HHS Public Access

Author manuscript

*Electron J Stat*. Author manuscript; available in PMC 2022 July 01.

Published in final edited form as:

*Electron J Stat*. 2021 ; 15(2): 4192–4235. doi:10.1214/21-ejs1887.

# Principal regression for high dimensional covariance matrices

**Yi Zhao**,

Department of Biostatistics and Health Data Science, Indiana University School of Medicine

**Brian Caffo**,

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

**Xi Luo**,

Department of Biostatistics and Data Science, The University of Texas Health Science Center at Houston

**Alzheimer's Disease Neuroimaging Initiative**[†]

## Abstract

This manuscript presents an approach to perform generalized linear regression with multiple high dimensional covariance matrices as the outcome. In many areas of study, such as resting-state functional magnetic resonance imaging (fMRI) studies, this type of regression can be utilized to characterize variation in the covariance matrices across units. Model parameters are estimated by maximizing a likelihood formulation of a generalized linear model, conditioning on a well-conditioned linear shrinkage estimator for multiple covariance matrices, where the shrinkage coefficients are proposed to be shared across matrices. Theoretical studies demonstrate that the proposed covariance matrix estimator is optimal achieving the uniformly minimum quadratic loss asymptotically among all linear combinations of the identity matrix and the sample covariance matrix. Under certain regularity conditions, the proposed estimator of the model parameters is consistent. The superior performance of the proposed approach over existing methods is illustrated through simulation studies. Implemented to a resting-state fMRI study acquired from the Alzheimer's Disease Neuroimaging Initiative, the proposed approach identified a brain network within which functional connectivity is significantly associated with Apolipoprotein E $\varepsilon 4$, a strong genetic marker for Alzheimer's disease.

### Keywords

Covariance matrix estimation; generalized linear regression; heteroscedasticity; shrinkage estimator

### MSC2020 subject classifications:

Primary 62J99; secondary 62H99

yz125@iu.edu .

[†]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete list of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

## 1. Introduction

In this manuscript, we study a regression problem with covariance matrices as the outcome under a high dimensional setting. Suppose $\mathbf{y}_{it} \in \mathbb{R}^p$ is a $p$-dimensional random vector, which is the $t$th acquisition from subject $i$, for $t = 1, \ldots, T_i$ and $i = 1, \ldots, n$, where $T_i$ is the number of observations of subject $i$ and $n$ is the number of subjects. Let $T_{\max} = \max_i T_i$. The term "high dimensionality" refers to the scenario when $T_{\max} \ll p$ and $p$ increases to infinity. The data, $\mathbf{y}_{it}$, are assumed to follow a normal distribution with covariance matrix $\Sigma_i$. Here, without loss of generality, it is assumed that the distribution mean is zero as the study interest focuses on the covariance matrices. Let $\mathbf{x}_i \in \mathbb{R}^q$ denote the $q$-dimensional covariates of interest acquired from subject $i$. For the covariance matrices, we assume the following regression model. For $i = 1, \ldots, n$, the data heteroscedasticity satisfies the following generalized linear regression model with a logarithmic link function,

$$\log\left(\boldsymbol{\gamma}^\top \Sigma_i \boldsymbol{\gamma}\right) = \mathbf{x}_i^\top \boldsymbol{\beta}, \tag{1.1}$$

where $\boldsymbol{\gamma} \in \mathbb{R}^p$ is a linear projection and $\boldsymbol{\beta} \in \mathbb{R}^q$ is the model coefficient. In $\mathbf{x}_i$, the first element is set to one to include the intercept term. Using a logarithmic link function, it is guaranteed that $\Sigma_i$'s are positive semi-definite. The goal is to estimate $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ using the observed data $\left\{\left(\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT_i}\right), \mathbf{x}_i\right\}_{i=1}^n$. In Model (1.1), $\boldsymbol{\gamma}$ is an unknown linear projection to be estimated such that the characteristic of the covariance matrices can be best captured by the covariates of interest.

One application of such a regression problem is to analyze covariate associated variations in brain coactivation in a functional magnetic resonance imaging (fMRI) study, where covariance/correlation matrices of the fMRI signals are generally utilized to reveal the coactivation patterns. Characterizing these patterns with population/individual covariates is of great interest in neuroimaging studies [34, 41]. Another example is the study of financial equities data. Considering a pool of stock values, covariance matrices over a period of time capture the comovement or synchronicity of the stocks. Firm and market-level information, such as industry type, firm's cash flow, stock size, and book-to-market ratio, plays an essential role in determining the synchronicity. Quantifying such association is an important topic in financial theory [43].

Assuming $T_{\min} = \min_i T_i > p$ and $p$ is fixed, Zhao et al. [41] first studied Model (1.1) and proposed to estimate $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ through a likelihood-based approach minimizing the negative log-likelihood function in the projection space. One sufficient condition to solve the likelihood-based criterion is that the sample covariance matrices are positive definite. Thus, the likelihood estimator is ill-posed when $T_{\max} < p$ as the sample covariance matrices are rank-deficient. Additionally, it has been shown that when $p$ increases, the sample covariance matrix performs poorly and can lead to invalid conclusions. For example, the largest eigenvalue of the sample covariance matrix is not a consistent estimator, and the eigenvectors can be nearly orthogonal to the truth [22]. To circumvent difficulties raised by the high dimensionality, one solution is to impose structural assumptions, such as bandable

covariance matrices, sparse covariance matrices, spiked covariance matrices, covariances with a tensor product structure, and latent graphical models [see a review of 6, and references therein]. Based on structural assumptions, many regularization-based methods have been developed. However, most of these methods produce covariance estimates that may not always be positive definite (numerically), and this creates subsequent numerical convergence issues when the quadratic product with $\Sigma_i$ is negative in (1.1). Moreover, most regularization methods can be computationally expensive on finding the solution and may require searching over different regularization parameters, not to mention the computational costs increase multiplicatively when computing over multiple covariance matrices. Further research is also needed to evaluate what structural assumptions are most appropriate for fMRI data. Another class of high-dimensional covariance matrix estimator is the shrinkage estimator. Daniels and Kass [11] considered two shrinkage estimators of the covariance matrix, a correlation shrinkage and a rotation shrinkage, offering a compromise between completely unstructured and structured estimators to improve the robustness. Ledoit and Wolf [24] introduced a well-conditioned estimator of the covariance matrix, which is an optimal linear combination of the identity matrix and the sample covariance matrix under squared error loss. This estimator is guaranteed to be positive definite and is easy to compute based on a simple and explicit formula. These advantages make it desirable for formulating the proposed estimator. Instead of a linear combination, Ledoit and Wolf [25] extended this work to nonlinear transformations of the sample eigenvalues and presented a way of finding the transformation that is asymptotically equivalent to the oracle linear combination. Based on Tyler's robust $M$-estimator [37] and the linear shrinkage estimator [24], Chen et al. [8] and Pascal et al. [28], in parallel, introduced robust estimators of covariance matrices for elliptical distributed samples.

To model multiple covariance matrices, procedures include regression-type approaches [1, 9, 21, 14, 43]; (common) principal component analysis related methods [13, 5, 20, 15]; and methods based on other types of matrix decomposition, such as the Cholesky decomposition [30]. Among these, Fox and Dunson [14] introduced a scalable nonparametric covariance regression model applying low-rank approximation. Franks and Hoff [15] generalized a Bayesian hierarchical model studying the heterogeneity in the covariance matrices to high dimensional settings. Assuming that the ideal covariance structure exists in the eigenspace of the data covariance matrix, Chen et al. [7] introduced a regression-based approach to remove the scanner effects in covariance achieving the goal of harmonization. Compared to the above-mentioned approaches, Model (1.1) offers higher flexibility in modeling the relationship with the covariates. For example, $x$ can be either continuous or categorical, and one can easily include interactions and/or polynomials of the covariates.

In the high dimensional setting considered in this study, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, as well as $n$ covariance matrices, will be estimated under Model (1.1). It is well known that the eigenvalues of the sample covariance matrix are more dispersed than the truth [24]. The class of linear combinations of the identity and sample covariance matrix corrects this dispersion issue by shrinking towards the identity matrix. The choice of the identity matrix can also be interpreted as a prior without strong structural assumptions or prior knowledge. Interestingly, it will be shown that estimating each covariance matrix separately, such

as using the shrinkage estimator proposed in Ledoit and Wolf [24], leads to suboptimal estimation accuracy for $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, and $\Sigma_i$'s. Thus, we propose a linear shrinkage estimator of all the covariance matrices jointly, of which the shrinkage coefficients are shared across matrices. In addition, it is shown that the proposed shrinkage estimator leads to a consistent estimator of model coefficients. We first replace the sample covariance formulation with the proposed shrinkage estimator, and then estimate ($\boldsymbol{\gamma}$, $\boldsymbol{\beta}$) through maximizing a plug-in likelihood evaluated at the shrinkage estimator. In fMRI studies, shrinkage is also a popular technique to improve the reliability of subject-level functional connectivity captured by the covariance matrix. In the technique, when estimating individual covariance matrix, population-level information is borrowed as prior knowledge [38, 31, 27, 32, 29].

The framework proposed in this manuscript has three major contributions.

1. This paper first studies a joint shrinkage estimator for multiple high dimensional covariance matrices, generalizing the linear shrinkage estimator for a single covariance matrix [24]. We show that the latter approach is suboptimal compared to the proposed joint covariance shrinkage estimator, where the shrinkage coefficients are shared across multiple matrices. Within this class of shrinkage estimators, we believe that this is among the first attempts to analyze the variations of a large number of covariance matrices associated with covariates in a regression setting under certain model assumptions.
2. The proposed shrinkage estimator of the covariance matrices is well conditioned and has uniformly minimum quadratic risk asymptotically among all linear combinations (Theorem 3.3).
3. Under certain regularity conditions, the proposed approach achieves consistent estimators of the parameters in Model (1.1) (Proposition 3.1).

The rest of the paper is organized as the following. Section 2 introduces the proposed shrinkage estimator of the covariance matrices and the pseudo-likelihood based method of estimating $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. Section 3 studies the asymptotic properties. In Section 4, the superior performance of the proposed approach over existing methods is demonstrated through simulation studies. Section 5 articulates an application to a resting-state fMRI data set acquired from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Section 6 concludes this paper with discussions. Technical proofs are collected in the appendix.

## 2. Methods

Considering the regression model (1.1), it is proposed to estimate the parameters by solving the following optimization problem.

$$\begin{aligned} \underset{(\boldsymbol{\beta},\boldsymbol{\gamma})}{\text{minimize}} \quad & \ell(\boldsymbol{\beta},\boldsymbol{\gamma}) = \frac{1}{2}\sum_{i=1}^{n} T_i\left\{\mathbf{x}_i^\top\boldsymbol{\beta} + \boldsymbol{\gamma}^\top\widehat{\Sigma}_i\boldsymbol{\gamma}\cdot\exp\left(-\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}, \\ \text{such that} \quad & \boldsymbol{\gamma}^\top\mathbf{H}\boldsymbol{\gamma} = 1, \end{aligned} \tag{2.1}$$

where $\widehat{\Sigma}_i$ is an estimator of the covariance matrix $\Sigma_i$ to be discussed later, which is positive definite, for $i = 1, \ldots, n$; and $\mathbf{H}$ is a positive definite matrix in $\mathbb{R}^{p\times p}$, which is set to be

the average of $\widehat{\Sigma}_i$'s, that is $\mathbf{H} = \sum_{i=1}^{n} T_i \widehat{\Sigma}_i / \sum_{i=1}^{n} T_i$. It is essential to impose a constraint on $\boldsymbol{\gamma}$, otherwise the objective function of (2.1) is minimized at $\boldsymbol{\gamma} = \mathbf{0}$ with fixed $\boldsymbol{\beta}$. When $\widehat{\Sigma}_i = \mathbf{S}_i = \sum_{t=1}^{T_i} \mathbf{y}_{it}\mathbf{y}_{it}^\top / T_i$ (i.e., the sample covariance matrix), which is the proposal in Zhao et al. [41], it is equivalent to minimize the negative log-likelihood function of $\{\boldsymbol{\gamma}^\top \mathbf{y}_{it}\}_{i,t}$ assuming the data are normally distributed. However, when $T_{\max} = \max_i T_i < p$, problem (2.1) is ill-posed as $\mathbf{S}_i$'s are rank-deficient. Thus, the goal of this manuscript is to propose a well-conditioned estimator of $\Sigma_i$ that yields optimal properties. To achieve this, a covariate-dependent linear shrinkage estimator, denoted as $\Sigma_i^*$, is proposed, which yields the minimum expected squared loss under regression model (1.1), where the expectation is taken over the sample covariance matrix $\mathbf{S}_i$.

$$\begin{aligned} &\underset{(\mu,\rho)}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left\{\boldsymbol{\gamma}^\top \Sigma_i^* \boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right\}^2, \\ &\text{such that} \quad \Sigma_i^* = \rho\mu\mathbf{I} + (1-\rho)\mathbf{S}_i, \quad \text{for } i = 1, \ldots, n. \end{aligned} \tag{2.2}$$

The following theorem gives the solution to (2.2).

**Theorem 2.1**. *For given* $(\boldsymbol{\gamma}, \boldsymbol{\beta})$, *the solution to optimization problem* (2.2) *is*

$$\Sigma_i^* = \frac{\psi^2}{\delta^2}\mu\mathbf{I} + \frac{\phi^2}{\delta^2}\mathbf{S}_i, \quad \text{for } i = 1, \ldots, n, \tag{2.3}$$

*and the minimum value is*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left\{\boldsymbol{\gamma}^\top \Sigma_i^* \boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right\}^2 = \frac{\phi^2\psi^2}{\delta^2}, \tag{2.4}$$

*where*

$$\begin{aligned} &\mu = \frac{1}{n\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right)}\sum_{i=1}^{n}\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right), \quad \phi^2 = \frac{1}{n}\sum_{i=1}^{n}\phi_i^2, \quad \psi^2 = \frac{1}{n}\sum_{i=1}^{n}\psi_i^2, \quad \delta^2 = \frac{1}{n}\sum_{i=1}^{n}\delta_i^2, \phi_i^2 \\ &= \left\{\mu\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right) - \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2, \quad \psi_i^2 = \mathbb{E}\left\{\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2, \delta_i^2 = \mathbb{E}\left\{\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma} - \mu\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right)\right\}^2; \end{aligned}$$

*and* Lemma 2.1 *shows that* $\psi^2/\delta^2 + \phi^2/\delta^2 = 1$.

**Lemma 2.1**. *For* $\forall\ i \in \{1, \ldots, n\}$, $\delta_i^2 = \phi_i^2 + \psi_i^2$, *and thus* $\delta^2 = \phi^2 + \psi^2$.

According to Theorem 2.1, parameters $\phi_i^2$, $\psi_i^2$ and $\delta_i^2$ are expected values as the objective is to minimize the expected squared loss. Thus, one cannot replace $\widehat{\Sigma}_i$ with $\sum_i^*$ in (2.1) and solve for solution using the data. For implementation in practice, the following sample counterparts are used to compute (2.3) and thus $\widehat{\Sigma}_i$ in (2.1). Let

$$\hat{\delta}_i^2 = \left\{\boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma} - \mu\left(\boldsymbol{\gamma}^\top \boldsymbol{\gamma}\right)\right\}^2, \quad \hat{\psi}_i^2 = \frac{1}{T_i}\left\{\boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right\}^2, \quad \hat{\phi}_i^2 = \hat{\delta}_i^2 - \hat{\psi}_i^2, \hat{\delta}^2 = \frac{1}{n}\sum_{i=1}^n \hat{\delta}_i^2, \quad \hat{\psi}^2 = \frac{1}{n}\sum_{i=1}^n \min\left(\hat{\psi}_i^2, \hat{\delta}_i^2\right), \quad \hat{\phi}^2 = \frac{1}{n}\sum_{i=1}^n \hat{\phi}_i^2,$$

and

$$\mathbf{S}_i^* = \frac{\hat{\psi}^2}{\hat{\delta}^2}\mu\mathbf{I} + \frac{\hat{\phi}^2}{\hat{\delta}^2}\mathbf{S}_i, \quad \text{for } i = 1, \ldots, n. \tag{2.5}$$

In Section 3, we show that $\mathbf{S}_i^*$ is a consistent estimator of $\Sigma_i^*$ and is uniformly optimal asymptotically among all the linear combinations of the sample covariance matrices and the identity matrix regarding the quadratic risk. The objective function $\ell(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is an approximation of the negative log-likelihood function if replacing $\hat{\Sigma}_i$ with the proposed shrinkage estimator $\mathbf{S}_i^*$. Thus, optimizing (2.1) can be considered as a pseudo-likelihood approach under the normality assumption.

The proof of Theorem 2.1 and Lemma 2.1 is presented in Appendix Section A.1. Formulation (2.2) introduces a shrinkage estimator of the covariance matrix, where the shrinkage is shared across subjects and is optimal under the squared error loss. For each subject, $\Sigma_i^*$ is a linear combination of the sample covariance matrix $\mathbf{S}_i$ and the identity matrix. The weighting parameters, $\rho$ and $\mu$, are population level parameters that are shared across subjects. This is equivalent to imposing a linear shrinkage on the sample eigenvalues. Assuming $\boldsymbol{\gamma}$ is a common eigenvector of all the covariance matrices, $\mu$ is the average eigenvalue corresponding to $\boldsymbol{\gamma}$. The level of shrinkage is determined by the leverage between the accuracy of $\mathbf{S}_i$'s and the variation in the eigenvalues. If $\mathbf{S}_i$'s are accurate or the errors are small relative to the variation in the eigenvalues, less shrinkage will be imposed; otherwise, if $\mathbf{S}_i$'s are inaccurate and the errors are comparable or even higher than the eigenvalue variability, the sample covariance matrices will be shrank more.

Algorithm 1 summarizes the optimization procedure. As problem (2.1) is nonconvex, a series of random initializations of $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ is considered and the one that achieves the minimum value of the objective function is the estimate. The initial values of $\boldsymbol{\gamma}$ can be set as the eigenvectors of the average sample covariance matrices, $\bar{\mathbf{S}} = \sum_{i=1}^n T_i \mathbf{S}_i / \sum_{i=1}^n T_i$; and the initial values of $\boldsymbol{\beta}$ is the corresponding solution to (2.1) by replacing $\hat{\Sigma}_i$ with a well-conditioned estimator, such as the estimator proposed in Ledoit and Wolf [24]. When $p < \sum_{i=1}^n T_i$, $\bar{\mathbf{S}}$ is of full rank, and the sample eigenvectors are consistent estimators assuming all the covariance matrices have the same eigendecomposition. Step 3 in the algorithm updates the covariance matrix estimators with a global shrinkage parameter. In Section 4, through simulation studies, we show that it improves the performance in estimating the covariance matrices and $\boldsymbol{\beta}$ with lower bias and higher stability. The details of updating $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ in Step 4 can be found in Algorithm 1 in Zhao et al. [41].

**Algorithm 1** The optimization algorithm for problems (2.1) and (2.2).

**Input**: $\{(\mathbf{y}_{i1},\ldots,\mathbf{y}_{iT_i}),\mathbf{x}_i\}_{i=1}^n$

1: **initialization**: $(\boldsymbol{\gamma}^{(0)},\boldsymbol{\beta}^{(0)})$

2: **repeat** for iteration $s=0,1,2,\ldots$

3: for $i=1,\ldots,n$, update

$$\mathbf{S}_i^{*(s+1)}=\frac{\hat{\psi}^{2(s)}}{\hat{\delta}^{2(s)}}\mu^{(s)}\mathbf{I}+\frac{\hat{\phi}^{2(s)}}{\hat{\delta}^{2(s)}}\mathbf{S}_i,$$

where $(\hat{\psi}^2,\hat{\phi}^2,\hat{\delta}^2,\mu)$ are set to the value with $\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(s)}$ and $\boldsymbol{\beta}=\boldsymbol{\beta}^{(s)}$,

4: update $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ by solving (2.1) with $\hat{\Sigma}_i=\mathbf{S}_i^{*(s+1)}$, denoted as $\boldsymbol{\gamma}^{(s+1)}$ and $\boldsymbol{\beta}^{(s+1)}$, respectively,

5: **until** the objective function in (2.1) converges;

6: consider a random series of initializations, repeat Steps 1–5, and choose the results with the minimum objective value.

**Output**: $(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\beta}})$

For higher-order components, one can first remove the identified components and use the new data to estimate the next with an additional orthogonality constraint, that is, the new component is orthogonal to the identified ones. Different from Algorithm 2 in Zhao et al. [41], there is no need to include a rank-completion step as $\mathbf{S}_i^*$ is introduced to render the rank-deficiency issue. To determine the number of components, the metric of average deviation from diagonality is adopted [41]. Let $\Gamma^{(k)}\in\mathbb{R}^{p\times k}$ denote the first $k$ estimated components, the average deviation from diagonality is defined as

$$\mathrm{DfD}\left(\Gamma^{(k)}\right)=\prod_{i=1}^{n}\left(\frac{\det\left\{\mathrm{diag}\left(\Gamma^{(k)\top}\mathbf{S}_i^*\Gamma^{(k)}\right)\right\}}{\det\left(\Gamma^{(k)\top}\mathbf{S}_i^*\Gamma^{(k)}\right)}\right)^{T_i/\sum_i T}, \tag{2.6}$$

where diag($\mathbf{A}$) is a diagonal matrix of the diagonal elements in a square matrix $\mathbf{A}$, and det($\mathbf{A}$) is the determinant of $\mathbf{A}$. If $\Gamma^{(k)}$ is a common diagonalization of $\mathbf{S}_i^*$'s, that is, $\Gamma^{(k)\top}\mathbf{S}_i^*\Gamma^{(k)}$ is a diagonal matrix, for $\forall\ i=1,\ldots,n$, then $\mathrm{DfD}(\Gamma^{(k)})=1$. In practice, $k$ can be chosen before DfD increases far away from one or before a sudden jump occurs.

## 3. Asymptotic Properties

In this section, we study the asymptotic properties of the proposed estimators. For $i=1,\ldots,n$, it is assumed that $\Sigma_i$ has the eigendecomposition of $\Sigma_i=\Pi_i\Lambda_i\Pi_i^\top$, where $\Lambda_i=\mathrm{diag}\{\lambda_{i1},\ldots,\lambda_{ip}\}$ is a diagonal matrix and $\Pi_i=(\boldsymbol{\pi}_{i1},\ldots,\boldsymbol{\pi}_{ip})$ is an orthonormal rotation matrix; $\{\lambda_{i1},\ldots,\lambda_{ip}\}$ are the eigenvalues and the columns of $\Pi_i$ are the corresponding eigenvectors. Let $\mathbf{Z}_i=\mathbf{Y}_i\Pi_i$, where $\mathbf{Y}_i=\left(\mathbf{y}_{i1},\ldots,\mathbf{y}_{iT_i}\right)^\top\in\mathbb{R}^{T_i\times p}$ is the data matrix of subject $i$. Under the normality assumption, the columns of $\mathbf{Z}_i=(z_{itj})_{t,j}$ are uncorrelated, and the rows, $\mathbf{z}_{it}=\left(z_{i1},\ldots,z_{ip}\right)\in\mathbb{R}^p$ for $t=1,\ldots,T_i$, are normally distributed with mean zero and covariance matrix $\Lambda_i$. The following assumptions are imposed.

**Assumption A1** There exists a constant $C_1$ independent of $T_{\max}$ such that $p/T_{\max}\le C_1$, where $T_{\max}=\max_i T_i$.

**Assumption A2** Let $N=\sum_{i=1}^n T_i$, $p/N\to\infty$ as $n$, $T_{\min}\to\infty$, where $T_{\min}=\min_i T_i$.

**Assumption A3** There exists a constant $C_2$ independent of $T_{\min}$ and $T_{\max}$ such that $\sum_{j=1}^p\mathbb{E}\left(z_{i1j}^8\right)/p\le C_2$, for $\forall\ i\in\{1,\ldots,n\}$.

**Assumption A4** Let $\mathcal{Q}$ denote the set of all the quadruples that are made of four distinct integers between 1 and $p$, for $\forall\ i \in \{1, \ldots, n\}$,

$$\lim_{T_i \to \infty} \frac{p^2}{T_i^2} \frac{\sum_{(j,k,l,m) \in \mathcal{Q}} \left\{ \mathrm{Cov}\left( z_{i1j} z_{i1k}, z_{i1l} z_{i1m} \right) \right\}^2}{|\mathcal{Q}|} = 0, \tag{3.1}$$

where $|\mathcal{Q}|$ is the cardinality of set $\mathcal{Q}$.

**Assumption A5** All the covariance matrices share the same set of eigenvectors, i.e., $\Pi_i = \Pi$, for $i = 1, \ldots, n$. For each $\Sigma_i$, there exists (at least) a column, indexed by $j_i$, such that $\boldsymbol{\gamma} = \boldsymbol{\pi}_{iji}$ and Model (1.1) is satisfied.

Assumption A1 allows the data dimension, $p$, to be greater than the (maximum) number of observations, $T_{\max}$, and to grow at the same rate as $T_{\max}$ does. This is a common regularity condition for shrinkage estimators [24]. Assumption A2 guarantees that the average sample covariance matrix $\overline{\mathbf{S}} = \sum_{i=1}^{n} T_i \mathbf{S}_i / N$ utilized in the initial step of Algorithm 1 is positive definite. Together with Assumption A5, the eigenvectors of $\overline{\mathbf{S}}$ are consistent estimators of $\Pi$ [2]. Assumptions A3 and A4 regulate $\mathbf{z}_{it}$ on higher-order moments, which is equivalent to imposing restrictions on the higher-order moments of $\mathbf{y}_{it}$. When the data are assumed to be normally distributed, both A3 and A4 are satisfied. Assumption A5 assumes that all the covariance matrices share the same eigenspace, though the ordering of the eigenvectors may differ. When $p/T_{\min} \to 0$, Zhao et al. [41] relaxed this assumption to partial common diagonalization and demonstrated the method robustness through numerical examples. Studying the asymptotic properties under the relaxation is difficult and not available in existing literature, especially when $p > T_{\max}$.

Taking the eigenvectors of $\overline{\mathbf{S}}$ as the initial values of $\boldsymbol{\gamma}$, the following proposition demonstrates the consistency of the proposed estimator.

**Proposition 3.1**. *Under Assumptions A1–A5, as* $n, T_{\min} \to \infty$, *the estimator of* $\boldsymbol{\gamma}$ *and* $\boldsymbol{\beta}$ *obtained by* Algorithm 1 *are asymptotically consistent.*

To prove Proposition 3.1, we first study the asymptotic properties of $\mathbf{S}_i^*$ and show that $\mathbf{S}_i^*$ is the optimal linear shrinkage estimator of the covariance matrix under the squared loss. This is accomplished under the assumption that $\boldsymbol{\gamma}$ is given. As the initialization of $\boldsymbol{\gamma}$ is already a consistent estimator, the consistency of the solution after iteration follows. For $\boldsymbol{\beta}$, it is firstly shown that the association between the shrinkage estimator, $\Sigma_i^*$, and the covariates is the same as the covariance matrix, $\Sigma_i$, does (Lemma 3.3). Thus, it is equivalent to optimize problems (2.1) and (2.2) to solve for $\boldsymbol{\beta}$, and the solution is a consistent estimator of $\boldsymbol{\beta}$ based on the pseudo-likelihood theory [16]. In the iteration step of Algorithm 1, $\mathbf{S}_i^*$ improves the estimation of the covariance matrices with lower squared loss, and in consequence, improves the estimation of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. In Section 4, the improvement is demonstrated through simulation studies.

In Section 2, the optimization problem (2.2) introduces a linear combination of the sample covariance matrix and the identity matrix, $\Sigma_i^*$, that achieves the minimum expected squared error. From Theorem 2.1, the solution has population-level parameters. Thus, the sample counterpart, $\mathbf{S}_i^*$, is introduced. The following Lemma 3.1 first shows that asymptotically, the weighting parameters in $\Sigma_i^*$ are well-behaved. Lemma 3.2 demonstrates that the corresponding sample counterpart of the weighting parameters are consistent estimators. Theorem 3.1 demonstrates that $\mathbf{S}_i^*$ performs as well as $\Sigma_i^*$ does asymptotically.

**Lemma 3.1**. *For given* $(\boldsymbol{\gamma}, \boldsymbol{\beta})$, *let* $T_{\min} = \min_i T_i$, *as* $T_{\min} \to \infty$, $\mu$, $\phi^2$, $\psi^2$ *and* $\delta^2$ *are bounded.*

**Lemma 3.2**. *For given* $(\boldsymbol{\gamma}, \boldsymbol{\beta})$, *as* $T_{\min} \to \infty$,

- **i.** $\mathbb{E}\left(\hat{\delta}_i^2 - \delta_i^2\right)^2 \to 0$, *for* $i = 1, \ldots, n$, *and thus* $\mathbb{E}\left(\hat{\delta}^2 - \delta^2\right)^2 \to 0$;
- **ii.** $\mathbb{E}\left(\hat{\psi}_i^2 - \psi_i^2\right)^2 \to 0$, *for* $i = 1, \ldots, n$, *and thus* $\mathbb{E}\left(\hat{\psi}^2 - \psi^2\right)^2 \to 0$;
- **iii.** $\mathbb{E}\left(\hat{\phi}_i^2 - \phi_i^2\right)^2 \to 0$, *for* $i = 1, \ldots, n$, *and thus* $\mathbb{E}\left(\hat{\phi}^2 - \phi^2\right)^2 \to 0$.

**Theorem 3.1**. *For* $\forall\ i \in \{1, \ldots, n\}$, $\mathbf{S}_i^*$ *is a consistent estimator of* $\Sigma_i^*$, *that is, as* $T_{\min} = \min_i T_i \to \infty$,

$$\mathbb{E}\|\mathbf{S}_i^* - \Sigma_i^*\|^2 \to 0. \tag{3.2}$$

*Thus, the asymptotic expected loss of* $\mathbf{S}_i^*$ *and* $\Sigma_i^*$ *are identical, that is,*

$$\mathbb{E}\left\{\boldsymbol{\gamma}^\top \mathbf{S}_i^* \boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right\}^2 - \mathbb{E}\left\{\boldsymbol{\gamma}^\top \Sigma_i^* \boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right\}^2 \to 0. \tag{3.3}$$

Next, we show that $\mathbf{S}_i^*$ uniformly achieves the minimum quadratic risk asymptotically over all linear combinations of the sample covariance matrix and the identity matrix. For given $(\boldsymbol{\gamma}, \boldsymbol{\beta})$, let $\Sigma_i^{**}$ denote the solution to the following optimization problem,

$$\begin{aligned} &\underset{\rho_1, \rho_2}{\text{minimize}} \frac{1}{n} \sum_{i=1}^{n} \left\{\boldsymbol{\gamma}^\top \Sigma_i^{**} \boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right\}^2, \\ &\text{such that} \quad \Sigma_i^{**} = \rho_1 \mathbf{I} + \rho_2 \mathbf{S}_i, \quad \text{for } i = 1, \ldots, n. \end{aligned} \tag{3.4}$$

**Theorem 3.2.** $\mathbf{S}_i^*$ *is a consistent estimator of* $\Sigma_i^{**}$, *that is, as* $T_{\min} = \min_i T_i \to \infty$, *for* $i = 1, \ldots, n$,

$$\mathbb{E}\|\mathbf{S}_i^* - \Sigma_i^{**}\|^2 \to 0. \tag{3.5}$$

*Then*, $\mathbf{S}_i^*$ has the same asymptotic expected loss as $\Sigma_i^{**}$ *does*, *that is*,

$$\mathbb{E}\left\{\boldsymbol{\gamma}^\top \mathbf{S}_i^* \boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right\}^2 - \mathbb{E}\left\{\boldsymbol{\gamma}^\top \Sigma_i^{**} \boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right\}^2 \to 0. \tag{3.6}$$

**Theorem 3.3**. *Assume* $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ *is given. With a fixed* $n \in \mathbb{N}^+$, *for any sequence of linear combinations* $\left\{\widehat{\Sigma}_i\right\}_{i=1}^n$ *of the identity matrix and the sample covariance matrix, where the combination coefficients are constant over* $i \in \{1, \ldots, n\}$, *the estimator* $\mathbf{S}_i^*$ *verifies:*

$$\lim_{T \to \infty} \inf_{T_i \geq T} \left[\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left\{\boldsymbol{\gamma}^\top \widehat{\Sigma}_i \boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right\}^2 - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left\{\boldsymbol{\gamma}^\top \mathbf{S}_i^* \boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right\}^2\right] \geq 0. \tag{3.7}$$

*In addition, every sequence of* $\left\{\widehat{\Sigma}_i\right\}_{i=1}^n$. *that performs as well as* $\{\mathbf{S}_i^*\}_{i=1}^n$ *identical to* $\{\mathbf{S}_i^*\}_{i=1}^n$ *in the limit:*

$$\lim_{T \to \infty} \left[\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left\{\boldsymbol{\gamma}^\top \widehat{\Sigma}_i \boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right\}^2 - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left\{\boldsymbol{\gamma}^\top \mathbf{S}_i^* \boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right\}^2\right] = 0 \tag{3.8}$$

$$\Leftrightarrow \quad \mathbb{E}\left\|\widehat{\Sigma}_i - \mathbf{S}_i^*\right\|^2 \to 0, \quad \textit{for } i = 1, \ldots, n. \tag{3.9}$$

The difference between $\Sigma_i^{**}$ and $\Sigma_i^*$ is that $\Sigma_i^{**}$ minimizes the squared loss instead of the expected loss, while asymptotically they are equivalent (Theorems 3.1 and 3.2). Theorem 3.3 presents the main result that, with a fixed sample size $n$, the proposed shrinkage estimator $\{\mathbf{S}_i^*\}_{i=1}^n$ achieves the uniformly minimum (average) quadratic risk asymptotically among all linear combinations of the identity matrix and the sample covariance matrix. Here, "average" implies an average over the subjects, and "asymptotically" refers to that the number of observations within each subject increases to infinity. Therefore, $\mathbf{S}_i^*$ is asymptotically optimal. In addition, it is guaranteed that $\mathbf{S}_i^*$ is positive definite (see a discussion in Appendix Section A.8). Thus, there exits unique solution to the optimization problem (2.1).

Next, we study the asymptotic properties of the model coefficient estimator. Let $\widehat{\boldsymbol{\beta}}$ denote the solution to the optimization problem (2.1).

**Lemma 3.3**. *For given* $\boldsymbol{\gamma}$, *assume the linear shrinkage estimator*, $\Sigma_i^*$, *satisfies*

$$\mathbb{E}\left(\boldsymbol{\gamma}^\top \Sigma_i^* \boldsymbol{\gamma}\right) = \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}^*\right), \quad \textit{for } i = 1, \ldots, n, \tag{3.10}$$

*then*

$$\boldsymbol{\beta}^* = \boldsymbol{\beta}. \tag{3.11}$$

**Theorem 3.4**. *For given $\boldsymbol{\gamma}$, assume Assumptions A1–A5 are satisfied, $\widehat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$ as $n$, $T_{\min} \to \infty$, where $T_{\min} = \min_i T_i$.*

Lemma 3.3 implies that under the rotation $\boldsymbol{\gamma}$, the expectation of the shrinkage estimator, $\Sigma_i^*$, has the same association with the covariates as the true covariance matrix, $\Sigma_i$, does. $\mathbf{S}_i^*$ is a consistent estimator of $\Sigma_i^*$ and is positive definite. This substantiates the choice of $\mathbf{S}_i^*$ replacing the sample covariance matrix $\mathbf{S}_i$ in the optimization problem. Theorem 3.4 states the consistency of $\widehat{\boldsymbol{\beta}}$.

## 4. Simulation Study

### 4.1. $\gamma$ is known

In this section, we focus on examining the performance of the proposed method in estimating the covariance matrices and model coefficients by assuming the projection $\boldsymbol{\gamma}$ is known. Three methods are compared. (i) Estimate each individual covariance matrix using the estimator proposed in Ledoit and Wolf [24] and replace $\widehat{\Sigma}_i$ with it in the optimization problem (2.1). We denote this approach as LW-CAP (**L**edoit and **W**olf based **C**ovariate **A**ssisted **P**rincipal regression), where the shrinkage is estimated on each individual covariance matrix. (ii) Estimate the covariance matrices using the proposed shrinkage estimator $\mathbf{S}_i^*$ in (2.5). We denote this approach as CS-CAP (**C**ovariate dependent **S**hrinkage CAP), where the shrinkage parameters are assumed to be shared across subjects. (iii) Estimate each individual covariance matrix using the sample covariance matrix and plug into the optimization problem (2.1). This is the CAP approach proposed in Zhao et al. [41], which is only applicable when $T_{\min} = \min_i T_i > p$.

The covariance matrices are generated using the eigendecomposition $\Sigma_i = \Pi \Lambda_i \Pi^\top$, where $\Pi = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_p)$ is an orthonormal matrix in $\mathbb{R}^{p \times p}$ and $\Lambda_i = \mathrm{diag}\{\lambda_{i1}, \ldots, \lambda_{ip}\}$ is a diagonal matrix with the diagonal elements to be the eigenvalues, for $i = 1, \ldots, n$. In $\Lambda_i$, the diagonal elements are exponentially decaying, where eigenvalues of the second and the fourth dimension (D2 and D4) satisfy the log-linear model in (1.1). We consider a case with a single predictor $X$ (thus $q = 2$), which is generated from a Bernoulli distribution with probability 0.5 to be one. For D2, the coefficient $\beta_1 = -1$; and for D4, $\beta_1 = 1$. For the rest dimensions, $\lambda_{ij}$, for $i = 1, \ldots, n$, is generated from a log-normal distribution, where the mean of the corresponding normal distribution decreases from 5 to −1 over $j$. Cases when $p = 20$, 50, 100 are considered.

We first compare the three approaches, LW-CAP, CS-CAP and CAP, under sample sizes $n = 50$ and $T_i = T = 50$ for all $i$ and present the result in Table 1. In the estimation, for dimension $j$, $\boldsymbol{\gamma}$ is set to be $\boldsymbol{\pi}_j$. In Table 1, we present the bias and the mean squared error (MSE) in estimating the eigenvalues and the model coefficient in D2 and D4. From the table, for both the eigenvalues and $\beta_1$, CS-CAP yields lower estimation bias and MSE than LW-CAP does. When $p < T$, CS-CAP achieves a similar estimation bias as the CAP approach does in estimating the covariance matrices, while the MSE is slightly lower. For the estimation of $\beta_1$, CS-CAP yields slightly lower bias. As the dimension $p$ increases, the bias and MSE of eigenvalue estimates from LW-CAP increase; while the bias and MSE of the estimates

from CS-CAP are similar at all $p$ settings. This demonstrates the superiority of the proposed estimator in estimating the covariance matrices. Figure 1 presents the estimation bias and MSE of CS-CAP estimator at various levels of $T$ when fixing $n = 50$ when $p = 20$. From the figure, as the number of observations within each subject increases, the estimates converge to the truth.

### 4.2. $\gamma$ is unknown

In this section, we evaluate the performance of the CS-CAP approach when $\boldsymbol{\gamma}$ is unknown and estimated by solving optimization problem (2.1) using Algorithm 1. The data are generated following the same procedure as in Section 4.1. To evaluate the performance in estimating the projection $\boldsymbol{\gamma}$, we consider a similarity metric measured by $|\langle \hat{\gamma}, \gamma \rangle|$, where $\langle \cdot, \cdot \rangle$ is the inner product of two vectors and $\hat{\gamma}$ denotes the estimate of $\boldsymbol{\gamma}$. When this metric is one, the two vectors are identical (up to sign flipping); and when this metric is zero, the two vectors are orthogonal. Case where $p = 100$ is studied. The performance of the CS-CAP approach is firstly compared to the LW-CAP approach with sample sizes $n = 100$ and $T_i = T = 100$. The results are presented in Table 2. From the table, the CS-CAP approach improves the performance with much lower MSE in estimating the eigenvalues, and lower MSE and higher coverage probability (CP) in estimating the $\boldsymbol{\beta}$ coefficient. After iterations, the CS-CAP approach yields an estimate of the projection with much higher similarity to the truth. To further examine the performance of the CS-CAP approach under finite sample size, combinations of sample sizes $n = 50, 100, 500, 1000$ and $T_i = T = 50, 100, 500, 1000$ are considered. Figure 2 presents the performance in estimating the second dimension (D2), including the bias, the MSE and the CP of $\hat{\beta}_1$, the MSE of $\hat{\lambda}_{ij}$, and the similarity of $\hat{\gamma}$ to the eigenvector of D2 (Appendix Section B.1 presents the results of the fourth dimension, D4). From the figure, as $n, T \to \infty$, all estimates converge to the truth.

In Appendix Section B.2, the robustness of the proposed method to model misspecification is examined. Two types of model misspecification are considered, model misspecification in $\boldsymbol{\beta}$ and model misspecification in $\boldsymbol{\gamma}$. When the log-linear model is misspecified, the proposed approach can correctly identify the linear projections under certain scenarios, while the estimate of model coefficients is biased. The proposed approach is robust to the setting that the eigenvectors of the covariance matrices are partially common, while it will not work when the eigenvectors are completely unique to each covariance matrix.

## 5. The Alzheimer's Disease Neuroimaging Initiative Study

Data used in this study are obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

We apply the proposed approach to ADNI resting-state functional MRI (fMRI) data acquired at the baseline screening. AD is an irreversible neurodegenerative disease that destroys memory and related brain functions causing problems in cognition and behavior. Apolipoprotein E $\varepsilon 4$ (APOE-$\varepsilon 4$) has been consistently identified as a strong genetic risk factor for AD. With an increasing number of APOE-$\varepsilon 4$ alleles, the lifetime risk of developing AD increases, and the age of onset decreases [10]. Thus, APOE-$\varepsilon 4$ is generally treated as a potential therapeutic target [33]. In AD studies, resting-state fMRI is another emerging biomarker for diagnosis [23]. It is important to articulate the genetic impact on brain functional architecture. In this study, $n = 194$ subjects diagnosed with either MCI or AD are analyzed. Resting-state fMRI data collected at the initial screening are preprocessed. Time courses are extracted from $p = 75$ brain regions, including 60 cortical and 15 subcortical regions grouped into 10 functional modules, using the Harvard-Oxford Atlas in FSL [35]. For each time course, a subsample is taken with an effective sample size of $T = 67$ to remove the temporal dependence. The resulting data, denoted as $\mathbf{y}_{it}$ (for $i = 1, \ldots, n$ and $t = 1, \ldots, T$), are assumed to follow a multivariate normal distribution with mean zero and covariance $\Sigma_i$. The off-diagonal elements in $\Sigma_i$ represent the pairwise functional connectivity between brain regions and $\Sigma_i$ represents the brain functional architecture of subject $i$. In the regression model, APOE-$\varepsilon 4$, sex, and age are entered as the covariates ($\mathbf{x}_i$'s). The validity of model assumptions is discussed in Appendix Section C.1.

The CS-CAP approach is applied to identify brain subnetworks within which the functional connectivity demonstrates a significant association with APOE-$\varepsilon 4$. Using the deviation from diagonality criterion, CS-CAP identifies three components denoted as C1, C2, and C3. The model coefficients and 95% bootstrap confidence interval from 500 bootstrap samples are presented in Table 3. From the table, C3 is significantly associated with APOE-$\varepsilon 4$ and age; C1 and C2 are significantly associated with sex and age. To better interpret C3, a fused lasso regression [36] is employed to sparsify the loading profile, similarly as in the sparse principal component analysis proposed in Zou et al. [42]. The fused lasso regularization is defined based on the modular information to impose local smoothness and consistency [19, 40]. Figure 3(a) presents the sparse loading profile colored by the corresponding functional module, and Figure 3(b) is the river plot illustrating the loading configuration. In C3, all regions with negative loadings are subcortical regions. Contributions to positive loadings are from regions in the default mode network (DMN), the ventral- and dorsal-attention networks, and the somato-motor network. Figure 3(c) presents these regions on a brain map. C3 is negatively associated with APOE-$\varepsilon 4$ indicating that functional connectivity between regions in the same sign among APOE-$\varepsilon 4$ carriers is lower, while connectivity between regions in the opposite signs among APOE-$\varepsilon 4$ carriers is higher. The findings are in line with existing knowledge about AD. Compared to APOE-$\varepsilon 4$ non-carriers, more functional connectivity between the left hippocampus and the insular/prefrontal cortex while more functional disconnection of the hippocampus has been observed in APOE-$\varepsilon 4$ carriers [12]. Alterations in DMN connectivity in cognitively normal APOE-$\varepsilon 4$ carriers have been reported across all age groups [3]. Increased connectivity in the limbic system, including the hippocampus, the amygdala, and the thalamus, has been detected in individuals with memory impairment [18, 17], though the effect of APOE-$\varepsilon 4$ carriage lacks consensus [3]. It was shown that the limbic hyperconnectivity is positively associated with the memory

performance, suggesting the preservation of brain function due to increased connectivity in the medial temporal lobe pathology [17].

## 6. Discussion

In this study, we introduce an approach to perform linear regression with multiple high dimensional covariance matrices as the outcome. A linear shrinkage estimator of the covariance matrix is firstly introduced, where the shrinkage coefficients are shared parameters across subjects. It is shown that the proposed estimator is optimal achieving the uniformly minimum quadratic loss asymptotically among all linear combinations of the identity matrix and the sample covariance matrix. Replacing the sample covariance matrices with the proposed well-conditioned estimator in the likelihood function, the linear projection parameter and the model coefficient are shown to be consistently estimated. Through simulation studies, the proposed approach demonstrates superior performance in estimating the covariance matrices and the model coefficients with lower estimation bias and variation over the existing methods. Applying to a resting-state fMRI data set acquired from the ADNI study, the findings are consistent with existing knowledge about AD.

The proposed framework extends the proposal in Zhao et al. [41] to high dimensional scenario. When $p$ is small, the proposed shrinkage estimator demonstrates lower squared loss than the sample covariance matrix as suggested in both theoretical results and simulation studies. Different from the linear shrinkage estimator introduced in Ledoit and Wolf [24], which was proposed for a single covariance matrix estimation, the shrinkage coefficients considered in this study are population level parameters shared across subjects. This is superior than the individual shrinkage as the proposed one leverages the accuracy of the sample covariance matrix and the variability in the eigenvalues across subjects.

In this study, the asymptotic properties are studied under the assumption that the covariance matrices have the same eigendecomposition. We leave the study of the consistency relaxing this assumption to future research. The proposed shrinkage estimator is optimal with respect to a squared risk. However, this may overshrink the small eigenvalues [11]. Other types of loss function, such as the Stein's loss, will be considered in the future. In the ADNI application, we included an *ad hoc* procedure to select important brain regions for interpretation. A next-step research is to include the regularization on $\boldsymbol{\gamma}$ into the optimization or to introduce an efficient approach to draw inference on the loadings (such as a bootstrap sampling procedure).

## Acknowledgments

Zhao was partially supported by NIH grant U54AG065181 and P30AG010133; Caffo by NIH grant R01EB029977 and P41EB031771; and Luo by NIH grant R01EB022911. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI, National Institutes of Health Grant U01 AG024904 and Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies;

Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

# Appendix A: Theory and Proof

## A.1. Proof of Theorem 2.1 and Lemma 2.1

*Proof.* Given $(\boldsymbol{\gamma}, \boldsymbol{\beta})$, $\mathbb{E}\left(\boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma}\right) = \boldsymbol{\gamma}^\top \Sigma_i \boldsymbol{\gamma} = \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)$. For the objective function in (2.2), under the constraint that $\Sigma_i^* = \rho\mu\mathbf{I} + (1-\rho)\mathbf{S}_i$, we have

$$\begin{aligned} & f(\mu, \rho) \\ & = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left\{\boldsymbol{\gamma}^\top \Sigma_i^* \boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right\}^2 \\ & = \frac{1}{n}\sum_{i=1}^{n}\left[\rho^2\left\{\mu\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right) - \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2 + (1-\rho)^2\mathbb{E}\left\{\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2\right]. \end{aligned}$$

In order to minimize the objective function, as the objective function is convex, derivatives are firstly taken over $\mu$ and $\rho$.

For $\mu$,

$$\begin{aligned} \frac{\partial f}{\partial \mu} &= \rho^2 \frac{1}{n}\sum_{i=1}^{n} 2\left\{\mu\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right) - \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right) = 0, \\ \Rightarrow \quad \mu &= \frac{1}{n\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right)}\sum_{i=1}^{n}\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right). \end{aligned}$$

For $\rho$, let $\phi_i^2 = \left\{\mu\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right) - \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2$ and $\psi_i^2 = \mathbb{E}\left\{\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2$,

$$\begin{aligned} \frac{\partial f}{\partial \rho} &= 2\rho\left(\frac{1}{n}\sum_{i=1}^{n}\phi_i^2\right) - 2(1-\rho)\left(\frac{1}{n}\sum_{i=1}^{n}\psi_i^2\right) = 0, \\ \Rightarrow \quad \rho &= \frac{\sum_{i=1}^{n}\psi_i^2}{\sum_{k=1}^{n}\phi_i^2 + \sum_{i=1}^{n}\psi_i^2}. \end{aligned}$$

Let $\delta_i^2 = \mathbb{E}\left\{\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma} - \mu\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right)\right\}^2$, then $\delta_i^2 = \phi_i^2 + \psi_i^2$. Let $\phi^2 = \sum_{i=1}^{n}\phi_i^2/n$, $\psi^2 = \sum_{i=1}^{n}\psi_i^2/n$, and $\delta^2 = \sum_{i=1}^{n}\delta_i^2/n$ (thus, $\delta^2 = \phi^2 + \psi^2$), the optimizer of problem (2.2) is

$$\Sigma_i^* = \frac{\psi^2}{\delta^2}\mu\mathbf{I} + \frac{\phi^2}{\delta^2}\mathbf{S}_i, \quad i = 1, \ldots, n.$$

The minimum value of the function is

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{\boldsymbol{\gamma}^\top\Sigma_i^*\boldsymbol{\gamma}-\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2\\
&=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{\frac{\psi^2}{\delta^2}\mu\boldsymbol{\gamma}^\top\boldsymbol{\gamma}+\frac{\phi^2}{\delta^2}\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}-\frac{\psi^2+\phi^2}{\delta^2}\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2\\
&=\frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{E}\left\{\frac{\psi^2}{\delta^2}\mu\boldsymbol{\gamma}^\top\boldsymbol{\gamma}-\frac{\psi^2}{\delta^2}\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2+\mathbb{E}\left\{\frac{\phi^2}{\delta^2}\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}-\frac{\phi^2}{\delta^2}\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2\right]\\
&=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\psi^4}{\delta^4}\phi_i^2+\frac{\phi^4}{\delta^4}\psi_i^2\right)\\
&=\frac{\psi^4\phi^2+\phi^4\psi^2}{\delta^4}\\
&=\frac{\phi^2\psi^2}{\delta^2}.
\end{aligned}
$$

□

## A.2. Proof of Proposition 3.1

*Proof.* Under Assumptions A2 and A5, the eigenvectors of $\overline{\mathbf{S}}$ are consistent estimators of $\Pi$. Replace $\boldsymbol{\gamma}$ with its estimate in Theorems 3.1–3.3 and Theorem 3.4, the consistency of $\boldsymbol{\beta}$ follows. □

## A.3. Proof of Lemma 3.1

*Proof.* (1) For $\mu$,

$$
\mu=\frac{1}{n\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right)}\sum_{i=1}^{n}\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)=\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{\gamma}^\top\Sigma_i\boldsymbol{\gamma}}{\boldsymbol{\gamma}^\top\boldsymbol{\gamma}}\le\frac{1}{n}\sum_{i=1}^{n}\|\Sigma_i\|_2^2.
$$

Under Assumption A2,

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\|\Sigma_i\|_2^2=\frac{1}{n}\sum_{i=1}^{n}\|\Lambda_i\|_2^2\\
&\le\frac{1}{n}\sum_{i=1}^{n}\|\Lambda_i\|_F^2\\
&=\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\left(z_{i1j}^2\right)^2\right\}\\
&=\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\left(z_{i1j}^4\right)\right\}\\
&\le\frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\left(z_{i1j}\right)^8}\\
&\le\frac{1}{n}\sum_{i=1}^{n}\sqrt{C_2}\\
&=\sqrt{C_2},
\end{aligned}
$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix.

(2) For $\phi^2$, upper limits of $\phi_i^2$ is derived first.

$$\begin{aligned}\phi_i^2 &= \left\{\mu\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right) - \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2 \\ &\le \mu^2\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right)^2 + \left\{\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2 \\ &= \mu^2\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right)^2 + \left(\boldsymbol{\gamma}^\top\Sigma_i\boldsymbol{\gamma}\right)^2 \\ &\le \left(\mu^2 + \|\Sigma_i\|_2^4\right)\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right)^2.\end{aligned}$$

From the above derivation, we have

$$\mu^2 \le C_2, \text{ and } \|\Sigma_i\|_2^2 = \|\Lambda_i\|_2^2 \le \|\Lambda_i\|_F^2 \le \sqrt{C_2}.$$

Since $\boldsymbol{\gamma}$ is given, without loss of generality, assume that $\|\boldsymbol{\gamma}\|_2 = 1$, i.e., $\boldsymbol{\gamma}^\top\boldsymbol{\gamma} = 1$. Then,

$$\phi_i^2 \le 2C_2\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right) = 2C_2.$$

Thus,

$$\phi^2 = \frac{1}{n}\sum_{i=1}^{n}\phi_i^2 \le 2C_2.$$

(3) For $\psi^2$, analogously, $\psi_i^2$ is considered first.

$$
\begin{aligned}
\psi_i^2 &= \mathbb{E}\left\{\boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right\}^2 = \mathbb{E}\left\{\boldsymbol{\gamma}^\top (\mathbf{S}_i - \Sigma_i)\boldsymbol{\gamma}\right\}^2 \le \left(\boldsymbol{\gamma}^\top \boldsymbol{\gamma}\right)^2 \mathbb{E}\|\mathbf{S}_i - \Sigma_i\|_2^2 \\
\mathbb{E}\|\mathbf{S}_i - \Sigma_i\|_F^2 &= \frac{1}{p}\sum_{j=1}^{p}\sum_{k=1}^{p} \mathbb{E}\left\{\left(\frac{1}{T_i}\sum_{t=1}^{T_i} y_{itj} y_{itk} - \sigma_{ijk}\right)^2\right\} \\
&= \frac{1}{p}\sum_{j=1}^{p}\sum_{k=1}^{p} \mathbb{E}\left\{\left(\frac{1}{T_i}\sum_{t=1}^{T_i} z_{itj} z_{itk} - \lambda_{ijk}\right)^2\right\} \\
&= \frac{1}{p}\sum_{j=1}^{p}\sum_{k=1}^{p} \mathrm{Var}\left(\frac{1}{T_i}\sum_{t=1}^{T_i} z_{itj} z_{itk}\right) \\
&= \frac{1}{p}\sum_{j=1}^{p}\sum_{k=1}^{p} \frac{1}{T_i}\mathrm{Var}\left(z_{i1j} z_{i1k}\right) \\
&\le \frac{1}{pT_i}\sum_{j=1}^{p}\sum_{k=1}^{p} \mathbb{E}\left(z_{i1j}^2 z_{i1k}^2\right) \\
&\le \frac{1}{pT_i}\sum_{j=1}^{p}\sum_{k=1}^{p} \sqrt{\mathbb{E} z_{i1j}^4}\sqrt{\mathbb{E} z_{i1k}^4} \\
&\le \frac{p}{T_i}\left(\frac{1}{p}\sum_{j=1}^{p}\sqrt{\mathbb{E} z_{i1j}^4}\right)^2 \\
&\le \frac{p}{T_i}\left(\frac{1}{p}\sum_{j=1}^{p} \mathbb{E} z_{i1j}^4\right) \\
&\le \frac{p}{T_i}\sqrt{\frac{1}{p}\sum_{j=1}^{p} \mathbb{E} z_{i1j}^8} \\
&\le C_1\sqrt{C_2}
\end{aligned}
$$

Thus, for $\psi^2$,

$$
\psi^2 = \frac{1}{n}\sum_{i=1}^{n}\psi_i^2 \le \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\gamma}^\top \boldsymbol{\gamma}\right)^2 C_1\sqrt{C_2} = C_1\sqrt{C_2}.
$$

(4) Finally, for $\delta^2$,

$$
\delta^2 = \phi^2 + \psi^2 \le 2C_2 + C_1\sqrt{C_2}.
$$

□

## A.4. Proof of Lemma 3.2

*Proof.* In the proof of Lemma 3.2, here, it is assumed that $\boldsymbol{\gamma}$ is a column of $\Pi_i$ indexed by $j_i$, for $i = 1, \ldots, n$ (Assumption A4).

(i) First, we prove the consistency of $\hat{\delta}_i^2$.

$$\begin{aligned}
&\hat{\delta}_i^2 - \delta_i^2 \\
&\quad = \left\{\gamma^\top \mathbf{S}_i \gamma - \mu\left(\gamma^\top \gamma\right)\right\}^2 - \mathbb{E}\left\{\gamma^\top \mathbf{S}_i \gamma - \mu\left(\gamma^\top \gamma\right)\right\}^2 \\
&\quad = \left\{\left(\gamma^\top \mathbf{S}_i \gamma\right)^2 - \mathbb{E}\left(\gamma^\top \mathbf{S}_i \gamma\right)^2\right\} - 2\mu\left(\gamma^\top \gamma\right)\left\{\left(\gamma^\top \mathbf{S}_i \gamma\right) - \mathbb{E}\left(\gamma^\top \mathbf{S}_i \gamma\right)\right\}
\end{aligned}$$

Under Assumption A4,

$$\gamma^\top \mathbf{S}_i \gamma = \frac{1}{T_i} \sum_{t=1}^{T_i} \gamma^\top \mathbf{y}_{it} \mathbf{y}_{it}^\top \gamma = \frac{1}{T_i} \sum_{t=1}^{T_i} z_{itj_i}^2 .$$

$$\left(\gamma^\top \mathbf{S}_i \gamma\right)^2 = \frac{1}{T_i^2}\left(\sum_{t=1}^{T_i} z_{itj_i}^2\right)^2 = \frac{1}{T_i^2} \sum_{t=1}^{T_i} z_{itj_i}^4 + \frac{1}{T_i^2} \sum_{t \neq s} z_{itj_i}^2 z_{isj_i}^2 .$$

$$\begin{aligned}
\mathbb{E}\left(\gamma^\top \mathbf{S}_i \gamma\right)^2 &= \frac{1}{T_i^2} T_i \mathbb{E} z_{i1j_i}^4 + \frac{1}{T_i} T_i (T_i - 1)\left(\mathbb{E} z_{itj_i}^2\right)^2 \\
&= \frac{1}{T_i} \mathbb{E} z_{i1j_i}^4 + \frac{T_i (T_i - 1)}{T_i^2}\left(\gamma^\top \Sigma_i \gamma\right)^2 .
\end{aligned}$$

For $\forall\ \epsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\left\{\left|\left(\gamma^\top \mathbf{S}_i \gamma\right) - \mathbb{E}\left(\gamma^\top \mathbf{S}_i \gamma\right)\right| \geq \epsilon\right\} \\
&\quad \leq \frac{1}{\epsilon^2} \mathrm{Var}\left(\gamma^\top \mathbf{S}_i \gamma\right) \\
&\quad = \frac{1}{\epsilon^2}\left[\mathbb{E}\left(\gamma^\top \mathbf{S}_i \gamma\right)^2 - \left\{\mathbb{E}\left(\gamma^\top \mathbf{S}_i \gamma\right)\right\}^2\right] \\
&\quad = \frac{1}{\epsilon^2}\left\{\frac{1}{T_i^2} \mathbb{E} z_{i1j_i}^4 + \frac{T_i (T_i - 1)}{T_i^2}\left(\gamma^\top \Sigma_i \gamma\right)^2 - \left(\gamma^\top \Sigma_i \gamma\right)^2\right\} \\
&\quad \xrightarrow{T_i \to \infty} 0.
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}\left(\gamma^\top \mathbf{S}_i \gamma\right)^4 \\
&\quad = \frac{1}{T_i^4} \mathbb{E}\left(\sum_{t=1}^{T_i} z_{itj_i}^2\right)^4 \\
&\quad = \frac{1}{T_i^4}\left\{\sum_t \mathbb{E} z_{itj_i}^8 + 2 \sum_{t \neq s} \mathbb{E} z_{itj_i}^4 z_{isj_i}^4 + 2 \sum_u \mathbb{E}\left(z_{iuj_i}^4 \sum_{t \neq s} z_{itj_i}^2 z_{isj_i}^2\right) + \sum_{u \neq v} \sum_{t \neq s} \mathbb{E}\left(z_{itj_i}^2 z_{isj_i}^2 z_{iuj_i}^2 z_{ivj_i}^2\right)\right\} \\
&\quad = \frac{1}{T_i^4}\left\{T_i \mathbb{E} z_{i1j_i}^8 + 2T_i (T_i - 1)\left(\mathbb{E} z_{i1j_i}^4\right)^2 + 2T_i^2 (T_i - 1) \mathbb{E} z_{i1j_i}^4 \left(\mathbb{E} z_{i1j_i}^2\right)^2 + T_i^2 (T_i - 1)^2 \left(\mathbb{E} z_{i1j_i}^2\right)^4\right\} .
\end{aligned}$$

$$\left\{\mathbb{E}\left(\boldsymbol{\gamma}^\top \mathbf{S}_i\boldsymbol{\gamma}\right)^2\right\}^2 = \frac{1}{T_i^2}\left(\mathbb{E}z_{i1j_i}^4\right)^2 + \frac{2T_i(T_i-1)}{T_i^3}\mathbb{E}z_{i1j_i}^4(\boldsymbol{\gamma}\Sigma_i\boldsymbol{\gamma})^2 + \frac{T_i^2(T_i-1)^2}{T_i^4}(\boldsymbol{\gamma}\Sigma_i\boldsymbol{\gamma})^4.$$

For $\forall\ \epsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\left\{\left|\left(\boldsymbol{\gamma}^\top \mathbf{S}_i\boldsymbol{\gamma}\right)^2 - \mathbb{E}\left(\boldsymbol{\gamma}^\top \mathbf{S}_i\boldsymbol{\gamma}\right)^2\right| \geq \epsilon\right\} \\
&\leq \frac{1}{\epsilon^2}\mathrm{Var}\left(\boldsymbol{\gamma}^\top \mathbf{S}_i\boldsymbol{\gamma}\right)^2 \\
&= \frac{1}{\epsilon^2}\left[\mathbb{E}\left(\boldsymbol{\gamma}^\top \mathbf{S}_i\boldsymbol{\gamma}\right)^4 - \left\{\mathbb{E}\left(\boldsymbol{\gamma}^\top \mathbf{S}_i\boldsymbol{\gamma}\right)^2\right\}^2\right] \\
&= \frac{1}{\epsilon^2}\left\{\frac{1}{T_i^3}\mathbb{E}z_{i1j_i}^8 + \frac{T_i-2}{T_i^3}\left(\mathbb{E}z_{i1j_i}^4\right)^2\right\} \\
&\xrightarrow{T_i\to\infty} 0.
\end{aligned}$$

Therefore, as $T_{\min} = \min_i T_i \to \infty$,

$$\mathbb{E}\left(\hat{\delta}_i^2 - \delta_i^2\right)^2 \to 0, \text{ for } i = 1, \ldots, n, \text{ and } \mathbb{E}\left(\hat{\delta}^2 - \delta^2\right)^2 \to 0.$$

(ii) Secondly, prove the consistency of $\hat{\psi}_i^2$, for $i = 1, \ldots, n$.

$$\hat{\psi}_i^2 - \psi_i^2 = \frac{1}{T_i}\left\{\boldsymbol{\gamma}^\top \mathbf{S}_i\boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2 - \mathbb{E}\left\{\boldsymbol{\gamma}^\top \mathbf{S}_i\boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2.$$

$$\begin{aligned}
\mathbb{E}\left\{\boldsymbol{\gamma}^\top \mathbf{S}_i\boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2 &= \mathbb{E}\left\{\frac{1}{T_i}\sum_t z_{itj_i}^2 - \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2 \\
&= \frac{1}{T_i^2}\sum_t \mathrm{Var}\left(z_{itj_i}^2\right) \\
&= \frac{1}{T_i}\mathrm{Var}\left(z_{i1j_i}^2\right).
\end{aligned}$$

$$\begin{aligned}
&\hat{\psi}_i^2 - \psi_i^2 \\
&= \frac{1}{T_i}\left[\left\{\boldsymbol{\gamma}^\top \mathbf{S}_i\boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2 - \mathrm{Var}\left(z_{i1j_i}^2\right)\right] \\
&= \frac{1}{T_i}\left[\left(\boldsymbol{\gamma}^\top \mathbf{S}_i\boldsymbol{\gamma}\right)^2 - \mathbb{E}z_{i1j_i}^4 - 2\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\left\{\boldsymbol{\gamma}^\top \mathbf{S}_i\boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}\right].
\end{aligned}$$

From above derivation and the fact that $\mathbb{E}\left(\boldsymbol{\gamma}^\top \mathbf{S}_i\boldsymbol{\gamma}\right) = \boldsymbol{\gamma}^\top \Sigma_i\boldsymbol{\gamma} = \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)$, as $T_i \to \infty$, for $\forall\ \epsilon > 0$,

$$\mathbb{P}\left\{\left|\left(\boldsymbol{\gamma}^\top \mathbf{S}_i\boldsymbol{\gamma}\right) - \mathbb{E}\left(\boldsymbol{\gamma}^\top \mathbf{S}_i\boldsymbol{\gamma}\right)\right| \geq \epsilon\right\} \to 0.$$

As both $(\boldsymbol{\gamma}^\top \mathbf{S}_i\boldsymbol{\gamma})^2$ and $\mathbb{E}z_{i1j_i}^4$ are bounded, then, as $T_{\min} = \min_i T_i \to \infty$,

$$\mathbb{E}\left(\widehat{\psi}_i^2 - \psi_i^2\right)^2 \to 0, \text{ for } i = 1, \ldots, n.$$

Let $\widetilde{\psi}_i^2 = \min\left(\widehat{\psi}_i^2, \widehat{\delta}_i^2\right)$.

$$\begin{aligned}
\widetilde{\psi}_i^2 - \psi_i^2 &= \min\left(\widehat{\psi}_i^2, \widehat{\delta}_i^2\right) - \psi_i^2 \\
&\leq \widehat{\psi}_i^2 - \psi_i^2 \\
&\leq \left|\widehat{\psi}_i^2 - \psi_i^2\right| \\
&\leq \max\left(\left|\widehat{\psi}_i^2 - \psi_i^2\right|, \left|\widehat{\delta}_i^2 - \delta_i^2\right|\right).
\end{aligned}$$

$\delta_i^2 = \phi_i^2 + \psi_i^2 \geq \psi_i^2$, then

$$\begin{aligned}
\widetilde{\psi}_i^2 - \psi_i^2 &= \min\left(\widehat{\psi}_i^2, \widehat{\delta}_i^2\right) - \psi_i^2 \\
&= \min\left(\widehat{\psi}_i^2 - \psi_i^2, \widehat{\delta}_i^2 - \psi_i^2\right) \\
&\geq \min\left(\widehat{\psi}_i^2 - \psi_i^2, \widehat{\delta}_i^2 - \delta_i^2\right) \\
&\geq \min\left(-\left|\widehat{\psi}_i - \psi_i^2\right|, -\left|\widehat{\delta}_i^2 - \delta_i^2\right|\right) \\
&\geq -\max\left(\left|\widehat{\psi}_i - \psi_i^2\right|, \left|\widehat{\delta}_i^2 - \delta_i^2\right|\right).
\end{aligned}$$

$$\mathbb{E}\left(\widetilde{\psi}_i^2 - \psi_i^2\right)^2 \leq \mathbb{E}\left\{\max\left(\left|\widehat{\psi}_i - \psi_i^2\right|, \left|\widehat{\delta}_i^2 - \delta_i^2\right|\right)^2\right\} \leq \mathbb{E}\left(\widehat{\psi}_i^2 - \psi_i^2\right)^2 + \mathbb{E}\left(\widehat{\delta}_i^2 - \delta_i^2\right)^2.$$

Therefore, as $T_{\min} = \min_i T_i \to \infty$,

$$\mathbb{E}\left(\widetilde{\psi}_i^2 - \psi_i^2\right)^2 \to 0, \text{ for } i = 1, \ldots, n, \text{ and } \mathbb{E}\left(\widehat{\psi}^2 - \psi^2\right)^2 \to 0.$$

(iii) Lastly, $\widehat{\phi}_i^2 = \widehat{\delta}_i^2 - \widehat{\psi}_i^2$. The consistency of $\widehat{\phi}_i^2$ (for $i = 1, \ldots, n$) and $\widehat{\phi}^2$ are straightforward. □

## A.5. Proof of Theorem 3.1

In order to prove Theorem 3.1, the following lemma is firstly introduced. This lemma is also used to prove Lemma A.2 in the next section.

**Lemma A.1**. *If $a_i^2$ is a sequence of nonnegative random variables (implicitly indexed by $T_i$) whose expectations converge to zero, for $i = 1, \ldots, n$, and $\kappa_1$, $\kappa_2$ are two nonrandom scalars, and*

$$\frac{a_i^2}{\hat{\delta}_i^{k_1}\delta_i^{\kappa_2}} \leq 2\left(\hat{\delta}_i^2 + \delta_i^2\right) a.s.,$$

*then, as $T_{\min} = \min_i T_i \to \infty$,*

$$\mathbb{E}\left(\frac{a_i^2}{\hat{\delta}_i^{\kappa_1}\delta_i^{\kappa_2}}\right) \to 0.$$

*Analogously, if $a^2$ is a sequence of nonnegative random variables (implicitly indexed by $T_{\min} = \min_i T_i$) whose expectations converge to zero, and $\kappa_1$, $\kappa_2$ are two nonrandom scalars, and*

$$\frac{a^2}{\hat{\delta}^{\kappa_1}\delta^{\kappa_2}} \leq 2\left(\hat{\delta}^2 + \delta^2\right) \quad a.s.,$$

*then, as $T_{\min} = \min_i T_i \to \infty$,*

$$\mathbb{E}\left(\frac{a^2}{\hat{\delta}^{\kappa_1}\delta^{\kappa_2}}\right) \to 0.$$

*Proof*. For a fixed $\epsilon > 0$, let $\mathcal{T}_i$ denote the set of indices $T_i$ such that $\delta_i^2 \leq \epsilon/8$. In Lemma 3.2, it is proved that $\mathbb{E}\left(\hat{\delta}_i^2 - \delta_i^2\right)^2 \to 0$. Thus, there exists an integer $T_{i1}$ such that $\forall\ T_i \geq T_{i1}$,

$$\mathbb{E}\left|\hat{\delta}_i^2 - \delta_i^2\right| \leq \epsilon/4.$$

For $\forall\ T_i \geq T_{i1}$ in the set $\mathcal{T}_i$,

$$\mathbb{E}\left(\frac{a_i^2}{\hat{\delta}_i^{\kappa_1}\delta_i^{\kappa_2}}\right) \leq 2\left(\mathbb{E}\hat{\delta}_i^2 + \delta_i^2\right) \leq 2\left(\mathbb{E}\left|\hat{\delta}_i^2 - \delta_i^2\right| + 2\delta_i^2\right) \leq 2\left(\frac{\epsilon}{4} + 2 \times \frac{\epsilon}{8}\right) = \epsilon.$$

Consider the complementary of set $\mathcal{T}_i$, since $\mathbb{E}a_i^2 \to 0$, there exists an integer $T_{i2}$ such that, $\forall$ $T_i \geq T_{i2}$,

$$\mathbb{E}a^2 \leq \frac{\epsilon^{\kappa_1 + \kappa_2 + 1}}{2^{4\kappa_1 + 3\kappa_2 + 1}}.$$

$\delta_i^2$ is bounded by $2C_2 + C_1\sqrt{C_2}$. Then, there exists an integer $T_{i3}$ such that, for $\forall\ Ti \geq Ti3$

$$\mathbb{P}\left(\left|\hat{\delta}_i^2 - \delta_i^2\right| \geq \frac{\epsilon}{16}\right) \leq \frac{4\epsilon}{16\left(2C_2 + C_1\sqrt{C_2}\right) + \epsilon}.$$

Let $\mathbf{1}_{\{\cdot\}}$ denote the indicator function. For $\forall\ T_i \geq \max(T_{i2},\ T_{i3})$ outside the set $\mathscr{T}_i$, then

$$\begin{aligned}
&\mathbb{E}\left(\frac{a_i^2}{\hat{\delta}_i^{\kappa 1}\delta_i^{\kappa 2}}\right) \\
&= \mathbb{E}\left(\frac{a_i^2}{\hat{\delta}_i^{\kappa 1}\delta_i^{\kappa 2}}\mathbf{1}\left\{\hat{\delta}_i^2 \leq \epsilon/16\right\}\right) + \mathbb{E}\left(\frac{a_i^2}{\hat{\delta}_i^{\kappa 1}\delta_i^{\kappa 2}}\mathbf{1}\left\{\hat{\delta}_i^2 > \epsilon/16\right\}\right) \\
&\leq \mathbb{E}\left\{2\left(\hat{\delta}_i^2 + \delta_i^2\right)\mathbf{1}\left\{\hat{\delta}_i^2 \leq \epsilon/16\right\}\right\} + \left(\frac{16}{\epsilon}\right)^{\kappa 1}\left(\frac{8}{\epsilon}\right)^{\kappa 2}\mathbb{E}\left(a_i^2\mathbf{1}\left\{\hat{\delta}_i^2 > \epsilon/16\right\}\right) \\
&\leq 2\left\{\left(2C_2 + C_1\sqrt{C_2}\right) + \frac{\epsilon}{16}\right\}\mathbb{P}\left(\left|\hat{\delta}_i^2 - \delta_i^2\right| \geq \frac{\epsilon}{16}\right) + \left(\frac{16}{\epsilon}\right)^{\kappa 1}\left(\frac{8}{\epsilon}\right)^{\kappa 2}\mathbb{E}\left(a_i^2\right) \\
&\leq 2\left\{\left(2C_2 + C_1\sqrt{C_2}\right) + \frac{\epsilon}{16}\right\}\frac{4\epsilon}{16\left(2C_2 + C_1\sqrt{C_2}\right) + \epsilon} + \left(\frac{16}{\epsilon}\right)^{\kappa 1}\left(\frac{8}{\epsilon}\right)^{\kappa 2}\frac{\epsilon^{\kappa 1 + \kappa 2 + 1}}{2^{4\kappa 1 + 3\kappa 2 + 1}} \\
&\leq \epsilon.
\end{aligned}$$

Bringing together the results inside and outside the set $\mathscr{T}_i$, for $\forall\ T_i \geq \max(T_{i1},\ T_{i2},\ T_{i3})$,

$$\mathbb{E}\left(\frac{a_i^2}{\hat{\delta}_i^{\kappa 1}\delta_i^{\kappa 2}}\right) \leq \epsilon.$$

The proof of the second part follows the same strategy. □

Now, we prove Theorem 3.1.

*Proof.* We first prove that $\mathbf{S}_i^*$ is a consistent estimator of $\Sigma_i^*$.

$$\begin{aligned}
\left\|\mathbf{S}_i^* - \Sigma_i^*\right\|^2 &= \max_{\boldsymbol{\gamma} \neq \mathbf{0}} \frac{\left\|\boldsymbol{\gamma}^\top(\mathbf{S}_i^* - \Sigma_i^*)\boldsymbol{\gamma}\right\|^2}{\boldsymbol{\gamma}^\top\boldsymbol{\gamma}} \\
&= \max_{\boldsymbol{\gamma} \neq \mathbf{0}} \frac{1}{\boldsymbol{\gamma}^\top\boldsymbol{\gamma}}\left\|\left(\frac{\hat{\phi}^2}{\hat{\delta}^2} - \frac{\phi^2}{\delta^2}\right)\left(\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma} - \mu\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right)\right\|^2 \\
&= \max_{\boldsymbol{\gamma} \neq \mathbf{0}} \frac{1}{\boldsymbol{\gamma}^\top\boldsymbol{\gamma}}\left(\frac{\hat{\phi}^2}{\hat{\delta}^2} - \frac{\phi^2}{\delta^2}\right)^2\hat{\delta}_i^2.
\end{aligned}$$

$$\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{S}_i^*-\Sigma_i^*\|^2=\max_{\boldsymbol{\gamma}\neq\mathbf{0}}\frac{1}{\boldsymbol{\gamma}^\top\boldsymbol{\gamma}}\frac{\left(\hat{\phi}^2\delta^2-\phi^2\hat{\delta}^2\right)^2}{\hat{\delta}^4\delta^4}\frac{1}{n}\sum_{i=1}^{n}\hat{\delta}_i^2$$
$$=\max_{\boldsymbol{\gamma}\neq\mathbf{0}}\frac{1}{\boldsymbol{\gamma}^\top\boldsymbol{\gamma}}\frac{\left(\hat{\phi}^2\delta^2-\phi^2\hat{\delta}^2\right)^2}{\hat{\delta}^2\delta^4}.$$

Using the fact that $\phi^2\leq\delta^2$ and $\hat{\phi}^2\leq\hat{\delta}^2$,

$$\frac{\left(\hat{\phi}^2\delta^2-\phi^2\hat{\delta}^2\right)^2}{\hat{\delta}^2\delta^4}\leq\hat{\delta}^2\leq 2\left(\hat{\delta}^2+\delta^2\right).$$

In Lemma 3.2, it is shown that $\mathbb{E}\left(\hat{\phi}^2-\phi^2\right)^2$ and $\mathbb{E}\left(\hat{\delta}^2-\delta^2\right)^2$ converge to zero. In addition, Lemma 3.1 shows that $\phi^2$ and $\delta^2$ are bounded. Thus,

$$\mathbb{E}\left(\hat{\phi}^2\delta^2-\phi^2\hat{\delta}^2\right)^2=\mathbb{E}\left\{\left(\hat{\phi}^2-\phi^2\right)\delta^2-\phi^2\left(\hat{\delta}^2-\delta^2\right)\right\}^2$$
$$\leq\delta^4\mathbb{E}\left(\hat{\phi}^2-\phi^2\right)^2+\phi^4\mathbb{E}\left(\hat{\delta}^2-\delta^2\right)^2$$
$$\to 0.$$

Let $a^2=\left(\hat{\phi}^2\delta^2-\phi^2\hat{\delta}^2\right)^2$, $\kappa_1=2$ and $\kappa_2=4$, then $\mathbb{E}a^2\to 0$, and using Lemma A.1,

$$\mathbb{E}\frac{\left(\hat{\phi}^2\delta^2-\phi^2\hat{\delta}^2\right)^2}{\hat{\delta}^2\delta^4}\to 0.$$

Thus,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{S}_i^*-\Sigma_i^*\|^2\to 0.$$

And therefore, for $\forall\ i$,

$$\mathbb{E}\|\mathbf{S}_i^*-\Sigma_i^*\|^2\to 0.$$

For the second statement,

$$\mathbb{E}\left|\|\mathbf{S}_i^*-\Sigma_i\|^2-\|\Sigma_i^*-\Sigma_i\|^2\right|=\mathbb{E}\left|\langle\mathbf{S}_i^*-\Sigma_i^*,\mathbf{S}_i^*+\Sigma_i^*-2\Sigma_i\rangle\right|$$
$$\leq\sqrt{\mathbb{E}\|\mathbf{S}_i^*-\Sigma_i^*\|^2}\sqrt{\mathbb{E}\|\mathbf{S}_i^*+\Sigma_i^*-2\Sigma_i\|^2}$$
$$\to\ 0.$$

Therefor,

$$\mathbb{E}\left\{\boldsymbol{\gamma}^\top \mathbf{S}_i^* \boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right\}^2 - \mathbb{E}\left\{\boldsymbol{\gamma}^\top \Sigma_i^* \boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right\}^2 \to 0.$$

□

## A.6. Proof of Theorem 3.2

Before proving Theorem 3.2, we first provide the solution to the optimization problem (3.4). Let

$$f(\rho_1, \rho_2) = \frac{1}{n}\sum_{i=1}^{n}\left\{\boldsymbol{\gamma}^\top(\rho_1 \mathbf{I} + \rho_2 \mathbf{S}_i)\boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right\}^2.$$

$$\frac{\partial f}{\partial \rho_1} = \frac{1}{n}\sum_{i=1}^{n} 2\left(\boldsymbol{\gamma}^\top \boldsymbol{\gamma}\right)\left\{\rho_1 \boldsymbol{\gamma}^\top \boldsymbol{\gamma} + \rho_2 \boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right\} = 0$$

$$\frac{\partial f}{\partial \rho_2} = \frac{1}{n}\sum_{i=1}^{n} 2\left(\boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma}\right)\left\{\rho_1\left(\boldsymbol{\gamma}^\top \boldsymbol{\gamma}\right) + \rho_2\left(\boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma}\right) - \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right\} = 0.$$

$$\Rightarrow \quad \rho_2 = \frac{\sum_i\left(\boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma}\right)\exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)/n - \left(\sum_i \boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma}/n\right)\left(\sum_i \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)/n\right)}{\sum_i\left(\boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma}\right)^2/n - \left(\sum_i \boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma}/n\right)^2}$$

$$\begin{aligned}\rho_1 &= \frac{1}{\boldsymbol{\gamma}^\top \boldsymbol{\gamma}}\left\{\frac{1}{n}\sum_{i=1}^{n}\exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right) - \frac{1}{n}\sum_{i=1}^{n}\rho_2\left(\boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma}\right)\right\} \\ &= \frac{1}{\boldsymbol{\gamma}^\top \boldsymbol{\gamma}}\left\{\frac{\left(\sum_i \boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma}/n\right)\left(\sum_i\left(\boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma}\right)\exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)/n\right)}{\sum_i\left(\boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma}\right)^2/n - \left(\sum_i \boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma}/n\right)^2} - \frac{\left(\sum_i \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)/n\right)\left(\sum_i\left(\boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma}\right)^2/n\right)}{\sum_i\left(\boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma}\right)^2/n - \left(\sum_i \boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma}/n\right)^2}\right\}.\end{aligned}$$

In order to prove Theorem 3.2, the following lemma is introduced.

**Lemma A.2**. *For given* $(\boldsymbol{\gamma}, \boldsymbol{\beta})$, *let* $T_{\min} = \min_i T_i$, *as* $T_{\min} \to \infty$, *for* $\forall\ i \in \{1, \ldots, n\}$,

$$\mathbb{E}\left(\left|\frac{\hat{\phi}_i^2 \hat{\psi}_i^2}{\hat{\delta}_i^2} - \frac{\phi_i^2 \psi_i^2}{\delta_i^2}\right|\right) \to 0.$$

*Then, as n,* $T_{\min} \to \infty$,

$$\mathbb{E}\left(\left|\frac{\hat{\phi}^2\hat{\psi}^2}{\hat{\delta}^2}-\frac{\phi^2\psi^2}{\delta^2}\right|\right)\to 0.$$

*Proof.*

$$\frac{\hat{\phi}_i^2\hat{\psi}_i^2}{\hat{\delta}_i^2}-\frac{\phi_i^2\psi_i^2}{\delta_i^2}=\frac{\hat{\phi}_i^2\hat{\psi}_i^2\delta_i^2-\phi_i^2\psi_i^2\hat{\delta}_i^2}{\hat{\delta}_i^2\delta_i^2}.$$

Let $a_i^2=\left|\hat{\phi}_i^2\hat{\psi}_i^2\delta_i^2-\phi_i^2\psi_i^2\hat{\delta}_i^2\right|$, $\kappa_1=2$ and $\kappa_2=2$. First need to verify the assumptions in Lemma A.1.

$$\left|\frac{\hat{\phi}_i^2\hat{\psi}_i^2}{\hat{\delta}_i^2}-\frac{\phi_i^2\psi_i^2}{\delta_i^2}\right|\le\frac{\hat{\phi}_i^2\hat{\psi}_i^2}{\hat{\delta}_i^2}+\frac{\phi_i^2\psi_i^2}{\delta_i^2}\le\hat{\phi}_i^2+\phi_i^2\le\hat{\delta}_i^2+\delta_i^2\le 2\left(\hat{\delta}_i^2+\delta_i^2\right),\quad \text{a.s..}$$

Furthermore,

$$\begin{aligned}
&\mathbb{E}\left(\left|\hat{\phi}_i^2\hat{\psi}_i^2\delta_i^2-\phi_i^2\psi_i^2\hat{\delta}_i^2\right|\right)\\
&=\mathbb{E}\left\{\left|\left(\hat{\phi}_i^2\hat{\psi}_i^2-\phi_i^2\psi_i^2\right)\delta_i^2-\phi_i^2\psi_i^2\left(\hat{\delta}_i^2-\delta_i^2\right)\right|\right\}\\
&=\mathbb{E}\left\{\left|\left(\hat{\phi}_i^2-\phi_i^2\right)\left(\hat{\psi}_i^2-\psi_i^2\right)\delta_i^2+\phi_i^2\left(\hat{\psi}_i^2-\psi_i^2\right)\delta_i^2+\left(\hat{\phi}_i^2-\phi_i^2\right)\psi_i^2\delta_i^2-\phi_i^2\psi_i^2\left(\hat{\delta}_i^2-\delta_i^2\right)\right|\right\}\\
&\le\sqrt{\mathbb{E}\left(\hat{\phi}_i^2-\phi_i^2\right)^2}\sqrt{\mathbb{E}\left(\hat{\psi}_i^2-\psi_i^2\right)^2}\delta_i^2+\phi_i^2\mathbb{E}\left|\hat{\psi}_i^2-\psi_i^2\right|\delta_i^2+\mathbb{E}\left|\hat{\phi}_i^2-\phi_i^2\right|\psi_i^2\delta_i^2-\phi_i^2\psi_i^2\mathbb{E}\left|\hat{\delta}_i^2-\delta_i^2\right|.
\end{aligned}$$

The right-hand side converges to zero. Therefore, $\mathbb{E}a_i^2\to 0$, conditions in Lemma A.1 are satisfied. Therefore,

$$\mathbb{E}\left|\frac{\hat{\phi}_i^2\hat{\psi}_i^2}{\hat{\delta}_i^2}-\frac{\phi_i^2\psi_i^2}{\delta_i^2}\right|\to 0.$$

Analogously, it can be shown that

$$\mathbb{E}\left|\frac{\hat{\phi}^2\hat{\psi}^2}{\hat{\delta}^2}-\frac{\phi^2\psi^2}{\delta^2}\right|\to 0.$$

□

Next, we prove Theorem

*Proof.* Let $\alpha_i=(\boldsymbol{\gamma}^\top\Sigma_i\boldsymbol{\gamma})(\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma})-\{\mu(\boldsymbol{\gamma}^\top\boldsymbol{\gamma})\}^2$ and $\alpha=\sum_{i=1}^n\alpha_i/n$.
$\mathbb{E}(\alpha_i)=\exp^2\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)-\mu^2\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right)^2$, then

$$\mathbb{E}\alpha = \frac{1}{n}\sum_{i=1}^{n} \exp^2\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right) - \mu^2\left(\boldsymbol{\gamma}^\top \boldsymbol{\gamma}\right) = \phi^2 .$$

First, need to prove that $\alpha - \phi^2$ converges to zero in quadratic mean.

$$\begin{aligned}
&\mathrm{Var}(\alpha_i) \\
&\quad = \mathrm{Var}\left\{\left(\boldsymbol{\gamma}^\top \Sigma_i \boldsymbol{\gamma}\right)\left(\boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma}\right) - \mu^2\left(\boldsymbol{\gamma}^\top \boldsymbol{\gamma}\right)^2\right\} \\
&\quad = \mathrm{Var}\left\{\left(\boldsymbol{\gamma}^\top \Sigma_i \boldsymbol{\gamma}\right)\left(\boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma}\right)\right\} + \mathrm{Var}\left\{\mu^2\left(\boldsymbol{\gamma}^\top \boldsymbol{\gamma}\right)^2\right\} - 2\mathrm{Cov}\left\{\left(\boldsymbol{\gamma}^\top \Sigma_i \boldsymbol{\gamma}\right)\left(\boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma}\right), \mu^2\left(\boldsymbol{\gamma}^\top \boldsymbol{\gamma}\right)^2\right\} \\
&\quad = \mathrm{Var}\left\{\left(\boldsymbol{\gamma}^\top \Sigma_i \boldsymbol{\gamma}\right)\left(\boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma}\right)\right\} .
\end{aligned}$$

$$\left(\boldsymbol{\gamma}^\top \Sigma_i \boldsymbol{\gamma}\right)\left(\boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma}\right) = \lambda_{ij_i}\left(\frac{1}{T_i}\sum_{t=1}^{T_i} z_{itj_i}^2\right).$$

$$\begin{aligned}
\mathrm{Var}\left\{\left(\boldsymbol{\gamma}^\top \Sigma_i \boldsymbol{\gamma}\right)\left(\boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma}\right)\right\} &= \mathrm{Var}\left\{\frac{1}{T_i}\sum_{t=1}^{T_i} \lambda_{ij_i} z_{itj_i}^2\right\} \\
&= \frac{1}{T_i}\mathrm{Var}\left(\lambda_{ij_i} z_{i1j_i}^2\right) \\
&\le \frac{1}{T_i}\mathbb{E}\left(\lambda_{ij_i} z_{i1j_i}^2\right)^2 \\
&\le \frac{1}{T_i}\mathbb{E}\lambda_{ij_i}^2 z_{i1j_i}^4 \\
&\le \frac{1}{T_i}\left(\mathbb{E} z_{i1j_i}^2\right)^2 \mathbb{E} z_{i1j_i}^4 \\
&\le \frac{1}{T_i}\left(\mathbb{E} z_{i1j_i}^4\right)^2 \\
&\le \frac{1}{T_i}\mathbb{E} z_{i1j_i}^8 \\
&\le \frac{C_2}{T_i} .
\end{aligned}$$

$$\mathrm{Var}(\alpha) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(\alpha_i) \le \frac{C_2}{n^2}\sum_{i=1}^{n}\frac{1}{T_i} \to 0, \text{ as } T_{\min} = \min_i T_i \to \infty .$$

This proves that $\alpha - \phi^2$ converges to 0 in quadratic mean. In the following, we prove that $\mathbf{S}_i^*$ is a consistent estimator of $\Sigma_i^{**}$.

$$\begin{aligned}
\mathbf{S}_i^* &= \frac{\hat{\psi}^2}{\hat{\delta}^2}\mu\mathbf{I} + \frac{\hat{\phi}^2}{\hat{\delta}^2}\mathbf{S}_i = \frac{\hat{\delta}^2 - \hat{\psi}^2}{\hat{\delta}^2}\mu\mathbf{I} + \frac{\hat{\phi}^2}{\hat{\delta}^2}\mathbf{S}_i = \mu\mathbf{I} + \frac{\hat{\phi}^2}{\hat{\delta}^2}(\mathbf{S}_i - \mu\mathbf{I}) . \\
\Sigma_i^{**} &= \rho_1\mathbf{I} + \rho_2\mathbf{S}_i = (\rho_1 + \rho_2\mu)\mathbf{I} + \rho_2(\mathbf{S}_i - \mu\mathbf{I}) .
\end{aligned}$$

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{S}_i^* - \Sigma_i^{**}\|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\|(\mu-\rho_1-\rho_2\mu)\mathbf{I} + \left(\frac{\hat{\phi}^2}{\hat{\delta}^2}-\rho_2\right)(\mathbf{S}_i-\mu\mathbf{I})\right\|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\max_{\boldsymbol{\gamma}\neq 0}\frac{1}{\boldsymbol{\gamma}^\top\boldsymbol{\gamma}}\left\|(\mu-\rho_1-\rho_2\mu)\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right) + \left(\frac{\hat{\phi}^2}{\hat{\delta}^2}-\rho_2\right)\left(\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}-\mu\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right)\right)\right\|^2\right\} \\
&\max_{\boldsymbol{\gamma}\neq 0}\left\{(\mu-\rho_1-\rho_2\mu)^2\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right) + \frac{1}{\boldsymbol{\gamma}^\top\boldsymbol{\gamma}}\left(\frac{\hat{\phi}^2}{\hat{\delta}^2}-\rho_2\right)^2\hat{\delta}_i^2 + 2(\mu-\rho_1-\rho_2\mu)\left(\frac{\hat{\phi}^2}{\hat{\delta}^2}-\rho_2\right)\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}-\mu\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right)\right)\right\}.
\end{aligned}
$$

$$
\begin{aligned}
&(\mu-\rho_1-\rho_2\mu)^2 \\
&= \frac{\left(\sum_i\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}/n - \sum_i\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)/n\right)^2}{\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right)^2\left\{\left(\sum_i\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}/n\right)^2 - \sum_i\left(\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}\right)^2/n\right\}^2}\cdot\frac{\left\{\left(\sum_i\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}/n\right)\left(\sum_i\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)/n\right) - \sum_i\left(\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}\right)\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)/n\right\}^2}{\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right)^2\left\{\left(\sum_i\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}/n\right)^2 - \sum_i\left(\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}\right)^2/n\right\}^2}
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}\left\{\frac{1}{n}\sum_i\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma} - \frac{1}{n}\sum_i\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2 \\
&= \frac{1}{n^2}\sum_i\mathbb{E}\left\{\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2 + \frac{1}{n^2}\sum_{i\neq i'}\mathbb{E}\left\{\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}\left\{\boldsymbol{\gamma}^\top\mathbf{S}_{i'}\boldsymbol{\gamma} - \exp\left(\mathbf{x}_{i'}^\top\boldsymbol{\beta}\right)\right\}.
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\left\{\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2 &= \mathbb{E}\left\{\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma} - \mathbb{E}\left(\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}\right)\right\}^2 \\
&= \mathrm{Var}\left(\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}\right) \\
&= \frac{1}{T_i}\mathbb{E}z_{i1}^4 j_i + \frac{T_i(T_i-1)}{T_i^2}\left(\boldsymbol{\gamma}^\top\Sigma_i\boldsymbol{\gamma}\right)^2 - \left(\boldsymbol{\gamma}^\top\Sigma_i\boldsymbol{\gamma}\right)^2 \\
&\xrightarrow{T_i\to\infty} 0.
\end{aligned}
$$

It is assumed that the samples/subjects are independent, therefore,

$$
\mathbb{E}\left\{\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma} - \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}\left\{\boldsymbol{\gamma}^\top\mathbf{S}_{i'}\boldsymbol{\gamma} - \exp\left(\mathbf{x}_{i'}^\top\boldsymbol{\beta}\right)\right\} = 0.
$$

Thus,

$$
\mathbb{E}\left\{\frac{1}{n}\sum_i\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma} - \frac{1}{n}\sum_i\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2 \to 0, \text{ as } T_{\min}\to\infty.
$$

$$
\begin{aligned}
&\mathbb{E}\left\{\left(\frac{1}{n}\sum_i\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}\right)\left(\frac{1}{n}\sum_i\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right) - \frac{1}{n}\sum_i\left(\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}\right)\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2 \\
&\leq \mathbb{E}\left(\frac{1}{n}\sum_i\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}\right)^2\left(\frac{1}{n}\sum_i\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right)^2 + \mathbb{E}\left\{\frac{1}{n}\sum_i\left(\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}\right)\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2.
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}\left(\frac{1}{n}\sum_i \gamma^\top \mathbf{S}_i\gamma\right)^2 \\
&= \frac{1}{n^2}\sum_i \mathbb{E}\left(\gamma^\top \mathbf{S}_i\gamma\right)^2 + \frac{1}{n^2}\sum_{i\neq i'} \mathbb{E}\left(\gamma^\top \mathbf{S}_i\gamma\right)\left(\gamma^\top \mathbf{S}_{i'}\gamma\right) \\
&= \frac{1}{n^2}\sum_i \left\{\frac{1}{T_i^2}\mathbb{E}z_{itj_i}^4 + \frac{1}{T_i^2}\sum_{t\neq s}\mathbb{E}z_{itj_i}^2 z_{isj_i}^2\right\} + \frac{1}{n^2}\sum_{i\neq i'}\left(\frac{1}{T_i^2}\sum_{t=1}^{T_i}\mathbb{E}z_{itj_i}^2\right)\left(\frac{1}{T_{i'}^2}\sum_{t=1}^{T_{i'}}\mathbb{E}z_{i't j_{i'}}^2\right) \\
&= \frac{1}{n^2}\sum_i \left\{\frac{1}{T_i}\mathbb{E}z_{i1j_i}^4 + \frac{T_i(T_i-1)}{T_i^2}\left(\gamma^\top \Sigma_i\gamma\right)^2\right\} + \frac{1}{n^2}\sum_{i\neq i'}\left(\frac{1}{T_i}\left(\gamma^\top \Sigma_i\gamma\right)\right)\left(\frac{1}{T_{i'}}\left(\gamma^\top \Sigma_{i'}\gamma\right)\right) \\
&\xrightarrow{T_{\min}\to\infty} \frac{1}{n^2}\sum_i \left(\gamma^\top \Sigma_i\gamma\right)^2.
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}\left\{\frac{1}{n}\sum_i \left(\gamma^\top \mathbf{S}_i\gamma\right)\exp\left(\mathbf{x}_i^\top\beta\right)\right\}^2 \\
&= \frac{1}{n^2}\sum_i\sum_i \mathbb{E}\left(\gamma^\top \mathbf{S}_i\gamma\right)^2\exp^2\left(\mathbf{x}_i^\top\beta\right) + \frac{1}{n^2}\sum_{i\neq i'}\mathbb{E}\left(\gamma^\top \mathbf{S}_i\gamma\right)\exp\left(\mathbf{x}_i^\top\beta\right)\mathbb{E}\left(\gamma^\top \mathbf{S}_{i'}\gamma\right)\exp\left(\mathbf{x}_{i'}^\top\beta\right) \\
&= \frac{1}{n^2}\sum_i \left\{\frac{1}{T_i}\mathbb{E}z_{itj_i}^4 + \frac{T_i(T_i-1)}{T_i^2}\left(\gamma^\top \Sigma_i\gamma\right)^2\right\}\left(\gamma^\top \Sigma_i\gamma\right)^2 + \frac{1}{n^2}\sum_{i\neq i'}\left(\gamma^\top \Sigma_i\gamma\right)^2\left(\gamma^\top \Sigma_{i'}\gamma\right)^2 \\
&\xrightarrow{T_{\min}\to\infty} \frac{1}{n^2}\sum_i \left(\gamma^\top \Sigma_i\gamma\right)^4 + \frac{1}{n^2}\sum_{i\neq i'}\left(\gamma^\top \Sigma_i\gamma\right)^2\left(\gamma^\top \Sigma_{i'}\gamma\right)^2.
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}\left\{\left(\frac{1}{n}\sum_i \gamma^\top \mathbf{S}_i\gamma\right)\left(\frac{1}{n}\sum_i \exp\left(\mathbf{x}_i^\top\beta\right)\right) - \frac{1}{n}\sum_i \left(\gamma^\top \mathbf{S}_i\gamma\right)\exp\left(\mathbf{x}_i^\top\beta\right)\right\}^2 \\
&\leq \mathbb{E}\left(\frac{1}{n}\sum_i \gamma^\top \mathbf{S}_i\gamma\right)^2\left(\frac{1}{n}\sum_i \exp\left(\mathrm{x}_i^\top\beta\right)\right)^2 + \mathbb{E}\left\{\frac{1}{n}\sum_i \left(\gamma^\top \mathbf{S}_i\gamma\right)\exp\left(\mathbf{x}_i^\top\beta\right)\right\}^2 \\
&\xrightarrow{T_{\min}\to\infty} \frac{1}{n^2}\sum_i \left(\gamma^\top \Sigma_i\gamma\right)^2\left(\frac{1}{n}\sum_i \left(\gamma^\top \Sigma_i\gamma\right)\right)^2 + \frac{1}{n^2}\sum_i \left(\gamma^\top \Sigma_i\gamma\right)^4 + \frac{1}{n^2}\sum_{i\neq i'}\left(\gamma^\top \Sigma_i\gamma\right)^2\left(\gamma^\top \Sigma_{i'}\gamma\right)^2.
\end{aligned}
$$

The above quantity on the right is bounded by a constant from above. Therefore, as $T_{\min} \to \infty$,

$$(\mu - \rho_1 - \rho_2\mu)^2 \to 0.$$

$$\left(\frac{\hat{\phi}^2}{\hat{\delta}^2} - \rho_2\right)^2 = \left(\frac{\hat{\phi}^2}{\hat{\delta}^2} - \frac{\phi^2}{\hat{\delta}^2}\right)^2 + \left(\frac{\phi^2}{\hat{\delta}^2} - \frac{\alpha}{\hat{\delta}^2}\right)^2 + \left(\frac{\alpha}{\hat{\delta}^2} - \rho_2\right)^2.$$

Since $\hat{\delta}^4$ is bounded,

$$\mathbb{E}\left(\hat{\phi}^2 - \phi^2\right)^2 \to 0 \quad \Rightarrow \quad \mathbb{E}\left(\frac{\hat{\phi}^2}{\hat{\delta}^2} - \frac{\phi^2}{\hat{\delta}^2}\right)^2 \to 0.$$

$$\mathbb{E}\left(\phi^2-\alpha\right)^2\to 0 \quad\Rightarrow\quad \mathbb{E}\left(\frac{\phi^2}{\hat{\delta}^2}-\frac{\alpha}{\hat{\delta}^2}\right)^2\to 0.$$

Let $\rho_2=\rho_2^{(1)}/\rho_2^{(2)}$, where

$$\rho_2^{(1)}=\frac{1}{n}\sum_i\left(\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}\right)\exp\left(\mathrm{x}_i^\top\boldsymbol{\beta}\right)-\left(\frac{1}{n}\sum_i\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}\right)\left(\frac{1}{n}\sum_i\exp\left(\mathrm{x}_i^\top\boldsymbol{\beta}\right)\right),$$
$$\rho_2^{(2)}=\frac{1}{n}\sum_i\left(\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}\right)^2-\left(\frac{1}{n}\sum_i\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}\right)^2.$$

$$\begin{aligned}&\mathbb{E}\left(\alpha-\rho_2^{(1)}\right)^2\\&\quad=\left(\frac{1}{n}\sum_i\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right)^2\mathbb{E}\left\{\frac{1}{n}\sum_i\left(\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}\right)-\frac{1}{n}\sum_i\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2\\&\quad\to 0.\end{aligned}$$

$$\begin{aligned}&\hat{\delta}^2\\&\quad=\frac{1}{n}\sum_{i=1}^n\left\{\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}-\mu\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right)\right\}^2\\&\quad=\frac{1}{n}\sum_{i=1}^n\left(\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}\right)^2-2\left(\frac{1}{n}\sum_{i=1}^n\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}\right)\left(\frac{1}{n}\sum_{i=1}^n\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right)+\left(\frac{1}{n}\sum_{i=1}^n\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right)^2.\end{aligned}$$

It can be concluded that as $T_{\min}\to\infty$,

$$\mathbb{E}\left(\hat{\delta}-\rho_2^2\right)^2=\mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^n\left(\boldsymbol{\gamma}^\top\mathbf{S}_i\boldsymbol{\gamma}\right)-\frac{1}{n}\sum_{i=1}^n\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2\to 0,$$

and

$$\mathbb{E}\left(\frac{\hat{\phi}^2}{\hat{\delta}^2}-\rho_2\right)^2\to 0.$$
$$\mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^n\left\|\mathbf{S}_i^*-\Sigma_i^{*\,*}\right\|^2\right\}\to 0,\quad\Rightarrow\quad\mathbb{E}\left\|\mathbf{S}_i^*-\Sigma_i^{*\,*}\right\|^2\to 0.$$

This implies that

$$\mathbb{E}\left\{\boldsymbol{\gamma}^\top\mathbf{S}_i^*\boldsymbol{\gamma}-\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2-\mathbb{E}\left\{\boldsymbol{\gamma}^\top\Sigma_i^{*\,*}\boldsymbol{\gamma}-\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2\to 0.$$

## A.7. Proof of Theorem 3.3

*Proof.* For the first statement,

$$\begin{aligned}
&\lim_{T_{\min}\to\infty}\inf_{T_i\geq T_{\min}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{\boldsymbol{\gamma}^\top\widehat{\Sigma}_i\boldsymbol{\gamma}-\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{\boldsymbol{\gamma}^\top\mathbf{S}_i^*\boldsymbol{\gamma}-\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2\right]\\
&\geq\inf\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{\boldsymbol{\gamma}^\top\widehat{\Sigma}_i\boldsymbol{\gamma}-\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{\boldsymbol{\gamma}^\top\Sigma_i^{**}\boldsymbol{\gamma}-\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2\right]\\
&+\lim\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{\boldsymbol{\gamma}^\top\Sigma_i^{**}\boldsymbol{\gamma}-\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{\boldsymbol{\gamma}^\top\mathbf{S}_i^*\boldsymbol{\gamma}-\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2\right].
\end{aligned}$$

By Theorem 3.2, the second term on the right converges to zero, and the first term is ≥ 0 by the definition of $\Sigma_i^{**}$.

For the second statement,

$$\begin{aligned}
&\lim_{T_{\min}\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{\boldsymbol{\gamma}^\top\widehat{\Sigma}_i\boldsymbol{\gamma}-\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{\boldsymbol{\gamma}^\top\mathbf{S}_i^*\boldsymbol{\gamma}-\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2\right]=0\\
&\Leftrightarrow\lim_{T_{\min}\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{\boldsymbol{\gamma}^\top\widehat{\Sigma}_i\boldsymbol{\gamma}-\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{\boldsymbol{\gamma}^\top\Sigma_i^{**}\boldsymbol{\gamma}-\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2\right]=0\\
&\Leftrightarrow\lim_{T_{\min}\to\infty}\mathbb{E}\left\{\boldsymbol{\gamma}^\top\widehat{\Sigma}_i\boldsymbol{\gamma}-\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2-\mathbb{E}\left\{\boldsymbol{\gamma}^\top\Sigma_i^{**}\boldsymbol{\gamma}-\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}^2=0\\
&\Leftrightarrow\lim_{T_{\min}\to\infty}\mathbb{E}\left\|\boldsymbol{\gamma}^\top\widehat{\Sigma}_i\boldsymbol{\gamma}-\boldsymbol{\gamma}^\top\Sigma_i^{**}\boldsymbol{\gamma}\right\|^2=0\\
&\Leftrightarrow\lim_{T_{\min}\to\infty}\mathbb{E}\left\|\boldsymbol{\gamma}^\top\widehat{\Sigma}_i\boldsymbol{\gamma}-\boldsymbol{\gamma}^\top\mathbf{S}_i^*\boldsymbol{\gamma}\right\|^2=0\\
&\Leftrightarrow\lim_{T_{\min}\to\infty}\mathbb{E}\left\|\widehat{\Sigma}_i-\mathbf{S}_i^*\right\|^2=0.
\end{aligned}$$

This finishes the proof of this theorem.□

## A.8. $\mathbf{S}_i^*$ is well-conditioned

In this section, we show that the proposed estimator $\mathbf{S}_i^*$ is well-conditioned and thus, invertible. This is achieved by two steps: for $i = 1, \ldots, n$, (1) prove that the largest eigenvalue of $\mathbf{S}_i^*$ is bounded in probability; (2) prove that the smallest eigenvalue of $\mathbf{S}_i^*$ is bounded away from zero in probability. The proof follows the same strategy as in Ledoit and Wolf [24], but considers the case with multiple covariance matrices.

The covariance matrix $\Sigma_i$ has the eigendecomposition as $\Sigma_i = \Pi_i\Lambda_i\Pi_i^\top$. Let $\mathbf{U}_i = \Lambda_i^{-1/2}\mathbf{Y}_i$. Denote $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ as the maximum and minimum eigenvalue of a matrix $\mathbf{A}$, respectively.

$$\lambda_{\max}(\mathbf{S}_i^*) = \lambda_{\max}\left(\frac{\hat{\psi}^2}{\hat{\delta}^2}\mu\mathbf{I} + \frac{\hat{\phi}^2}{\hat{\delta}^2}\mathbf{S}_i\right) = \frac{\hat{\psi}^2}{\hat{\delta}^2}\mu + \frac{\hat{\phi}^2}{\hat{\delta}^2}\lambda_{\max}(\mathbf{S}_i).$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n}\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right) = \frac{1}{n}\sum_{i=1}^{n}\lambda_{ij_i} \le \max_i \lambda_{\max}(\Lambda_i).$$

$$\begin{aligned}\lambda_{\max}(\mathbf{S}_i) &= \lambda_{\max}\left(\frac{1}{T_i}\Lambda^{1/2}\mathbf{U}_i\mathbf{U}_i^\top\Lambda_i^{1/2}\right)\\ &\le \lambda_{\max}\left(\frac{1}{T_i}\mathbf{U}_i\mathbf{U}_i^\top\right)\lambda_{\max}(\Lambda_i)\\ &\le \lambda_{\max}\left(\frac{1}{T_i}\mathbf{U}_i\mathbf{U}_i^\top\right)\max_i \lambda_{\max}(\Lambda_i).\end{aligned}$$

Assume that $p/T_{\max}$ converges to a limit, denoted as $c$. Based on Assumption A1, $c \le C_1$. Based on the results in Yin et al. [39], as $T_{\min} = \min_i T_i \to \infty$, for $i = 1, \ldots, n$,

$$\lim \lambda_{\max}\left(\frac{1}{T_i}\mathbf{U}_i\mathbf{U}_i^\top\right) = (1+\sqrt{c})^2, \text{ a.s.}$$

This implies that

$$\mathbb{P}\left\{\lambda_{\max}(\mathbf{S}_i^*) \le (1+\sqrt{c})^2 \max_i \lambda_{\max}(\Lambda_i)\right\} \to 1,$$

and

$$\mathbb{P}\left\{\lambda_{\max}(\mathbf{S}_i^*) \le \left(1+\sqrt{C_1}\right)^2 \max_i \lambda_{\max}(\Lambda_i)\right\} \to 1.$$

Therefore, if $p/T_{\max}$ converges to a constant, the largest eigenvalue of $\mathbf{S}_i^*$ is bounded in probability. If $p/T_{\max}$ has no limit, under Assumption A1, there exists a subsequence such that $p/T_{\max}$ converges. Along this sequence, the largest eigenvalue of $\mathbf{S}_i^*$ is bounded in probability. This is true for any converging sequence, and in addition, the upper bound is independent of the particular subsequence. As a result, it holds for the whole sequence.

Next, we show that the smallest eigenvalue of $\mathbf{S}_i^*$ is bounded away from zero in probability. Analogously, we have

$$\begin{aligned}\lambda_{\min}(\mathbf{S}_i) &= \lambda_{\min}\left(\frac{1}{T_i}\Lambda^{1/2}\mathbf{U}_i\mathbf{U}_i^\top\Lambda_i^{1/2}\right)\\ &\geq \lambda_{\min}\left(\frac{1}{T_i}\mathbf{U}_i\mathbf{U}_i^\top\right)\lambda_{\min}(\Lambda_i)\\ &\geq \lambda_{\min}\left(\frac{1}{T_i}\mathbf{U}_i\mathbf{U}_i^\top\right)\min_i\lambda_{\min}(\Lambda_i).\end{aligned}$$

First, assume $p/T_{\max}$ converges to a constant $c$. If $c \in (0, 1)$, based on the results in Bai and Yin [4],

$$\lim\ \lambda_{\min}\left(\frac{1}{T_i}\mathbf{U}_i\mathbf{U}_i^\top\right) = (1-\sqrt{c})^2,\quad \text{a.s.}$$

Assume $c \leq 1 - \kappa$ for some $\kappa \in (0, 1)$. One can conclude that

$$\mathbb{P}\left\{\lambda_{\min}(\mathbf{S}_i^*) \geq (1-\sqrt{1-\kappa})^2\min_i\lambda_{\min}(\Lambda_i)\right\} \to 1.$$

When $c > 1 - \kappa$, we propose to identify a lower bound from the following

$$\lambda_{\min}(\mathbf{S}_i^*) = \lambda_{\min}\left(\frac{\hat{\psi}^2}{\hat{\delta}^2}\mu\mathbf{I} + \frac{\hat{\phi}^2}{\hat{\delta}^2}\mathbf{S}_i\right) \geq \frac{\hat{\psi}^2}{\hat{\delta}^2}\mu.$$

Compare the right-hand side in the above to it population counterpart,

$$\frac{\hat{\psi}^2}{\hat{\delta}^2}\mu - \frac{\psi^2}{\delta^2}\mu = \mu\left\{\frac{\hat{\psi}^2-\psi^2}{\delta^2} + \hat{\psi}^2\left(\frac{1}{\hat{\delta}^2} - \frac{1}{\delta^2}\right)\right\}.$$

From Lemmas 3.1 and 3.2, we can show that the above converges to zero in probability. First, consider $\psi^2 = \sum_{i=1}^n \psi_i^2/n$, where $\psi_i^2 = \mathbb{E}\left\{\boldsymbol{\gamma}^\top(\mathbf{S}_i - \Sigma_i)\boldsymbol{\gamma}\right\}^2$. From the proof of Lemma 3.1,

$$\begin{aligned}\mathbb{E}\|\mathbf{S}_i - \Sigma_i\|^2 &= \frac{1}{pT_i}\sum_{j=1}^p\sum_{k=1}^p \mathbb{E}\left(z_{i1j}^2 z_{i1k}^2\right) - \frac{1}{pT_i}\sum_{j=1}^p\sum_{k=1}^p \lambda_{ijk}^2\\ &= \frac{p}{T_i}\left\{\frac{1}{p^2}\sum_{j=1}^p\sum_{k=1}^p \mathbb{E}\left(z_{i1j}^2 z_{i1k}^2\right)\right\} - \frac{1}{pT_i}\sum_{j=1}^p \lambda_{ijj}^2.\end{aligned}$$

As $T_{\min} \to \infty$, the second term on the right-hand side converges to zero. For $\epsilon > 0$, there exists a constant $M > 0$ such that when $T_{\min} > M$, $\sum_{j=1}^p \lambda_{ijj}^2/(pT_i) < \epsilon$. Thus, $\psi_i^2 \geq (1-\kappa) - \epsilon$ and $\psi^2 \geq (1-\kappa) - \epsilon$.

$$\begin{aligned}\lambda_{\min}(\mathbf{S}_i^*) &\geq \frac{\widehat{\psi}^2}{\hat{\delta}^2}\mu \\ &= \frac{\psi^2}{\delta^2}\mu + \left(\frac{\widehat{\psi}^2}{\hat{\delta}^2}\mu - \frac{\psi^2}{\delta^2}\mu\right) \\ &\geq \frac{\psi^2}{\delta^2}\mu - \epsilon \\ &\geq \frac{\psi^2}{2C_2 + C_1\sqrt{C_2}} - \epsilon \\ &\geq \frac{(1-\kappa) - \epsilon}{2C_2 + C_1\sqrt{C_2}} - \epsilon .\end{aligned}$$

For a choice of $\epsilon$, we have

$$\mathbb{P}\left\{\lambda_{\min}(\mathbf{S}_i^*) \geq \frac{1-\kappa}{2\left(2C_2 + C_1\sqrt{C_2}\right)}\right\} \to 1.$$

Therefore, for both $c \leq 1 - \kappa$ and $c > 1 - \kappa$, the smallest eigenvalue of $\mathbf{S}_i^*$ is bounded away from zero. Analogous to the proof of the largest eigenvalue, for the case that $p/T_{\max}$ does not have a limit, we can also have the conclusion for the whole sequence. Since both the largest and the smallest eigenvalues are bounded, $\mathbf{S}_i^*$ is well-conditioned and invertible.

## A.9. Proof of Lemma 3.3 and Theorem 3.4

We first proof Lemma 3.3.

Proof.

$$\begin{aligned}\mathbb{E}\left(\boldsymbol{\gamma}^\top \Sigma_i^* \boldsymbol{\gamma}\right) &= \frac{\psi^2}{\delta^2}\mu\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right) + \frac{\phi^2}{\delta^2}\mathbb{E}\left(\boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma}\right) \\ &= \frac{\psi^2}{\delta^2}\mu\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right) + \frac{\phi^2}{\delta^2}\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right) \\ &= \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}^*\right).\end{aligned}$$

$$\frac{\sum_i \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}^*\right)/n}{\sum_i \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)/n} = \frac{\psi^2}{\delta^2}\frac{\mu\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right)}{\sum_i \exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)/n} + \frac{\phi^2}{\delta^2} = \frac{\psi^2}{\delta^2} + \frac{\phi^2}{\delta^2} = 1.$$

$$\Rightarrow \quad \frac{1}{n}\sum_{i=1}^{n}\exp\left(\mathrm{x}_i^\top\boldsymbol{\beta}^*\right) = \frac{1}{n}\sum_{i=1}^{n}\exp\left(\mathrm{x}_i^\top\boldsymbol{\beta}\right).$$

Therefore,

$$\boldsymbol{\beta}^* = \boldsymbol{\beta}.$$

Next, we prove that the proposed estimator $\boldsymbol{\beta}$ is a consistent estimator (Theorem 3.4).

*Proof.* Using the consistency of pseudo-likelihood estimator [16] and the conclusion in Lemma 3.3, $\hat{\beta}$ is a consistent estimator of $\boldsymbol{\beta}$. □

# Appendix B: Additional Simulation Results

## B.1. $\gamma$ unknown

Here, we present the performance of estimating the fourth dimension (D4) when $\boldsymbol{\gamma}$ is unknown (Figure B.1). From the figures, as $n$ and $T$ increase, the estimate of the covariance matrices, the projection and the model coefficient converge to the truth.

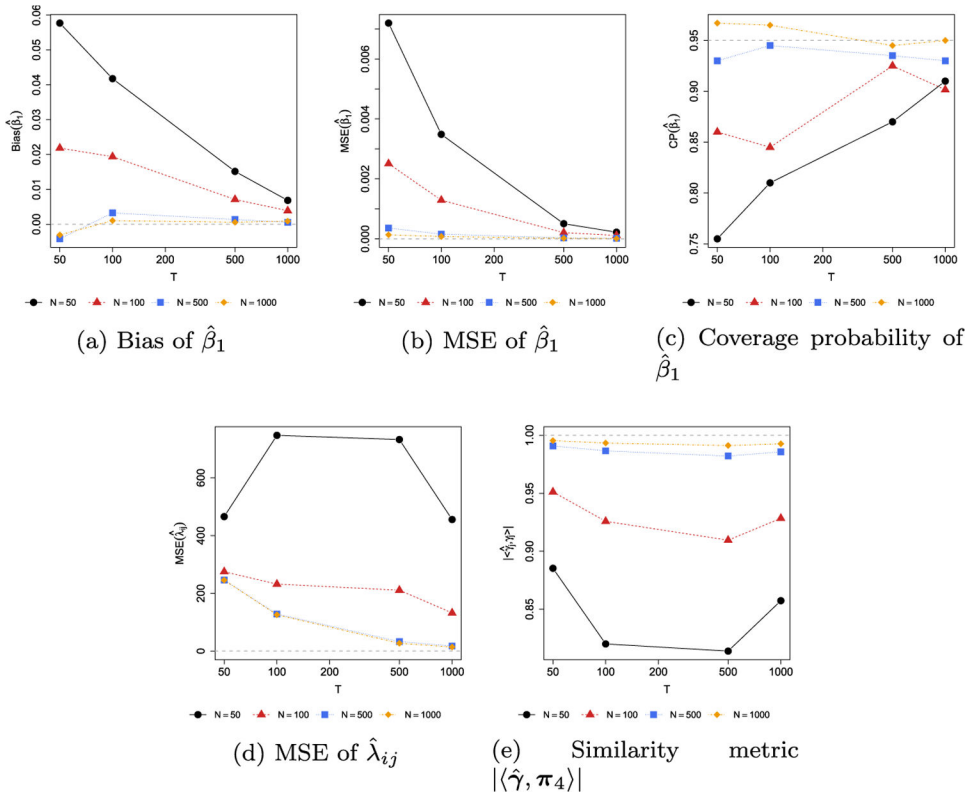
## B.2. Model misspecification

### B.2.1. Model misspecification in β

In this section, we examine the performance of the proposed approach when the log-linear model (1.1) is misspecified. We consider the case when the data dimension is $p = 20$ and sample size $n = 100$ and $T_i = T = 100$ for illustration. Two scenarios are considered. In the first scenario, the true model has two correlated covariates generated from a bivariate normal distribution with mean zero, standard deviation 0.5, and correlation 0.2:

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}\,. \tag{B.1}$$

In D2, $|\beta_1| = |\beta_2|$ and in D4, $|\beta_1| = 2|\beta_2|$. Under the misspecified case, the second covariate, $x_{i2}$, is ignored. Table B.1 presents the results using the proposed approach.




(a) Bias of $\hat{\beta}_1$ (b) MSE of $\hat{\beta}_1$ (c) Coverage probability of $\hat{\beta}_1$

(d) MSE of $\hat{\lambda}_{ij}$ (e) Similarity metric $|\langle\hat{\boldsymbol{\gamma}}, \boldsymbol{\pi}_4\rangle|$

**Fig B.1.**

Estimation performance of PS-CAP in estimating the fourth dimension (D4) when $\boldsymbol{\gamma}$ is unknown. For $\hat{\beta}_1$, (a) bias, (b) mean squared error (MSE) and (c) coverage probability (CP) are presented, where CP is obtained from 500 bootstrap samples. For the eigenvalues $\hat{\boldsymbol{\lambda}}_{ij}$, (d) MSE is presented. For $\hat{\boldsymbol{\lambda}}$, (e) similarity to $\boldsymbol{\pi}_4$ is presented. Data dimension p = 100. Sample sizes vary from n = 50, 100, 500, 100 and $T_i$ = T = 50, 100, 500, 1000.

The second scenario considers the following log-linear model for the eigenvalues is considered, which includes an interaction between the covariates:

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 (x_{i1} \times x_{i2}), \tag{B.2}$$

where $x_{i1}$ is generated from a Bernoulli distribution with probability 0.5 to be one and $x_{i2}$ is generated from a normal distribution with mean zero and standard deviation 0.5. Table B.2 presents the estimation results using the proposed method. Under the misspecified case, the interaction between the two covariates is ignored. Thus, in the table, it has no estimate of $\boldsymbol{\beta}_3$.

From both tables, under either correctly specified or misspecified model, the proposed approach correctly identifies the components related to the covariates. Under the misspecified model, the estimate of $\boldsymbol{\beta}$ is biased.

### B.2.2. Model misspecification in $\gamma$

In this section, we discuss the robustness of the proposed approach to the violation of the assumption that all the covariance matrices share the same eigenspace. One advantage of the proposed shrinkage estimator of the covariance matrix is that it will not change the eigenvectors compared to the sample covariance matrix. In Section E.6 of the supplementary materials of Zhao et al. [41], the performance under the partial common diagonalization assumption was examined through a simulation study. In the setting, two eigencomponents are set to be the same across subjects and the rest are unique to each subject. The method can correctly identify the common component across subjects that is related to the covariates. As the proposed approach in this study has the property of preserving the eigenstructure, under the setting of partial common diagonalization, it will also correctly identify the common component that is related to the covariates.

Here, we also consider a case that each covariance matrix has a unique eigenspace, that is, the covariance matrices are generated using the eigendecomposition $\Sigma_i = \Pi_i \Lambda_i \Pi_i^\top$, where $\Pi_i$ is an orthonormal matrix in $\mathbb{R}^{p \times p}$, for $i = 1, \ldots, n$. The rest simulation settings are the same as in Section 4.2. Table B.3 presents the results when $p = 20$, $n = 100$, and $T_i = T = 100$. For the estimation of $\boldsymbol{\gamma}$, we compare with an average of the eigenvectors (after scaling to unit $\ell_2$-norm). For D2, through the correlation between the estimated $\boldsymbol{\gamma}$ and the average if 0.915, the estimation of $\beta_1$ is off. For D4, both the estimate of $\boldsymbol{\gamma}$ and $\beta_1$ are away from the truth. Therefore, an assumption of partial common diagonalization is essential for the proposed framework.

## Appendix C: Additional Analysis of the ADNI Study

## C.1. Validity of model assumptions

In resting-state fMRI studies, the output data are generally considered normally distributed. For each time course, data are temporally correlated of at most lag two [26]. Thus, we subsample the data to remove the temporal correlation. Figure C.1 presents the normal Q-Q plot and the histogram of the data extracted from one brain region of one subject. From the figure, the marginal distribution is close to normal. Thus, the normality assumption is satisfied.

In Section 3, five assumptions are imposed to achieve estimation consistency of the parameters. By setting $C_1 = 2$, Assumption A1 is satisfied. In the fMRI dataset, the increase of the total number of observations across subjects (i.e. $N = \sum_{i=1}^{n} T_i$) can be faster than the number of variables ($p$). Thus, Assumption A2 can be satisfied. Under the normality assumption, the eighth order moment exits, and Assumptions A3 and A4 are valid. Assumption A5 concerns the population eigenvalues. We cannot easily assess this assumption using sample covariance matrices due to the large bias under the high-dimensional setting [22]. We thus can only provide some empirical examination while noting that the empirical results should be evaluated with caution due to this bias issue. To empirically assess the validity of Assumption A5, we first calculate the average sample covariance matrix and then compare the eigenvectors of the average covariance matrix with

the eigenvectors of each individual's sample covariance matrix. When the correlation of two eigenvectors is greater than 0.5, we say there is a high similarity, allowing variability and bias in sample eigenvectors. About 67% of the eigenvectors have a high similarity across multiple subjects. Since the individual sample covariance matrix is rank-deficient, the eigenvectors are not unique. With about 67% overlapping, the assumption of common eigenstructure is partially satisfied. In addition, as discussed in Section B.2.2, when the eigenstructure is partially common across subjects, it will not impact the identification of the common components that are related to the covariates. The proposed approach identifies three components based on the metric of average deviation from diagonality suggesting that these three components commonly diagonalize the covariance matrices. Here the assumption that the log-linear model is correctly specified is challenging to validate using data alone. The current model is considered based on the domain knowledge and the study interest of AD research.

**Table B.1**

Bias and mean squared error (MSE) in estimating $\boldsymbol{\beta}$, and the similarity of $\widehat{\boldsymbol{\lambda}}$ to $\boldsymbol{\pi}_j$ and the standard error (SE), for j = 2, 4, under the misspecified and correctly specified models for model (B.1). Data dimension p = 20, sample size n = 100 and $T_i = T = 100$.

| | | $\widehat{\beta}_1$ | | $\widehat{\beta}_2$ | | $\widehat{\boldsymbol{\lambda}}$ |
|---|---|---|---|---|---|---|
| | | **Bias** | **MSE** | **Bias** | **MSE** | $\lvert\langle\widehat{\boldsymbol{\gamma}}, \boldsymbol{\pi}_j\rangle\rvert$ **(SE)** |
| D2 | Misspecified | 0.105 | 0.014 | - | - | 0.994 (0.003) |
| | Correctly specified | 0.002 | 0.001 | < 0.001 | 0.001 | 0.993 (0.003) |
| D4 | Misspecified | −0.081 | 0.008 | - | - | 0.991 (0.004) |
| | Correctly specified | −0.015 | 0.001 | 0.008 | 0.001 | 0.983 (0.009) |

**Table B.2**

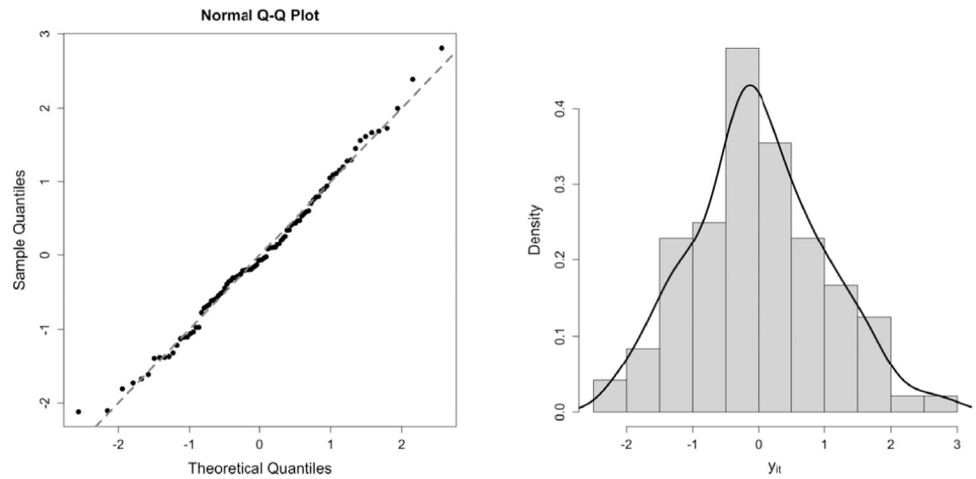
Bias and mean squared error (MSE) in estimating $\boldsymbol{\beta}$, and the similarity of $\widehat{\boldsymbol{\lambda}}$ to $\boldsymbol{\pi}_j$ and the standard error (SE), for j = 2, 4, under the misspecified and correctly specified models for model (B.2). Data dimension p = 20, sample size n = 100 and $T_i = T = 100$.

| | | $\widehat{\beta}_1$ | | $\widehat{\beta}_2$ | | $\widehat{\beta}_3$ | | $\widehat{\boldsymbol{\lambda}}$ |
|---|---|---|---|---|---|---|---|---|
| | | **Bias** | **MSE** | **Bias** | **MSE** | **Bias** | **MSE** | $\lvert\langle\widehat{\boldsymbol{\gamma}}, \boldsymbol{\pi}_j\rangle\rvert$ **(SE)** |
| D2 | Misspecified | < 0.001 | 0.002 | −0.252 | 0.066 | - | - | 0.993 (0.003) |
| | Correctly specified | 0.002 | 0.001 | −0.002 | 0.002 | −0.001 | 0.003 | 0.993 (0.003) |
| D4 | Misspecified | −0.010 | 0.001 | −0.113 | 0.014 | - | - | 0.987 (0.006) |
| | Correctly specified | −0.010 | 0.001 | 0.015 | 0.002 | −0.005 | 0.003 | 0.988 (0.006) |

**Table B.3**

Bias and mean squared error (MSE) in estimating $\beta_1$ and the similarity of $\widehat{\boldsymbol{\lambda}}$ to the average of $\boldsymbol{\pi}_{ij}$ (denoted as $\bar{\pi}_j$) and the standard error (SE), for j = 2, 4, when each covariance matrix has a unique eigenspace. Data dimension p = 20, sample size n = 100 and $T_i = T = 100$.

| | $\widehat{\beta}_1$ | | | $\widehat{\boldsymbol{\lambda}}$ |
|---|---|---|---|---|
| | **Truth** | **Bias** | **MSE** | $\lvert\langle\widehat{\boldsymbol{\gamma}}, \bar{\boldsymbol{\pi}}_j\rangle\rvert$ **(SE)** |
| D2 | −1 | 0.980 | 0.971 | 0.915 (0.031) |
| D4 | 1 | −0.523 | 0.293 | 0.571 (0.060) |




(a) Normal Q-Q plot (b) Marginal histogram

**Fig C.1.**
Normal Q-Q plot and histogram of the data extracted from one brain region of one subject.

## References

[1]. Anderson T (1973). Asymptotically efficient estimation of covariance matrices with linear structure. The Annals of Statistics 1 135–141.

[2]. Anderson TW (1963). Asymptotic theory for principal component analysis. The Annals of Mathematical Statistics 34 122–148.

[3]. Badhwar A, Tam A, Dansereau C, Orban P, Hoffstaedter F and Bellec P (2017). Resting-state network dysfunction in Alzheimer's disease: a systematic review and meta-analysis. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 8 73–85.

[4]. Bai Z and Yin Y (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. The annals of Probability 1275–1294.

[5]. Boik RJ (2002). Spectral models for covariance matrices. Biometrika 89 159–182.

[6]. Cai TT, Ren Z and Zhou HH (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. Electronic Journal of Statistics 10 1–59.

[7]. Chen AA, Beer JC, Tustison NJ, Cook PA, Shinohara RT and Shou H (2020). Removal of scanner effects in covariance improves multivariate pattern analysis in neuroimaging data. bioRxiv 858415.

[8]. Chen Y, Wiesel A and Hero AO (2011). Robust shrinkage estimation of high-dimensional covariance matrices. IEEE Transactions on Signal Processing 59 4097–4107.

[9]. Chiu TY, Leonard T and Tsui K-W (1996). The matrix-logarithmic covariance model. Journal of the American Statistical Association 91 198–210.

[10]. Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small G, Roses AD, Haines J and Pericak-Vance MA (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 261 921–923. [PubMed: 8346443]

[11]. Daniels MJ and Kass RE (2001). Shrinkage estimators for covariance matrices. Biometrics 57 1173–1184. [PubMed: 11764258]

[12]. De Marco M and Venneri A (2017). ApoE-dependent differences in functional connectivity support memory performance in early-stage Alzheimer's disease (P4. 094). Neurology 88.

[13]. Flury BN (1984). Common principal components in k groups. Journal of the American Statistical Association 79 892–898.

[14]. Fox EB and Dunson DB (2015). Bayesian nonparametric covariance regression. Journal of Machine Learning Research 16 2501–2542.

[15]. Franks AM and Hoff P (2019). Shared Subspace Models for Multi-Group Covariance Estimation. Journal of Machine Learning Research 20 1–37.

[16]. Gong G and Samaniego FJ (1981). Pseudo maximum likelihood estimation: theory and applications. The Annals of Statistics 861–869.

[17]. Gour N, Felician O, Didic M, Koric L, Gueriot C, Chanoine V, Confort-Gouny S, Guye M, Ceccaldi M and Ranjeva JP (2014). Functional connectivity changes differ in early and late-onset Alzheimer's disease. Human Brain Mapping 35 2978–2994. [PubMed: 24123475]

[18]. Gour N, Ranjeva J-P, Ceccaldi M, Confort-Gouny S, Barbeau E, Soulier E, Guye M, Didic M and Felician O (2011). Basal functional connectivity within the anterior temporal network is associated with performance on declarative memory tasks. Neuroimage 58 687–697. [PubMed: 21722740]

[19]. Grosenick L, Klingenberg B, Katovich K, Knutson B and Taylor JE (2013). Interpretable whole-brain prediction analysis with GraphNet. NeuroImage 72 304–321. [PubMed: 23298747]

[20]. Hoff PD (2009). A hierarchical eigenmodel for pooled covariance estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71 971–992.

[21]. Hoff PD and Niu X (2012). A covariance regression model. Statistica Sinica 22 729–753.

[22]. Johnstone IM and Lu AY (2009). On consistency and sparsity for principal components analysis in high dimensions. Journal of the American Statistical Association 104.

[23]. Koch W, Teipel S, Mueller S, Benninghoff J, Wagner M, Bokde AL, Hampel H, Coates U, Reiser M and Meindl T (2012). Diagnostic power of default mode network resting state fMRI in the detection of Alzheimer's disease. Neurobiology of Aging 33 466–478. [PubMed: 20541837]

[24]. Ledoit O and Wolf M (2004). A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis 88 365–411.

[25]. Ledoit O and Wolf M (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. The Annals of Statistics 40 1024–1060.

[26]. Lindquist MA (2008). The statistical analysis of fMRI data. Statistical Science 23 439–464.

[27]. Mejia AF, Nebel MB, Barber AD, Choe AS, Pekar JJ, Caffo BS and Lindquist MA (2018). Improved estimation of subject-level functional connectivity using full and partial correlation with empirical Bayes shrinkage. NeuroImage 172 478–491. [PubMed: 29391241]

[28]. Pascal F, Chitour Y and Quek Y (2014). Generalized robust shrinkage estimator and its application to STAP detection problem. IEEE Transactions on Signal Processing 62 5640–5651.

[29]. Pervaiz U, Vidaurre D, Woolrich MW and Smith SM (2020). Optimising network modelling methods for fMRI. Neuroimage 211 116604. [PubMed: 32062083]

[30]. Pourahmadi M, Daniels MJ and Park T (2007). Simultaneous modelling of the Cholesky decomposition of several covariance matrices. Journal of Multivariate Analysis 98 568–587.

[31]. Rahim M, Thirion B and Varoquaux G (2017). Population-shrinkage of covariance to estimate better brain functional connectivity. In International Conference on Medical Image Computing and Computer-Assisted Intervention 460–468. Springer.

[32]. Rahim M, Thirion B and Varoquaux G (2019). Population shrinkage of covariance (PoSCE) for better individual brain functional-connectivity estimation. Medical Image Analysis 54 138–148. [PubMed: 30903965]

[33]. Safieh M, Korczyn AD and Michaelson DM (2019). ApoE4: an emerging therapeutic target for Alzheimer's disease. BMC Medicine 17 1–17. [PubMed: 30651111]

[34]. Seiler C and Holmes S (2017). Multivariate heteroscedasticity models for functional brain connectivity. Frontiers in Neuroscience 11 696. [PubMed: 29311777]

[35]. Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TE, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage 23 S208–S219. [PubMed: 15501092]

[36]. Tibshirani R, Saunders M, Rosset S, Zhu J and Knight K (2005). Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 91–108.

[37]. Tyler DE (1987). A distribution-free M-estimator of multivariate scatter. The Annals of Statistics 234–251.

[38]. Varoquaux G, Gramfort A, Poline J-B and Thirion B (2010). Brain covariance selection: better individual functional connectivity models using population prior. In Advances in neural information processing systems 2334–2342.

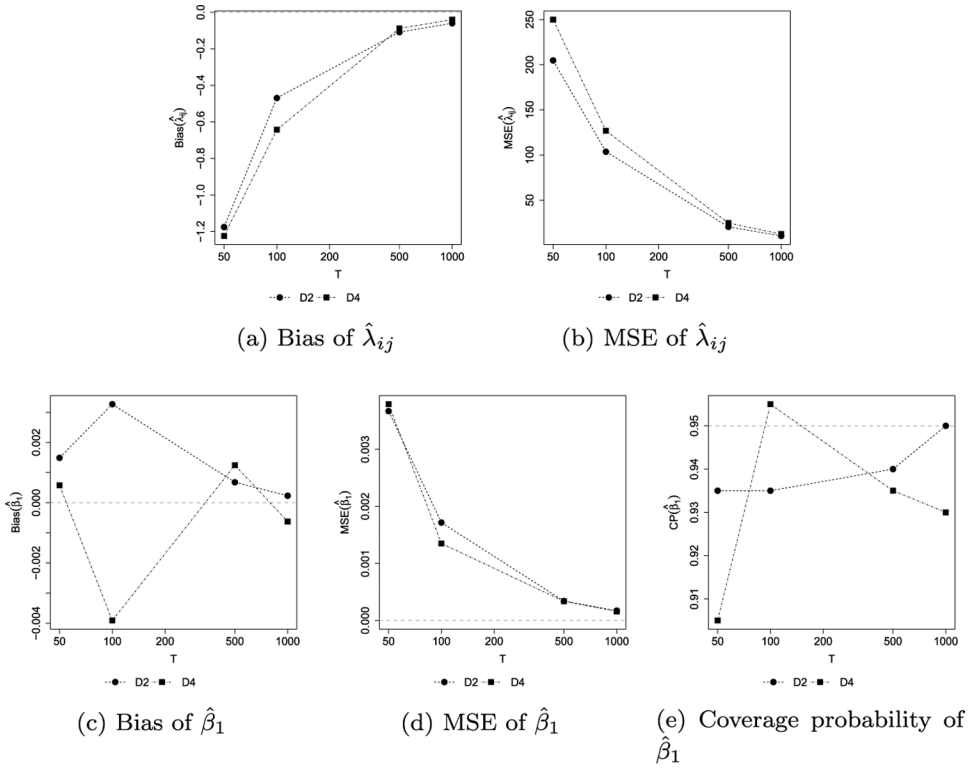
[39]. Yin Y-Q, Bai Z-D and Krishnaiah PR (1988). On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. Probability theory and related fields 78 509–521.

[40]. Zhao Y, Lindquist MA and Caffo BS (2020). Sparse principal component based high-dimensional mediation analysis. Computational Statistics & Data Analysis 142 106835. [PubMed: 32863492]

[41]. Zhao Y, Wang B, Mostofsky SH, Caffo BS and Luo X (2021). Covariate assisted principal regression for covariance matrix outcomes. Biostatistics 22 629–645. [PubMed: 31851318]

[42]. Zou H, Hastie T and Tibshirani R (2006). Sparse principal component analysis. Journal of computational and graphical statistics 15 265–286.

[43]. Zou T, Lan W, Wang H and Tsai C-L (2017). Covariance Regression Analysis. Journal of the American Statistical Association 112 266–281.

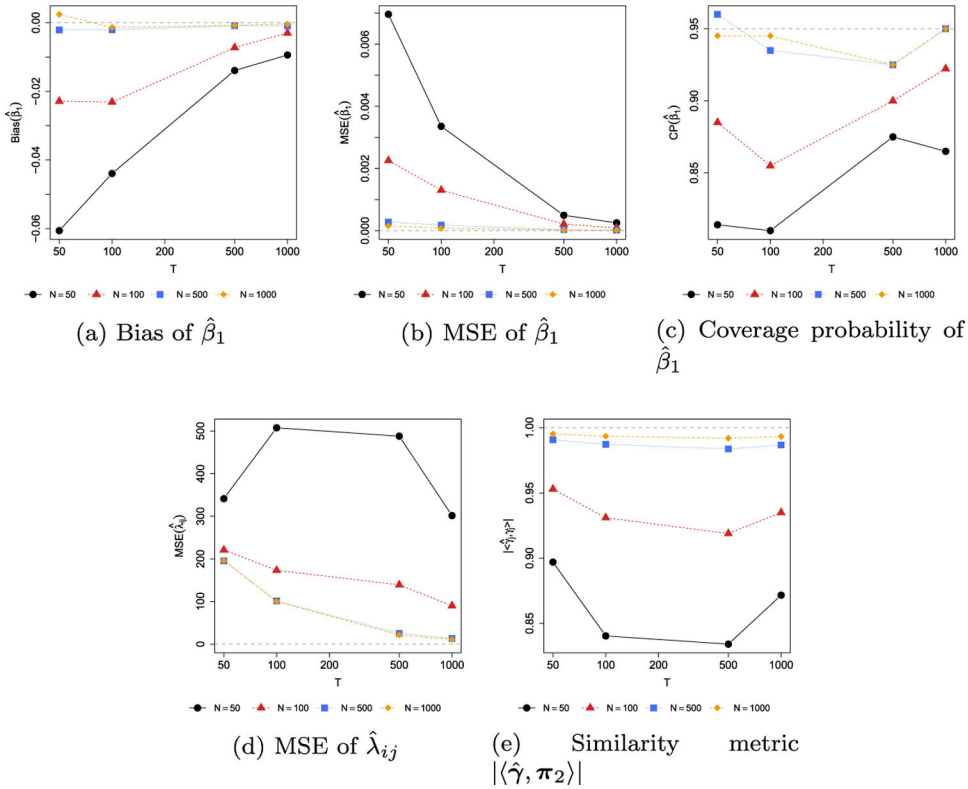



(a) Bias of $\hat{\lambda}_{ij}$ (b) MSE of $\hat{\lambda}_{ij}$

(c) Bias of $\hat{\beta}_1$ (d) MSE of $\hat{\beta}_1$ (e) Coverage probability of $\hat{\beta}_1$

**Fig 1.**

Bias and mean squared error (MSE) in estimating the eigenvalues of the covariance matrices and bias, MSE, and coverage probability in estimating $\beta_1$ coefficient using CS-CAP with the number of subjects n = 50 at various numbers of observations from each subject with p = 20 when $\boldsymbol{\gamma}$ is known.

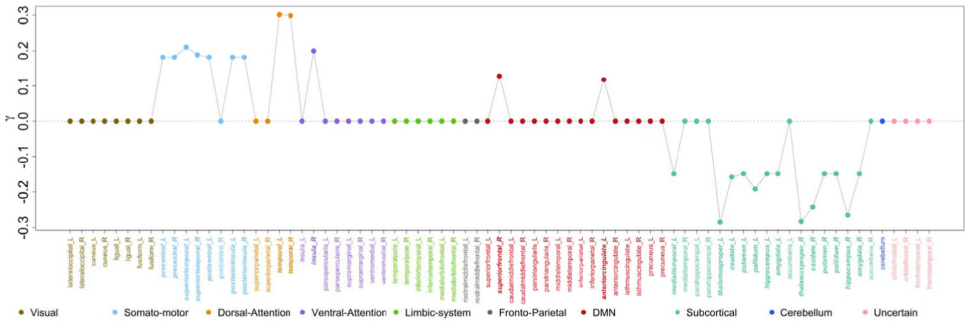



(a) Bias of $\hat{\beta}_1$ (b) MSE of $\hat{\beta}_1$ (c) Coverage probability of $\hat{\beta}_1$

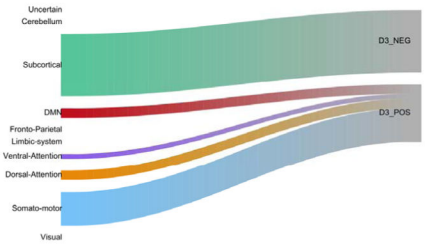
(d) MSE of $\hat{\lambda}_{ij}$ (e) Similarity metric $|\langle\hat{\boldsymbol{\gamma}}, \boldsymbol{\pi}_2\rangle|$

**Fig 2.**

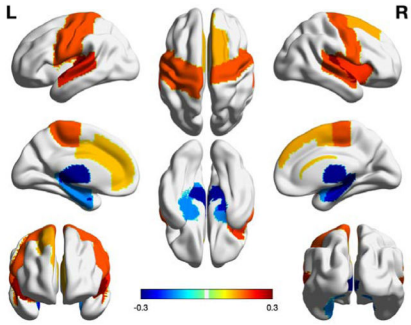
Estimation performance of CS-CAP in estimating the second dimension (D2) when $\boldsymbol{\gamma}$ is unknown. For $\hat{\beta}_1$, (a) bias, (b) mean squared error (MSE) and (c) coverage probability (CP) are presented, where CP is obtained from 500 bootstrap samples. For the eigenvalues $\hat{\lambda}_{ij}$, (d) MSE is presented. For $\hat{\boldsymbol{\gamma}}$, (e) similarity to $\boldsymbol{\pi}_2$ is presented. Data dimension p = 100. Sample sizes vary from n = 50, 100, 500, 100 and $T_i$ = T = 50, 100, 500, 1000.




(a) Sparse loading profile of C3.




(b) River plot of C3 loading.




(c) Brain map of C3.

**Fig 3.**

(a)The sparsified loading profile, (b) the module river plot, and (c) regions with nonzero loadings in a brain map of C3. In (a) and (b), the figure and the legend are colored by brain functional modules. In (c), the brain maps are colored by the loading weights.

**Table 1**

Bias and mean squared error (MSE) in estimating the eigenvalues of the covariance matrices and bias, MSE, and coverage probability (CP) in estimating $\beta_1$ coefficient with sample sizes n = 50 and $T_i$ = T = 50, for i = 1, …, n, when $\boldsymbol{\gamma}$ is known.

| | | Method | $\widehat{\lambda}_{ij}$ | | $\widehat{\beta}_1$ | | |
|---|---|---|---|---|---|---|---|
| | | | **Bias** | **MSE** | **Bias** | **MSE** | **CP** |
| p = 20 | D2 | LW-CAP | −6.520 | 225.360 | 0.053 | 0.006 | 0.795 |
| | | CS-CAP | −1.175 | 204.686 | 0.001 | 0.004 | 0.935 |
| | | CAP | −1.175 | 206.117 | −0.003 | 0.004 | 0.935 |
| | D4 | LW-CAP | −7.422 | 277.888 | −0.040 | 0.005 | 0.860 |
| | | CS-CAP | −1.223 | 249.881 | 0.001 | 0.004 | 0.905 |
| | | CAP | −1.223 | 251.595 | 0.005 | 0.004 | 0.910 |
| $p$ = 50 | D2 | LW-CAP | −7.975 | 244.326 | 0.028 | 0.004 | 0.915 |
| | | CS-CAP | −1.428 | 202.141 | 0.008 | 0.003 | 0.935 |
| | | CAP | - | - | - | - | - |
| | D4 | LW-CAP | −8.641 | 295.221 | −0.012 | 0.004 | 0.915 |
| | | CS-CAP | −1.242 | 248.254 | 0.001 | 0.004 | 0.925 |
| | | CAP | - | - | - | - | - |
| $p$ = 100 | D2 | LW-CAP | −8.924 | 260.268 | 0.010 | 0.004 | 0.915 |
| | | CS-CAP | −0.973 | 203.151 | −0.001 | 0.003 | 0.930 |
| | | CAP | - | - | - | - | - |
| | D4 | LW-CAP | −10.487 | 331.864 | −0.011 | 0.003 | 0.940 |
| | | CS-CAP | −1.705 | 245.754 | −0.007 | 0.003 | 0.940 |
| | | CAP | - | - | - | - | - |

**Table 2**

Bias, mean squared error (MSE), and coverage probability (CP) from 500 bootstrap samples in estimating the $\beta_1$ coefficient, the similarity of $\hat{\boldsymbol{\gamma}}$ to $\boldsymbol{\pi}_j$ and the standard error (SE), and the MSE in estimating the eigenvalues $\hat{\lambda}_{ij}$, for j = 2, 4. Data dimension p = 100, sample size n = 100 and $T_i = T = 100$.

| | Method | $\hat{\beta}_1$ | | | $\hat{\boldsymbol{\lambda}}$ | $\hat{\boldsymbol{\lambda}}_{ij}$ |
|---|---|---|---|---|---|---|
| | | **Bias** | **MSE** | **CP** | $\lvert\langle\hat{\boldsymbol{\gamma}}, \boldsymbol{\pi}_j\rangle\rvert$ **(SE)** | **MSE** |
| D2 | LW-CAP | −0.027 | 0.002 | 0.782 | 0.653 (0.033) | 1812.091 |
| | CS-CAP | −0.023 | 0.001 | 0.855 | 0.931 (0.012) | 173.225 |
| D4 | LW-CAP | 0.018 | 0.002 | 0.770 | 0.666 (0.027) | 2186.265 |
| | CS-CAP | 0.019 | 0.001 | 0.845 | 0.926 (0.011) | 231.856 |

**Table 3**

Model coefficient estimate and 95% bootstrap confidence interval using the PS-CAP approach. The intervals are obtained over 500 bootstrap samples.

| | APOE-$\varepsilon$4 | Sex | Age |
|---|---|---|---|
| C1 | 0.012 (−0.031, 0.263) | −0.431 (−0.636, −0.230) | −0.227 (−0.319, −0.129) |
| C2 | 0.049 (−0.191, 0.309) | −0.544 (−0.867, −0.186) | −0.232 (−0.383, −0.066) |
| 03 | −0.156 (−0.270, −0.045) | −0.061 (−0.201, 0.075) | −0.241 (−0.328,−0.172) |