CSNet: Estimating cell-type-specific gene co-expression networks from bulk gene expression data

Chang Su[1], Jingfei Zhang[2], Hongyu Zhao[3]

[1,3] Department of Biostatistics, Yale University

[2] Department of Management Science, University of Miami

Abstract

Inferring and characterizing gene co-expression networks have led to important insights on the molecular mechanisms and functional pathways in healthy and diseased individuals. Most co-expression analyses to date have been performed on gene expression data collected from bulk tissues with different cell type compositions across samples, resulting in co-expression estimates confounded by heterogeneity in cell type proportions. To address this limitation in co-expression analysis, we propose a flexible framework that estimates cell-type-specific gene co-expressions from bulk sample data, where the cell-type-specific distributions of gene expression levels are not assumed known. To overcome the computational challenge in estimating covariances and correlations from a convolution of high dimensional densities, we develop a novel thresholded least squares estimator, named `CSNet`, that is efficient to implement and has good theoretical properties. We further investigate the convergence rate of `CSNet`. The utility and efficacy of `CSNet` is demonstrated through simulation studies and an application to a gene co-expression study with bulk samples from Alzheimer's disease patients, where our analysis identified new cell-type-specific modules of AD risk genes.

Keywords: approximate sparsity; cell-type-specific co-expression network; deconvolution; high-dimensional covariance estimation; SCAD.

1 Introduction

Gene co-expression networks, which characterize correlations of gene expression levels across biological samples, capture the complex interactions among genes whose expressions are regulated by transcriptional factors, chromatin conformations and epigenetic modifications (Gaiteri et al., 2014). Statistical network analyses, such as gene clustering (Langfelder and Horvath, 2008; Gaiteri et al., 2015), have been successful in identifying co-expressed modules underlying biological processes, which can be further associated with disease phenotypes (Mostafavi et al., 2018; Wang et al., 2016). For example, an integrated network analysis in Zhang et al. (2013) identified gene 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 (Haure-Mirande et al., 2019).

The co-expression analyses to date have mostly been performed on bulk tissue samples, that may have different cell type compositions across these samples. As a result, the co-expressions may be confounded with cell type proportions. For example, consider two genes that are both highly expressed in one cell type without correlation 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, rendering an inaccurate view of the gene co-expression network. Although cell-type-specific networks can possibly be estimated from single cell RNA-seq data that have expression profiles in single cells, these data are much more limited in the number of biological samples analyzed, have high noises due to low coverage and biological noises at the single-cell level, and may be biased due to cell isolation and sequencing protocols. While the literature on co-expression network analyses is steadily increasing (e.g., Zhang et al., 2013; Mostafavi et al., 2018; Wan et al., 2020), very few methods allow estimating cell-type-specific co-expression networks from bulk samples.

In this work, we propose a flexible framework for estimating cell-type-specific gene co-expressions (i.e., correlation matrices) using bulk sample data. Our proposal does not require the distributions of expression levels from different cell types to be known. To overcome the computational challenge in estimating the covariances in a convolution of densities especially when the dimension p (i.e., number of genes) is large, we propose a novel least squares approach that is efficient to implement and has good theoretical properties. We further propose a sparse estimator with SCAD penalty, named `CSNet`, that allows the number of genes p to far exceed the sample size n. In theory, assuming the cell-type-specific gene expression profile (marginally) is sub-exponential, we derive the non-asymptotic convergence rate in spectral norm of `CSNet` and establish variable selection consistency, ensuring that the edges in the cell-type specific networks can be correctly identified with probability tending to 1. Our theoretical analysis is nontrivial, challenged by the sub-exponential tail of the cell-type-specific expression profile and the uncertainty from estimating the unknown cell-type-specific mean parameters. Our procedure also considers correlation matrices, normalized with estimated variances, which adds additional complexity to our theoretical analysis.

We note that there is a related literature on decomposing bulk gene expression profiles into cell-type-specific profiles (Cobos et al., 2020). Great progress have been made recently (e.g., Abbas et al., 2009; Wang et al., 2019; Newman et al., 2019; Jew et al., 2020a) on inferring cell type proportions or cell-type-specific mean expressions using the bulk data. Given the inferred cell type proportions, estimating the large cell-type-specific gene co-expression networks is challenging as it requires estimating the cell-type-specific covariance (or correlation) from a convolution of unknown densities; see Section 2.2.

The rest of the paper is organized as follows. Section 2 introduces the model and proposes `CSNet` for cell-type-specific correlation estimation from bulk samples. Section 3 investigates theoretical properties of `CSNet`. Section 4 reports the simulation results, and Section 5

conducts an analysis of cell-type-specific Alzheimer's disease gene co-expression networks using bulk samples from the ROSMAP project. Section 6 concludes the paper with a brief discussion. All proofs and technical lemmas are relegated to the Supplementary Materials.

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 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 gene expression and co-expression profiles. 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 written 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 any positive integer m. We shall assume the cell-type proportions $\{\pi_{ik}\}_{i \in [n], k \in [K]}$ in (2) are given in the ensuing development (see more discussions in Section 6). Many methods have been developed for inferring π_{ik}'s from bulk samples using signature genes (i.e., a small set of cell-type-specific genes) with expressions profiles gathered

from pure cell types (Newman et al., 2015; Li et al., 2016) or scRNA-seq data (Wang et al., 2019; Newman et al., 2019; Jew et al., 2020b; Dong 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 cell-type proportions π_{ik}'s, our goal is to estimate the cell-type-specific correlations $\boldsymbol{R}^{(k)}$'s from bulk samples.

2.2 Element-wise estimation

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 $\boldsymbol{\Sigma}^{(k)}$'s 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,

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

where $\operatorname{tr}(\cdot)$ 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 new least squares approach that is efficient to implement and flexible, in that it allows heterogeneous distributions in the K cell types. Next, we discuss this estimation procedure.

Letting $\boldsymbol{\Sigma}^{(k)} = (\sigma_{jj'}^{(k)})_{p\times p}$ and $z_{ij} = x_{ij} - \sum_{k=1}^{K} \pi_{ik}\mu_j^{(k)}$, (1) and (2) together imply

$$z_{ij} z_{ij'} = \sum_{k=1}^{K} \pi_{ik}^2 \sigma_{jj'}^{(k)} + \epsilon_{ijj'}, \quad j, j' \in [p], \tag{3}$$

where $\mathbb{E}(\epsilon_{ijj'}) = 0$. This formulation facilitates a least squares estimation procedure to be detailed in the next paragraph. Note that (3) holds generally without requiring the

distributions from the K cell types to be known. To ease the exposition of key ideas, we first assume $\boldsymbol{\mu}^{(k)}$'s are known in the ensuing development, and focus on the estimation of $\boldsymbol{\Sigma}^{(k)}$'s. In Section 3.1, we drop this assumption, develop an estimation procedure and derive theory when $\boldsymbol{\mu}^{(k)}$'s are unknown.

Recall $\boldsymbol{z}_i = \boldsymbol{x}_i - \mathbb{E}(\boldsymbol{x}_i)$. Denote $\boldsymbol{H} = (\pi_{ik}^2)_{n\times K}$ and $\boldsymbol{y}_{jj'} = (z_{1j}z_{1j'}, \ldots, z_{nj}z_{nj'}) \in \mathbb{R}^n$. Then, the least squares solution to $\sigma_{jj'}^{(k)}$'s in (3) is given as

$$\hat{\sigma}_{jj'}^{(k)} = [\left(\boldsymbol{H}^\top \boldsymbol{H}\right)^{-1} \boldsymbol{H}^\top \boldsymbol{y}_{jj'}]_k, \quad j, j' \in [p], k \in [K], \tag{4}$$

where $[\boldsymbol{x}]_k$ is the kth entry in $\boldsymbol{x} \in \mathbb{R}^K$, and the cell-type-specific correlation estimate $\hat{\boldsymbol{R}}^{(k)}$ is

$$\hat{R}_{jj'}^{(k)} = \hat{\sigma}_{jj'}^{(k)} / \sqrt{\hat{\sigma}_{jj}^{(k)} \hat{\sigma}_{j'j'}^{(k)}}, \quad j, j' \in [p], k \in [K]. \tag{5}$$

In Section A3.1, we show that $\hat{R}_{jj'}^{(k)}$ is consistent with a $\sqrt{n}$-convergence rate when the profile $x_{ij}^{(k)}$ from each cell type follows a sub-exponential distribution, negative binomial.

2.3 `CSNet`: sparse estimation with large p

In co-expression network analysis, the number of genes p often far exceeds n, the number of bulk RNA-seq samples. Though correlation entries estimated in (5) have a $\sqrt{n}$-convergence rate, the accumulated errors across $O(p^2)$ entries in $\hat{\boldsymbol{R}}^{(k)}$ can be excessive, negatively impacting 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 definition of sparsity in Assumption 2. Sparsity is a commonly employed dimension reduction assumption and is plausible in our data problem,

as gene co-expressions are expected to be sparse when p is large (Zhang and Horvath, 2005).

To estimate sparse cell-type-specific correlation matrices, we consider a thresholding procedure. 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. We propose `CSNet`, a thresholded estimator under the SCAD penalty, defined as $\mathcal{T}_{\lambda_k}(\hat{\boldsymbol{R}}^{(k)})$ with its (j, j')th entry calculated as

$$[\mathcal{T}_{\lambda_k}(\hat{\boldsymbol{R}}^{(k)})]_{jj'} = \text{sign}(\hat{R}^{(k)}_{jj'}) \times \begin{cases} (|\hat{R}^{(k)}_{jj'}| - \lambda_k)_+ & |\hat{R}^{(k)}_{jj'}| \leq 2\lambda_k \\ \lambda_k + \dfrac{a-1}{a-2}(|\hat{R}^{(k)}_{jj'}| - 2\lambda_k) & |\hat{R}^{(k)}_{jj'}| \in (2\lambda_k, a\lambda_k) \\ |\hat{R}^{(k)}_{jj'}| & \text{otherwise,} \end{cases} \tag{6}$$

where λ_k is a tuning parameter and we discuss its selection in the next paragraph. 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). We allow cell-type-specific threshold λ_k's to accommodate varying sparsity among different cell types, though the λ_k's can be selected separately in our tuning procedure (see next paragraph), which is an attractive computational property of our procedure. In Section 3.1, we show the error rates of $\mathcal{T}_{\lambda_k}(\hat{\boldsymbol{R}}^{(k)})$ 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 (PD), though it is asymptotically PD as ensured by Theorems 3.1-3.3; see more discussions in Section 6 on considerations for PD. To ensure the finite sample validity of correlation estimates, we threshold the correlation estimates to be within [-1,1] in our experiments.

In our procedure, the tuning parameters λ_k's are selected using cross validation or, if available, an independent validation data set. Next, we will explain 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 (5). 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 gives a good performance in our numerical experiments.

3 Theory

We derive the non-asymptotic convergence rate of the sparse cell-type-specific correlation estimator `CSNet` and establish variable selection consistency. Our result is first derived assuming mean parameters $\boldsymbol{\mu}^{(k)}$'s are known and then extended to the more challenging case where $\boldsymbol{\mu}^{(k)}$'s are unknown and need to be estimated. The theoretical analysis is challenged by 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 $\boldsymbol{y}_{jj'}$ in (6) is the product of two sub-exponential random variables, which requires a new concentration result (see Lemma A2.2 and its proof in Section A2.1). *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*, when $\boldsymbol{\mu}^{(k)}$'s are unknown, $\boldsymbol{y}_{jj'}$ in (6) is noisy with errors from estimating the mean parameter $\boldsymbol{\mu}^{(k)}$'s; see more discussions in Section 3.1. Next, we state a few regularity conditions.

Assumption 1. *Suppose $x_{ij}^{(k)}$ follows 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 $\boldsymbol{x}_i^{(k)}$ (i.e., the expression profile from the kth cell type) in that each element $x_{ij}^{(k)}$, $j \in [p]$ 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. *Assume* $\boldsymbol{R}^{(k)} \in \mathcal{U}(q, s_p)$ *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,*

$$\mathcal{U}(q, s_p) = \left\{ \boldsymbol{R} : \boldsymbol{R} \succ 0,\ \max_{j'} \sum_{j=1}^{p} |R_{jj'}|^q \leq s_p \right\}. \tag{7}$$

This assumption stipulates that the cell-type-specific correlation matrices are approximately sparse. The definition of approximate sparsity in (7) 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 $s_p = \max_k s_p^{(k)}$.

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

Assumption 3. *Assume there exist constants* $\phi_1, \phi_2 > 0$ *such that 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 matrix $\boldsymbol{H} \in \mathbb{R}^{n\times K}$,

which can be viewed as a fixed design matrix in (3). The bounded eigenvalue condition on $\mathbf{\Phi}_{K\times K}$ is not restrictive as the number of cell types K is treated as fixed. The condition $\min_k \mathbf{\Phi}_{kk} \geq \phi_2$ 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 a underlying multinomial or Dirichlet distribution with fixed parameters, both conditions in Assumption 3 are expected to hold (see proof in Section A4).

Theorem 3.1. *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* $k \in [K]$*,*

$$\left\|\mathcal{T}_\lambda(\hat{\boldsymbol{R}}^{(k)}) - \boldsymbol{R}^{(k)}\right\| \leq M s_p \left(\frac{\log p}{n}\right)^{(1-q)/2} \tag{8}$$

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

Theorem 3.1 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 (8) 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 (8) 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 3.2. *Suppose Assumptions 1 and 3 hold. Let* $\log p = o(n^{1/3})$*,* $\lambda = M\sqrt{\log p/n}$ *for a sufficiently large constant* $M > 0$*, and* $|R_{jj'}^{(k)}| > \rho_{\min}$ *for some* $\rho_{\min} > 0$ *such that* $\sqrt{n}(\rho_{\min} - \lambda) \to \infty$*, then we have* $\mathbb{P}\left\{\Omega(\mathcal{T}_\lambda(\hat{\boldsymbol{R}}^{(k)})) = \Omega(\boldsymbol{R}^{(k)})\right\} \to 1$ *for* $k \in [K]$*.*

Theorem 3.2 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 3.2 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.

3.1 Estimation with unknown means: a two-step procedure

In our previous analysis, $\boldsymbol{\mu}^{(k)}$'s are assumed known. Next, we drop this assumption and present a two-step estimation procedure in the case of unknown $\boldsymbol{\mu}^{(k)}$'s, followed by its theoretical properties. Denoting $\boldsymbol{y}_j = (x_{1j}, \ldots, x_{nj})$ and $\boldsymbol{D} = (\pi_{ik})_{n\times K}$, Step 1 estimates $\boldsymbol{\mu}^{(k)}$ via $\mu_j^{(k)} = [(\boldsymbol{D}^\top \boldsymbol{D})^{-1}\boldsymbol{D}^\top \boldsymbol{y}_j]_k$ for all j, k by noting

$$x_{ij} = \sum_{k=1}^{K} \pi_{ik}\mu_j^{(k)} + z_{ij}, \quad j \in [p], \tag{9}$$

where $\mathbb{E}(z_{ij}) = 0$; in Step 2, we approximate each $\boldsymbol{z}_i$ with $\hat{\boldsymbol{z}}_i = \boldsymbol{x}_i - \sum_{k=1}^{K} \pi_{ik}\hat{\boldsymbol{\mu}}^{(k)}$, where $\hat{\boldsymbol{\mu}}^{(k)}$ is estimated from the first step, and estimate $\boldsymbol{R}$ based on $\hat{\boldsymbol{z}}_1, \ldots, \hat{\boldsymbol{z}}_n$ using the procedure described in Section 2.2. In practice, to ensure $\mu^{(k)}$'s are valid, a nonnegative least squares estimator can be considered for step 1. For simplicity, we consider the least squares estimator in our theoretical analysis and note that it is asymptotically nonnegative.

Theorem 3.3. *Assume all conditions in Theorem 3.1 hold and the eigenvalues of $\boldsymbol{D}^\top\boldsymbol{D}/n$ are lower bounded by a constant $\phi_1' > 0$. Letting $\hat{\boldsymbol{R}}^{(k)}$ be the estimator from the two-step procedure, we have,*

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

with probability at least $1 - C_3 \exp(-C_4 \log p)$, where C_3 and C_4 are some positive constants.

Moreover, under all conditions in Theorem 3.2, we have, for all $k \in [K]$

$$\mathbb{P}\left\{\Omega(\mathcal{T}_\lambda(\hat{\boldsymbol{R}}^{(k)})) = \Omega(\boldsymbol{R}^{(k)})\right\} \to 1. \tag{11}$$

When $\boldsymbol{\mu}^{(k)}$'s are unknown, as opposed to the oracle regression equation $z_{ij}z_{ij'} = \sum_{k=1}^{K} \pi_{ik}^2 \sigma_{jj'}^{(k)} + \epsilon_{ijj'}$, we only have access to the noisy equation $\hat{z}_{ij}\hat{z}_{ij'} = \sum_{k=1}^{K} \pi_{ik}^2 \sigma_{jj'}^{(k)} + e_{ijj'}$, where $e_{ijj'} = \epsilon_{ijj'} + (\hat{z}_{ij}\hat{z}_{ij'} - z_{ij}z_{ij'})$. It is seen that the rate in (10) is the same as that derived in the oracle case (if $\boldsymbol{\mu}^{(k)}$'s were known and did not need to be estimated) in (8), which is a desirable property of our proposed two-step estimation method.

4 Simulation Studies

We investigate the finite sample performance of our proposed method and compare it with two competing methods. Specifically, we consider the proposed sparse cell-type-specific estimator in (6), referred to as `CSNet`, the estimator in (5) without thresholding, referred to as `d-CSNet`, the sparse estimator in Rothman et al. (2009) for bulk samples (i.e., not cell-type-specific), referred to as `Bulk`, and the estimator computed from cell-type-specific expressions estimated by CIBERSORTx high-resolution mode (Newman et al., 2019), referred to as `CIBERSORTx`. Both `CSNet` and `d-CSNet` are computed using the two-step procedure in Section 3.1 with nonnegative least squares to estimate the unknown $\boldsymbol{\mu}^{(k)}$'s. A notable contribution from our analysis is a new and efficient method for simulating from multivariate negative binomial distributions with large and structured covariances (see Algorithm 1), which can be of independent interest.

We simulate n bulk samples of dimension p following $\boldsymbol{x}_i = \sum_{k=1}^{K} \pi_{ik}\boldsymbol{x}_i^{(k)}$ in (1) with $K = 2$ cell types and π_{i1}'s i.i.d from Beta$(2, 1)$. Correspondingly, cell type 1 ($m = 2/3$) is twice as abundant as cell type 2 ($m = 1/3$), where m denotes the average cell type proportions.

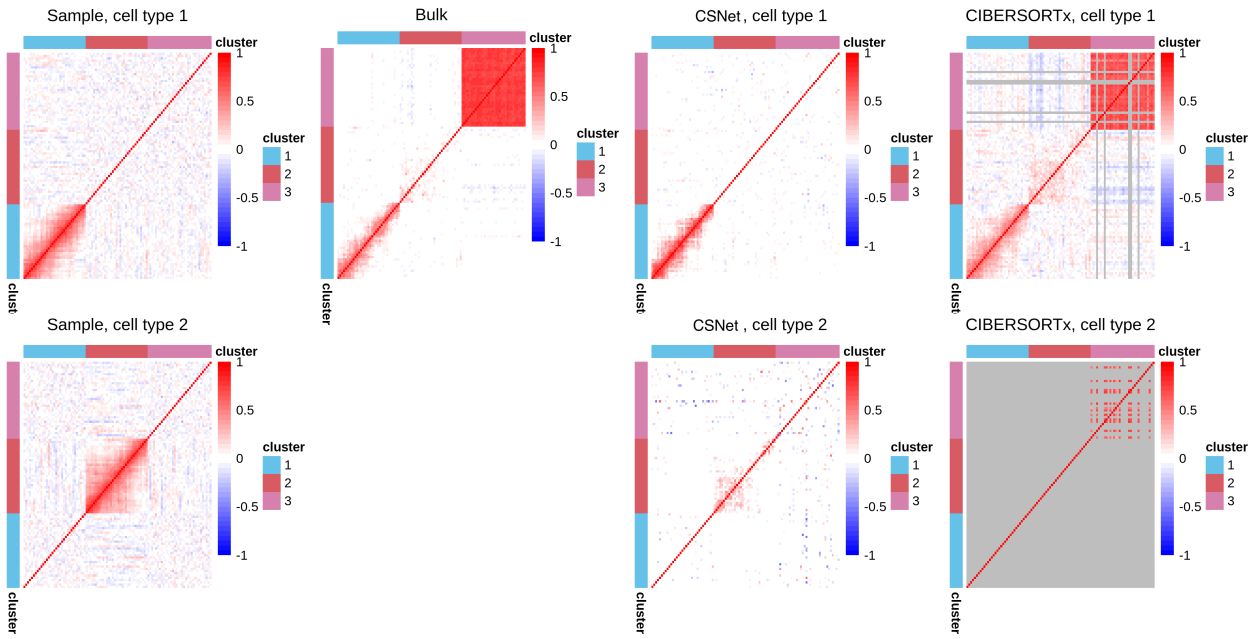


Figure 1: Correlation matrix estimates in one data replicate with $p = 100$ and $n = 150$, under a negative binomial distribution with an AR(1) correlation structure in V_1 or V_2 (See section 4). From left to right: sample correlation of simulated cell-type-specific negative binomial gene expression data, thresholded estimates using bulk samples (`Bulk`), the proposed sparse cell-type-specific estimates (`CSNet`) and the estimates computed from gene expressions estimated by CIBERSORTx high-resolution mode (`CIBERSORTx`). In `CIBERSORTx`, genes with constant cell-type-specific expression estimates across all samples are marked in grey. There is only one correlation matrix estimate from `Bulk`, as `Bulk` estimates are not cell-type-specific. The estimates from `d-CSNet` show similar patterns as `CSNet` plus additional noises, and are omitted due to space limitation.

We simulate $\boldsymbol{x}_i^{(k)}$'s from multivariate negative binomial distributions 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 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

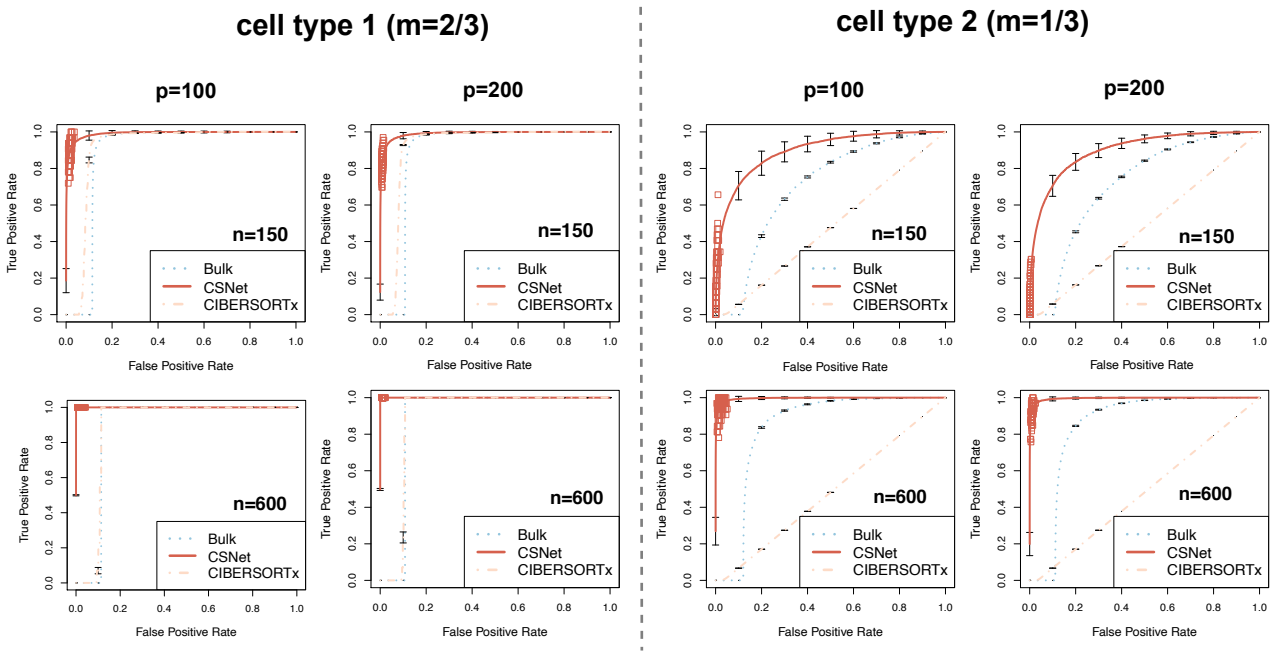


Figure 2: ROC curves with varying sample size n and network size p under a negative binomial distribution with an MA(1) correlation structure in V_1 or V_2 (See section 4). Cell type 1 is more abundant with $m = 2/3$ and cell type 2 is less abundant with $m = 1/3$. The three methods under comparison are thresholded estimates using bulk samples (`Bulk`), the proposed sparse cell-type-specific estimates (`CSNet`) and the estimates computed from gene expressions estimated by CIBERSORTx high-resolution mode (`CIBERSORTx`). The thresholds selected by our proposed procedure in Section 2.3 are marked on the curves (in red squares).

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 not reported for `d-CSNet` and `CIBERSORTx` as these two estimators are not sparse. Table 1 reports the average criteria under the MA(1) model for four methods,

			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	24.64	0.16	**0.97**	27.82	27.20	0.16	**0.69**
			(0.17)	(0.17)	(0.02)	**(0.02)**	(0.14)	(0.13)	(0.02)	**(0.04)**
		`d-CSNet`	11.91	3.82	-	-	21.39	8.81	-	-
			(0.16)	(0.21)	-	-	(0.47)	(0.66)	-	-
		`CSNet`	**2.35**	**0.63**	**0.03**	0.96	**3.06**	**0.78**	**0.01**	0.66
			(0.11)	**(0.04)**	**(0.01)**	(0.02)	**(0.06)**	**(0.03)**	**(0.01)**	(0.05)
		`CIBERSORTx`	20.33	18.97	-	-	15.79	15.10	-	-
			(1.28)	(1.33)	-	-	(1.92)	(1.96)	-	-
	200	`Bulk`	49.06	48.54	0.14	**0.96**	54.21	53.58	0.14	**0.66**
			(0.20)	(0.20)	(0.01)	**(0.02)**	(0.18)	(0.18)	(0.01)	**(0.02)**
		`d-CSNet`	23.90	6.59	-	-	43.02	16.46	-	-
			(0.16)	(0.30)	-	-	(0.67)	(0.96)	-	-
		`CSNet`	**3.58**	**0.68**	**0.02**	0.93	**4.43**	**0.79**	**0.00**	0.63
			(0.10)	**(0.02)**	**(0.00)**	(0.02)	**(0.04)**	**(0.02)**	**(0.00)**	(0.03)
		`CIBERSORTx`	39.03	36.50	-	-	30.33	29.60	-	-
			(1.93)	(2.02)	-	-	(2.68)	(2.71)	-	-
600	100	`Bulk`	25.17	24.77	0.16	**1.00**	27.94	27.29	0.16	0.86
			(0.08)	(0.08)	(0.02)	**(0.00)**	(0.06)	(0.06)	(0.02)	(0.04)
		`d-CSNet`	5.87	1.53	-	-	10.37	3.26	-	-
			(0.06)	(0.07)	-	-	(0.13)	(0.21)	-	-
		`CSNet`	**0.91**	**0.27**	**0.04**	**1.00**	**2.13**	**0.58**	**0.04**	**0.98**
			(0.08)	**(0.04)**	**(0.01)**	**(0.00)**	**(0.13)**	**(0.04)**	**(0.02)**	**(0.02)**
		`CIBERSORTx`	24.22	23.64	-	-	14.99	14.26	-	-
			(0.68)	(0.69)	-	-	(2.24)	(2.29)	-	-
	200	`Bulk`	49.18	48.77	0.14	**1.00**	54.42	53.76	0.14	0.82
			(0.12)	(0.12)	(0.01)	**(0.00)**	(0.08)	(0.08)	(0.01)	(0.04)
		`d-CSNet`	11.79	2.46	-	-	20.81	5.65	-	-
			(0.07)	(0.09)	-	-	(0.17)	(0.31)	-	-
		`CSNet`	**1.47**	**0.32**	**0.02**	**1.00**	**3.24**	**0.64**	**0.02**	**0.97**
			(0.08)	**(0.03)**	**(0.01)**	**(0.00)**	**(0.12)**	**(0.03)**	**(0.01)**	**(0.02)**
		`CIBERSORTx`	47.33	46.52	-	-	29.95	29.22	-	-
			(0.90)	(0.90)	-	-	(3.11)	(3.15)	-	-

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 4). The four methods under comparison are thresholded estimates using bulk samples (`Bulk`), the cell-type-specific estimates without thresholding (`d-CSNet`), the proposed sparse cell-type-specific estimates (`CSNet`), and the estimates computed from gene expressions estimated by CIBERSORTx high-resolution mode (`CIBERSORTx`). 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.

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. From Figure 1, it is clearly seen that both `Bulk` and `CIBERSORTx` give a less accurate view of the true co-expressions. `Bulk` estimates high co-expressions in V_3 while genes V_3 are not co-coexpressed in either cell type. We also present the ROC curves in Figure 2, where we evaluate our proposed procedure for selecting the thresholds. It is seen 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 Figure 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 the bulk samples.

5 Co-expressions of Alzheimer's Disease Risk Genes

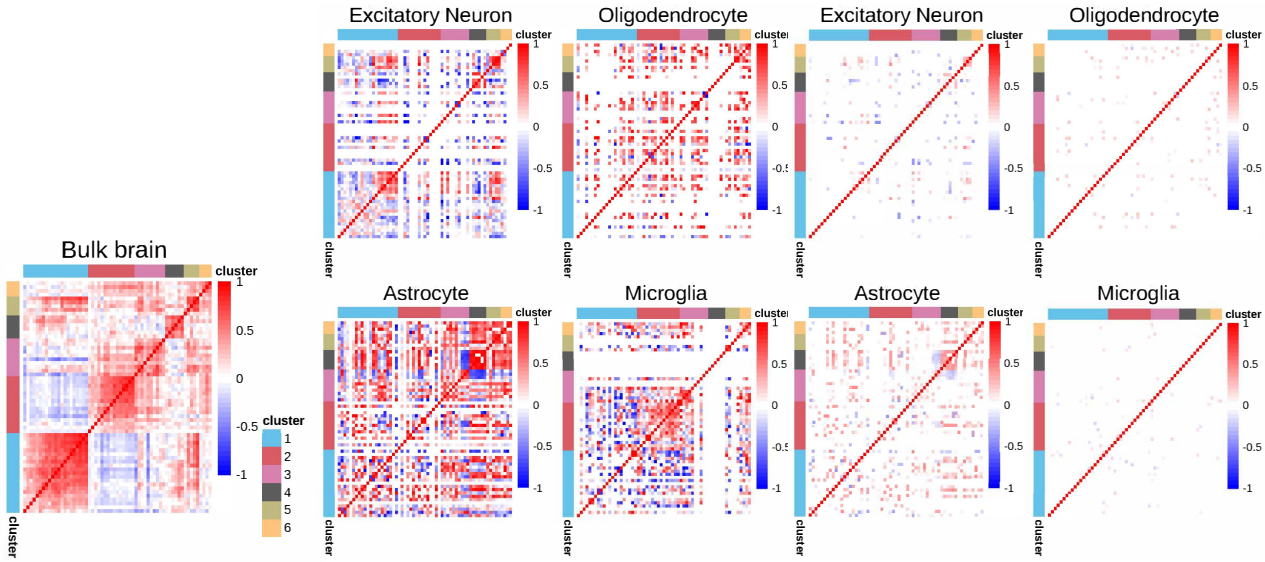
We demonstrate the efficacy of our approach by estimating cell-type-specific co-expression networks using the bulk RNA-seq data from the Religious Orders Study and Rush Memory and Aging Project (ROSMAP) (Bennett et al., 2018), a clinical-pathologic cohort study of Alzheimer's disease (AD). AD, the most common form of dementia, is a neuro-degenerative disorder that causes progressive and irreversible loss of neurons in the brain (Winblad et al., 2016). Genetic factors are known to play an important part in AD, with an estimated heritability of 58–79% for late-onset AD (Sims et al., 2020). To date, effective therapeutic targets remain elusive and understanding the molecular underpinning of the disease is a key step towards better prevention and treatment.

In the ROSMAP study, postmortem brain samples from 541 subjects were collected from

the grey matter of dorsolateral prefrontal cortex (DLPFC), a brain region heavily implicated in AD pathology. We use expression unit FPKM (Trapnell et al., 2010) to quantify gene expressions with no notable batch effects observed in these samples (see Figure S1). The cell type proportions 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 the samples. The obtained cell type proportions for major cell types are then compared to an independent, immunohistochemistry-based quantification of cell type abundance (Patrick et al., 2020), which showed good agreement. Specifically, the Pearson correlations are $\rho = 0.35$ (p-value 0.014) for neurons and microglia, and $\rho = 0.3$ (p-value 0.035) for astrocytes; see Figure S2 for details.

We focused on 61 AD risk genes from Sims et al. (2020) (see gene names in Table S3) and applied the two-step procedure in Section 3.1 with the tuning parameter selected using cross validation as in Section 2.3. Figure 3 shows the estimates from `Bulk`, `d-CSNet` and `CSNet` in four most abundant cell types including excitatory neurons, oligodendrocytes, astrocytes, and microglia. The average proportions for these four cell types are 0.48, 0.21, 0.18 and 0.08, respectively. The genes are ordered into clusters (or modules), obtained by applying the clustering method WGCNA (Langfelder and Horvath, 2008) to bulk samples.

It is seen that some within cluster co-expressions from bulk samples are no longer seen in the cell-type-specific estimates, likely due to the confounding effect of cell type proportions. The `CSNet` estimates in Figure 3(c) show that Cluster 4 (colored in black) were highly co-expressed in astrocytes. This gene cluster includes APOE, a major AD 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 AD (Yamazaki et al., 2019). Additionally, the `d-CSNet` estimates in Figure 3(b) suggest Cluster 2 (colored in red) were co-expressed in microglia, though the signals are



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

Figure 3: Co-expression networks of AD risk genes inferred from the ROSMAP data. From left to right: sample correlation matrix estimated from bulk RNA-seq data (panel (a)), the cell-type-specific estimates without thresholding `d-CSNet` (panel (b)), and the sparse cell-type-specific estimates `CSNet` (panel (c)).

relatively weak and the thresholded estimator for microglia is nearly diagonal. Nevertheless, the identified Cluster 2 is likely microglia specific as it includes TREM2, another major AD risk gene and two genes from the MS4A gene cluster, a key modulator of TREM2 expression in microglia (Deming et al., 2019). Moreover, an enrichment analysis shows that Cluster 2 is enriched for immune related functions (GO enrichment test adjusted p-value < 0.05), while microglia is the primary immune cell type in the central nervous system (Lehnardt, 2010). We have also conducted sensitivity analyses with small noises added to the cell type proportions, and the findings of our analysis remain similar. The estimated co-expressions in gene Clusters 1 and 2 reveal previously unknown cell-type-specific co-expressions among AD risk genes, and may suggest cell-type-specific disease pathology for AD.

6 Discussion

In our analysis, we had assumed that the cell-type proportions π_{ik}'s are given. For example, they can be inferred using existing methods such as CIBERSORTx (Newman et al., 2019). Our empirical investigations showed that our method is not overly sensitive to errors in π_{ik}'s (see sensitive analysis in Section A1.2). 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 PD. However, as the true correlation matrices are PD, it follows from Theorems 3.1-3.3 that the estimated correlation matrices are asymptotically PD. 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 PD) to find the nearest correlation matrix in Frobenius norm. This can be carried out efficiently using existing solvers (e.g., Higham, 2002).

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. However, platform differences and other factors that can affect the measured expression levels need to be carefully modeled for such integrated analyses.

Acknowledgements

We thank the ROSMAP project for sharing the bulk RNA-seq and single nueclues RNA-seq data. ROSMAP resources can be requested at https://www.radc.rush.edu. 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.

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.

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.

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.

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., Chen, M., Szymanski, B., Kuzmin, K., Xie, J., Lee, C., Blanche, T., Neto, E. C., Huang, S.-C., Grabowski, T., et al. (2015), "Identifying robust communities and multi-community nodes by combining top-down and bottom-up approaches to clustering," *Scientific reports*, 5, 1–14.

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.

Haure-Mirande, J.-V., Wang, M., Audrain, M., Fanutza, T., Kim, S. H., Heja, S., Readhead, B., Dudley, J. T., Blitzer, R. D., Schadt, E. E., et al. (2019), "Integrative approach to sporadic Alzheimer's disease: deficiency of TYROBP in cerebral Aβ amyloidosis mouse normalizes clinical phenotype and complement subnetwork molecular pathology without reducing Aβ burden," *Molecular psychiatry*, 24, 431–446.

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

Jew, B., Alvarez, M., Rahmani, E., Miao, Z., Ko, A., Garske, K. M., Sul, J. H., Pietiläinen, K. H., Pajukanta, P., and Halperin, E. (2020a), "Accurate estimation of cell composition in bulk expression through robust integration of single-cell information," *Nature communications*, 11, 1–11.

— (2020b), "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.

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

Lehnardt, S. (2010), "Innate immunity and neuroinflammation in the CNS: The role of microglia in Toll-like receptor-mediated neuronal injury," *Glia*, 58, 253–263.

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.

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.

Patrick, E., Taga, M., Ergun, A., Ng, B., Casazza, W., Cimpean, M., Yung, C., Schneider, J. A., Bennett, D. A., Gaiteri, C., et al. (2020), "Deconvolving the contributions of cell-type heterogeneity on cortical gene expression," *PLOS Computational Biology*, 16, e1008120.

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.

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, 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.

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.

Supplementary Materials for "CSNet: Estimating cell-type-specific gene co-expression networks from bulk gene expression data"

A1 Computational Details and Sensitivity Analysis

A1.1 Simulate multivariate NB samples with large and structured covariances

Consider a negative binomial (NB) random variable

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

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 gives the distribution in (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.

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, and draw $T_{jj'}$ samples $A_{jj'} = \{\eta_h\}_{h\in[T_{jj'}]}$ i.i.d. from Exp(1),

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: $(X_1, \ldots, X_p)$

Before we proceed, we first discuss a useful property of Gamma random variables which entails that if g_1, g_2 are two independent samples from Gamma(T, ϕ), then $g_1 + g_2 \sim$ Gamma$(2T, \phi)$. This property facilitates an effective procedure for generating correlated Gamma random variables. For example, if g_1, g_2, 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 $\mathrm{Var}[X_j] = \sigma^2$ 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 4 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 $0.70 \approx 0.9 \times b$. As a validation, the sample correlation matrices of simulated multivariate negative binomial random variables in Figure 1 indeed show the expected AR(1) structure for genes in V_1 and V_2.

A1.2 Sensitivity Analysis

We conduct a sensitivity analysis to examine the performance of our method when π_{ik}'s are observed with noises. We consider the same simulation setting as in Table 1 with $n = 150$, $p = 100$ and $K = 2$. To add noises to π_{ik}'s, we simulate $\{e_i\}_{i \in [n]}$ independently from $\kappa \times N(0, 0.04)$, where κ is set to $\{0, 0.25, 0.5, 1\}$. Correspondingly, when $\kappa = 0$, π_{ik}'s are observed with no noises. For $i \in [n]$, let $\tilde{\pi}_{i1} = \pi_{i1} + e_i$ and then thresholded to be within $[0, 1]$. The estimation is then carried out following the two-step procedure in Section A3.3, such that both $\hat{\mu}_j^{(k)}$ and $\hat{\sigma}_{jj'}^{(k)}$ in (3) are estimated with noisy cell type proportions $\tilde{\pi}_{ik}$'s.

Table S1 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 misspecified setting with noisy cell type proportions, our method still performs reasonably well.

	$m = 2/3$				$m = 1/3$			
	F-norm	Op.-norm	FPR	TPR	F-norm	Op.-norm	FPR	TPR
$\kappa = 0$	2.35	0.63	0.03	0.96	3.06	0.78	0.01	0.66
	(0.11)	(0.04)	(0.01)	(0.02)	(0.06)	(0.03)	(0.01)	(0.05)
$\kappa = 0.25$	2.38	0.63	0.03	0.95	3.07	0.78	0.01	0.66
	(0.11)	(0.04)	(0.01)	(0.02)	(0.06)	(0.03)	(0.01)	(0.05)
$\kappa = 0.5$	2.72	0.68	0.03	0.89	3.08	0.77	0.00	0.63
	(0.10)	(0.03)	(0.01)	(0.04)	(0.04)	(0.01)	(0.01)	(0.03)
$\kappa = 1$	3.09	0.77	0.00	0.62	3.10	0.77	0.00	0.61
	(0.01)	(0.00)	(0.00)	(0.01)	(0.02)	(0.00)	(0.01)	(0.00)

Table S1: Sensitivity analysis. `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 $[0, 1]$.

A2 Technical Lemmas

We first present a set of technical lemmas. The proofs for Lemmas A2.2 and A2.5 are presented in Sections A2.1 and A2.2, respectively.

Lemma A2.1. *[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 A2.2. *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 $t \geq d_2 \times \gamma_2 \times \|\boldsymbol{a}\|_\infty$,

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

Lemma A2.3. *[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* $i \in [n]$, $\|X_i\|_{\psi_2} \leq \gamma_3$. *Denote the d-th derivative of f by* $\mathcal{D}^d f$. *Then for every polynomial* $f\colon \mathbb{R}^n \to \mathbb{R}$ *of degree D and every* $p \geq 2$,

$$\|f(X) - \mathbb{E}f(X)\|_p \leq C_D \sum_{d=1}^{D} \gamma_3^d \sum_{\mathcal{J}\in\mathcal{P}_d} p^{|\mathcal{J}|/2} \|\mathbb{E}\mathcal{D}^d f(X)\|_{\mathcal{J}},$$

where $\|\cdot\|_{\mathcal{J}}$ *denotes a tensor norm whose definition is relegated to Section A2.1.*

As a consequence, for 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 $\eta_f(t) = \min_{1\leq d\leq D}\min_{\mathcal{J}\in\mathcal{P}_d}\left(\frac{t}{\gamma_3^d\|\mathbb{E}\mathcal{D}^d f(X)\|_{\mathcal{J}}}\right)^{2/|\mathcal{J}|}$ *and* C_D *is some positive constant depending on D.*

Lemma A2.4. *[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 (6) 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*

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

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 A2.5. *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 $t \leq \mu$,*

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

A2.1 Proof of Lemma A2.2

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

$$\|A\|_{\mathcal{J}} = \sup\{\sum_{\boldsymbol{i} \in [n]^d} a_{\boldsymbol{i}} \prod_{j=1}^{k} \boldsymbol{x}^{(j)}_{\boldsymbol{i}_{J_j}} : \|\boldsymbol{x}^{(j)}\|_2 \leq 1, \boldsymbol{x}^{(j)} \in \mathbb{R}^{n^{|J_j|}}, 1 \leq j \leq k\}, \tag{S10}$$

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:

$$F : \mathbb{R}^n \to \mathbb{R}, F(\boldsymbol{u}) = \sum_{i=1}^{n} a_i f(u_i). \tag{S11}$$

Let $\boldsymbol{Z} = [Z_1, \ldots, Z_n]^\top$ be a vector of independent components with $\|Z_i\| \leq \gamma'^{1/4}$, where γ' is some positive constant. Plugging in $F(Z)$, letting $D = 4$, Lemma A2.3 gives

$$\mathbb{P}\left(|F(\boldsymbol{Z}) - \mathbb{E}[F(\boldsymbol{Z})]| \geq t\right) \leq 2\exp(-c_0 \times \eta_F(t)), \tag{S12}$$

and

$$\eta_F(t) = \min_{1\leq d\leq 4} \min_{\mathcal{J}\in\mathcal{P}_d} \left[\frac{t}{\gamma'^{d/4} \times \|\mathbb{E}[\mathcal{D}^d F(\boldsymbol{Z})]\|_{\mathcal{J}}}\right]^{2/|\mathcal{J}|}. \tag{S13}$$

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

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

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

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

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

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

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

$$\gamma'^{d/4} \times \|\mathbb{E}[D^\alpha F(Z)]\|_{\mathcal{J}} \leq \begin{cases} \tilde{c}_d \gamma' \|\boldsymbol{a}\|_2 & \text{if } |\mathcal{J}| = 1 \\ \tilde{c}_d \gamma' \|\boldsymbol{a}\|_{\max} & \text{if } |\mathcal{J}| \geq 2 \end{cases} \tag{S14}$$

Combining (S13) and (S14), we have

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

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

$$\begin{aligned} \eta_F(t) &\geq \min \left[\min_{1 \leq d \leq 4} \left(\frac{t}{\tilde{c}_d \gamma' \|\boldsymbol{a}\|_2} \right)^2, \min_{2 \leq d \leq 4} \frac{t}{\tilde{c}_d \gamma' \|\boldsymbol{a}\|_{\max}} \right] \\ &\geq \min \left[\left(\frac{t}{c_{\max} \gamma' \|\boldsymbol{a}\|_2} \right)^2, \frac{t}{c_{\max} \gamma' \|\boldsymbol{a}\|_{\max}} \right]. \end{aligned} \tag{S15}$$

where $c_{\max} = \max_{d \in \{1,\ldots,4\}} \{\tilde{c}_d\}$. Plugging it in Lemma A2.3, we obtain

$$\mathbb{P}\left\{ \left| \sum_{i=1}^n a_i Z_i^4 - a_i \mathbb{E}(Z_i^4) \right| > t \right\} \leq 2 \exp \left\{ -c_0 \min \left[\left(\frac{t}{c_{\max} \gamma' \|\boldsymbol{a}\|_2} \right)^2, \frac{t}{c_{\max} \gamma' \|\boldsymbol{a}\|_{\max}} \right] \right\}. \tag{S16}$$

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

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

Similarly, we obtain

$$\mathbb{P}\left\{ \left| \sum_{i=1}^n a_i Z_i^4 - a_i \mathbb{E}(Z_i^4) \right| > t \right\} \leq 2 \exp \left\{ -c_0 \times \min \left[\left(\frac{t}{c_{\max} \gamma' \|\boldsymbol{a}\|_2} \right)^2, \left(\frac{t}{c_{\max} \gamma' \|\boldsymbol{a}\|_{\max}} \right)^{1/2} \right] \right\}. \tag{S17}$$

Step 2: Next, we show how a product of sub-exponential random variables can be repre-

sented 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

$$\|Z_i\|_{\psi_2} \leq \max_i \frac{1}{\sqrt{2}}\left(\|X_i\|_{\psi_1}\|Y_i\|_{\psi_1}\right)^{1/4} = \frac{1}{\sqrt{2}}\gamma_2^{1/4}. \tag{S18}$$

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 S16, we obtain

$$\mathbb{P}\left\{\left|\sum_{i=1}^n a_iZ_i^4 - a_i\mathbb{E}(Z_i^4)\right| > t/2\right\} \leq 2\exp\left\{-c_0\min\left[\left(\frac{t}{c_{\max}\gamma_2\|\boldsymbol{a}\|_2/8}\right)^2, \frac{t}{c_{\max}\gamma_2\|\boldsymbol{a}\|_\infty/8}\right]\right\}. \tag{S19}$$

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

$$\mathbb{P}\left\{\left|\sum_{i=1}^n a_iB_i^+ - a_i\mathbb{E}(B_i^+)\right| > t/2\right\} \leq 2\exp\left\{-c_0\min\left[\left(\frac{t}{c_{\max}\gamma_2\|\boldsymbol{a}\|_2/8}\right)^2, \frac{t}{c_{\max}\gamma_2\|\boldsymbol{a}\|_\infty/8}\right]\right\}. \tag{S20}$$

By symmetry, similar derivations give the same bound for B_i^- as S20. Then

$$\begin{aligned}
&\mathbb{P}\left\{\left|\sum_{i=1}^n a_i X_i Y_i - a_i \mathbb{E}(X_i Y_i)\right| > t\right\} \\
&\leq \mathbb{P}\left\{\left|\sum_{i=1}^n a_i B_i^+ - a_i \mathbb{E}(B_i^+)\right| > t/2\right\} + \mathbb{P}\left\{\left|\sum_{i=1}^n a_i B_i^- - a_i \mathbb{E}(B_i^-)\right| > t/2\right\} \\
&\leq 4\exp\left\{-c_0 \min\left[\left(\frac{t}{c_{\max}\gamma_2\|\boldsymbol{a}\|_2/8}\right)^2, \frac{t}{c_{\max}\gamma_2\|\boldsymbol{a}\|_\infty/8}\right]\right\}.
\end{aligned}$$

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

$$\mathbb{P}\left\{\left|\sum_{i=1}^n a_i X_i Y_i - a_i \mathbb{E}(X_i Y_i)\right| > t\right\} \leq 4\exp\left\{-c_0 \min\left[\left(\frac{t}{c_{\max}\gamma_2\|\boldsymbol{a}\|_2/8}\right)^2, \left(\frac{t}{c_{\max}\gamma_2\|\boldsymbol{a}\|_\infty/8}\right)^{1/2}\right]\right\}.$$

Then, setting $d_1 = c_{\min}/2$, $c_3 = 8c_0/c_{\max}^2$, $d_2 = c_{\max}/2$ gives the desired inequalities.

A2.2 Proof of Lemma A2.5

Given $\mu = \sqrt{\frac{1}{n}\mathbb{E}\|X\|_2^2}$, by Lemma A2.2, we have

$$\mathbb{P}\left[\left|\frac{1}{n}\|X\|_2^2 - \mu^2\right| \geq \delta\right] \leq 4\exp\left\{-C_1' \times \min\left[\left(\frac{\sqrt{n}\delta}{\gamma_4}\right)^2, \left(\frac{n\delta}{\gamma_4}\right)^{1/2}\right]\right\}, \tag{S21}$$

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

$$\mathbb{P}\left[\left|\frac{\|X\|_2^2}{n\mu^2} - 1\right| \geq u\right] \leq 4\exp\left\{-C_1' \times \min\left[\left(\frac{\sqrt{n}u\mu^2}{\gamma_4}\right)^2, \left(\frac{nu\mu^2}{\gamma_4}\right)^{1/2}\right]\right\}. \tag{S22}$$

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

$$\begin{aligned}
\mathbb{P}\left[\left|\frac{\|X\|_2}{\sqrt{n\mu^2}} - 1\right| \geq u\right] &\leq \mathbb{P}\left[\left|\frac{\|X\|_2^2}{n\mu^2} - 1\right| \geq \max(u, u^2)\right] \\
&\leq 4\exp\left\{-C_1' \times \min\left[\left(\frac{\sqrt{n}u\mu^2}{\gamma_4}\right)^2, \left(\frac{nu\mu^2}{\gamma_4}\right)^{1/2}\right]\right\}
\end{aligned} \tag{S23}$$

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

A3 Proofs of Main Results

A3.1 Proof of Theorem 3.1

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 $\hat{\sigma}_{jj'}^{(k)}$ defined in (4) satisfies

$$\mathbb{P}\left(\left|\hat{\sigma}_{jj'}^{(k)} - \sigma_{jj'}^{(k)}\right| > t\right) \leq 4K \exp\left[-c_1' \min\left(\frac{nt^2}{\gamma^2}, \sqrt{\frac{nt}{\gamma}}\right)\right], \tag{S24}$$

where c_1' is a positive constant, $\|x_{ij}^{(k)}\|_{\psi_1} < \gamma_0$ and $\gamma = (2K_0\gamma_0)^2$.

Let $\boldsymbol{\beta}_{jj'} = (\sigma_{jj'}^{(1)}, \ldots, \sigma_{jj'}^{(K)})$, $\boldsymbol{y}_{jj'} = \boldsymbol{H}\boldsymbol{\beta}_{jj'} + \boldsymbol{\epsilon}_{jj'}$, $\boldsymbol{\Delta}_{jj'} = \hat{\boldsymbol{\beta}}_{jj'} - \boldsymbol{\beta}_{jj'}$. Under Assumption 3, it holds that

$$\phi_1 \|\boldsymbol{\Delta}_{jj'}\|_2^2 \leq \frac{1}{n}\|\boldsymbol{H}\boldsymbol{\Delta}_{jj'}\|_2^2. \tag{S25}$$

As $\hat{\boldsymbol{\beta}}_{jj'}$ is the minimizer of the loss function, we have $\|\boldsymbol{y}_{jj'} - \boldsymbol{H}\hat{\boldsymbol{\beta}}_{jj'}\|_2^2 \leq \|\boldsymbol{y}_{jj'} - \boldsymbol{H}\boldsymbol{\beta}_{jj'}\|_2^2$. As $\|\boldsymbol{y}_{jj'} - \boldsymbol{H}\hat{\boldsymbol{\beta}}_{jj'}\|_2^2 = \|\boldsymbol{\epsilon}_{jj'} - \boldsymbol{H}\boldsymbol{\Delta}_{jj'}\|_2^2$ and $\|\boldsymbol{y}_{jj'} - \boldsymbol{H}\boldsymbol{\beta}_{jj'}\|_2^2 = \|\boldsymbol{\epsilon}_{jj'}\|_2^2$, it then follows that

$$\|\boldsymbol{H}\boldsymbol{\Delta}_{jj'}\|_2^2 \leq 2\langle \boldsymbol{\epsilon}_{jj'}, \boldsymbol{H}\boldsymbol{\Delta}_{jj'}\rangle. \tag{S26}$$

By Hölder's inequality, we have

$$|\langle \boldsymbol{\epsilon}_{jj'}, \boldsymbol{H}\boldsymbol{\Delta}_{jj'}\rangle| \leq \|\boldsymbol{H}^\top \boldsymbol{\epsilon}_{jj'}\|_\infty \|\boldsymbol{\Delta}_{jj'}\|_1 \leq \sqrt{K}\|\boldsymbol{H}^\top \boldsymbol{\epsilon}_{jj'}\|_\infty \|\boldsymbol{\Delta}_{jj'}\|_2. \tag{S27}$$

Putting (S25)-(S27) together, we arrive at that

$$\|\boldsymbol{\Delta}_{jj'}\|_2 \leq \frac{2\sqrt{K}}{\phi_1} \frac{\|\boldsymbol{H}^\top \boldsymbol{\epsilon}_{jj'}\|_\infty}{n}. \tag{S28}$$

Letting $\boldsymbol{H} = (\boldsymbol{\pi}_1^2, \ldots, \boldsymbol{\pi}_k^2) \in \mathbb{R}^{n\times K}$, then $\|\boldsymbol{H}^\top \boldsymbol{\epsilon}_{jj'}\|_\infty = \max_k |\langle \boldsymbol{\pi}_k^2, \boldsymbol{\epsilon}_{jj'}\rangle| = \max_k |\sum_{i=1}^n \pi_{ik}^2 \epsilon_{ijj'}|$. Therefore, by (S28) and the union sum inequality, we have

$$\mathbb{P}\{\|\boldsymbol{\Delta}_{jj'}\|_2 \geq t\} \leq K\mathbb{P}\left\{\frac{1}{c_2' n} \max_k \left|\langle \boldsymbol{\pi}_k^2, \boldsymbol{\epsilon}_{jj'}\rangle\right| \geq t\right\} \tag{S29}$$

where $c_2' = \dfrac{\phi_1}{2\sqrt{K}}$.

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 $X_i = z_{ij}, Y_i = z_{ij'}, a_i = \dfrac{1}{c_2' n}\pi_{ik}$ for $i \in [n]$ and $\gamma_X = \gamma_Y = 2K_0\gamma_0$ in Lemma A2.2, we have

$$\begin{aligned}\mathbb{P}\{\frac{1}{c_2' n}\left|\langle \boldsymbol{\pi}_k^2, \boldsymbol{\epsilon}_{jj'}\rangle\right| \geq t\} &\leq 4\exp\left\{-c_3' \min\left[\left(\frac{c_2' nt}{\gamma\|\boldsymbol{\pi}_k^2\|_2}\right)^2, \sqrt{\frac{c_2' nt}{\gamma\|\boldsymbol{\pi}_k^2\|_\infty}}\right]\right\} \\ &\leq 4\exp\left[-c_4' \min\left(\frac{nt^2}{\gamma^2}, \sqrt{\frac{nt}{\gamma}}\right)\right]\end{aligned} \tag{S30}$$

for $k \in [K]$ and some positive constant c_4', where the last inequality is implied by $\frac{1}{n}\|\boldsymbol{\pi}_k\|_2^2 = O(1)$ (Assumption 3) and $\gamma = (2K_0\gamma_0)^2$. Note both cases in Lemma A2.2 are considered to give the result in (S30). Hence, by (S29), we have

$$\mathbb{P}\{\|\boldsymbol{\Delta}_{jj'}\|_2 \geq t\} \leq 4K\exp\left[-c_4' \min\left(\frac{nt^2}{\gamma^2}, \sqrt{\frac{nt}{\gamma}}\right)\right].$$

Finally, as $\left|\hat{\sigma}_{jj'}^{(k)} - \sigma_{jj'}^{(k)}\right| \leq \|\boldsymbol{\Delta}_{jj'}\|_2$, letting $c_1' = c_4'$, we achieve the desired result in (S24).

Step 2. In this step, we show that, suppose the variance and covariance estimates satisfy

$$\mathbb{P}\left(|\hat{\sigma}_{jj'} - \sigma_{jj'}| > t\right) \leq c_5' \exp(-c_6' t^{\alpha}) \tag{S31}$$

for $j, j' \in [p]$ and some positive constants c_5', c_6' and α. Then, the corresponding correlation estimate $\hat{R}_{jj'} = \dfrac{\hat{\sigma}_{jj'}}{\sqrt{\hat{\sigma}_{jj}\hat{\sigma}_{jj'}}}$ satisfies $\mathbb{P}(|\hat{R}_{jj'} - R_{jj'}| > t) \leq c_7' \exp\left(-c_8' \times t^{\alpha}\right)$, where $c_8' = c_6'[\min(\sigma_{jj}, \sigma_{j'j'})/8]^{\alpha}$. Note that the superscript (k) is dropped in (S31) 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

$$\mathbb{P}(|\hat{R}_{jj'} - R_{jj'}| > t) \leq \underbrace{\mathbb{P}(|\tilde{R}_{jj'} - R_{jj'}| > t/2)}_{\text{(I)}} + \underbrace{\mathbb{P}(|\hat{R}_{jj'} - \tilde{R}_{jj'}| > t/2)}_{\text{(II)}} \tag{S32}$$

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

$$(I) = \mathbb{P}(|\tilde{R}_{jj'} - R_{jj'}| > t/2) = \mathbb{P}(|\hat{\sigma}_{jj'} - \sigma_{jj'}| > \sqrt{\sigma_{jj}\sigma_{j'j'}} \times t/2). \tag{S33}$$

For term (II) in (S32), consider

$$a = \frac{t}{2(|R_{jj'}| + t)} \tag{S34}$$

and

$$\begin{aligned}
\text{(II)} = \mathbb{P}(|\hat{R}_{jj'} - \tilde{R}_{jj'}| > t/2) &= \mathbb{P}(|\tilde{R}_{jj'}(s_{jj'} - 1)| > t/2) \\
&\leq \mathbb{P}(|\tilde{R}_{jj'}| > \frac{t}{2a}) + \mathbb{P}(|s_{jj'} - 1| > a) \\
&\leq \mathbb{P}(|\tilde{R}_{jj'} - R_{jj'}| > t) + \mathbb{P}(|s_{jj'} - 1| > a).
\end{aligned} \tag{S35}$$

The term $\mathbb{P}(|\tilde{R}_{jj'} - R_{jj'}| > t)$ can again be bounded using (S33). Note that $a < 1$, the

following holds for the second term

$$\begin{aligned}
\mathbb{P}(|s_{jj'}-1|>a) &= \mathbb{P}(s_{jj'}>a+1)+\mathbb{P}(s_{jj'}<1-a) \\
&\le \mathbb{P}\left\{\frac{1}{s_{jj'}^2}<\frac{1}{(a+1)^2}\right\}+\mathbb{P}\left\{\frac{1}{s_{jj'}^2}>\frac{1}{(1-a)^2}\right\} \\
&\le \mathbb{P}\left\{\hat{\sigma}_{jj}<(1-\frac{a}{a+1})\times\sigma_{jj}\right\}+\mathbb{P}\left\{\hat{\sigma}_{j'j'}<(1-\frac{a}{a+1})\times\sigma_{j'j'}\right\}+ \\
&\quad \mathbb{P}\left\{\hat{\sigma}_{jj}>(1+\frac{a}{1-a})\times\sigma_{jj}\right\}+\mathbb{P}\left\{\hat{\sigma}_{j'j'}>(1+\frac{a}{1-a})\times\sigma_{j'j'}\right\} \\
&\le \mathbb{P}(|\hat{\sigma}_{jj}-\sigma_{jj}|>\frac{a}{a+1}\sigma_{jj})+\mathbb{P}(|\hat{\sigma}_{j'j'}-\sigma_{j'j'}|>\frac{a}{a+1}\sigma_{j'j'}).
\end{aligned} \tag{S36}$$

Finally, by (S34) 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

$$\begin{aligned}
&\mathbb{P}(|\hat{R}_{jj'}-R_{jj'}|>t)\le 2\mathbb{P}(|\tilde{R}_{jj'}-R_{jj'}|>t/2)+\mathbb{P}(|\hat{\sigma}_{jj}-\sigma_{jj}|>\frac{t}{8}\sigma_{jj})+\mathbb{P}(|\hat{\sigma}_{j'j'}-\sigma_{j'j'}|>\frac{t}{8}\sigma_{j'j'}) \\
&\le 2c_5'\exp\left[-c_6'\left(\frac{\sqrt{\sigma_{jj}\sigma_{j'j'}}}{2}\right)^{\alpha}t^{\alpha}\right]+c_5'\exp(-c_6'\left(\frac{\sigma_{jj}}{8}\right)^{\alpha}t^{\alpha})+c_5'\exp(-c_6'\left(\frac{\sigma_{j'j'}}{8}\right)^{\alpha}t^{\alpha}) \\
&\le c_7'\exp\left[-c_8't^{\alpha}\right],
\end{aligned}$$

where $c_7'=4c_5', c_8'=c_6'[\min(\sigma_{jj},\sigma_{j'j'})/8]^{\alpha}$. Thus, we have finished Step 2.

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

$$\mathbb{P}\{\max_{j,j'}|\hat{R}_{jj'}^{(k)}-R_{jj'}^{(k)}|>t\}\le c_9'p^2\exp\left[-c_{10}'\min\left(\frac{nt^2}{\gamma^2},\frac{\sqrt{nt}}{\gamma^{1/2}}\right)\right], \tag{S37}$$

where c_9' and c_{10}' 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, (S37) can be written as

$$\max_{j,j'}|\hat{R}_{jj'}^{(k)}-R_{jj'}^{(k)}|\le\sqrt{\frac{\log p}{n}}\times M, \tag{S38}$$

holds with probability at least $1-c_9'exp(-c_{11}'\log p)$, where $c_{11}'=\frac{c_{10}'}{\gamma^2}M^2-2$. Combined with

Lemma A2.4, we have

$$\|\mathcal{T}_\lambda(\hat{\boldsymbol{R}}^{(k)}) - \boldsymbol{R}^{(k)}\| \leq s_p \left(\frac{\log p}{n}\right)^{(1-q)/2} M \tag{S39}$$

with probability at least $1 - c'_{12}\exp(-c'_{13}\log p)$ for some positive constants c'_{12} and c'_{13}.

A3.2 Proof of Theorem 3.2

As it holds that

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

we have by (S39) that

$$\text{(I)} = \mathbb{P}\left(\sum_{j,j'}\mathbf{1}\left\{\mathcal{T}_\lambda(\hat{R}^{(k)}_{jj'}) \neq 0\right\}\mathbf{1}\left\{R^{(k)}_{jj'} = 0\right\} > 0\right) \leq \tilde{C}_1\exp(-\tilde{C}_2\log p) \tag{S40}$$

for some positive constants $\tilde{C}_1$ and $\tilde{C}_2$.

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

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

Combined with (S39), we obtain

$$\text{(II)} = \mathbb{P}\left(\sum_{i,j}\mathbf{1}\left\{\mathcal{T}_\lambda(\hat{R}^{(k)}_{jj'})R^{(k)}_{jj'} < 0\right\} > 0\right) \leq \tilde{C}_3 p^2\exp\left(-\frac{\tilde{C}_4}{\gamma^2} \times n(\rho_{\min} - \lambda)^2\right), \tag{S41}$$

for some positive constants $\tilde{C}_3$ and $\tilde{C}_4$, which tends to 0 as $\sqrt{n}(\rho_{\min} - \lambda) \to \infty$. Finally, as $1 - \mathbb{P}\left\{\Omega(\tilde{\boldsymbol{R}}^{(k)}) = \Omega(\boldsymbol{R}^{(k)})\right\} \leq \text{(I)} + \text{(II)}$, we obtain $\mathbb{P}\left\{\Omega(\tilde{\boldsymbol{R}}^{(k)}) = \Omega(\boldsymbol{R}^{(k)})\right\} \to 1$.

A3.3 Proof of Theorem 3.3

As defined in Section 3.1, when $\boldsymbol{\mu}^{(k)}$'s are unknown, as opposed to the oracle regression equation $\boldsymbol{y}_{jj'} = \boldsymbol{H}\boldsymbol{\beta}_{jj'} + \boldsymbol{\epsilon}_{jj'}$, we only have access to the noisy equation $\hat{\boldsymbol{y}}_{jj'} = \boldsymbol{H}\boldsymbol{\beta}_{jj'} + \hat{\boldsymbol{\epsilon}}_{jj'}$, where $\hat{\boldsymbol{\epsilon}}_{jj'} = \boldsymbol{\epsilon}_{jj'} + \hat{\boldsymbol{y}}_{jj'} - \boldsymbol{y}_{jj'}$.

In the two-step procedure, we estimate $\boldsymbol{\beta}_{jj'}$ with $\tilde{\boldsymbol{\beta}}_{jj'} = (\boldsymbol{H}^\top \boldsymbol{H})^{-1}\boldsymbol{H}^\top \hat{\boldsymbol{y}}_{jj'}$, which satisfies

$$\|\hat{\boldsymbol{y}}_{jj'} - \boldsymbol{H}\tilde{\boldsymbol{\beta}}_{jj'}\|_2^2 \leq \|\hat{\boldsymbol{y}}_{jj'} - \boldsymbol{H}\boldsymbol{\beta}_{jj'}\|_2^2 . \tag{S42}$$

Recall that $\boldsymbol{\epsilon}_{jj'} = \boldsymbol{y}_{jj'} - \boldsymbol{H}\boldsymbol{\beta}_{jj'}$. Letting $\boldsymbol{d}_{jj'} = \hat{\boldsymbol{y}}_{jj'} - \boldsymbol{y}_{jj'}$ and $\tilde{\boldsymbol{\Delta}}_{jj'} = \tilde{\boldsymbol{\beta}}_{jj'} - \boldsymbol{\beta}_{jj'}$, we have $\hat{\boldsymbol{y}}_{jj'} - \boldsymbol{H}\tilde{\boldsymbol{\beta}}_{jj'} = \boldsymbol{\epsilon}_{jj'} + \boldsymbol{d}_{jj'} - \boldsymbol{H}\tilde{\boldsymbol{\Delta}}_{jj'}$ and $\hat{\boldsymbol{y}}_{jj'} - \boldsymbol{H}\boldsymbol{\beta}_{jj'} = \boldsymbol{\epsilon}_{jj'} + \boldsymbol{d}_{jj'}$. There, (S42) gives that

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

This is in parallel with (S26) in the proof of Theorem 3.1. Then, we apply similar arguments as (S25) and (S27) and obtain

$$\begin{aligned} \|\tilde{\boldsymbol{\Delta}}_{jj'}\|_2 &\leq \frac{2\sqrt{K}}{\phi_1}\frac{1}{n}\|\boldsymbol{H}^\top(\boldsymbol{\epsilon}_{jj'} + \boldsymbol{d}_{jj'})\|_\infty \\ &\leq \frac{2\sqrt{K}}{\phi_1}\frac{1}{n}\Big(\underbrace{\|\boldsymbol{H}^\top\boldsymbol{\epsilon}_{jj'}\|_\infty}_{\text{(I)}} + \underbrace{\|\boldsymbol{H}^\top\boldsymbol{d}_{jj'}\|_\infty}_{\text{(II)}}\Big). \end{aligned} \tag{S43}$$

The first term (I) can be bounded using the same argument as (S29)-(S30). Therefore, we focus on the second term (II).

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:

$$|\langle \boldsymbol{\pi}_k^2, \boldsymbol{\eta}_j \boldsymbol{\eta}_{j'} \rangle| \leq \frac{1}{2}\|\boldsymbol{\pi}_k^2 \circ \boldsymbol{\eta}_j\|_2^2 + \frac{1}{2}\|\boldsymbol{\eta}_{j'}\|_2^2 \leq \frac{1}{2}\|\boldsymbol{\eta}_j\|_2^2 + \frac{1}{2}\|\boldsymbol{\eta}_{j'}\|_2^2 \tag{S44}$$

$$|\langle \boldsymbol{\pi}_k^2, \boldsymbol{z}_j \boldsymbol{\eta}_{j'} \rangle| \leq \|\boldsymbol{\pi}_k^2 \circ \boldsymbol{z}_j\|_2 \|\boldsymbol{\eta}_{j'}\|_2 \leq \|\boldsymbol{z}_j\|_2 \|\boldsymbol{\eta}_{j'}\|_2. \tag{S45}$$

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 the least squares estimate of $\boldsymbol{\mu}_j$. Assuming the eigenvalues of $\frac{1}{n}\boldsymbol{D}^\top \boldsymbol{D}$ are lower bounded by a constant $\phi_1' > 0$, and using similar arguments as in (S25)-(S27), we have

$$\|\boldsymbol{\eta}_j\|_2^2 \leq 2\langle \boldsymbol{D}(\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j), \boldsymbol{z}_j \rangle, \quad \|\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j\|_2 \leq \frac{2\sqrt{K}}{\phi_1'} \frac{1}{n} \left\| \boldsymbol{D}^\top \boldsymbol{z}_j \right\|_\infty. \tag{S46}$$

By Hölder's inequality, we can further combine the equations in (S46) and get

$$\|\boldsymbol{\eta}_j\|_2^2 \leq \frac{4K}{\phi_1'} \frac{1}{n} \left\| \boldsymbol{D}^\top \boldsymbol{z}_j \right\|_\infty^2. \tag{S47}$$

Next, we rewrite term (II) as

$$\begin{aligned}
\text{(II)} &= \frac{2\sqrt{K}}{\phi_1} \frac{1}{n} \|\boldsymbol{D}^\top \boldsymbol{d}_{jj'}\|_\infty \\
&\leq c_1'' \frac{1}{n^2} \left(\left\| \boldsymbol{D}^\top \boldsymbol{z}_j \right\|_\infty^2 + \left\| \boldsymbol{D}^\top \boldsymbol{z}_{j'} \right\|_\infty^2 \right) + c_2'' \frac{1}{n^{3/2}} \left(\left\| \boldsymbol{D}^\top \boldsymbol{z}_j \right\|_\infty \|\boldsymbol{z}_{j'}\|_2 + \left\| \boldsymbol{D}^\top \boldsymbol{z}_{j'} \right\|_\infty \|\boldsymbol{z}_j\|_2 \right).
\end{aligned} \tag{S48}$$

for some positive constants c_1'' and c_2''.

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 $\|z_{ij}\| \leq 2K\gamma_0$. Therefore, let $a_i = \pi_{ik}, X_i = z_{ij}$ for $i \in [n]$ and $\gamma = (2K_0\gamma_0)^2$, Lemma A2.1

implies

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n}\pi_{ik}z_{ij}\right|\geq t\right)\leq 2\exp\left\{-c_3''\min\left[\left(\frac{nt}{\gamma^{1/2}\|\boldsymbol{\pi}_k\|_2}\right)^2,\frac{nt}{\gamma^{1/2}\|\boldsymbol{\pi}_k\|_\infty}\right]\right\}$$
$$\leq 2\exp\left[-c_4''\min\left(\frac{nt^2}{\gamma^{1/2}},\frac{nt}{\gamma^{1/2}}\right)\right]$$

for $k\in[K]$ and some positive constants c_3'',c_4'', where the last inequality is implied by Assumption 3 (i.e., $\frac{1}{n}\sum_{i=1}^n\pi_{ik}^2>\frac{1}{n}\sum_{i=1}^n\pi_{ik}^4>\phi_2$). Then, we have

$$\mathbb{P}\left(\left\|\boldsymbol{D}^\top\boldsymbol{z}_j\right\|_\infty>t\right)\leq 2K\exp\left[-c_4''\min\left(\frac{nt^2}{\gamma^2},\frac{nt}{\gamma}\right)\right]. \tag{S49}$$

By Lemma A2.5, we have the following result for $\frac{1}{\sqrt{n}}\|\boldsymbol{z}_j\|_2$:

$$\mathbb{P}\left[\left|\frac{1}{\sqrt{n}}\|\boldsymbol{z}_j\|_2-m_j\right|\geq t\right]\leq 4\exp\left[-c_5''\min\left(\frac{m_j^2}{\gamma^2}nt^2,\sqrt{\frac{m_j}{\gamma}nt}\right)\right], \tag{S50}$$

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_5'' is a positive constant.

Putting (S30) for term (I) and (S49), (S50) for term (II) together, we have

$$\begin{aligned}\mathbb{P}\{\|\tilde{\boldsymbol{\Delta}}\|_2\geq t\}&\leq\mathbb{P}\{(\mathrm{I})>t/2\}+\mathbb{P}\{(\mathrm{II})>t/2\}\\&\leq c_6''\exp\left[-c_7''\min(nt^2,\sqrt{nt},nt,n\sqrt{t})\right],\end{aligned} \tag{S51}$$

for some positive constants c_6'' and c_7''.

In comparison to Section A3.1, this result is in parallel to step 1, and it can be shown that step 2 follows with

$$\mathbb{P}\{\|\hat{R}_{jj'}-R_{jj'}\|\geq t\}\leq c_8''\exp\left[-c_9''\min(nt^2,\sqrt{nt},nt,n\sqrt{t})\right] \tag{S52}$$

for some positive constants c_8'' and c_9''. Finally, assuming $\log p=o(n^{1/3})$ and setting $\lambda=\sqrt{\frac{\log p}{n}}\times M'$ will again lead to the desired rate as (S39). To show that the result for the

selection consistency also holds as if the truth $\boldsymbol{\mu}_j$'s were known, we may plug (S52) into (S40) and (S41) with $\lambda = \sqrt{\frac{\log p}{n}} \times M'$. Detailed steps in the proofs are very similar and thus omitted here.

A4 Extended discussions on Assumptions 3

Recall

$$\frac{1}{n}\boldsymbol{H}^\top \boldsymbol{H} = \frac{1}{n}\begin{pmatrix} \sum_{i=1}^n \pi_{i1}^4 & \sum_{i=1}^n \pi_{i1}^2\pi_{i2}^2 & \cdots & \sum_{i=1}^n \pi_{i1}^2\pi_{iK}^2 \\ \sum_{i=1}^n \pi_{i1}^2\pi_{i2}^2 & \sum_{i=1}^n \pi_{i2}^4 & \cdots & \sum_{i=1}^n \pi_{i2}^2\pi_{iK}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n \pi_{iK}^2\pi_{i1}^2 & \sum_{i=1}^n \pi_{iK}^2\pi_{i2}^2 & \cdots & \sum_{i=1}^n \pi_{iK}^4 \end{pmatrix}. \tag{S53}$$

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 (S53). Specifically, the smallest eigenvalue of

$$\mathbb{E}\left(\frac{1}{n}\boldsymbol{H}^\top \boldsymbol{H}\right) = \text{Diag}(\boldsymbol{\theta}) \tag{S54}$$

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

Next, consider a more challenging model $(\pi_{i1}, \ldots, \pi_{iK}) \sim \text{Dirichlet}(\boldsymbol{\alpha})$, 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,

$$\mathbb{E}\left(\frac{1}{n}\boldsymbol{H}^\top \boldsymbol{H}\right) = \frac{(\alpha_0 - 1)!}{(\alpha_0 + 3)!}\left[\boldsymbol{\gamma}\boldsymbol{\gamma}^\top + \text{diag}(\boldsymbol{\Delta})\right], \tag{S55}$$

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 $\mathbb{E}(\frac{1}{n}\boldsymbol{H}^\top \boldsymbol{H})$ is lower bounded by $\frac{(\alpha_0 - 1)!}{(\alpha_0 + 3)!}\min_k\{\Delta_k\}$, which is a positive constant.

The above derivation demonstrates that under reasonable probabilistic models, the eigenvalue condition in Assumption 3 is satisfied at the population level. This bounded norm

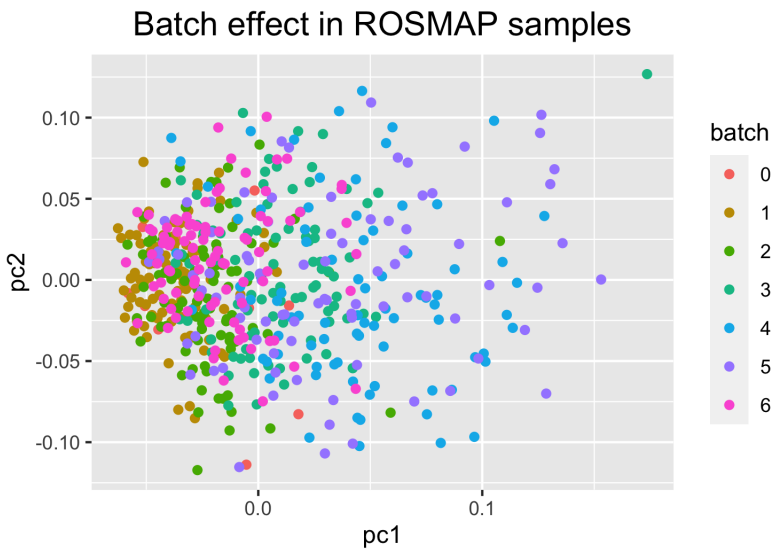


Figure S1: Top two PCs from the principal component analysis of the 541 ROSMAP samples using protein coding genes with average FPKM> 0.25, with points colored by processing batches.

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 $\mathbb{E}[\pi_{ik}^2] > C_k^2$.

A5 Additional Numerical Results

			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	24.67	0.32	**0.75**	30.83	27.23	0.37	0.45
			(0.21)	(0.17)	(0.05)	**(0.06)**	(0.22)	(0.13)	(0.04)	(0.07)
		`d-CSNet`	11.82	4.78	-	-	21.23	9.54	-	-
			(0.28)	(0.63)	-	-	(0.68)	(1.03)	-	-
		`CSNet`	**5.39**	**3.58**	**0.16**	0.72	**9.16**	**6.91**	**0.12**	**0.52**
			(0.59)	**(0.73)**	**(0.04)**	(0.07)	**(0.69)**	**(0.86)**	**(0.03)**	**(0.08)**
		`CIBERSORTx`	20.75	18.96	-	-	19.55	15.10	-	-
			(1.28)	(1.34)	-	-	(1.55)	(1.96)	-	-
	200	`Bulk`	49.91	48.58	0.26	**0.48**	57.68	53.62	0.29	0.28
			(0.24)	(0.21)	(0.03)	**(0.04)**	(0.22)	(0.18)	(0.03)	(0.04)
		`d-CSNet`	23.76	8.15	-	-	42.76	17.68	-	-
			(0.30)	(0.74)	-	-	(0.90)	(1.71)	-	-
		`CSNet`	**9.19**	**5.82**	**0.11**	0.46	**14.68**	**9.57**	**0.07**	**0.30**
			(0.57)	**(0.65)**	**(0.02)**	(0.04)	**(0.67)**	**(0.72)**	**(0.02)**	**(0.05)**
		`CIBERSORTx`	39.54	36.51	-	-	34.87	29.60	-	-
			(1.90)	(2.00)	-	-	(2.33)	(2.71)	-	-
600	100	`Bulk`	25.54	24.78	0.37	**0.90**	30.69	27.30	0.40	0.67
			(0.10)	(0.08)	(0.05)	**(0.04)**	(0.14)	(0.06)	(0.05)	(0.05)
		`d-CSNet`	5.82	2.17	-	-	10.28	3.89	-	-
			(0.12)	(0.29)	-	-	(0.22)	(0.43)	-	-
		`CSNet`	**2.73**	**1.63**	**0.13**	0.86	**4.93**	**3.10**	**0.15**	**0.75**
			(0.34)	**(0.42)**	**(0.07)**	(0.05)	**(0.43)**	**(0.74)**	**(0.06)**	**(0.07)**
		`CIBERSORTx`	24.57	23.64	-	-	19.02	14.42	-	-
			(0.70)	(0.71)	-	-	(1.81)	(2.26)	-	-
	200	`Bulk`	49.66	48.79	0.30	**0.63**	57.71	53.77	0.32	0.44
			(0.13)	(0.12)	(0.03)	**(0.04)**	(0.13)	(0.08)	(0.03)	(0.04)
		`d-CSNet`	11.73	3.54	-	-	20.71	6.78	-	-
			(0.14)	(0.37)	-	-	(0.28)	(0.67)	-	-
		`CSNet`	**4.79**	**2.82**	**0.10**	0.57	**8.35**	**5.02**	**0.11**	**0.49**
			(0.32)	**(0.49)**	**(0.04)**	(0.05)	**(0.52)**	**(0.75)**	**(0.04)**	**(0.05)**
		`CIBERSORTx`	47.81	46.59	-	-	34.41	29.06	-	-
			(0.92)	(0.94)	-	-	(2.65)	(3.10)	-	-

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 4). The four methods under comparison are thresholded estimates using bulk samples (`Bulk`), the cell-type-specific estimates without thresholding (`d-CSNet`), the proposed sparse cell-type-specific estimates (`CSNet`), and the estimates computed from gene expressions estimated by CIBERSORTx high-resolution mode (`CIBERSORTx`). 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.

Cluster	Gene
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, CLU, SPPL2A, ECHDC3, TP53INP1, CD2AP, ZNF423, FERMT2, IQCK, RORA
4	APOE, MS4A4E, ZCWPW1, ABCA7, BIN1, SHARPIN
5	HESX1, APH1B, ADAM10, PICALM, ZNF655
6	CR1, HBEGF, ADAMTS1, ADAMTS4

Table S3: The list of Alzheimer's Disease risk genes used in Section 5. Gene names are displayed by cluster and in the same order as they appear in Figure 3. This gene set is obtained from Table 1 in Sims et al. (2020). Our analysis included only the sequenced genes with a FPKM greater than 0.1 in at least 50 ROSMAP samples.

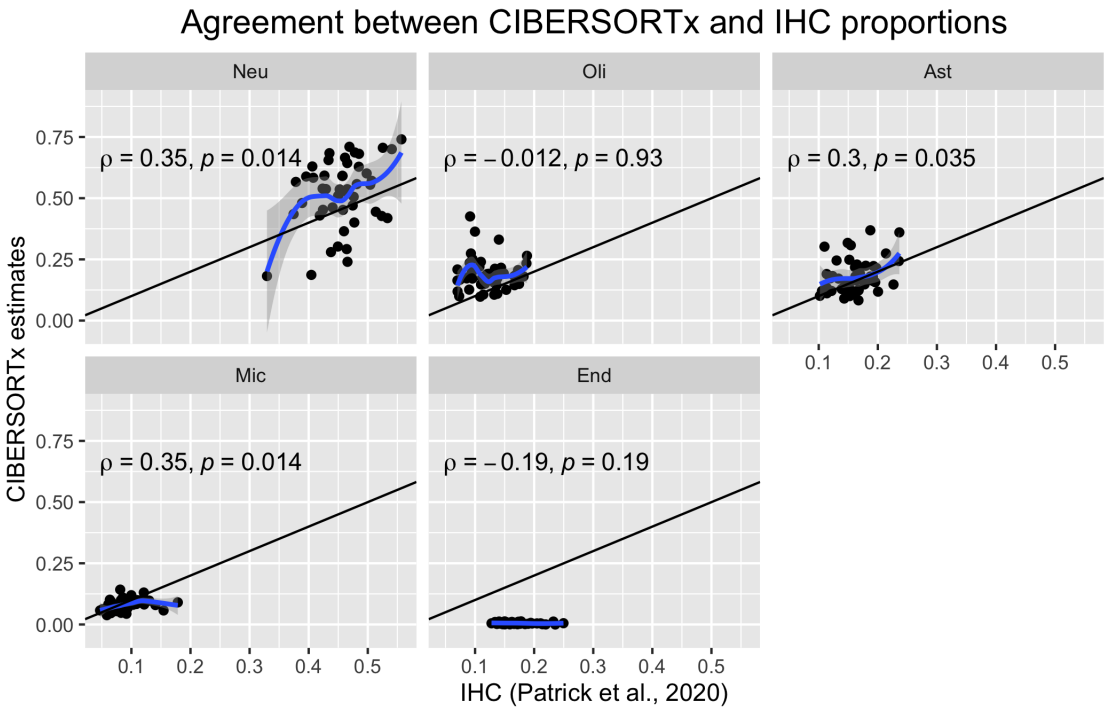


Figure S2: Agreement between the CIBERSORTx estimates and IHC proportion measurements from Patrick et al. (2020). Pearson correlations and associated p-values are provided for each cell type.