi_{i2} & \sum_{i=1}^n \pi_{i2}^2 & \cdots & \sum_{i=1}^n \pi_{i2}\pi_{iK} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n \pi_{i1}\pi_{iK} & \sum_{i=1}^n \pi_{i2}\pi_{iK} & \cdots & \sum_{i=1}^n \pi_{iK}^2 \end{pmatrix}, \tag{S51}$$

$$\Phi = \frac{1}{n}\boldsymbol{H}^\top \boldsymbol{H} = \frac{1}{n}\begin{pmatrix} \sum_{i=1}^n \pi_{i1}^4 & \sum_{i=1}^n \pi_{i1}^2\pi_{i2}^2 & \cdots & \sum_{i=1}^n \pi_{i1}^2\pi_{iK}^2 \\ \sum_{i=1}^n \pi_{i1}^2\pi_{i2}^2 & \sum_{i=1}^n \pi_{i2}^4 & \cdots & \sum_{i=1}^n \pi_{i2}^2\pi_{iK}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n \pi_{i1}^2\pi_{iK}^2 & \sum_{i=1}^n \pi_{i2}^2\pi_{iK}^2 & \cdots & \sum_{i=1}^n \pi_{iK}^4 \end{pmatrix}. \tag{S52}$$

First, suppose $(\pi_{i1}, \ldots, \pi_{iK}) \sim \text{Multinomial}(1, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \in \mathbb{R}^K$ is a vector of positive parameters summed to 1. Though this assumption on cell type proportions implies that each tissue sample can have cells from only one of the K cell types, it nevertheless provides some insight into (S51) and (S52). Specifically, the smallest eigenvalue of

$$\mathbb{E}\left(\boldsymbol{\Psi}\right) = \mathbb{E}\left(\boldsymbol{\Phi}\right) = \text{Diag}(\boldsymbol{\theta}) \tag{S53}$$

is $\min\{\theta_k\}$, a positive constant that does not depend on n.

Next, consider a less simplified model with $(\pi_{i1}, \ldots, \pi_{iK}) \sim \text{Dirichlet}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K) \in \mathbb{R}^K$ is a vector of postive parameters. Let $\alpha_0 = \sum_k \alpha_k$. Then,

$$\mathbb{E}\left(\boldsymbol{\Psi}\right) = \frac{(\alpha_0 - 1)!}{(\alpha_0 + 1)!}\left[\boldsymbol{\alpha}\boldsymbol{\alpha}^\top + \text{diag}(\boldsymbol{\alpha})\right], \quad \mathbb{E}\left(\boldsymbol{\Phi}\right) = \frac{(\alpha_0 - 1)!}{(\alpha_0 + 3)!}\left[\boldsymbol{\gamma}\boldsymbol{\gamma}^\top + \text{diag}(\boldsymbol{\Delta})\right] \tag{S54}$$

where $\boldsymbol{\gamma} = (\boldsymbol{\alpha} + 1) \circ \boldsymbol{\alpha} \in \mathbb{R}^K, \boldsymbol{\Delta} = \boldsymbol{\alpha} \circ (\boldsymbol{\alpha} + 1) \circ (4\boldsymbol{\alpha} + 6) \in \mathbb{R}^K$. The smallest eigenvalue of $\mathbb{E}(\boldsymbol{\Psi})$ and $\mathbb{E}(\boldsymbol{\Phi})$ are lower bounded by positive constants, $\frac{(\alpha_0 - 1)!}{(\alpha_0 + 1)!}\min_k\{\boldsymbol{\alpha}_k\}$ and $\frac{(\alpha_0 - 1)!}{(\alpha_0 + 3)!}\min_k\{\boldsymbol{\Delta}_k\}$, respectively.

The above derivation demonstrates that under reasonable probabilistic models, the eigenvalue conditions in Assumption 3 are satisfied at the population level. This bounded norm condition in Assumption 3 requires the sum of the squared proportions to grow linearly with the size of the population. This is plausible when considering any common cell types: a cell type is defined as "common" if $\mathbb{E}[\pi_{ik}] > C_k$ for some constant $C_k > 0$, which also implies $\mathbb{E}[\pi_{ik}^2] > C_k^2$.

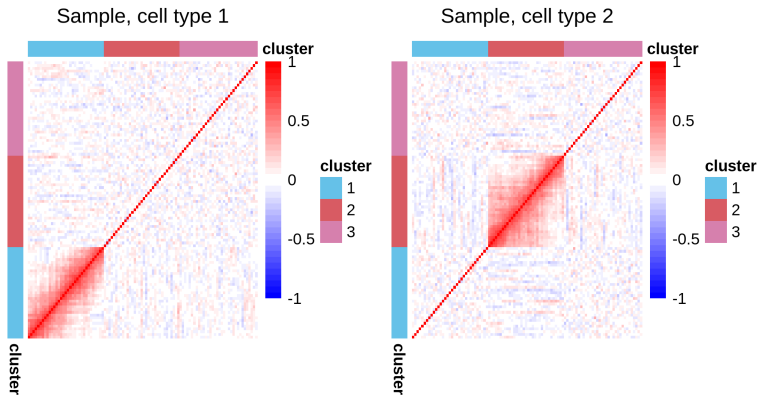


Figure S9: Sample correlations from $n = 150$ samples of $p = 100$ genes, with data simulated from Algorithm 1 under a multivariate negative binomial distribution with an AR(1)-type correlation structure in V_1 for cell type 1 and V_2 for cell type 2 (See Section 3).

A4 More Results from Simulation studies in Section 3

In Section 4, we followed the `bMIND` tutorial at https://htmlpreview.github.io/?https://github.com/randel/MIND/blob/master/bMIND_tutorial.html to estimate cell-type-specific expression levels in each sample using `bMIND`. There are two steps provided by the `bMIND` implementation: first, use single cell data to infer priors for each gene; second, fit the Bayesian mixed effects model to infer posterior means in each sample. However, due to the high dropout rates in single cell data and/or the low expression levels, some genes may have zero UMI counts detected, which is infeasible to infer priors. Therefore, only a subset of genes would have valid priors inferred by `bMIND`. For those genes, we supplied the valid priors as informative priors to `bMIND` for posterior inference, as recommended by the tutorial. For the other genes, we used the default non-informative priors in `bMIND`.

			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	`bMIND (inf)`	9.05	2.54	-	-	11.76	3.82	-	-
			(0.10)	(0.12)	-	-	(0.16)	(0.23)	-	-
	200	`bMIND (inf)`	18.02	4.20	-	-	23.41	6.70	-	-
			(0.10)	(0.15)	-	-	(0.19)	(0.35)	-	-
600	100	`bMIND (inf)`	4.72	1.29	-	-	6.41	1.77	-	-
			(0.05)	(0.01)	-	-	(0.08)	(0.12)	-	-
	200	`bMIND (inf)`	9.17	1.72	-	-	12.53	3.15	-	-
			(0.05)	(0.05)	-	-	(0.09)	(0.20)	-	-

Table S1: Evaluation criteria under the same setting as Table 1. `bMIND (inf)` denotes the co-expression estimates computed with `bMIND` given informative priors. We use F-norm to denote the Frobenius norm and Op.-norm to denote the operator norm.

			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`	25.72 (0.21)	24.67 (0.17)	0.32 (0.05)	**0.75 (0.06)**	28.55 (0.26)	24.67 (0.17)	0.36 (0.04)	**0.43 (0.07)**
		`d-CSNet`	11.82 (0.28)	4.78 (0.63)	- -	- -	25.87 (0.98)	10.13 (1.08)	- -	- -
		`CSNet`	**5.39 (0.59)**	**3.58 (0.73)**	**0.16 (0.04)**	0.72 (0.07)	**10.85 (0.54)**	**8.43 (0.69)**	**0.08 (0.02)**	0.37 (0.09)
		`bMIND`	12.07 (0.26)	7.22 (0.43)	- -	- -	16.48 (0.42)	9.48 (0.19)	- -	- -
	200	`Bulk`	49.91 (0.24)	48.58 (0.21)	0.26 (0.03)	**0.48 (0.04)**	52.98 (0.23)	48.58 (0.21)	0.28 (0.03)	**0.27 (0.04)**
		`d-CSNet`	23.76 (0.30)	8.15 (0.74)	- -	- -	51.50 (1.39)	18.56 (1.76)	- -	- -
		`CSNet`	**9.19 (0.57)**	**5.82 (0.65)**	**0.11 (0.02)**	0.46 (0.04)	**17.14 (0.40)**	**11.41 (0.49)**	**0.04 (0.02)**	0.15 (0.05)
		`bMIND`	20.99 (0.27)	8.97 (0.39)	- -	- -	27.84 (0.48)	11.61 (0.37)	- -	- -
600	100	`Bulk`	25.54 (0.10)	24.78 (0.08)	0.37 (0.05)	**0.90 (0.04)**	28.42 (0.17)	24.78 (0.08)	0.40 (0.05)	0.67 (0.05)
		`d-CSNet`	5.82 (0.12)	2.17 (0.29)	- -	- -	12.08 (0.28)	4.20 (0.53)	- -	- -
		`CSNet`	**2.73 (0.34)**	**1.63 (0.42)**	**0.13 (0.07)**	0.86 (0.05)	**5.72 (0.53)**	**3.68 (0.77)**	**0.16 (0.05)**	**0.71 (0.07)**
		`bMIND`	8.89 (0.10)	6.77 (0.18)	- -	- -	12.53 (0.31)	8.82 (0.21)	- -	- -
	200	`Bulk`	49.66 (0.13)	48.79 (0.12)	0.30 (0.03)	**0.63 (0.04)**	53.03 (0.17)	48.79 (0.12)	0.32 (0.03)	**0.43 (0.05)**
		`d-CSNet`	11.73 (0.14)	3.54 (0.37)	- -	- -	24.18 (0.37)	7.19 (0.71)	- -	- -
		`CSNet`	**4.79 (0.32)**	**2.82 (0.49)**	**0.10 (0.04)**	0.57 (0.05)	**9.71 (0.54)**	**6.10 (0.71)**	**0.11 (0.03)**	**0.43 (0.04)**
		`bMIND`	13.42 (0.11)	8.42 (0.21)	- -	- -	19.58 (0.40)	10.46 (0.24)	- -	- -

Table S2: Evaluation criteria with varying sample size n and network size p under a negative binomial distribution with an AR(1)-type correlation structure in V_1 or V_2 (See Section 3). The four methods under comparison are `Bulk`, `d-CSNet`, `CSNet`, and `bMIND`. 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.

			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	`bMIND (inf)`	11.78	7.20	-	-	14.81	8.12	-	-
			(0.16)	(0.30)	-	-	(0.24)	(0.37)	-	-
	200	`bMIND (inf)`	20.32	9.09	-	-	26.39	10.10	-	-
			(0.18)	(0.32)	-	-	(0.31)	(0.39)	-	-
600	100	`bMIND (inf)`	8.89	6.95	-	-	10.89	7.72	-	-
			(0.08)	(0.14)	-	-	(0.12)	(0.21)	-	-
	200	`bMIND (inf)`	13.16	8.70	-	-	17.17	9.45	-	-
			(0.08)	(0.17)	-	-	(0.17)	(0.25)	-	-

Table S3: Evaluation criteria under the same setting as Table S2. `bMIND (inf)` denotes the co-expression estimates computed with `bMIND` given informative priors. We use F-norm to denote the Frobenius norm and Op.-norm to denote the operator norm.