Provable Guarantees for Estimating Covariances between Latent Variables with Application to Precision Matrix Estimation

Haichi Long
Computing and Information Systems
The University of Melbourne

Qifan Song
Statistics
Purdue University

Jean Honorio
Computing and Information Systems
The University of Melbourne, and
ARC OPTIMA

Abstract

In many scientific fields, key variables of interest are latent—either because they cannot be measured directly or because doing so is prohibitively expensive. As a result, researchers often rely on high-dimensional surrogate observations and must infer relationships among the unobserved quantities. In this work, we address a fundamental challenge: How can one estimate the covariance between variables that are not directly observable? We consider a model where each latent variable elicits high-dimensional observable covariates. Under our model, we propose a method that estimates several spiked covariances from the observed variables and then reconstructs the covariance matrix among the latent variables. Our estimator achieves quadratic-time complexity with respect to the number of latent variables and only requires the sample size to be logarithmic in the number of latent variables. As an immediate application, our procedure can be leveraged to recover the conditional independence structure among the latent variables, providing interpretable insights. Extensive synthetic experiments validate our theory, demonstrating accurate estimation in practice.

1 INTRODUCTION

In statistical machine learning, covariance and precision matrix estimation plays a pivotal role in uncovering relationships among variables, particularly for high-dimensional data. The precision matrix, the inverse of the covariance matrix, offers insights into the conditional dependencies among variables, helping researchers to isolate and understand direct interactions without confounding effects (Fan et al., 2016). Such a feature is critically useful in various applications, such as identifying key factors in biological studies, where understanding conditional independence is essential. As datasets grow larger and more complex, covariance and precision matrix estimation has become increasingly vital for exploratory research with high-dimensional data.

Unfortunately, in many scientific fields, key variables of interest are latent. In neuroscience and social sciences for instance, it is sometimes impossible to directly measure complex cognitive processes such as inhibitory control, cue reactivity or self-awareness. Instead, researchers carefully design tasks to collect some surrogate data, such as recording brain activity through functional magnetic resonance imaging.[1] We argue that these high-dimensional surrogate observations can be used to infer relationships among the unobserved quantities. In this paper, we address a fundamental challenge of estimating the covariance and precision matrix between variables that are not directly observable.

However, there is a notable gap in the current literature, as existing theoretical work focuses on fully observed data for estimating covariance matrices (Bickel and Levina, 2008; Lam and Fan, 2009) and precision matrices (Ravikumar et al., 2011), leaving the case of latent variables largely unexplored. This gap highlights the need for novel methods with provable guarantees, i.e., methods that can accurately estimate covariance and precision matrices of latent variables. Moreover, the high-dimensional regime poses addi-

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

[1] Please see Section 5.3 for some results on understanding cocaine addiction.

tional challenges in terms of the required computation (Banerjee et al., 2008; Guillot et al., 2012; Hsieh et al., 2013, 2014; Johnson et al., 2012; Kambadur and Lozano, 2013; Treister and Turek, 2014; Wang et al., 2013) and number of samples (Bickel and Levina, 2008; Lam and Fan, 2009; Ravikumar et al., 2011).

Contributions. Our contributions are fivefold. First, we address the challenge of estimating the covariance matrix of latent variables. We propose a model in which latent variables are correlated and where each latent variable elicits high-dimensional observable covariates. Second, we propose an empirical estimator by leveraging the observed data to estimate the underlying covariance matrix of latent variables. In order to do this, we estimate several spiked covariance matrices from the observed variables, which enables us to use sparse principal component analysis. Third, we provide theoretical guarantees for our latent covariance estimator. For n samples, p latent variables, and dimension of the observed variables q, the sample complexity of our estimator is $n = O(\log p + \log q)$. The logarithmic dependence makes our proposed estimator suitable for high-dimensions. Fourth, we go one step further by considering latent-variable precision matrix estimation as one of our applications. Finally, extensive synthetic experiments validate our theory, demonstrating the practical effectiveness of our method.

2 NOTATION

In this section, we introduce the notations that will be used throughout the paper. We use lowercase bold letters for vectors (e.g. $\mathbf{a}$) and uppercase bold letters for matrices (e.g., $\mathbf{A}$), and subscripts of non-bold letters to denote respective entries of a vector or matrix (e.g., a_i and A_{ij}). We also index vectors or matrices and use subscripts, but in this case we retain the bold letter. That is, $\mathbf{a}_1, \mathbf{a}_2, \ldots$ are vectors and $\mathbf{A}_1, \mathbf{A}_2, \ldots$ are matrices. For a set S, we denote its complement by S^c. Given a matrix $\mathbf{A}$ and sets S, T, we denote $\mathbf{A}_{ST}$ as the submatrix of $\mathbf{A}$ containing the rows in S and columns in T. For a matrix $\mathbf{A}$, its trace is denoted by $\mathrm{tr}(\mathbf{A})$ and its determinant is denoted by $\det(\mathbf{A})$. $\mathbf{A} \succ \mathbf{0}$ (or $\mathbf{A} \succeq \mathbf{0}$) denote matrix $\mathbf{A}$ is positive definite (or positive semi-definite). We denote the inner product of two matrices $\mathbf{A}$ and $\mathbf{B}$ as $\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{ij} = a_{ij} b_{ij}$. For a vector $\mathbf{a}$, $\|\mathbf{a}\|_1$, $\|\mathbf{a}\|_\infty$ and $\|\mathbf{a}\|_2$ denote the ℓ_1 norm, ℓ_∞ norm and Euclidean norm respectively. For a matrix $\mathbf{A}$, $\|\mathbf{A}\|_1$, $\|\mathbf{A}\|_\infty$ and $\|\mathbf{A}\|_{\mathcal{F}}$ denote the entrywise ℓ_1 norm, entrywise ℓ_∞ norm and Frobenius norm respectively. For a matrix $\mathbf{A}$, we denote its eigenvalues by $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \ldots$ The identity matrix with k rows and k columns is denoted as $\mathbf{I}_k$.

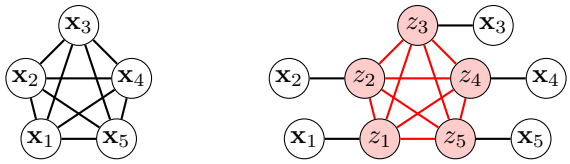

Figure 1: Left: A model in which all variables $\mathbf{x}_i$'s are observed and we can easily compute covariances among those variables. Right: Our model in which the we are interested on computing covariances between latent (i.e., unobserved) variables z_i's in red, but we can only observe the variables $\mathbf{x}_i$'s

For a matrix $\mathbf{A}$, $\mathrm{diag}(\mathbf{A})$ is the vector containing the diagonal entries of $\mathbf{A}$. For a vector $\mathbf{a}$, the support $\mathrm{supp}(\mathbf{a})$ is the set of indices with non-zero entries, i.e., $\mathrm{supp}(\mathbf{a}) = \{i \mid a_i \neq 0\}$. For a matrix $\mathbf{A}$, the support $\mathrm{supp}(\mathbf{A})$ is the set of indices with non-zero entries, i.e., $\mathrm{supp}(\mathbf{A}) = \{(i,j) \mid A_{ij} \neq 0\}$. We use superscript to denote the index of i.i.d. samples, e.g., the vector $\mathbf{x}^{(i)}$ denotes the i-th sample in the dataset.

3 MAIN RESULTS

In this section, we address the challenge of estimating the covariance matrix for a set of latent variables. We first describe our generative model assumptions. We then present some properties of the population quantities, which in term motivate our proposed empirical estimator. We finish the section by providing theoretical guarantees, where all proofs are postponed to Appendix B.

3.1 Generative Model

First, we define our model of correlated latent random variables, where each of them is associated to a observable random vector. For a quick graphical presentation, please see Figure 1.

In our definition, we use a general class of distributions called sub-Gaussian. The class of sub-Gaussian variables includes for instance, Gaussian variables, any bounded random variable (such as Bernoulli, multinomial, or uniform), any random variable with a strictly log-concave density, and any finite mixture of sub-Gaussian variables.

Definition 1. *For $i = 1, \ldots, p$, the observable random vector $\mathbf{x}_i \in \mathbb{R}^{q_i}$ is associated with a latent random variable $z_i \in \mathbb{R}$, where the dimension $q_i > 1$. We follow a linear model in which*

$$\mathbf{x}_i = \mathbf{c}_i z_i + \boldsymbol{\epsilon}_i \ ,$$

for some sparse constant vector $\mathbf{c}_i \in \mathbb{R}^{q_i}$ satisfying $\|\mathbf{c}_i\|_2 = 1$ and $|\mathrm{supp}(\mathbf{c}_i)| \leq s$, and random noise

$\boldsymbol{\epsilon}_i \in \mathbb{R}^{q_i}$ *satisfying* $\mathbb{E}[\boldsymbol{\epsilon}_i] = \mathbf{0}$ *and* $\mathbb{E}[\boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^T] = \sigma_i^2 \mathbf{I}_{q_i}$.[2] *The latent variables* z_i*'s are zero mean and have covariance matrix* $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$*, that is,* $\mathbb{E}[z_i] = 0$ *and* $\Sigma_{ij} = \mathbb{E}[z_i z_j]$*. The random noise vectors* $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_p$ *are independent. Furthermore,* $z_i / \sqrt{\Sigma_{ii}}$ *and* $\boldsymbol{\epsilon}_i / \sigma_i$ *are sub-Gaussian with parameter* γ*. As an important quantity in our analysis, we define*

$$\mathbf{M}_{ij} = \mathbb{E}[\mathbf{x}_i \mathbf{x}_j^T] \,,$$

for all $i, j = 1, \ldots, p$*. Note that the constant matrix* $\boldsymbol{\Sigma}$*, vectors* $\mathbf{c}_i$*'s and scalars* σ_i*'s are unknown. For clarity, we define the support set* $J_i = \operatorname{supp}(\mathbf{c}_i)$ *and the full observed dimension* $Q = \sum_{i=1}^{p} q_i$*.*

Before we continue further, we prove that the linearity assumption in Definition 1 is necessary. Note that a main requirement for estimating covariances of latent variables is to be able to observe some function of such latent variables. For clarity of exposition, we consider a one-dimensional noiseless case, in which we have an observed variable x and latent variable is z. Assume that $x = g(z)$ for some function g. Furthermore, assume we observe some function f of the variance of z. Next we show that both f and g have to be linear, providing theoretical evidence for our choice.

Theorem 1. *Let* $z \in \mathbb{R}$ *be a random variable. Let* $f, g : \mathbb{R} \to \mathbb{R}$ *be two functions where* g *is monotonic and differentiable. We have that:* $f(\mathbb{E}\, z^2) = \mathbb{E}\, g(z)^2$ *for any probability distribution of* z *if and only if* $f(t) = c^2 t$ *and* $g(z) = cz$ *for some constant* $c \in \mathbb{R}$*.*

Unfortunately, as we are dealing with latent variables, there is some expected uncertainties in the sign of the non-diagonal entries of the covariance matrices. Despite these inherent issues, we will be able to propose a reliable estimator. We formally introduce a new concept of equivalence this in what follows.

Definition 2. *A* sign-flip *matrix is a diagonal matrix* $\mathbf{F} \in \mathbb{R}^{p \times p}$ *where each entry in its diagonal is either* 1 *or* -1*. That is,* $F_{ii} \in \{-1, +1\}$ *for all* $i = 1, \ldots, p$*. A* sign-flip transform *of matrix* $\mathbf{A} \in \mathbb{R}^{p \times p}$ *is the product* $\mathbf{FAF}$ *where* $\mathbf{F}$ *is a sign-flip matrix. We say that two matrices* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p \times p}$ *are* sign-flip equivalent*, if* $\mathbf{B}$ *is a sign-flip transform of* $\mathbf{A}$*, or equivalently if* $\mathbf{A}$ *is a sign-flip transform of* $\mathbf{B}$*.*

Next, we discuss some useful properties regarding sign-flip equivalence. We will use the first claim in this section since we want to ensure that the estimated covariance matrix is still positive semidefinite. The other three properties will be useful for our application in Section 4.

Lemma 1. *Under Definition 2, we claim that:*

1. *Let matrices* $\mathbf{A}$ *and* $\mathbf{B}$ *be sign-flip equivalent. We have that* $|A_{ij}| = |B_{ij}|$ *for all* i, j*.*
2. *Let matrices* $\mathbf{A}$ *and* $\mathbf{B}$ *be sign-flip equivalent. We have that* $\mathbf{A} \succeq \mathbf{0}$ *if and only if* $\mathbf{B} \succeq \mathbf{0}$*. Similarly, we have that* $\mathbf{A} \succ \mathbf{0}$ *if and only if* $\mathbf{B} \succ \mathbf{0}$*.*
3. *Let matrices* $\mathbf{A}$ *and* $\mathbf{B}$ *be sign-flip equivalent. We have that* $\det(\mathbf{A}) = \det(\mathbf{B})$*.*
4. *Let matrices* $\mathbf{A}$ *and* $\mathbf{B}$ *be two arbitrary matrices that undergo the same sign-flip transform. That is,* $\mathbf{A}' = \mathbf{FAF}$ *and* $\mathbf{B}' = \mathbf{FBF}$*. We have that* $\langle \mathbf{A}', \mathbf{B}' \rangle = \langle \mathbf{A}, \mathbf{B} \rangle$*.*

3.2 Population Properties

Here we study properties of the population quantities in our generative model. This will later motivate a particular empirical estimator in the next subsection. The following lemma reveals the relationship between the population covariance of the latent variables and the population covariance of the observed variables.

Lemma 2. *Under Definition 1, we have that* $\Sigma_{ij} = \mathbb{E}[z_i z_j] = \mathbf{c}_i^T \mathbb{E}[\mathbf{x}_i \mathbf{x}_j^T] \mathbf{c}_j - \sigma_i^2 1[i = j]$ *for all* $i, j = 1, \ldots, p$*.*

The lemma below shows that we can express the population covariance of the observed variables in terms of the sparse constant vectors $\mathbf{c}_i$'s and the noise variances σ_i^2's. Further, we show that the population covariance of the observed variables is a spiked covariance.

Lemma 3. *Under Definition 1, we have that* $\mathbf{M}_{ij} = \mathbb{E}[\mathbf{x}_i \mathbf{x}_j^T] = \Sigma_{ij} \mathbf{c}_i \mathbf{c}_j^T + 1[i = j] \sigma_i^2 \mathbf{I}_{q_i}$ *for all* $i, j = 1, \ldots, p$*.*

Armed with this result, we show that the constant sparse vector $\mathbf{c}_i$ and noise variance σ_i^2 are quantities involved in the eigen-decomposition of the population covariance of the observed variables.

Lemma 4. *Under Definition 1, the maximum eigenvector of* $\mathbf{M}_{ii} = \mathbb{E}[\mathbf{x}_i \mathbf{x}_i^T]$ *is either* $\mathbf{c}_i$ *or* $-\mathbf{c}_i$ *with corresponding eigenvalue* $\Sigma_{ii} + \sigma_i^2$*. The second maximum eigenvalue of* $\mathbf{M}_{ii}$ *is* σ_i^2*.*

The fact that the main eigenvector of $\mathbf{M}_{ii}$ is either $\mathbf{c}_i$ or $-\mathbf{c}_i$ (i.e., $\mathbf{c}_i$ is identifiable up to its sign) is one reason for introducing the concept of sign-flip equivalence.

3.3 Empirical Estimator

Assume we receive n samples, each of them of the form $(\mathbf{x}_1^{(k)}, \mathbf{x}_2^{(k)}, \ldots, \mathbf{x}_p^{(k)})$ for $k = 1, \ldots, n$ and where each $\mathbf{x}_i^{(k)} \in \mathbb{R}^{q_i}$ for $i = 1, \ldots, p$. Inspired by the population properties in Lemmas 2 and 3, we propose the following empirical estimator by leveraging the observed

[2] We argue for this assumption in Appendix A.

data to estimate the underlying covariance matrix:

$$\widehat{\Sigma}_{ij} = \widehat{\mathbf{c}}_i^T \widehat{\mathbf{M}}_{ij} \widehat{\mathbf{c}}_j - \widehat{\sigma}_i^2 1[i=j] \ ,$$
$$\widehat{\mathbf{M}}_{ij} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_i^{(k)} (\mathbf{x}_j^{(k)})^T \ . \quad (1)$$

The equations above bring the important question of how to estimate the sparse constant vectors $\mathbf{c}_i$ as well as the noise variance σ_i^2. Noting the eigen-decomposition properties of the population covariance of the observed variables and the fact that $\mathbf{c}_i$'s are sparse, suggests the use of sparse principal component analysis (PCA). Thus, we propose the following optimization problem from (Lei and Vu, 2015) to estimate the sparse constant vectors $\mathbf{c}_i$'s which we found are the maximum eigenvectors of $\mathbf{M}_{ii}$. More formally

$$\widehat{\mathbf{H}}_i = \underset{\mathbf{0} \preceq \mathbf{H}_i \preceq \mathbf{I},\ \mathrm{tr}(\mathbf{H}_i)=1}{\arg\max} \langle \widehat{\mathbf{M}}_{ii}, \mathbf{H}_i \rangle - \rho \left\| \mathbf{H}_i \right\|_1 \ , \quad (2)$$

where ρ is a regularization parameter.[3] As shown in Theorem 2 of (Lei and Vu, 2015), with a proper regularization parameter $\rho = O(\sqrt{(\log q_i)/n})$, the solution to the above optimization problem $\widehat{\mathbf{H}}_i = \widehat{\mathbf{c}}_i \widehat{\mathbf{c}}_i^T$ is a good estimator for $\mathbf{H}_i = \mathbf{c}_i \mathbf{c}_i^T$, in the sense that the support recovery is correct, i.e., $\mathrm{supp}(\widehat{\mathbf{c}}_i) = J_i \equiv \mathrm{supp}(\mathbf{c}_i)$. Further, Lemma 2 in (Lei and Vu, 2015) shows that $\left\| (\mathbf{c}_i \mathbf{c}_i^T - \widehat{\mathbf{c}}_i \widehat{\mathbf{c}}_i^T)_{J_i J_i} \right\|_{\mathcal{F}} \leq \frac{4\rho s}{\lambda_1(\mathbf{M}_{ii}) - \lambda_2(\mathbf{M}_{ii})}$.

Next, we discuss the estimation of the noise variances σ_i^2's. Let $\phi_i = \lambda_1(\mathbf{M}_{ii})$. Note that by Lemma 4, we have $\phi_i = \Sigma_{ii} + \sigma_i^2$. Furthermore by Lemma 2, we have $\phi_i = \mathbf{c}_i^T \mathbf{M}_{ii} \mathbf{c}_i$. Furthermore, by Lemma 4 we have that $\sigma_i^2 = \lambda_2(\mathbf{M}_{ii})$, therefore $\sigma_i^2 = \lambda_1((\mathbf{M}_{ii} - \phi_i \mathbf{c}_i \mathbf{c}_i^T)_{J_i J_i})$. The above provides a well-grounded motivation of the following two empirical estimators:

$$\widehat{\phi}_i = \widehat{\mathbf{c}}_i^T \widehat{\mathbf{M}}_{ii} \widehat{\mathbf{c}}_i \ ,$$
$$\widehat{\sigma}_i^2 = \lambda_1((\widehat{\mathbf{M}}_{ii} - \widehat{\phi}_i \widehat{\mathbf{c}}_i \widehat{\mathbf{c}}_i^T)_{J_i J_i}) \ . \quad (3)$$

3.4 Provable Guarantees

Here we provide theoretical guarantees for our estimator proposed in the previous section. Note that the solution of the sparse PCA problem in eq.(2) has another valid factorization of the form $\widehat{\mathbf{H}}_i = (-\widehat{\mathbf{c}}_i)(-\widehat{\mathbf{c}}_i)^T$. This is another reason for proposing the concept of sign-flip equivalence in Definition 2. In our analysis, we define a *sign-flip* matrix $\widehat{\mathbf{F}}$ as in Definition 2 as follows. For each $i = 1, \ldots, p$ we set $\widehat{F}_{ii} = 1$ if $\widehat{\mathbf{c}}_i$ and $\mathbf{c}_i$ point in the same direction, and $\widehat{F}_{ii} = -1$ if $\widehat{\mathbf{c}}_i$ and $\mathbf{c}_i$ point in opposite directions.

Note that the matrix $\widehat{\mathbf{F}}$ is only of theoretical use, and does not affect our proposed estimator. In the next lemma, we show that we can decompose the bound of the error of the covariance of latent variables into three terms, which we bound later.

Lemma 5. *Under Definition 1 and estimators in eq.(1), we have*

$$\begin{aligned} &|\widehat{F}_{ii} \widehat{\Sigma}_{ij} \widehat{F}_{jj} - \Sigma_{ij}| \\ &\leq s \underbrace{\left\| \widehat{\mathbf{M}}_{ij} - \mathbf{M}_{ij} \right\|_\infty}_{\text{Lemma 6}} \\ &\quad + \left(\Sigma_{ij} + 1[i=j] \sigma_i^2 \right) \underbrace{\left\| (\widehat{F}_{ii} \widehat{F}_{jj} \widehat{\mathbf{c}}_i \widehat{\mathbf{c}}_j^T - \mathbf{c}_i \mathbf{c}_j^T)_{J_i J_j} \right\|_1}_{\text{Lemma 7}} \\ &\quad + \underbrace{|\widehat{\sigma}_i^2 - \sigma_i^2|}_{\text{Lemma 8}} 1[i=j] \ . \end{aligned}$$

for all $i, j = 1, \ldots, p$.

Note that the above lemma provides a bound for $\|\widehat{\mathbf{F}} \widehat{\mathbf{\Sigma}} \widehat{\mathbf{F}} - \mathbf{\Sigma}\|_\infty$. Intuitively speaking, this shows an "approximate" *sign-flip equivalence* between $\widehat{\mathbf{F}} \widehat{\mathbf{\Sigma}} \widehat{\mathbf{F}}$ and $\mathbf{\Sigma}$.

In the next lemma we provide a bound for the error of the covariance of observed variables. The results follows from the fact that the observed variable $\mathbf{x}_i$ is sub-Gaussian given that the latent variable z_i and the random noise $\boldsymbol{\epsilon}_i$ are sub-Gaussian. We then apply concentration inequalities and a union bound across all matrix entries.

Lemma 6. *Under Definition 1 and estimators in eq.(1), with probability at least $1 - \delta$ we have*

$$\begin{aligned} &\left\| \widehat{\mathbf{M}}_{ij} - \mathbf{M}_{ij} \right\|_\infty \\ &\leq O\left(\frac{(1 + 8\gamma^2)(\max_i \Sigma_{ii} + \sigma_i^2)}{\sqrt{n}} \sqrt{\log Q + \log(1/\delta)} \right) \ , \end{aligned}$$

for all $i, j = 1, \ldots, p$.

Therefore, the sample complexity for estimating the covariance matrix of the observed variables is logarithmic in the full observed dimension $Q = \sum_{i=1}^{p} q_i$. That is, we require $n = O(\log Q)$ samples. This makes our partial result suitable for the high-dimensional regime.

In the following lemma, we bound the second term encountered in Lemma 5 which pertains to how well we estimate the sparse constant vectors $\mathbf{c}_i$'s when using sparse PCA as in eq.(2). Next, we show that beyond properly estimating the support, the estimated vectors $\widehat{\mathbf{c}}_i$'s are "close" to the true sparse constant vectors $\mathbf{c}_i$'s in a carefully defined distance.

[3] The use of the nondiagonal entries in $\widehat{\mathbf{M}}_{ii}$ (and not just its diagonal) is important because otherwise, we could not disambiguate between two vectors $\mathbf{c}_i = (1, 1, 1, 0, \ldots, 0)^T / \sqrt{3}$ and $\mathbf{c}_i = (1, -1, 1, 0, \ldots, 0)^T / \sqrt{3}$ for instance. This motivates further the use of sparse PCA.

Lemma 7. *Under Definition 1 and by setting* $\rho = O(\sqrt{(\log q_i)/n})$ *in the sparse PCA problem in eq.*(2)*, we have that*

$$\text{supp}(\widehat{\mathbf{c}}_i) = J_i \equiv \text{supp}(\mathbf{c}_i)$$

$$\left\|(\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_i^T - \mathbf{c}_i\mathbf{c}_i^T)_{J_iJ_i}\right\|_{\mathcal{F}} \leq O\left(\frac{s\sqrt{\log q_i}}{\Sigma_{ii}\sqrt{n}}\right) ,$$

$$\left\|(\widehat{F}_{ii}\widehat{F}_{jj}\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_j^T - \mathbf{c}_i\mathbf{c}_j^T)_{J_iJ_j}\right\|_1 \leq O\left(\frac{s^{3/2}}{\sqrt{n}}\max_i \frac{\sqrt{\log q_i}}{\sqrt{\Sigma_{ii}}}\right) ,$$

for all $i, j = 1, \ldots, p$.

Next, we bound the third term in Lemma 5 which pertains to how well we estimate the noise variances σ_i^2's with our proposed method in eq.(3).

Lemma 8. *Under Definition 1 and estimators in eq.*(3)*, with probability at least* $1-\delta$ *we have*

$$\begin{aligned}
&|\widehat{\sigma}_i^2 - \sigma_i^2| \\
&\leq O\left(\gamma(\Sigma_{ii}+\sigma_i^2)\min\left(\sqrt{\frac{s+\log p+\log(1/\delta)}{n}}, \right.\right. \\
&\qquad\qquad \left.\left.\frac{s+\log p+\log(1/\delta)}{n}\right)\right) \\
&\quad + \underbrace{|\phi_i - \widehat{\phi}_i|}_{Lemma\ 9} + O\left(\phi_i\frac{s\sqrt{\log q_i}}{\Sigma_{ii}\sqrt{n}}\right) .
\end{aligned}$$

for all $i, j = 1, \ldots, p$.

This lead to the next lemma to find the bound for how close the estimated $\widehat{\phi}_i$ is to the true value ϕ_i, which is needed for bounding the error of the estimator $\widehat{\sigma}_i^2$ above.

Lemma 9. *Under Definition 1 and estimators in eq.*(3)*, we have*

$$\begin{aligned}
|\widehat{\phi}_i - \phi_i| &\leq s\left\|\widehat{\mathbf{M}}_{ii} - \mathbf{M}_{ii}\right\|_\infty \\
&\quad + O\left((\Sigma_{ii}+\sigma_i^2)\frac{s^{3/2}}{\sqrt{n}}\max_i\frac{\sqrt{\log q_i}}{\sqrt{\Sigma_{ii}}}\right) .
\end{aligned}$$

for all $i, j = 1, \ldots, p$.

Combining all the results above, we are now ready to state our main result, which we state in order notation and less formally than the above results for clarity of exposition.

Theorem 2. *Under Definition 1 and estimators in eq.*(1)*, eq.*(2) *and eq.*(3)*, with probability at least* $1-\delta$ *we have*

$$\begin{aligned}
&\left\|\widehat{\mathbf{F}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{F}} - \mathbf{\Sigma}\right\|_\infty \\
&\leq O\left(\frac{s\sqrt{\log Q + \log(1/\delta)} + s^{3/2}\sqrt{\max_i \log q_i}}{\sqrt{n}}\right. \\
&\left. + \min\left(\sqrt{\frac{s+\log p+\log(1/\delta)}{n}}, \frac{s+\log p+\log(1/\delta)}{n}\right)\right) .
\end{aligned}$$

Proof. By invoking Lemmas 5, 6, 7, 8 and 9. □

To have a better more intuitive grasp of the above bound, lets consider a case in which the observed dimension q_i for all variables i are the same. Given the above, the sample complexity for estimating a covariance matrix of latent variables $\widehat{\mathbf{\Sigma}}$ that is *sign-flipped equivalent* to the true covariance matrix $\mathbf{\Sigma}$ is $n = O(\max(s^2(\log p + \log q_i), s^3 \log q_i)$, which is logarithmic in the number of latent variables p and the observed dimension q_i, making our proposed estimator suitable for high-dimensions.

4 APPLICATION: LATENT-VARIABLE PRECISION MATRIX ESTIMATION

In this section, we make use of our estimated covariance matrix $\widehat{\mathbf{\Sigma}}$ of latent variables to go one step further: to estimate the conditional dependence between those latent variables. The above relates to estimate a precision matrix (i.e., inverse covariance matrix) that is sparse (i.e., having few nonzero entries.) The precision matrix is denoted as $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$. A zero entry in the precision matrix, i.e., $\Omega_{ij} = 0$, implies that variables i and j are conditionally independent given all others. A non-zero entry implies that those two variables are conditionally dependent given all others. Thus, the literature in precision matrix estimation guarantees (Ravikumar et al., 2011) focuses on estimating the support $\text{supp}(\mathbf{\Omega})$.

While precision matrix estimation is typically done for a covariance of observed variables, here we focus on a more challenging setting by using our covariance of latent variables. Given our empirical covariance matrix $\widehat{\mathbf{\Sigma}}$, we can estimate $\mathbf{\Omega}$ by solving the following ℓ_1-norm regularized maximum likelihood estimation problem as done in (Banerjee et al., 2008; Ravikumar et al., 2011)

$$\widehat{\mathbf{\Omega}} = \arg\min_{\mathbf{\Omega}\succ\mathbf{0}} \langle\mathbf{\Omega}, \widehat{\mathbf{\Sigma}}\rangle - \log\det(\mathbf{\Omega}) + \nu\left\|\mathbf{\Omega}\right\|_{1,\text{off}} , \quad (4)$$

where ν is a regularization parameter and the $\ell_{1,\text{off}}$ norm $\|\mathbf{\Omega}\|_{1,\text{off}} = \sum_{i\neq j}|\Omega_{ij}|$. At this point, the reader might be asking whether it makes sense to use the above optimization problem given that we can only guarantee recovery of a *sign-flip equivalent* covariance matrix. Fortunately, in Lemma 1 we showed that for sign-flip equivalent matrices, the magnitudes of entries are the same and therefore their $\ell_{1,\text{off}}$ norm. Similarly, positive definiteness is preserved, and their determinants are the same. Finally, the inner product of two matrices that undergo the same sign-flip transform is the same as the original inner product. All of the above guarantees that the constraint (i.e., positive

definiteness) and objective function (i.e., inner product, determinant and $\ell_{1,\text{off}}$ norm) in the optimization problem in eq.(4) is preserved.

Next, we provide a corollary for the estimation of a precision matrix of latent variables. There are three differences in our statement with respect to Theorem 1 and Corollary 1 in (Ravikumar et al., 2011). First, the result in (Ravikumar et al., 2011) pertains only to the precision matrix of observed variables. Second, we use of the sign-flip matrix $\widehat{\mathbf{F}}$. Third, while (Ravikumar et al., 2011) sets the regularization parameter $\nu = O(\sqrt{(\log p)/n})$, our setting is different as it relates to our provable guarantee for the covariance matrix of latent variables.

Corollary 1. *Under the same assumptions as in Theorem 1 and Corollary 1 in (Ravikumar et al., 2011) and by setting the regularization parameter* $\nu = O((s\sqrt{\log Q + \log(1/\delta)} + s^{3/2}\sqrt{\max_i \log q_i})/\sqrt{n} + \min(\sqrt{s + \log p + \log(1/\delta)}/\sqrt{n}, (s + \log p + \log(1/\delta))/n))$*, we have*

$$\text{supp}(\widehat{\mathbf{\Omega}}) = \text{supp}(\mathbf{\Omega}),$$

and

$$\left\|\widehat{\mathbf{F}}\widehat{\mathbf{\Omega}}\widehat{\mathbf{F}} - \mathbf{\Omega}\right\|_\infty \leq O\left((1+8\gamma^2)\sqrt{\frac{\log p}{n}}\right) .$$

where γ is the sub-Gaussianity parameter as in Definition 1.

Proof. We invoke Theorem 1 and Corollary 1 in (Ravikumar et al., 2011) where $\nu = O\left(\left\|\widehat{\mathbf{F}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{F}} - \mathbf{\Sigma}\right\|_\infty\right)$. Thus, by Theorem 2 we require $\nu = O((s\sqrt{\log Q + \log(1/\delta)} + s^{3/2}\sqrt{\max_i \log q_i})/\sqrt{n} + \min(\sqrt{s + \log p + \log(1/\delta)}/\sqrt{n}, (s + \log p + \log(1/\delta))/n))$. □

5 EXPERIMENTAL VALIDATION

In this section, we validate our theoretical results through various synthetic experiments on covariance and precision matrix estimation. We also perform experiments on real-world data.

To the best of our knowledge, there are no prior methods for empirical comparison. Thus, we consider a reasonable covariance baseline $\widehat{\mathbf{B}}$ that ignores the latent variable structure (i.e., constant sparse vectors $\mathbf{c}_i$'s and noise variances σ_i^2's) and directly uses the observed variables $\mathbf{x}_i$'s as follows. Recall that we receive n samples, each of them of the form $(\mathbf{x}_1^{(k)}, \mathbf{x}_2^{(k)}, \ldots, \mathbf{x}_p^{(k)})$ for $k = 1, \ldots, n$ and where each $\mathbf{x}_i^{(k)} \in \mathbb{R}^{q_i}$ for $i = 1, \ldots, p$. Given the above, the covariance baseline is computed as $\widehat{B}_{ij} = \frac{1}{n}\sum_{k=1}^n (\mathbf{x}_i^{(k)})^T \mathbf{x}_j^{(k)}$. Of course, this would not make sense unless all q_i are equal, which is the setup we consider for our experiments. To do a fair comparison, we also confirmed that the scale of the entry magnitudes in $\widehat{\mathbf{B}}$ and our estimator were similar.

5.1 Synthetic Data

Our experiments aim to assess how our estimator performs with finite and varying sample sizes, particularly focusing on the speed of convergence. We consider the number of latent variables p in $\{10, 20, 50\}$. The dimension of the observed variables q_i is the same for all i. We set q_i in $\{20, 50, 100\}$. Note that under this setting the full observed dimension $Q = pq_i$.

We generate a random covariance matrix of latent variables $\mathbf{\Sigma}$ such that its inverse is sparse (i.e., have at most 3 nonzeros per row) and has minimum eigenvalue 0.1. The above allows us to perform experiments of support recovery of the precision matrix. To generate the sparse vectors $\mathbf{c}_i$'s, we first choose 3 entries out of q_i uniformly at random to be nonzero, we then sample Gaussian distributed values for those entries, and normalize the vector so that $\|\mathbf{c}_i\|_2 = 1$. The noise variance $\sigma_i^2 = 0.1$ for all i.

We create n random samples as follows. For each sample, first sample the latent random vector $(z_1, \ldots, z_p)$ from a multivariate Gaussian distribution with zero mean and covariance matrix $\mathbf{\Sigma}$. For each latent variable i, we sample a noise vector $\boldsymbol{\epsilon}_i$ from a multivariate Gaussian distribution with zero mean and covariance matrix $\sigma_i^2 \mathbf{I}_{q_i}$. The observed variable i then follows Definition 1, i.e., $\mathbf{x}_i = \mathbf{c}_i z_i + \boldsymbol{\epsilon}_i$.

For optimization problems eq.(2) and eq.(4) set the regularization parameters ρ and ν according to our theory, by using factors 0.34 and 0.74 respectively. We perform 30 repetitions to ensure statistical reliability. We vary the number of samples n and analyze how this affects our estimator. We choose the number of samples $n = T \log Q$ where T is either $\{10, 20, 50, 100, 200, 500, 1000\}$.

In Figure 2 we aim to validate our main result in Theorem 2. Thus, we report the average relative ℓ_∞ distance $\left\|\widehat{\mathbf{F}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{F}} - \mathbf{\Sigma}\right\|_\infty / \|\mathbf{\Sigma}\|_\infty$. We observe that more samples allows to obtain better estimates for the covariance matrix $\mathbf{\Sigma}$ of latent variables. This is consistent with Theorem 2.

In Figure 3 we aim to validate our application result in Corollary 1. Thus, we report the empirical probability of correct support recovery, i.e., $\text{supp}(\widehat{\mathbf{\Omega}}) = \text{supp}(\mathbf{\Omega})$. We observe that more samples allows to obtain a better support recovery of the precision matrix $\mathbf{\Omega}$. This

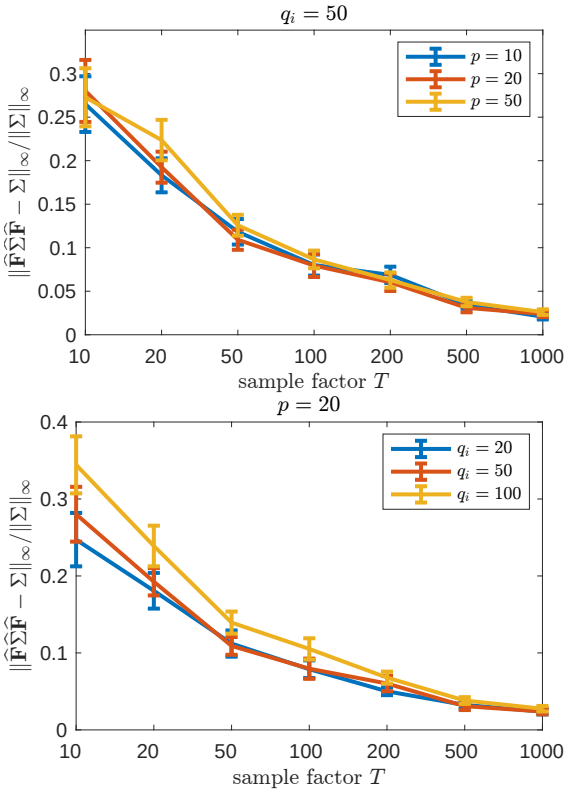

Figure 2: Average relative ℓ_∞ distance $\left\|\widehat{\mathbf{F}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{F}} - \mathbf{\Sigma}\right\|_\infty / \|\mathbf{\Sigma}\|_\infty$ with varying number of latent variables p (Top) and varying observed dimensions q_i (Bottom). We show standard error bars at 95% confidence level. Note that in both cases, the more samples the better the estimates of the covariance matrix $\mathbf{\Sigma}$ of latent variables is.

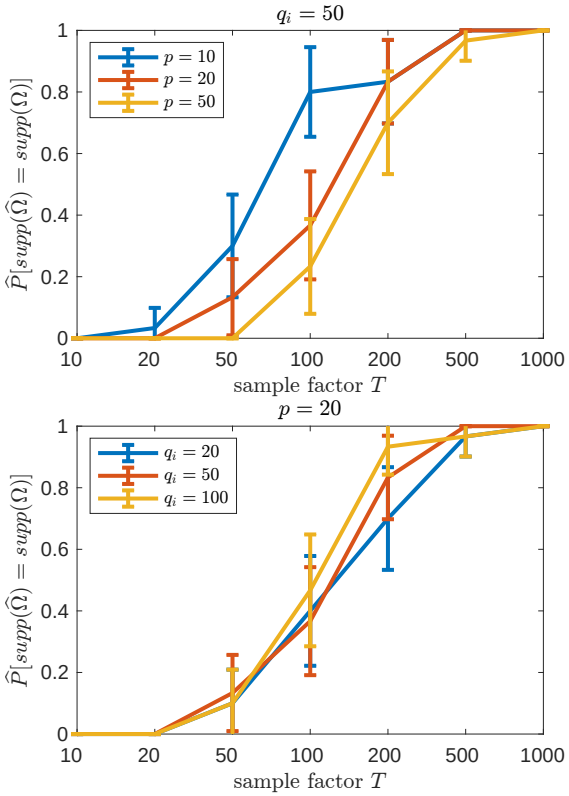

Figure 3: Empirical probability of correct support recovery $\text{supp}(\widehat{\mathbf{\Omega}}) = \text{supp}(\mathbf{\Omega})$ with varying number of latent variables p (Top) and varying observed dimensions q_i (Bottom). We show standard error bars at 95% confidence level. Note that in both cases, the more samples the better the support recovery of the precision matrix $\mathbf{\Omega}$ is.

is consistent with Corollary 1. The chart also shows the characteristic phase transition behavior: the probability of exact recovery is essentially 0% at very small sample sizes, but climbs rapidly to 100% once the sample size exceeds a critical threshold.

Figure 4 compares our estimator against a the baseline described at the beginning of this section, for $p = 20$ latent variables and observed dimension $q_i = 50$. Motivated by both Theorem 2 and Corollary 1, we report the average relative ℓ_∞ distance $\left\||\widehat{\mathbf{\Sigma}}| - |\mathbf{\Sigma}|\right\|_\infty / \|\mathbf{\Sigma}\|_\infty$ where $|\widehat{\mathbf{\Sigma}}|$ and $|\mathbf{\Sigma}|$ are matrices containing the entries' magnitudes of $\widehat{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}$ respectively. We also report the empirical probability of correct support recovery of $\mathbf{\Omega}$. We observe that the baseline performs extremely poorly under both metrics.

Additional experiments showing the evaluation of the intermediate steps for estimating the constant sparse vectors $\mathbf{c}_i$'s and noise variances σ_i^2's are shown in Appendix C.1.

5.2 Real-World Data

The *1000 functional connectomes* dataset contains resting-state functional magnetic resonance images (fMRI) of 1128 subjects collected on 41 sites around the world. The data set is publicly available at `http://www.nitrc.org/projects/fcon_1000/`. Resting-state fMRI is a procedure that captures brain function of a subject that is at wakeful rest (i.e., not focused on the outside world). The dataset contains 84 Brodmann areas including left and right side of the brain.

Our experimental setup is semi-synthetic. The true precision matrix $\mathbf{\Omega}$ of latent variables is estimated by using eq.(4) on the 84×84 covariance matrix of the dataset. The true covariance matrix is computed as $\mathbf{\Sigma} = \mathbf{\Omega}^{-1}$. After this, we consider $p = 84$ latent variables, observed dimension $q_i = 10$ and noise variance $\sigma_i^2 = 0.1$. Note that the real-world dataset becomes the latent variables z_i's. We then randomly generate sparse vectors $\mathbf{c}_i$'s, noise vectors $\boldsymbol{\epsilon}_i$'s and observed variables $\mathbf{x}_i$'s as done in the previous subsection. Es-

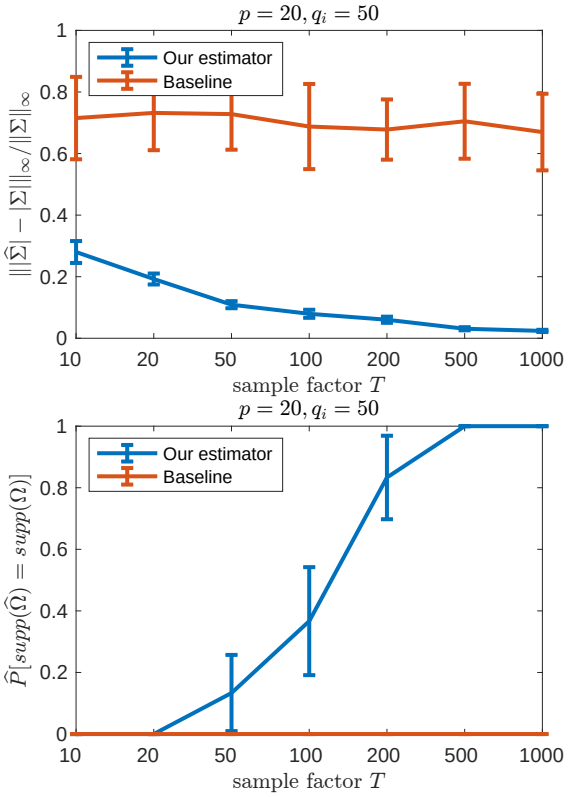

Figure 4: Comparison between our estimator and a reasonable baseline. Top: Average relative ℓ_∞ distance $\left\|\|\widehat{\mathbf{\Sigma}}| - |\mathbf{\Sigma}|\right\|_\infty / \|\mathbf{\Sigma}\|_\infty$ where $|\widehat{\mathbf{\Sigma}}|$ and $|\mathbf{\Sigma}|$ are matrices containing the entries' magnitudes of $\widehat{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}$ respectively. Bottom: Empirical probability of correct support recovery $\text{supp}(\widehat{\mathbf{\Omega}}) = \text{supp}(\mathbf{\Omega})$. We show standard error bars at 95% confidence level. Note that the baseline method performs very poorly, while our estimator succeeds in both performance measures.

timating the precision matrix $\widehat{\mathbf{\Omega}}$ with our covariance estimator $\widehat{\mathbf{\Sigma}}$ obtains 98.1% precision and 98.1% recall for support recovery. The baseline method did not produce any true or false positive, and therefore precision and recall were undefined.

Additional experiments showing our estimated precision matrix can be found in Appendix C.2.

5.3 Understanding Cocaine Addiction

The problem we studied was born after several discussions with a neuroscientist who studies cocaine addiction. Understanding the relationship between inhibitory control, cue reactivity and self-awareness would allow for more informed treatments.

We would like to figure out the covariances between the latent variables: inhibitory control (which allows us to suppress inappropriate thoughts/actions/impulses) and cue reactivity (i.e., reaction to visual stimuli related to drugs).

These variables are impossible to be measured directly, but we have functional magnetic resonance imaging, giving us oxygen level in each of different 3D locations inside the brain. Oxygen level is related to brain activity through the haemodynamic response (rapid delivery of oxygenated blood to active neuronal tissue).

Each latent variable (inhibitory control and cue reactivity) is associated with an observed variable. Previous studies on independent samples allowed to find out which 3D brain locations are relevant. For the first latent variable, the associated observed variable is 50-dimensional, i.e. $q_1 = 50$. For the second latent variable, the associated observed variable is 15-dimensional, i.e. $q_2 = 15$.

While some brain 3D locations might exhibit a higher or lower activity (modulated by c_i), it is reasonable to assume that the variance is the same across all brain 3D locations activity.

Furthermore, there are two groups: 12 cocaine use disorder (CUD) subjects, and 16 healthy control (HC) subjects. The few subjects are due to the high cost of collecting the data.

With our method, we found that the magnitude of the correlation in the CUD group was smaller than that in the HC group, which was a very meaningful finding for the neuroscientist.

In the future, we plan to add more latent variables such as self-awareness (ability to understand own thoughts/feelings/behaviors, and their impact in ourselves and others) among others to have a more complete model of addiction.

6 CONCLUDING REMARKS

To the best of our knowledge, we are the first to provide formal guarantees for the estimation of covariance matrices of latent variables. Having said that, there are several exciting ways to extend our initial result. It is known that in the case of fully observed data, precision matrix estimation is a blackbox method used for the estimation structural equation models. We wonder whether such extension is possible for latent variables. Finally, while we show that linearity is necessary for covariance estimation, it would be very interesting to study which possible modeling assumptions are required to allow for both non-linearity and provable guarantees.

References

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9(Mar):485–516, 2008.

Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

Jianqing Fan, Yuan Liao, and Han Liu. An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19(1):C1–C32, 03 2016.

D. Guillot, B. Rajaratnam, B. Rolfs, A. Maleki, and I. Wong. Iterative thresholding algorithm for sparse inverse covariance estimation. *Neural Information Processing Systems*, 25:1574–1582, 2012.

C. Hsieh, M. Sustik, I. Dhillon, P. Ravikumar, and R. Poldrack. BIG & QUIC: Sparse inverse covariance estimation for a million variables. *Neural Information Processing Systems*, 26:3165–3173, 2013.

C. Hsieh, M. Sustik, I. Dhillon, and P. Ravikumar. QUIC: Quadratic approximation for sparse inverse covariance estimation. *Journal of Machine Learning Research*, 15(Oct):2911–2947, 2014.

C. Johnson, A. Jalali, and P. Ravikumar. High-dimensional sparse inverse covariance estimation using greedy methods. *International Conference on Artificial Intelligence and Statistics*, 22:574–582, 2012.

P. Kambadur and A. Lozano. A parallel, block greedy method for sparse inverse covariance estimation for ultra-high dimensions. *International Conference on Artificial Intelligence and Statistics*, 23:351–359, 2013.

Clifford Lam and Jianqing Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37(6B):4254–4278, 2009.

Jing Lei and Vincent Q. Vu. Sparsistency and agnostic inference in sparse PCA. *The Annals of Statistics*, 43(1):299 – 322, 2015.

P. Ravikumar, M. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Alessandro Rinaldo. Advanced statistical theory. In-class lecture, 2017. URL `https://www.stat.cmu.edu/~arinaldo/Teaching/36755/F17/Scribed_Lectures/F17_0925.pdf`. Lecture notes.

E. Treister and J. Turek. A block-coordinate descent approach for large-scale sparse inverse covariance estimation. *Neural Information Processing Systems*, 27:927–935, 2014.

H. Wang, A. Banerjee, C. Hsieh, P. Ravikumar, and I. Dhillon. Large scale distributed sparse precision estimation. *Neural Information Processing Systems*, 26:584–592, 2013.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [No, but we provide all the experimental details to reproduce the main results]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable, no special computing infrastructure was used.]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]

(b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Material: Provable Guarantees for Estimating Covariances between Latent Variables with Application to Precision Matrix Estimation

A Necessity of $\sigma_i^2\mathbf{I}$

Here we argue that assuming that $\mathbb{E}[\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T] = \sigma_i^2\mathbf{I}$ is a reasonable assumption, and why assuming $\mathbb{E}[\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T]$ being any arbitrary diagonal matrix leads to a problem that is *impossible to be solved*. We show this by counterexample.

Recall that in our model $\mathbf{x}_i = \mathbf{c}_i\mathbf{z}_i + \boldsymbol{\epsilon}_i$ where $\mathbf{x}_i$ is the observed variable/vector, z_i is the latent variable with $\mathbb{V}\text{ar}[z_i] = \Sigma_{ii}$ and $\mathbf{c}_i$ is an unknown sparse vector. Remember that one of our main goals is to estimate Σ_{ii} from empirical samples of $\mathbf{x}_i$.

Consider a 2-dimensional case ($\mathbf{x}_i, \boldsymbol{\epsilon}_i \in \mathbb{R}^2$). To simplify our counterexample, consider we know $\mathbf{c}_i = \begin{bmatrix}1\\0\end{bmatrix}$. In this paper, we do not know $\mathbf{c}_i$, but we want to argue that even with this knowledge the "arbitrary diagonal" assumption leads to an impossibility.

Let $\mathbb{E}[\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T] = \begin{bmatrix}\alpha_{i1} & 0\\ 0 & \alpha_{i2}\end{bmatrix}$. Note that:

$$\mathbb{V}\text{ar}[x_{i1}] = c_{i1}^2\Sigma_{ii} + \mathbb{V}\text{ar}[\epsilon_{i1}] = \Sigma_{ii} + \alpha_{i1}$$

$$\mathbb{V}\text{ar}[x_{i2}] = c_{i2}^2\Sigma_{ii} + \mathbb{V}\text{ar}[\epsilon_{i2}] = \alpha_{i2}$$

While we might be able to figure out α_{i2} (which is not our goal), it will be impossible to figure out Σ_{ii} since infinitely many values lead to the same sum $\Sigma_{ii} + \alpha_{i1}$. For instance, consider: $\Sigma_{ii} + \alpha_{i1} = 10$. How could we know what is the value of Σ_{ii} and α_{i1}? It might be $\Sigma_{ii} = 9$ and $\alpha_{i1} = 1$, or it might be $\Sigma_{ii} = 6$ and $\alpha_{i1} = 4$, or it can be any $\Sigma_{ii} > 0$ and $\alpha_{i1} = 10 - \Sigma_{ii} > 0$, and there are *infinitely many solutions*.

B Proofs

B.1 Proof of Theorem 1

Proof. First we focus on the following direction. If $f(\mathbb{E}\, z^2) = \mathbb{E}\, g(z)^2$ for any probability distribution of z, then $f(t) = c^2 t$ and $g(z) = cz$ for some constant $c \in \mathbb{R}$. Since the above holds for any probability distribution of z, we choose a particular random variable z. Let z be Bernoulli distributed with probability q. That is, $z = 1$ with probability q, and $z = 0$ with probability $1 - q$ for some constant $q \in (0, 1)$. Since $\mathbb{E}\, z^2 = q$, we have

$$\begin{aligned} f(q) &= f(\mathbb{E}\, z^2)\\ &= \mathbb{E}\, g(z)^2\\ &= q\; g(1)^2 + (1-q)\; g(0)^2\\ &= (g(1)^2 - g(0)^2)\; q + g(0)^2 \end{aligned}$$

Thus, f is a linear function of its parameter. Going back to the general case, for any probability distribution of z, assume $f(t) = c^2 t + d$ for some $c, d \in \mathbb{R}$. We have

$$\begin{aligned} \mathbb{E}\, g(z)^2 &= f(\mathbb{E}\, z^2)\\ &= c^2(\mathbb{E}\, z^2) + d\\ &= \mathbb{E}((cz)^2 + d) \end{aligned}$$

Assume general distributions with support on $[a, b]$ and probability density function p. The above implies

$$\int_a^b g(z)^2 p(z) dz = \int_a^b ((cz)^2 + d) p(z) dz$$

for all $a, b \in \mathbb{R}$ and probability density functions p with support on $[a, b]$. Next we use the fundamental theorem of calculus, i.e., for any function h we have $h(b) = \frac{\partial}{\partial b} \int_a^b h(z) dz)$. By deriving both sides we have $g(b)^2 p(b) = (c^2 b^2 + d) p(b)$ for all $b \in \mathbb{R}$ and probability density functions p with support on $[a, b]$. Since this holds for all p, we chose distributions where $p(b) > 0$ leading to $g(b)^2 = c^2 b^2 + d$ for all b. Thus, either $g(b) = \sqrt{c^2 b^2 + d}$ or $g(b) = -\sqrt{c^2 b^2 + d}$. Those two cases are non-monotonic unless $d = 0$ and thus $f(b) = c^2 b + d$. This leads to two cases, either $g(b) = cb$ or $g(b) = c|b|$. Note that the second case is non-differentiable. Therefore $g(b) = cb$.

The opposite direction can be trivially shown. If $f(t) = c^2 t$ and $g(z) = cz$ for some constant $c \in \mathbb{R}$, then $f(\mathbb{E}\, z^2) = \mathbb{E}\, g(z)^2$ for any probability distribution of z. Note that $f(\mathbb{E}\, z^2) = c^2\, \mathbb{E}\, z^2 = \mathbb{E}(cz)^2 = \mathbb{E}\, g(z)^2$. □

B.2 Proof of Lemma 1

Proof. To prove claim 1, note that $B_{ij} = F_{ii} A_{ij} F_{jj}$. Thus, $|B_{ij}| = |F_{ii} A_{ij} F_{jj}| = |F_{ii}||A_{ij}||F_{jj}| = |A_{ij}|$ since $|F_{ii}| = 1$ for all $i = 1, \ldots, p$. To prove claim 2, we use the definition of positive definitiveness Given $\mathbf{A} \succeq \mathbf{0}$, we have $(\forall \mathbf{y})\ \mathbf{y}^T \mathbf{A} \mathbf{y} \geq 0$. Note that

$$\begin{aligned}
\mathbf{y}^T \mathbf{A} \mathbf{y} &= \mathbf{y}^T \mathbf{F} \mathbf{B} \mathbf{F} \mathbf{y} \\
&= (\mathbf{F}\mathbf{y})^T \mathbf{B} (\mathbf{F}\mathbf{y}) \\
&= \mathbf{z}^T \mathbf{B} \mathbf{z} \\
&\geq 0
\end{aligned}$$

which holds $\forall \mathbf{z}$ since $\mathbf{F}$ is full rank and we made $\mathbf{z} = \mathbf{F}\mathbf{y}$. Thus $\mathbf{B} \succeq \mathbf{0}$. The proof for positive definiteness follows the same argument. To prove claim 3, we use the multiplicative property of the determinant

$$\begin{aligned}
\det(\mathbf{A}) &= \det(\mathbf{F}\mathbf{B}\mathbf{F}) \\
&= \det(\mathbf{F}) \det(\mathbf{B}) \det(\mathbf{F}) \\
&= \det(\mathbf{B})
\end{aligned}$$

since $\det(\mathbf{F})^2 = 1$. To prove claim 4, we use the properties the inner product and trace

$$\begin{aligned}
\langle \mathbf{A}', \mathbf{B}' \rangle &= \operatorname{tr}((\mathbf{A}')^T \mathbf{B}') \\
&= \operatorname{tr}((\mathbf{F}\mathbf{A}\mathbf{F})^T \mathbf{F}\mathbf{B}\mathbf{F}) \\
&= \operatorname{tr}(\mathbf{F}\mathbf{A}^T \mathbf{F}\mathbf{F}\mathbf{B}\mathbf{F}) \\
&= \operatorname{tr}(\mathbf{A}^T \mathbf{F}\mathbf{F}\mathbf{B}\mathbf{F}\mathbf{F}) \\
&= \operatorname{tr}(\mathbf{A}^T \mathbf{B}) \\
&= \langle \mathbf{A}, \mathbf{B} \rangle
\end{aligned}$$

since $\mathbf{F}\mathbf{F} = \mathbf{I}$. □

B.3 Proof of Lemma 2

Proof. We start from the definition of $\mathbf{x}_i = \mathbf{c}_i z_i + \boldsymbol{\epsilon}_i$,

$$\begin{aligned}
\mathbf{c}_i^T \mathbf{x}_i &= \mathbf{c}_i^T (\mathbf{c}_i z_i + \boldsymbol{\epsilon}_i) \\
&= \mathbf{c}_i^T \mathbf{c}_i z_i + \mathbf{c}_i^T \boldsymbol{\epsilon}_i \\
&= z_i + \mathbf{c}_i^T \boldsymbol{\epsilon}_i
\end{aligned}$$

Next we can take the expectation

$$\begin{aligned}
\mathbb{E}[\mathbf{c}_i^T \mathbf{x}_i \mathbf{x}_j^T \mathbf{c}_i] &= \mathbb{E}[(z_i + \mathbf{c}_i^T \boldsymbol{\epsilon}_i)(z_i + \boldsymbol{\epsilon}_j^T \mathbf{c}_j)] \\
&= \mathbb{E}[z_i z_j + z_i \boldsymbol{\epsilon}_j^T \mathbf{c}_j + z_j \mathbf{c}_i^T \boldsymbol{\epsilon}_i + \mathbf{c}_i^T \boldsymbol{\epsilon}_j \boldsymbol{\epsilon}_i^T \mathbf{c}_j] \\
&= \mathbb{E}[z_i z_j] + c_j \, \mathbb{E}(z_i) \, \mathbb{E}[\boldsymbol{\epsilon}_j^T] + \mathbf{c}_i^T \, \mathbb{E}[z_j] \, \mathbb{E}[\boldsymbol{\epsilon}_i] + \mathbf{c}_i^T \, \mathbb{E}[\boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_j^T] \mathbf{c}_j \\
&= \mathbb{E}[z_i z_j] + \mathbf{c}_i^T \, \mathbb{E}[\boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_j^T] \mathbf{c}_j \\
&= \mathbb{E}[z_i z_j] + \sigma_i^2 1[i = j]
\end{aligned}$$

Thus,

$$\mathbb{E}[z_i z_j] = \mathbf{c}_i^T \, \mathbb{E}[\mathbf{x}_i \mathbf{x}_j^T] \mathbf{c}_j - \sigma_i^2 1[i = j]$$

for all $i, j = 1, \ldots, p$, which proves our claim. □

B.4 Proof of Lemma 3

Proof. By expanding the covariance matrix of the observable $\mathbf{x}_i$

$$\begin{aligned}
\mathbb{E}[\mathbf{x}_i \mathbf{x}_j^T] &= \mathbb{E}[(\mathbf{c}_i z_i + \boldsymbol{\epsilon}_i)(\mathbf{c}_j z_j + \boldsymbol{\epsilon}_j)^T] \\
&= \mathbb{E}[(\mathbf{c}_i z_i)(\mathbf{c}_j z_j)^T + \boldsymbol{\epsilon}(\mathbf{c}_i z_i)^T + (\mathbf{c}_i z_i) \boldsymbol{\epsilon}_j^T + \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_j^T] \\
&= \mathbb{E}[z_i z_j (\mathbf{c}_i \mathbf{c}_j^T)] + \mathbb{E}[\boldsymbol{\epsilon}_i] \, \mathbb{E}[(\mathbf{c}_j z_j)^T] + \mathbb{E}[\mathbf{c}_i z_i] \, \mathbb{E}[\boldsymbol{\epsilon}_j^T] + \mathbb{E}[\boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_j^T] \\
&= \mathbb{E}[z_i z_j (\mathbf{c}_i \mathbf{c}_j^T)] + \mathbb{E}[\boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_j^T] \\
&= \mathbb{E}[z_i z_j] \mathbf{c}_i \mathbf{c}_j^T + \sigma_i^2 \mathbf{I}_{q_i} 1[i = j] \\
&= \Sigma_{ij} \mathbf{c}_i \mathbf{c}_j^T + \sigma_i^2 \mathbf{I}_{q_i} 1[i = j]
\end{aligned}$$

which proves our claim. □

B.5 Proof of Lemma 4

Proof. Let λ and $\mathbf{y}$ be an eigenvalue and eigenvector of $\mathbf{M}_{ii}$. By the definition of eigenvalue and eigenvector, and the fact that any non-zero vectors are eigenvectors of the identity matrix (i.e., $\mathbf{c}_i$ can be an eigenvector of $\mathbf{I}$), we have:

$$\begin{aligned}
\lambda \mathbf{y} &= \mathbf{M}_{ii} \mathbf{y} \\
&= (\Sigma_{ii} \mathbf{c}_i \mathbf{c}_i^T + \sigma_i^2 \mathbf{I}_{q_i}) \mathbf{y}.
\end{aligned}$$

where the last step follows from Lemma 3. Clearly, if $\mathbf{y} \perp \mathbf{c}_i$ (or equivalently $\mathbf{y} \perp -\mathbf{c}_i$) then $\lambda = \sigma_i^2$. If $\mathbf{y} = \mathbf{c}_i$ or $\mathbf{y} = -\mathbf{c}_i$ then $\lambda = \Sigma_{ii} + \sigma_i^2$. Thus, the maximum eigenvector of $\mathbf{M}_{ii}$ is $\mathbf{c}_i$ or $-\mathbf{c}_i$ with corresponding eigenvalue $\Sigma_{ii} + \sigma_i^2$. The second maximum eigenvalue of $\mathbf{M}_{ii}$ is σ_i^2. □

B.6 Proof of Lemma 5

Proof. We can breakdown the divergence of sample covariance by adding and subtracting the same term and applying Holder's inequality, also bound them by the support so the constant does not scale with the number of

samples.

$$
\begin{aligned}
&|\widehat{F}_{ii}\widehat{\Sigma}_{ij}\widehat{F}_{jj} - \Sigma_{ij}| \\
&= |(\widehat{F}_{ii}\widehat{F}_{jj}\widehat{\mathbf{c}}_i^T\widehat{\mathbf{M}}_{ij}\widehat{\mathbf{c}}_j - \widehat{\sigma}_i^2 1[i=j]) - (\mathbf{c}_i^T\mathbf{M}_{ij}\mathbf{c}_j - \sigma_i^2 1[i=j])| \\
&\leq |\langle \widehat{\mathbf{M}}_{ij}, \widehat{F}_{ii}\widehat{F}_{jj}\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_j^T\rangle - \langle \mathbf{M}_{ij}, \mathbf{c}_i\mathbf{c}_j^T\rangle| + |\widehat{\sigma}_i^2 - \sigma_i^2|1[i=j] \\
&= |\langle \widehat{\mathbf{M}}_{ij}, \widehat{F}_{ii}\widehat{F}_{jj}\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_j^T\rangle - \langle \mathbf{M}_{ij}, \widehat{F}_{ii}\widehat{F}_{jj}\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_j^T\rangle + \langle \mathbf{M}_{ij}, \widehat{F}_{ii}\widehat{F}_{jj}\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_j^T\rangle - \langle \mathbf{M}_{ij}, \mathbf{c}_i\mathbf{c}_j^T\rangle| + |\widehat{\sigma}_i^2 - \sigma_i^2|1[i=j] \\
&= |\langle \widehat{\mathbf{M}}_{ij} - \mathbf{M}_{ij}, \widehat{F}_{ii}\widehat{F}_{jj}\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_j^T\rangle| + |\langle \mathbf{M}_{ij}, \widehat{F}_{ii}\widehat{F}_{jj}\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_j^T - \mathbf{c}_i\mathbf{c}_j^T\rangle| + |\widehat{\sigma}_i^2 - \sigma_i^2|1[i=j] \\
&\leq |\langle (\widehat{\mathbf{M}}_{ij} - \mathbf{M}_{ij})_{J_iJ_j}, \widehat{F}_{ii}\widehat{F}_{jj}(\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_j^T)_{J_iJ_j}\rangle| + |\langle (\mathbf{M}_{ij})_{J_iJ_j}, \widehat{F}_{ii}\widehat{F}_{jj}(\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_j^T - \mathbf{c}_i\mathbf{c}_j^T)_{J_iJ_j}\rangle| + |\widehat{\sigma}_i^2 - \sigma_i^2|1[i=j] \\
&\leq \left\|(\widehat{\mathbf{M}}_{ij} - \mathbf{M}_{ij})_{J_iJ_j}\right\|_\infty \left\|\widehat{F}_{ii}\widehat{F}_{jj}(\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_j^T)_{J_iJ_j}\right\|_1 + \left\|(\mathbf{M}_{ij})_{J_iJ_j}\right\|_\infty \left\|(\widehat{F}_{ii}\widehat{F}_{jj}\widehat{\mathbf{c}}\widehat{\mathbf{c}}_j^T - \mathbf{c}_i\mathbf{c}_j^T)_{J_iJ_j}\right\|_1 + |\widehat{\sigma}_i^2 - \sigma_i^2|1[i=j] \\
&\leq \left\|\widehat{\mathbf{M}}_{ij} - \mathbf{M}_{ij}\right\|_\infty \left\|(\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_j^T)_{J_iJ_j}\right\|_1 + \left\|(\mathbf{M}_{ij})_{J_iJ_j}\right\|_\infty \left\|(\widehat{F}_{ii}\widehat{F}_{jj}\widehat{\mathbf{c}}\widehat{\mathbf{c}}_j^T - \mathbf{c}_i\mathbf{c}_j^T)_{J_iJ_j}\right\|_1 + |\widehat{\sigma}_i^2 - \sigma_i^2|1[i=j]
\end{aligned}
$$

where the second to the last step follows by Holder inequality. To bound the factor for the first term, recall that $|J_i| \leq s$ and $\|\widehat{\mathbf{c}}_i\|_2 = 1$ for all $i = 1, \ldots, p$. Thus

$$
\begin{aligned}
\left\|(\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_j^T)_{J_iJ_j}\right\|_1 &\leq s\left\|(\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_j^T)_{J_iJ_j}\right\|_{\mathcal{F}} \\
&\leq s\left\|\widehat{\mathbf{c}}_i\right\|_2\left\|\widehat{\mathbf{c}}_j\right\|_2 \\
&= s
\end{aligned}
$$

From Lemma 3 and since and $\|\mathbf{c}_i\|_2 = 1$ for all $i = 1, \ldots, p$, we can bound the constant for the second term.

$$
\begin{aligned}
\left\|(\mathbf{M}_{ij})_{J_iJ_j}\right\|_\infty &= \left\|\Sigma_{ij}(\mathbf{c}_i\mathbf{c}_j^T)_{J_iJ_j} + 1[i=j]\sigma_i^2\mathbf{I}_{|J_i|}\right\|_\infty \\
&\leq \Sigma_{ij}\left\|(\mathbf{c}_i\mathbf{c}_j^T)_{J_iJ_j}\right\|_\infty + 1[i=j]\sigma_i^2 \\
&\leq \Sigma_{ij}\left\|(\mathbf{c}_i\mathbf{c}_j^T)_{J_iJ_j}\right\|_{\mathcal{F}} + 1[i=j]\sigma_i^2 \\
&\leq \Sigma_{ij}\left\|\mathbf{c}_i\right\|_2\left\|\mathbf{c}_j\right\|_2 + 1[i=j]\sigma_i^2 \\
&\leq \Sigma_{ij} + 1[i=j]\sigma_i^2
\end{aligned}
$$

and we prove our claim. □

B.7 Proof of Lemma 6

Proof. Let

$$
\mathcal{M} = \begin{bmatrix} \mathbf{M}_{11} & \ldots & \mathbf{M}_{1p} \\ \vdots & \ddots & \vdots \\ \mathbf{M}_{p1} & \ldots & \mathbf{M}_{pp} \end{bmatrix} \quad , \quad \widehat{\mathcal{M}} = \begin{bmatrix} \widehat{\mathbf{M}}_{11} & \ldots & \widehat{\mathbf{M}}_{1p} \\ \vdots & \ddots & \vdots \\ \widehat{\mathbf{M}}_{p1} & \ldots & \widehat{\mathbf{M}}_{pp} \end{bmatrix}
$$

and $X = (\mathbf{x}_i, ..., \mathbf{x}_p) \in \mathbb{R}^Q$ where $Q = \sum_{i=1}^p q_i$. If a random vector Y is said to be sub-Gaussian with parameter γ, then $\mathbb{E}[e^{tY}] \leq e^{\frac{\gamma^2t^2}{2}}$, for all $t \in \mathbb{R}$ Let X_l be the l-th element of X, by definition 1 we have $X_l = (\mathbf{x}_i)_m = (\mathbf{c}_i)_m z_i + (\boldsymbol{\epsilon}_i)_m$, where m is the corresponding index for lth element of X in the observed variable i.

$$
\begin{aligned}
\mathbb{E}[e^{tX_l}] &= \mathbb{E}[e^{t(\mathbf{x}_i)_m}] \\
&= \mathbb{E}[e^{t(\mathbf{c}_i)_m z_i + (\boldsymbol{\epsilon}_i)_m}] \\
&= \mathbb{E}[e^{t(\mathbf{c}_i)_m z_i} e^{t(\boldsymbol{\epsilon}_i)_m}] \\
&= \mathbb{E}[e^{t(\mathbf{c}_i)_m z_i}]\,\mathbb{E}[e^{t(\boldsymbol{\epsilon}_i)_m}] \text{ since } z_i, \boldsymbol{\epsilon}_i \text{ are independent} \\
&\leq e^{\frac{(\gamma\Sigma_{ll})^2(t(\mathbf{c}_i)_m)^2}{2}} e^{\frac{(\gamma\sigma_i^2)^2t^2}{2}} \\
&\leq e^{\frac{(\gamma\Sigma_{ll})^2t^2}{2}} e^{\frac{(\gamma\sigma_i^2)^2t^2}{2}} \text{ since } \left\|\mathbf{c}_i\right\|_2 = 1 \\
&\leq e^{\frac{(\gamma(\Sigma_{ll}+\sigma_i^2))^2t^2}{2}}
\end{aligned}
$$

We can say that $X_l/\sqrt{\mathcal{M}_{ll}}$ is sub-Gaussian with parameter γ, then by invoking Lemma 1 in (Ravikumar et al., 2011), we have for any entries i, j

$$\mathbb{P}\left[|\widehat{\mathcal{M}}_{lm} - \mathcal{M}_{lm}| > \Delta\right] \leq 4\exp\left(-\frac{n\Delta^2}{128(1+8\gamma^2)^2 \max_l(\mathcal{M}_{ll})^2}\right)$$

for all $\Delta \in (0, \max_l(\mathcal{M}_{ll})8(1+8\gamma^2))$. Since $\mathcal{M}, \widehat{\mathcal{M}} \in \mathbb{R}^{Q\times Q}$, by union bound

$$\mathbb{P}\left[(\exists i,j)\ \left\|\widehat{\mathbf{M}}_{ij} - \mathbf{M}_{ij}\right\|_\infty > \Delta\right] \leq 4Q^2\exp\left(-\frac{n\Delta^2}{128(1+8\gamma^2)^2 \max_l(\mathcal{M}_{ll})^2}\right)$$

Since $\mathbf{M}_{ij}$ and $\widehat{\mathbf{M}}_{ij}$ are submatrices of $\mathcal{M}$ and $\widehat{\mathcal{M}}$ respectively, the above bound can be applied to $\mathbf{M}_{ij}$ and $\widehat{\mathbf{M}}_{ij}$. Furthermore, from Lemma 3, we $\mathbf{M}_{ii} = \Sigma_{ii}\mathbf{c}_i\mathbf{c}_i^T + \sigma_i^2\mathbf{I}_{q_i}$ which has a maximum possible diagonal entry $\Sigma_{ii} + \sigma_i^2$. Thus

$$\mathbb{P}\left[(\exists i,j)\ \left\|\mathbf{M}_{ij} - \widehat{\mathbf{M}}_{ij}\right\|_\infty > \Delta\right] \leq 4Q^2\exp\left(-\frac{n\Delta^2}{128(1+8\gamma^2)^2(\max_i \Sigma_{ii} + \sigma_i^2)^2}\right) = \delta$$

or equivalently

$$\mathbb{P}\left[(\forall i,j)\ \left\|\mathbf{M}_{ij} - \widehat{\mathbf{M}}_{ij}\right\|_\infty \leq \Delta\right] \geq 1 - 4Q^2\exp\left(-\frac{n\Delta^2}{128(1+8\gamma^2)^2(\max_i \Sigma_{ii} + \sigma_i^2)^2}\right) = 1-\delta$$

Solving for Δ above, we get $\Delta = O\left(\frac{(1+8\gamma^2)(\max_i \Sigma_{ii}+\sigma_i^2)}{\sqrt{n}}(\log Q + \log(1/\delta))\right)$. □

B.8 Proof of Lemma 7

Proof. As shown in Theorem 2 of (Lei and Vu, 2015), with a proper regularization setting $\rho = O(\sqrt{(\log q_i)/n})$, the solution to sparse PCA in eq.(2) is $\widehat{\mathbf{H}}_i = \widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_i^T$ fulfills $\text{supp}(\widehat{\mathbf{c}}_i) = J_i \equiv \text{supp}(\mathbf{c}_i)$. This proves our first claim. Further, Lemma 2 in (Lei and Vu, 2015) shows that

$$\left\|(\mathbf{c}_i\mathbf{c}_i^T - \widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_i^T)_{J_iJ_i}\right\|_{\mathcal{F}} \leq \frac{4\rho s}{\lambda_1(\mathbf{M}_{ii}) - \lambda_2(\mathbf{M}_{ii})}$$

From Lemma 4 we have that $\lambda_1(\mathbf{M}_{ii}) = \Sigma_{ii} + \sigma_i^2$ and $\lambda_2(\mathbf{M}_{ii}) = \sigma_i^2$ and thus (Lei and Vu, 2015), we have

$$\begin{aligned}\left\|(\mathbf{c}_i\mathbf{c}_i^T - \widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_i^T)_{J_iJ_i}\right\|_{\mathcal{F}} &\leq \frac{4\rho s}{\Sigma_{ii}} \\ &= O\left(\frac{s\sqrt{\log q_i}}{\Sigma_{ii}\sqrt{n}}\right)\end{aligned}$$

which proves our second claim. For the third claim, by invoking Lemma 2 in (Lei and Vu, 2015) again, we have

$$\begin{aligned}\left\|(\widehat{F}_{ii}\widehat{\mathbf{c}}_i - \mathbf{c}_i)_{J_i}\right\|_2^2 &= \left\|(\mathbf{c}_i\mathbf{c}_i^T - \widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_i^T)_{J_iJ_i}\right\|_{\mathcal{F}} \\ &\leq \frac{4\rho s}{\lambda_1(\mathbf{M}_{ii}) - \lambda_2(\mathbf{M}_{ii})}\end{aligned}$$

From the above and since $|J_i| \leq s$ and $\|\mathbf{c}_i\|_2 = 1$ for all $i = 1, \ldots, p$. Thus

$$\begin{aligned}
\left\|(\widehat{F}_{ii}\widehat{F}_{jj}\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_j^T - \mathbf{c}_i\mathbf{c}_j^T)_{J_iJ_j}\right\|_1 &= \left\|(\widehat{F}_{ii}\widehat{F}_{jj}\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_j^T - \widehat{F}_{ii}\widehat{\mathbf{c}}_i\mathbf{c}_j^T + \widehat{F}_{ii}\widehat{\mathbf{c}}_i\mathbf{c}_j^T - \mathbf{c}_i\mathbf{c}_j^T)_{J_iJ_j}\right\|_1 \\
&\leq \left\|(\widehat{F}_{ii}\widehat{F}_{jj}\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_j^T - \widehat{F}_{ii}\widehat{\mathbf{c}}_i\mathbf{c}_j^T)_{J_iJ_j}\right\|_1 + \left\|(\widehat{F}_{ii}\widehat{\mathbf{c}}_i\mathbf{c}_j^T - \mathbf{c}_i\mathbf{c}_j^T)_{J_iJ_j}\right\|_1 \\
&= \left\|(\widehat{F}_{ii}\widehat{\mathbf{c}}_i(\widehat{F}_{jj}\widehat{\mathbf{c}}_j^T - \mathbf{c}_j^T))_{J_iJ_j}\right\|_1 + \left\|((\widehat{F}_{ii}\widehat{\mathbf{c}}_i - \mathbf{c}_i)\mathbf{c}_j^T)_{J_iJ_j}\right\|_1 \\
&\leq s\left\|(\widehat{F}_{ii}\widehat{\mathbf{c}}_i(\widehat{F}_{jj}\widehat{\mathbf{c}}_j^T - \mathbf{c}_j^T))_{J_iJ_j}\right\|_{\mathcal{F}} + s\left\|((\widehat{F}_{ii}\widehat{\mathbf{c}}_i - \mathbf{c}_i)\mathbf{c}_j^T)_{J_iJ_j}\right\|_{\mathcal{F}} \\
&\leq s\left\|\widehat{\mathbf{c}}_i\right\|_2\left\|(\widehat{F}_{jj}\widehat{\mathbf{c}}_j^T - \mathbf{c}_j^T))_{J_j}\right\|_2 + s\left\|(\widehat{F}_{ii}\widehat{\mathbf{c}}_i - \mathbf{c}_i)_{J_i}\right\|_2\left\|\mathbf{c}_j\right\|_2 \\
&\leq O\left(\frac{s^{3/2}}{\sqrt{n}}\left(\frac{\sqrt{\log q_i}}{\sqrt{\lambda_1(\mathbf{M}_{ii}) - \lambda_2(\mathbf{M}_{ii})}} + \frac{\sqrt{\log q_i}}{\sqrt{\lambda_1(\mathbf{M}_{jj}) - \lambda_2(\mathbf{M}_{jj})}}\right)\right) \\
&= O\left(\frac{s^{3/2}}{\sqrt{n}}\left(\frac{\sqrt{\log q_i}}{\sqrt{\Sigma_{ii}}} + \frac{\sqrt{\log q_i}}{\sqrt{\Sigma_{jj}}}\right)\right) \\
&\leq O\left(\frac{s^{3/2}}{\sqrt{n}}\max_i\frac{\sqrt{\log q_i}}{\sqrt{\Sigma_{ii}}}\right)
\end{aligned}$$

which proves our claim. □

B.9 Proof of Lemma 8

Proof. Since the maximum eigenvalue of a matrix is also the spectral norm for the matrix, then by triangular inequality

$$\begin{aligned}
|\widehat{\sigma}_i^2 - \sigma_i^2| &= |\lambda_1((\widehat{\mathbf{M}}_{ii})_{J_iJ_i} - \widehat{\phi}_i(\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_i^T)_{J_iJ_i}) - \lambda_1((\mathbf{M}_{ii})_{J_iJ_i} - \phi_i(\mathbf{c}_i\mathbf{c}_i^T)_{J_iJ_i})| \\
&\leq |\lambda_1((\widehat{\mathbf{M}}_{ii})_{J_iJ_i} - \widehat{\phi}_i(\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_i^T)_{J_iJ_i} - ((\mathbf{M}_{ii})_{J_iJ_i} - \phi_i(\mathbf{c}_i\mathbf{c}_i^T)_{J_iJ_i}))|
\end{aligned}$$

Futhermore,

$$\begin{aligned}
&\lambda_1((\widehat{\mathbf{M}}_{ii})_{J_iJ_i} - \widehat{\phi}_i(\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_i^T)_{J_iJ_i} - ((\mathbf{M}_{ii})_{J_iJ_i} - \phi_i(\mathbf{c}_i\mathbf{c}_i^T)_{J_iJ_i})) \\
&= \lambda_1((\widehat{\mathbf{M}}_{ii} - \mathbf{M}_{ii})_{J_iJ_i} + (\phi_i\mathbf{c}_i\mathbf{c}_i^T - \widehat{\phi}_i\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_i^T)_{J_iJ_i}) \\
&\leq \lambda_1((\widehat{\mathbf{M}}_{ii} - \mathbf{M}_{ii})_{J_iJ_i}) + \lambda_1((\phi_i\mathbf{c}_i\mathbf{c}_i^T - \widehat{\phi}_i\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_i^T + \phi_i\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_i^T - \phi_i\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_i^T)_{J_iJ_i}) \\
&= \lambda_1((\widehat{\mathbf{M}}_{ii} - \mathbf{M}_{ii})_{J_iJ_i}) + \lambda_1(((\phi_i - \widehat{\phi}_i)\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_i^T + \phi_i(\mathbf{c}_i\mathbf{c}_i^T - \widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_i^T))_{J_iJ_i}) \\
&\leq \lambda_1((\widehat{\mathbf{M}}_{ii} - \mathbf{M}_{ii})_{J_iJ_i}) + \lambda_1((\phi_i - \widehat{\phi}_i)(\widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_i^T)_{J_iJ_i}) + \lambda_1(\phi_i(\mathbf{c}_i\mathbf{c}_i^T - \widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_i^T)_{J_iJ_i}) \\
&\leq \lambda_1((\widehat{\mathbf{M}}_{ii} - \mathbf{M}_{ii})_{J_iJ_i}) + \max(\phi_i - \widehat{\phi}_i, 0) + \phi_i\lambda_1((\mathbf{c}_i\mathbf{c}_i^T - \widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_i^T)_{J_iJ_i}) \\
&\leq O\left(\gamma(\Sigma_{ii} + \sigma_i^2)\min\left(\sqrt{\frac{s + \log p + \log(1/\delta)}{n}}, \frac{s + \log p + \log(1/\delta)}{n}\right)\right) + |\phi_i - \widehat{\phi}_i| + \phi_i\left\|(\mathbf{c}_i\mathbf{c}_i^T - \widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_i^T)_{J_iJ_i}\right\|_{\mathcal{F}}
\end{aligned}$$

where the last step holds with probability at least $1 - p$, from Theorem 7.1 from Rinaldo (Rinaldo, 2017) and a union bound across all $i = 1, \ldots, p$. Finally,

$$\begin{aligned}
&|\widehat{\sigma}_i^2 - \sigma_i^2| \\
&\leq O\left(\gamma(\Sigma_{ii} + \sigma_i^2)\min\left(\sqrt{\frac{s + \log p + \log(1/\delta)}{n}}, \frac{s + \log p + \log(1/\delta)}{n}\right)\right) + |\phi_i - \widehat{\phi}_i| + \phi_i\left\|(\mathbf{c}_i\mathbf{c}_i^T - \widehat{\mathbf{c}}_i\widehat{\mathbf{c}}_i^T)_{J_iJ_i}\right\|_{\mathcal{F}} \\
&\leq O\left(\gamma(\Sigma_{ii} + \sigma_i^2)\min\left(\sqrt{\frac{s + \log p + \log(1/\delta)}{n}}, \frac{s + \log p + \log(1/\delta)}{n}\right)\right) + |\phi_i - \widehat{\phi}_i| + O\left(\phi_i\frac{s\sqrt{\log q_i}}{\Sigma_{ii}\sqrt{n}}\right)
\end{aligned}$$

where we invoked Lemma 7 in the last step. □

B.10 Proof of Lemma 9

Proof. We have

$$\begin{aligned}
|\widehat{\phi}_i - \phi_i| &= |\widehat{\mathbf{c}}_i \widehat{\mathbf{M}}_{ii} \widehat{\mathbf{c}}_i^T - \mathbf{c}_i \mathbf{M}_{ii} \mathbf{c}_i^T| \\
&= |\langle \widehat{\mathbf{M}}_{ii}, \widehat{\mathbf{c}}_i \widehat{\mathbf{c}}_i^T \rangle - \langle \mathbf{M}_{ii}, \mathbf{c}_i \mathbf{c}_i^T \rangle| \\
&= |\langle \widehat{\mathbf{M}}_{ii}, \widehat{\mathbf{c}}_i \widehat{\mathbf{c}}_i^T \rangle - \langle \mathbf{M}_{ii}, \mathbf{c}_i \mathbf{c}_i^T \rangle + \langle \mathbf{M}_{ii}, \widehat{\mathbf{c}}_i \widehat{\mathbf{c}}_i^T \rangle - \langle \mathbf{M}_{ii}, \widehat{\mathbf{c}}_i \widehat{\mathbf{c}}_i^T \rangle| \\
&= |\langle \widehat{\mathbf{M}}_{ii} - \mathbf{M}_{ii}, \widehat{\mathbf{c}}_i \widehat{\mathbf{c}}_i^T \rangle + \langle \mathbf{M}_{ii}, \widehat{\mathbf{c}}_i \widehat{\mathbf{c}}_i^T - \mathbf{c}_i \mathbf{c}_i^T \rangle| \\
&\leq |\langle (\widehat{\mathbf{M}}_{ii} - \mathbf{M}_{ii})_{J_i J_i}, (\widehat{\mathbf{c}}_i \widehat{\mathbf{c}}_i^T)_{J_i J_i} \rangle + \langle (\mathbf{M}_{ii})_{J_i J_i}, (\widehat{\mathbf{c}}_i \widehat{\mathbf{c}}_i^T - \mathbf{c}_i \mathbf{c}_i^T) J \rangle| \\
&\leq \left\| \widehat{\mathbf{M}}_{ii} - \mathbf{M}_{ii} \right\|_\infty \left\| (\widehat{\mathbf{c}}_i \widehat{\mathbf{c}}_i^T)_{J_i J_i} \right\|_1 + \left\| (\mathbf{M}_{ii})_{J_i J_i} \right\|_\infty \left\| (\widehat{\mathbf{c}}_i \widehat{\mathbf{c}}_i^T - \mathbf{c}_i \mathbf{c}_i^T)_{J_i J_i} \right\|_1 \\
&\leq \left\| \widehat{\mathbf{M}}_{ii} - \mathbf{M}_{ii} \right\|_\infty \left\| (\widehat{\mathbf{c}}_i \widehat{\mathbf{c}}_i^T)_{J_i J_j} \right\|_1 + + O\left(\left\| (\mathbf{M}_{ii})_{J_i J_i} \right\|_\infty \frac{s^{3/2}}{\sqrt{n}} \max_i \frac{\sqrt{\log q_i}}{\sqrt{\Sigma_{ii}}} \right)
\end{aligned}$$

where we invoked Lemma 7 in the last step. To bound the factor for the first term, recall that $|J_i| \leq s$ and $\left\| \widehat{\mathbf{c}}_i \right\|_2 = 1$ for all $i = 1, \ldots, p$. Thus

$$\begin{aligned}
\left\| (\widehat{\mathbf{c}}_i \widehat{\mathbf{c}}_i^T)_{J_i J_i} \right\|_1 &\leq s \left\| (\widehat{\mathbf{c}}_i \widehat{\mathbf{c}}_i^T)_{J_i J_i} \right\|_{\mathcal{F}} \\
&\leq s \left\| \widehat{\mathbf{c}}_i \right\|_2^2 \\
&= s
\end{aligned}$$

From Lemma 3 and since and $\left\| \mathbf{c}_i \right\|_2 = 1$ for all $i = 1, \ldots, p$, we can bound the constant for the second term.

$$\begin{aligned}
\left\| (\mathbf{M}_{ii})_{J_i J_i} \right\|_\infty &= \left\| \Sigma_{ii} (\mathbf{c}_i \mathbf{c}_i^T)_{J_i J_i} + \sigma_i^2 \mathbf{I}_{|J_i|} \right\|_\infty \\
&\leq \Sigma_{ii} \left\| (\mathbf{c}_i \mathbf{c}_i^T)_{J_i J_i} \right\|_\infty + \sigma_i^2 \\
&\leq \Sigma_{ii} \left\| (\mathbf{c}_i \mathbf{c}_i^T)_{J_i J_i} \right\|_{\mathcal{F}} + \sigma_i^2 \\
&\leq \Sigma_{ii} \left\| \mathbf{c}_i \right\|_2^2 + \sigma_i^2 \\
&\leq \Sigma_{ii} + \sigma_i^2
\end{aligned}$$

and we prove our claim. □

C Additional Experiments

Here we show additional experimental results that were not included in the main text due to space constraints.

C.1 Synthetic Data

In Figure 5 we aim to validate our intermediate result in Lemma 7. Thus, we report the empirical probability of correct support recovery, i.e., $\text{supp}(\widehat{\mathbf{c}}_i) = \text{supp}(\mathbf{c}_i)$. We observe that more samples allows to obtain a better support recovery of the sparse vectors $\mathbf{c}_i$'s. This is consistent with Lemma 7. The chart also shows the characteristic phase transition behavior: the probability of exact recovery is essentially 0% at very small sample sizes, but climbs rapidly to 100% once the sample size exceeds a critical threshold.

In Figure 6 we aim to validate our intermediate result in Lemma 8. Thus, we report the average relative deviation $|\widehat{\sigma}_i^2 - \sigma_i^2| / \sigma_i^2$. We observe that more samples allows to obtain better estimates for the noise variances σ_i^2's. This is consistent with Lemma 8.

C.2 Real-World Data

Figure 7 shows our estimated precision matrix from the *1000 functional connectomes* dataset. We remind the reader that our method obtained 98.1% precision and 98.1% recall for support recovery.

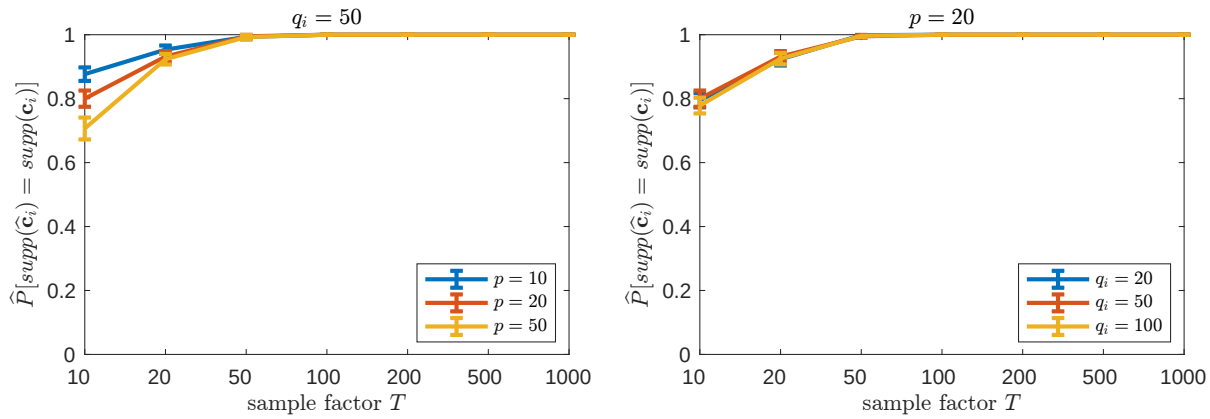

Figure 5: Empirical probability of correct support recovery $\text{supp}(\widehat{\mathbf{c}}_i) = \text{supp}(\mathbf{c}_i)$ with varying number of latent variables p (Left) and varying observed dimensions q_i (Right) versus number of samples $n = T \log Q$. We show standard error bars at 95% confidence level. Note that in both cases, the more samples the better the support recovery of the sparse vectors $\mathbf{c}_i$'s is.

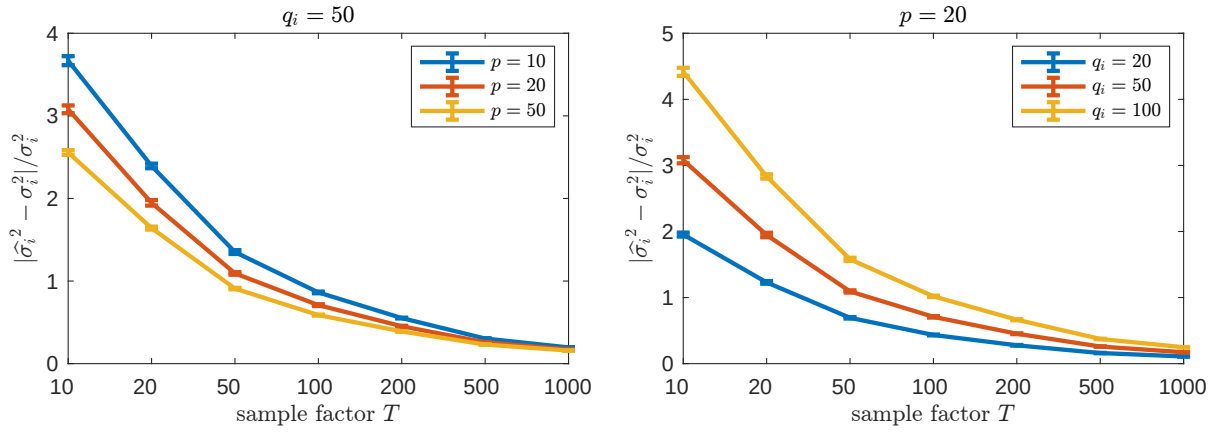

Figure 6: Average relative deviation $|\widehat{\sigma}_i^2 - \sigma_i^2|/\sigma_i^2$ with varying number of latent variables p (Left) and varying observed dimensions q_i (Right) versus number of samples $n = T \log Q$. We show standard error bars at 95% confidence level. Note that in both cases, the more samples the better the estimates of the noise variances σ_i^2's are.

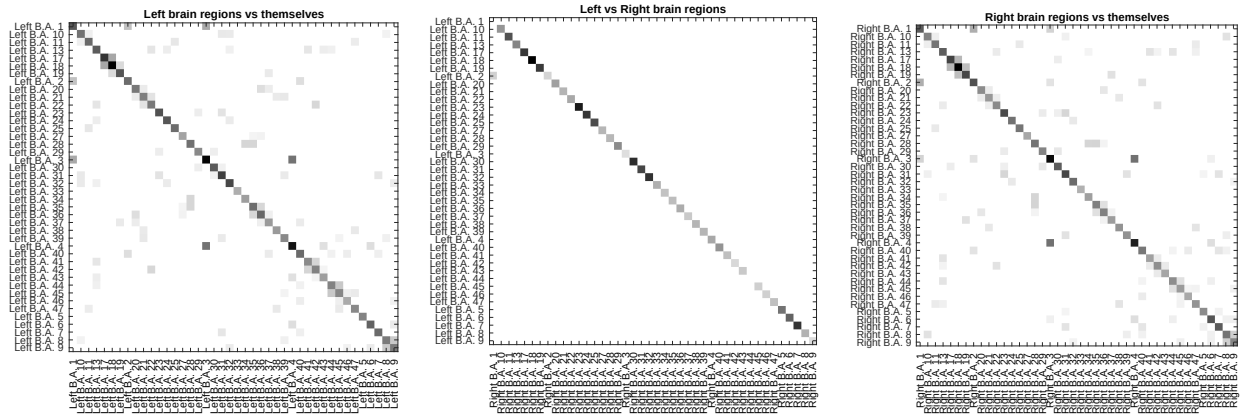

Figure 7: Our estimated precision matrix from the *1000 functional connectomes* dataset from 84 Brodmann areas which act as latent variables. Left: precision submatrix showing left brain regions versus themselves. Center: precision submatrix showing left versus right brain regions. Right: precision submatrix showing right brain regions versus themselves.