Covariate Assisted Principal Regression for Covariance Matrix Outcomes

Yi Zhao
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
and
Bingkai Wang
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
and
Stewart H. Mostofsky
Center for Neurodevelopmental and Imaging Research (CNIR)
at Kennedy Krieger Institute
Johns Hopkins University School of Medicine
and
Brian S. Ca
o
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
and
Xi Luo
Department of Biostatistics
School of Public Health, Brown University

September 13, 2018

This is the author's manuscript of the article published in final edited form as:

Zhao, Y., Wang, B., Mostofsky, S. H., Caffo, B. S., & Luo, X. (2019). Covariate Assisted Principal regression for covariance matrix outcomes. Biostatistics. https://doi.org/10.1093/biostatistics/kxz057

Abstract

Modeling variances in data has been an important topic in many fields, including in financial and neuroimaging analysis. We consider the problem of regressing covariance matrices on a vector covariates, collected from each observational unit. The main aim is to uncover the variation in the covariance matrices across units that are explained by the covariates. This paper introduces *Covariate Assisted Principal* (CAP) regression, an optimization-based method for identifying the components predicted by (generalized) linear models of the covariates. We develop computationally efficient algorithms to jointly search the projection directions and regression coefficients, and we establish the asymptotic properties. Using extensive simulation studies, our method shows higher accuracy and robustness in coefficient estimation than competing methods. Applied to a resting-state functional magnetic resonance imaging study, our approach identifies the human brain network changes associated with age and sex.

Keywords: Common diagonalization; Heteroscedasticity; Linear projection

1 Introduction

Modeling variances is an important topic in the statistics and financial literature. In linear regression with heterogeneous errors, various (generalized) linear models have been proposed to model the error variances using the covariates directly or indirectly as a function of the mean (see for example Box and Cox (1964); Carroll et al. (1982); Smyth (1989); Cohen et al. (1993)). These models use separate regression models of the covariates to predict a scalar variance parameter of the error, as well as the mean of the response. Usually, the goal is to improve the efficiency of estimating the mean regression model, while the variance regression model is of less interest.

Regression models for covariance matrices were studied before under different settings. For time series, the "autoregressive conditionally heteroscedastic" (ARCH) models (Engle and Kroner, 1995) were developed to model temporal heteroscedasticity. Anderson (1973) proposed an asymptotically efficient estimator for a class of covariance matrices, where the covariance matrix is modeled as a linear combination of symmetric matrices. Chiu et al. (1996) proposed to model the elements of the logarithm of the covariance matrix as a linear function of the covariates. Pourahmadi (1999) considered another type of matrix decomposition, where the covariates predict linearly the unconstrained elements in the Cholesky decomposition. However, this approach is not order invariant, and requires the matrix columns/rows follow a meaningful ordering. These matrix regression models usually require a large number of parameters to be estimated.

Several approaches were proposed to extend matrix outcome regression models to high dimensions. Hoff and Niu (2012) introduced a regression model, where the covariance matrix is a parsimonious quadratic function of the explanatory variables. Applying low-rank approximation techniques, Fox and Dunson (2015) generalized the framework to a nonpara-

metric covariance regression model and enabled scaling up to high dimensions. In a recent paper, Zou et al. (2017) linked the matrix outcome to a linear combination of similarity matrices of covariates, and studied the asymptotic properties of various estimators under this model. These approaches again model the whole covariance matrix as outcomes, and thus the interpretation could be challenging for large matrices.

Closely related to covariance matrices, principal component analysis (PCA) and related methods are widely used to generate interpretable results for large dimensional data. These methods have been extended to model multiple covariance matrices. Flury (1984) and Flury (1988) introduced a class of models, called common principal components models, to uncover the shared covariance structures. Boik (2002) generalized these models using spectral decompositions. Hoff (2009) developed a Bayesian hierarchical model and estimation procedure to study the heterogeneity in both the eigenvectors and eigenvalues of covariance matrices. Assuming that the eigenvectors span the same subspace, Franks and Hoff (2016) extended this to the so-called high dimensional setting with large p and small n. It is unclear, however, how these methods can be extended to incorporate multiple covariates.

In the application area of neuroimaging analysis, PCA-type methods are becoming increasingly popular for modeling covariance matrices, partly because of their desirable interpretability and computational capability for analyzing large and multilevel observations. Covariance matrices (or correlation matrices after standardization) of multiple brain regions are also commonly known as functional connectivity analysis (Friston, 2011). Decomposing the covariance matrices into separate components enable identification of coherently active brain subnetworks (Poldrack et al., 2011), and usually a few principal components are needed to explain the variation in neuroimaging data (Friston et al., 1993). As before, it is unclear how these methods can be extended to include multiple covariates.

Indeed, modeling the covariate-related alterations in covariance matrices is an important topic in neuroimaging analysis, because changes in functional connectivity have been found to be associated with various demographic and clinical factors, such as age, gender, and cognitive behavioral functions including developmental and mental health capacities (Just et al., 2006; Wang et al., 2007; Luo et al., 2011; Mennes et al., 2012; Hafkemeijer et al., 2015; Park et al., 2016). A commonly implemented method to analyze the covariance changes is to regress one matrix entry on the covariates, and this model is repeatedly fitted for each matrix element (see, for example, Wang et al. (2007) and Lewis et al. (2009)). Though this approach has good interpretability and is scalable, it suffers from the multiplicity issues, because of the large number of regressions involved. For example, $p(p-1)/2$ regressions for p brain regions. Adapting the covariance regression model proposed in Hoff and Niu (2012), Seiler and Holmes (2017) introduced a simplified model to analyze a large and multilevel neuroimaging dataset.

In this paper, we propose a *Covariate Assisted Principal (CAP) regression* model for multiple covariance matrix outcomes. This model integrates the PCA principle with a generalized linear model of multiple covariates. Analogous to PCA, our model aims to identify linear projections to allow for interpretability of the covariance matrices, while being computationally feasible for large data. Unlike PCA, our method targets the projections that are associated with the covariates. This enables us to study the changes in covariance matrices associated with subject-specific factors, such as individual demographic or disease information.

This paper is organized as follows. In Section 2, we introduce our proposed CAP regression model. Section 3 presents the estimation and computation algorithms in identifying the proposed principal projection directions. We compare the performance of our proposed methods with competing approaches through simulation studies in Section 4. We then

apply our methods to a real fMRI dataset in Section 5. Section 6 summarizes the paper with a summary and discussion of future directions.

2 Model

For each $i \in \{1, \ldots, n\}$, let $\mathbf{y}_{it} \in \mathbb{R}^p$, $t = 1, \ldots, T_i$, be independent and identically distributed random samples from a multivariate normal distribution with mean zero and covariance matrix, Σ_i, where Σ_i may depend on explanatory variables, $\mathbf{x}_i \in \mathbb{R}^{q-1}$. In our application example, $\mathbf{y}_{it}$ is a sample of brain fMRI measurements of p regions, and $\mathbf{x}_i$ is a vector of covariates postulated to be related to fMRI measurements, both collected from subject i. We assume that there exists a vector $\boldsymbol{\gamma} \in \mathbb{R}^p$ such that $z_{it} \triangleq \boldsymbol{\gamma}^\top \mathbf{y}_{it}$ satisfies the following multiplicative heteroscedasticity model:

$$\log\{\mathrm{Var}(z_{it})\} = \log(\boldsymbol{\gamma}^\top \Sigma_i \boldsymbol{\gamma}) = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}_1, \tag{1}$$

where β_0 and $\boldsymbol{\beta}_1$ are model coefficients. The logarithmic linear model follows from Harvey (1976) in which Σ_i is a scalar.

A toy example of this model ($p = 2$) is shown in Figure 1. The covariance matrices, represented by the contour plot ellipses, vary as the covariate x varies. On the first projection direction (PD1) with the largest variability, there is no variation under different x values. However, the variance in the second direction (PD2) decreases as x increases. Our proposed model (1) thus aims to identify the second projection direction. In other words, the objective is to discover the rotation such that the data variation in the new space can be best characterized by the explanatory variables.

Compared with existing methods, our proposed model has two main advantages. First, different from the model proposed by Hoff and Niu (2012) and Zou et al. (2017), which

directly model Σ_i by linear combinations of symmetric matrices constructed out of $\mathbf{x}_i$, we assume a log-linear model for the variance component after rotation. The linear form allows easy interpretation of the regression coefficient and provides the modeling flexibility shared by all other (generalized) linear models, such as interactions. The projection enables computational scalability similar to PCA. The common principal component approach, studied in Flury (1984), only allows the eigenvalues to vary across a group indicator, our model (1) provides a direct model of multiple covariates, including continuous ones. This enables studying the covariate-related changes in covariances in our fMRI experiment. Second, our model relaxes the standard complete common principal component assumption imposed in Flury (1984) and Boik (2002), and we assume that there exists at least one projection direction such that model (1) is satisfied. This partial common diagonalization assumption is more realistic for data with higher dimensions.

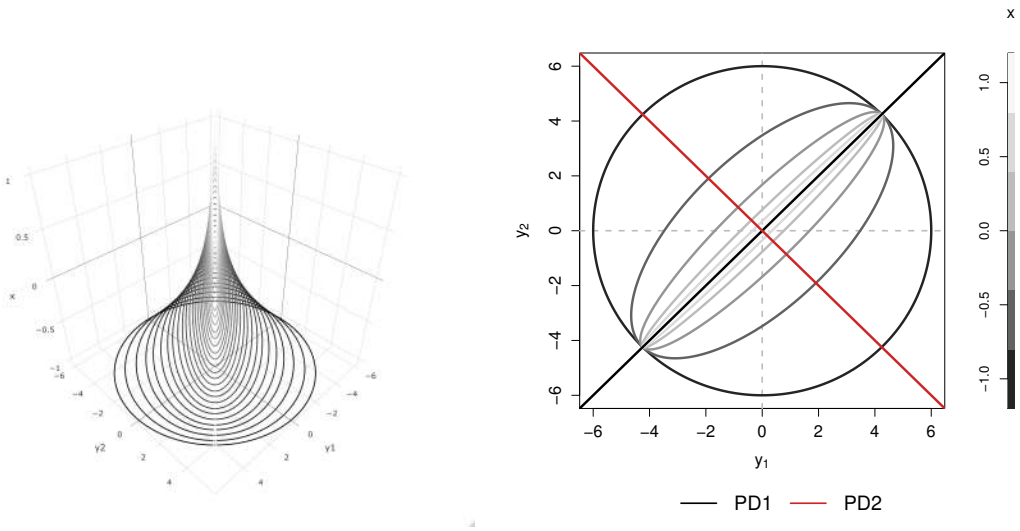
3 Method

We propose to estimate the model parameters by maximizing the likelihood function under a quadratic constraint:

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

where $\ell(\boldsymbol{\beta},\boldsymbol{\gamma})$ is the negative log-likelihood function (ignoring constants), $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^\top)^\top \in \mathbb{R}^q$, $\mathbf{S}_i = \sum_{t=1}^{T_i}\mathbf{y}_{it}\mathbf{y}_{it}^\top$, and $\mathbf{H}$ is a positive definite matrix in $\mathbb{R}^{p\times p}$. Without the constraint, $\ell(\boldsymbol{\beta},\boldsymbol{\gamma})$, for any fixed $\boldsymbol{\beta}$, is minimized by $\boldsymbol{\gamma} = \mathbf{0}$. Thus the constraint is critical.

Two natural choices of $\mathbf{H}$ in the constraint are:



(a) 3D contour plot.

(b) 2D contour plot.

Figure 1: Covariance matrices, shown as contour plot ellipses when $p = 2$, vary as a continuous X varies (z-axis in (a) and gray/color scales in (b)).

(C1) $\mathbf{H} = \mathbf{I}$ which is equivalent to a unit constraint under ℓ_2-norm, i.e.,

$$\boldsymbol{\gamma}^\top \boldsymbol{\gamma} = 1; \tag{3}$$

(C2) $\mathbf{H} = \bar{\Sigma}$ which is equivalent to a unit constraint with respect to the average sample covariance, i.e.,

$$\boldsymbol{\gamma}^\top \bar{\Sigma} \boldsymbol{\gamma} = 1, \quad \bar{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{T_i} \mathbf{S}_i. \tag{4}$$

(C1) is inspired by standard PCA. The second one is by common principal component analysis. We show in the following proposition that (C1) will lead to a solution that is less appealing in certain situations.

Proposition 1. *When* $\mathbf{H} = \mathbf{I}$ *in the optimization problem* (2)*, for any fixed* $\boldsymbol{\beta}$*, the solution of* $\boldsymbol{\gamma}$ *is the eigenvector corresponding to the minimum eigenvalue of matrix*

$$\sum_{i=1}^{n} \frac{\mathbf{S}_i}{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}.$$

The matrix $\mathbf{S}_i / \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$ can be regarded as a normalization on the covariance matrices based on the explanatory variables. Thus, constraint (C1) achieves the projection direction with the lowest normalized data variation. In the Appendix Section D, we further discuss the property of these two constraints using examples. We will focus on constraint (C2) in this paper because the signals are usually not associated with the smallest eigenvalue in most scenarios.

3.1 Algorithm

The optimization problem (2) is biconvex. We propose to solve the optimization problem by block coordinate descent. For given $\boldsymbol{\gamma}$, the update of $\boldsymbol{\beta}$ is obtained by the Newton-Raphson algorithm. For given $\boldsymbol{\beta}$, the solving for $\boldsymbol{\gamma}$ requires quadratic programming. Though generic quadratic programming packages could be used, we derive the explicit solution in Proposition A.1 in the supplementary material. The algorithm is summarized in Algorithm 1. This algorithm works for any positive definite $\mathbf{H}$. To obtain robustness against obtaining a solution in a local minimum, we propose to randomly choose a series of initial values and take the estimate with the lowest objective function value.

3.2 Extension for finding multiple projection directions

It is possible that more than one projection direction is associated with the covariates. We propose to find these directions sequentially. This is modified from the strategy of finding

Algorithm 1 A block coordinate descent algorithm for solving optimization problem (2).

Input:

$\mathbf{Y}$: a list of data where the ith element is a $T_i \times p$ data matrix

$\mathbf{X}$: a $n \times q$ matrix of covariate variables with the first column of ones

$\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$: initial values

Output: $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}$

Given $(\boldsymbol{\beta}^{(s)}, \boldsymbol{\gamma}^{(s)})$ from the sth step, for the $(s+1)$th step:

(i) update $\boldsymbol{\beta}$ by

$$\boldsymbol{\beta}^{(s+1)} = \boldsymbol{\beta}^{(s)} - \left(\sum_{i=1}^{n} \exp(-\mathbf{x}_i^\top \boldsymbol{\beta}^{(s)}) \boldsymbol{\gamma}^{(s)\top} \mathbf{S}_i \boldsymbol{\gamma}^{(s)} \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \left(\sum_{i=1}^{n} (T_i - \exp(-\mathbf{x}_i^\top \boldsymbol{\beta}^{(s)}) \boldsymbol{\gamma}^{(s)\top} \mathbf{S}_i \boldsymbol{\gamma}^{(s)}) \mathbf{x}_i \right), \tag{5}$$

where $\mathbf{S}_i = \sum_{t=1}^{T_i} \mathbf{y}_{it} \mathbf{y}_{it}^\top$;

(ii) update $\boldsymbol{\gamma}$ by solving

$$\begin{aligned} \underset{\boldsymbol{\gamma}}{\text{minimize}} \quad & \frac{1}{2} \sum_{i=1}^{n} \exp(-\mathbf{x}_i^\top \boldsymbol{\beta}^{(s)}) \boldsymbol{\gamma}^\top \mathbf{S}_i \boldsymbol{\gamma}, \\ \text{such that} \quad & \boldsymbol{\gamma}^\top \mathbf{H} \boldsymbol{\gamma} = 1, \end{aligned}$$

where $\mathbf{H} = \mathbf{I}$ under (C1) and $\mathbf{H} = \bar{\Sigma} = (\sum_{i=1}^{n} \mathbf{S}_i / T_i)/n$ under (C2), using Proposition A.1.

Repeat steps (i)-(ii) until convergence.

multiple principal components one by one.

Suppose $\Gamma^{(k-1)} = (\boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(k-1)})$ contains the first $(k-1)$ components (for $k \geq 2$), and let $\hat{\mathbf{Y}}_i^{(k)} = \mathbf{Y}_i - \mathbf{Y}_i \Gamma^{(k-1)} \Gamma^{(k-1)\top}$, where $\mathbf{Y}_i = (\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT_i})^\top$ for $i = 1, \ldots, n$. We

cannot directly apply Algorithm 1 to $\hat{\mathbf{Y}}_i^{(k)}$ as in PCA algorithms, since $\hat{\mathbf{Y}}_i^{(k)}$ is not of full rank. We introduce a rank-completion step. The whole algorithm is summarized in Algorithm 2. In the algorithm, step (iii) completes the data to full rank by adding nonzero positive eigenvalues to those zero eigencomponents, which are the exponential of model intercept of the corresponding directions. This step also guarantees that there are no identical eigenvalues in the covariance matrix of $\tilde{\mathbf{Y}}_i$, which is a necessary condition for unique eigenvector identification.

Analogous to the PCA approach, step (iv) is an orthogonal constraint to ensure that the kth direction is orthogonal to the previous ones, which is equivalent to the following optimization problem, for $k \geq 2$,

$$\begin{aligned} \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\text{minimize}} \quad & \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i^\top \boldsymbol{\beta}^{(k)}) \cdot T_i + \frac{1}{2}\sum_{i=1}^{n} \boldsymbol{\gamma}^{(k)\top}\mathbf{S}_i^{(k)}\boldsymbol{\gamma}^{(k)} \cdot \exp(-\mathbf{x}_i^\top \boldsymbol{\beta}^{(k)}), \\ \text{such that} \quad & \boldsymbol{\gamma}^{(k)\top}\mathbf{H}\boldsymbol{\gamma}^{(k)} = 1, \\ \text{and} \quad & \Gamma^{(k-1)\top}\boldsymbol{\gamma}^{(k)} = \mathbf{0}. \end{aligned} \tag{6}$$

For any fixed $\boldsymbol{\beta}^{(k)}$, we derive an explicit formula for solving $\boldsymbol{\gamma}^{(k)}$, see Section A.2 of the supplementary material. The proof is adapted from Rao (1964, 1973).

3.3 Choosing the number of projection directions

We propose a data-driven approach to choose the number of projection directions. Extending the common principal component model, Flury and Gautschi (1986) introduced a metric to quantify the "deviation from diagonality". Suppose $\mathbf{A}$ is a positive definite symmetric matrix, the "deviation from diagonality" is defined as

$$\nu(\mathbf{A}) = \frac{\det\{\text{diag}(\mathbf{A})\}}{\det(\mathbf{A})}, \tag{7}$$

Algorithm 2 An algorithm for finding the kth projection direction under constraint (C2).

Input:

$\mathbf{Y}$: a list of data where the ith element is a $T_i \times p$ data matrix

$\mathbf{X}$: a $n \times q$ matrix of covariate variables with the first column of ones

$\Gamma^{(k-1)}$: a $p \times (k-1)$ matrix contains the first $(k-1)$ directions

$\mathbf{B}^{(k-1)}$: a $q \times (k-1)$ matrix contains the model coefficients of the first $(k-1)$ directions

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

(i) For $i = 1, \ldots, n$, let $\hat{\mathbf{Y}}_i^{(k)} = \mathbf{Y}_i - \mathbf{Y}_i \Gamma^{(k-1)} \Gamma^{(k-1)\top}$.

(ii) Apply singular value decomposition (SVD) on $\hat{\mathbf{Y}}_i^{(k)}$, such that $\hat{\mathbf{Y}}_i^{(k)} = U_i D_i V_i^\top$.

(iii) Let $\tilde{\mathbf{Y}}_i^{(k)} = U_i \tilde{D}_i V_i^\top$ with

$$\tilde{D}_i = \text{diag}\{D_{i1}, \ldots, D_{i(p-(k-1))}, \exp(\beta_{10}), \ldots, \exp(\beta_{(k-1)0})\},$$

where $\{D_{i1}, \ldots, D_{i(p-(k-1))}\}$ are the first $(p-(k-1))$ diagonal elements of matrix D_i, and $\beta_{10}, \ldots, \beta_{(k-1)0}$ are the intercept of the first $(k-1)$ directions (first row of $\mathbf{B}^{(k-1)}$).

(iv) Treat $\tilde{\mathbf{Y}}_i^{(k)}$ $(i = 1, \ldots, n)$ as the new data, and apply Algorithm 1 under constraint (C2) with an additional orthogonal constraint

$$\Gamma^{(k-1)\top} \boldsymbol{\gamma}^{(k)} = \mathbf{0}.$$

---

where $\text{diag}(\mathbf{A})$ is a diagonal matrix with the diagonal elements the same as matrix $\mathbf{A}$, and $\det(\mathbf{A})$ is the determinant of matrix $\mathbf{A}$. From Hadamard's inequality, we have that $\nu(\mathbf{A}) \geq 1$, where equality is achieved if and only if $\mathbf{A}$ is a diagonal matrix.

To adapt this metric in our model, we let $\Gamma^{(k)} \in \mathbb{R}^{p\times k}$ denote the matrix containing the first k projection directions. We define the average deviation from diagonality as

$$\mathrm{DfD}(\Gamma^{(k)}) = \left(\prod_{i=1}^{n} \nu(\Gamma^{(k)\top} \mathbf{S}_i \Gamma^{(k)}/T_i)^{T_i} \right)^{1/\sum_i T_i}, \tag{8}$$

which is the weighted geometric mean of each subject's deviation from diagonality. As k increases, the requirement for $\Gamma^{(k)\top}\mathbf{S}_i\Gamma^{(k)}$ to be a diagonal matrix, as in Flury and Gautschi (1986), may become more stringent. In practice, we can plot the average deviation from diagonality and choose the first few projection directions with DfD value close to one or choose a suitable number right before a sudden jump in the plot. See an example in Section E of the supplementary material.

3.4 Analysis under a Common Principal Component Model

We need additional assumptions to perform theoretical analysis. Following Flury (1986), we assume that the covariance matrices $\Sigma_1, \ldots, \Sigma_n$ can be diagonalized by the same orthogonal matrix. That is, there exists an orthogonal matrix Γ, such that

$$\Sigma_i = \Gamma \Lambda_i \Gamma^\top, \quad \text{for } i = 1, \ldots, n, \tag{9}$$

where $\Gamma = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p)$ and $\Lambda_i = \mathrm{diag}\{\lambda_{i1}, \ldots, \lambda_{ip}\}$. Suppose the eigenvalues are ordered as $\bar{\lambda}_1 > \cdots > \bar{\lambda}_p$, where $\bar{\lambda}_j = \sum_{i=1}^n \lambda_{ij}/n$. Let $\hat{\Sigma}_i = \mathbf{S}_i/T_i$ denote the sample covariance matrix. Suppose $\Phi = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_p)$ and $\Delta_i = \mathrm{diag}\{\delta_{i1}, \ldots, \delta_{ip}\}$ are the maximum likelihood estimator of Γ and Λ_i $(i = 1, \ldots, n)$, respectively, using the method proposed in Flury (1984). Flury (1986) showed that they are both consistent estimators, and thus $\tilde{\Sigma}_i = \Phi\Delta_i\Phi^\top$ is a consistent estimator of Σ_i. Therefore, we have

$$\| \hat{\Sigma}_i - \tilde{\Sigma}_i \| \to 0, \quad \text{as } \min_i T_i \to \infty \text{ and } n \to \infty. \tag{10}$$

Based on (10), we replace $\mathbf{S}_i$ by the consistent estimator $T_i\tilde{\Sigma}_i$ in our optimization problem (2). Since Φ is the orthonormal eigenbasis, $\boldsymbol{\gamma}$ can be represented by the linear combination of the columns in Φ, i.e., $\boldsymbol{\gamma} = \Phi\mathbf{a} = \sum_{j=1}^{p} a_j\boldsymbol{\phi}_j$, where $\mathbf{a} = (a_1, \ldots, a_p)^\top$. The optimization problem (2) is reformulated as:

$$\begin{aligned} \underset{\boldsymbol{\beta},\mathbf{a}}{\text{minimize}} \quad & \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i^\top\boldsymbol{\beta})\cdot T_i + \frac{1}{2}\mathbf{a}^\top \left(\sum_{i=1}^{n} T_i \exp(-\mathbf{x}_i^\top\boldsymbol{\beta})\Delta_i\right)\mathbf{a}, \\ \text{such that} \quad & \mathbf{a}^\top\mathbf{H}\mathbf{a} = 1, \end{aligned} \tag{11}$$

where $\mathbf{H} = \mathbf{I}$ under (C1) and $\mathbf{H} = \bar{\Delta} = \sum_{i=1}^{n}\Delta_i/n$ under (C2). With given $\boldsymbol{\beta}$, under constraint (C1), it is equivalent to solve

$$\min_{j\in\{1,\ldots,p\}} \sum_{i=1}^{n} T_i \exp(-\mathbf{x}_i^\top\boldsymbol{\beta})\delta_{ij}. \tag{12}$$

Suppose the eigenvectors are ordered based on the average eigenvalues (i.e., $\bar{\delta}_1 > \cdots > \bar{\delta}_p$, $\bar{\delta}_j = \sum_i \delta_{ij}/n$), we have $\hat{\mathbf{a}} = \boldsymbol{\phi}_p$. Therefore, under constraint (C1), the method yields the common eigenvector with the lowest average eigenvalue. Now consider the constraint (C2). Let $\mathbf{b} = (b_1, \ldots, b_p)^\top$ with $b_j = a_j\sqrt{\delta_j}$. Minimizing the objective function in (11) under constraint (C2) is equivalent to solving the following problem,

$$\min_{j\in\{1,\ldots,p\}} \frac{\sum_{i=1}^{n} T_i\delta_{ij}/\exp(\mathbf{x}_i^\top\boldsymbol{\beta})}{\sum_{i=1}^{n}\delta_{ij}}. \tag{13}$$

Suppose $\lambda_{ik} = \exp(\mathbf{x}_i^\top\boldsymbol{\beta})$ satisfies the model assumption, with $T_i = T$, the minimizer of above optimization problem is k if

$$\bar{\delta}_k > \frac{1}{\bar{\pi}_{jk}}\bar{\delta}_j, \quad \text{for } j \neq k, \tag{14}$$

where $\bar{\delta}_j = \sum_{i=1}^{n}\delta_{ij}/n$, $\bar{\pi}_{jk} = \sum_{i=1}^{n}\pi_{ijk}/n$, and $\pi_{ijk} = \delta_{ij}/\delta_{ik}$, $j = 1, \ldots, p$. Since δ_{ij} is a consistent estimator of λ_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, p$, we impose the following condition.

Condition 1 (Eigenvalue condition). *Assume* $\Sigma_i = \Gamma\Lambda_i\Gamma^\top$ *is the eigendecomposition of* Σ_i *with* $\Gamma = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p)$ *an orthogonal matrix and* $\Lambda_i = \text{diag}\{\lambda_{i1}, \ldots, \lambda_{ip}\}$ *a diagonal matrix, for* $i = 1, \ldots, n$. *The eigenvalues are ordered as* $\bar{\lambda}_1 > \cdots > \bar{\lambda}_p$, *where* $\bar{\lambda}_j = \sum_{i=1}^n \lambda_{ij}/n$. *Suppose there exists* $k \in \{1, \ldots, p\}$ *such that* $\lambda_{ik} = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$ *satisfies the model assumption, and assume*

$$\bar{\lambda}_k > \frac{1}{\bar{\tau}_{jk}} \bar{\lambda}_j, \quad \text{for } j \neq k, \tag{15}$$

where $\bar{\tau}_{jk} = \sum_{i=1}^n \tau_{ijk}/n$, *and* $\tau_{ijk} = \lambda_{ij}/\lambda_{ik}$.

Under this condition, we propose a min-max algorithm (Algorithm 3) to identify the common principal component with eigenvalues that fit the log-regression model (1) and meanwhile explain large variations in the data. We call this algorithm a min-max approach as it contains a minimization (of the objective function) and maximization (of data variation) steps. To acquire the first k ($k \geq 2$) directions, we propose to order those p_s components that satisfy $\hat{s}(j) = j$ in step (iv) by the average eigenvalues and return the first $\min\{k, p_s\}$ components. Thus this algorithm also provides an estimate of the number of components.

3.4.1 Asymptotic properties

We first discuss the asymptotic property of $\boldsymbol{\beta}$ estimator given the true $\boldsymbol{\gamma}$. As $\boldsymbol{\beta}$ is estimated by maximizing the log-likelihood function, we have the following theorem.

Theorem 1. *Assume* $\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top/n \to \mathbf{Q}$ *as* $n \to \infty$. *Let* $T = \min_i T_i$, $M_n = \sum_{i=1}^n T_i$, *under the true* $\boldsymbol{\gamma}$, *we have*

$$\sqrt{M_n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, 2\mathbf{Q}^{-1}\right), \quad \text{as } n, T \to \infty, \tag{16}$$

where $\hat{\boldsymbol{\beta}}$ *is the maximum likelihood estimator when the true* $\boldsymbol{\gamma}$ *is known.*

Algorithm 3 A common principal component based method for solving optimization problem (2) under constraint (C2).

Input:

Y: a list of data where the ith element is a $T_i \times p$ data matrix

X: a $n \times q$ matrix of covariate variables with the first column of ones

Output: $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}$

(i) Use Flury (1984) method to estimate Φ and Δ_i $(i = 1, \ldots, n)$ for $\mathbf{Y}$ with n groups.

(ii) For $j = 1, \ldots, p$, estimate $\boldsymbol{\beta}$ with $\boldsymbol{\gamma} = \boldsymbol{\phi}_j$, denoted by $\boldsymbol{\beta}^{(j)}$.

(iii) For each j, minimize the objective function with

$$\hat{s}(j) = \underset{s \in \{1,\ldots,p\}}{\arg\min} \frac{\sum_{i=1}^{n} T_i \delta_{ij} / \exp(\mathbf{x}_i^\top \boldsymbol{\beta}^{(j)})}{\sum_{i=1}^{n} \delta_{ij}}.$$

(iv) For those $\hat{s}(j) = j$, maximize the variance with

$$\hat{k} = \underset{k \in \{k : \hat{s}(k) = k\}}{\arg\max} \sum_{i=1}^{n} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}^{(k)}).$$

(v) Estimate $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(\hat{k})}$ and $\hat{\boldsymbol{\gamma}} = \boldsymbol{\phi}_{\hat{k}}$.

When $p = 1$ and $T_i = 1$ (for $i = 1, \ldots, n$), our proposed model (1) degenerates to a multiplicative heteroscedastic regression model. The asymptotic distribution of $\hat{\boldsymbol{\beta}}$ in Theorem 1 is the same as in Harvey (1976). We now establish the asymptotic theory when $\boldsymbol{\gamma}$ is estimated from the common principal component approach (Flury, 1984).

Theorem 2. *Assume* $\Sigma_i = \Gamma \Lambda_i \Gamma^\top$, *where* $\Gamma = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p)$ *is an orthogonal matrix and* $\Lambda_i = \mathrm{diag}\{\lambda_{i1}, \ldots, \lambda_{ip}\}$ *with* $\lambda_{ik} \neq \lambda_{il}$ $(k \neq l)$, *for at least one* $i \in \{1, \ldots, n\}$. *There*

exists $k \in \{1, \ldots, p\}$ *such that for* $\forall\ i \in {}_k\{1, \ldots, n\}_i^\top\ \boldsymbol{\gamma}^\top \Sigma_i \boldsymbol{\gamma}_k = \exp(\mathbf{x}$ *). Let* $\boldsymbol{\gamma}\hat{}$ *be the maximum likelihood estimator of* $\boldsymbol{\gamma}_k$ *in Flury (1984). Then assuming that the assumptions*

in Theorem 1 are satisfied,

$\hat{}$ *from Algorithm 3 is* $\sqrt{M_n}$*-consistent estimator of*

4 Simulation Study

In the simulation study, we generate data from a multivariate normal distribution with $p = 5$ and covariance Σ_i for sample i. We assume the covariance matrices satisfy the common diagonalization assumption, i.e., $\Sigma_i = \Gamma \Lambda_i \Gamma^\top$, where

$$\Gamma = \begin{pmatrix} 0.447 & 0.447 & 0.447 & 0.447 & 0.447 \\ 0.447 & -0.862 & 0.138 & 0.138 & 0.138 \\ 0.447 & 0.138 & -0.862 & 0.138 & 0.138 \\ 0.447 & 0.138 & 0.138 & -0.862 & 0.138 \\ 0.447 & 0.138 & 0.138 & 0.138 & -0.862 \end{pmatrix} \tag{17}$$

is an orthogonal matrix, and Λ_i is a diagonal matrix with diagonal elements $\{\lambda_{i1}, \ldots, \lambda_{ip}\}$. For the log-linear model, X_i $(i = 1, \ldots, n)$ is generated from a Bernoulli distribution with probability 0.5 to be one. Thus $q = 2$ because of the additional intercept column. Two scenarios are tested: (i) the null case with $\beta_1 = 0$ and (ii) the alternative case with the second and third eigenvalues satisfying the regression model. For the first cases with $\beta_1 = 0$, λ_{ij} is generated from a log-normal distribution with mean β_0 and variance 0.5^2; and for the second case with $\beta_1 \neq 0$, $\lambda_{ij} = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$ with $\mathbf{x}_i = (1\ X_i)^\top$. The simulation is repeated 200 times.

As demonstrated in Section 3, constraint (C1) yields the component with the lowest normalized data variation. Thus, in this section, we only present the performance under constraint (C2). Under constraint (C2), for higher-order directions, enforcing the orthogo-

nality constraint reduces the parameter search space and increases computation complexity. In this simulation study, we implement both cases with and without the orthogonality constraint. We compare the following methods:

(1) our proposed block coordinate descent method (Algorithms 1 and 2) under constraint (C2) without the orthogonality constraint, denoted as CAP;

(2) our proposed block coordinate descent method (Algorithms 1 and 2) under constraint (C2) with the orthogonality constraint, denoted as CAP-OC;

(3) our method under the complete common principal component model (Algorithm 3) for finding the first k projection directions, denoted as CAP-C;

(4) a principal component analysis (PCA) based method, where we apply PCA on each subject and regress each of the first k eigenvalues on the covariates, denoted as PCA;

(5) a common principal component method, where we apply common PCA on all subjects using the method in Flury (1984) and regress each of the first k eigenvalues on the covariates, denoted as CPCA.

We first evaluate the performance under the null case, i.e., $\beta_1 = 0$. β_0's are set to be $\boldsymbol{\beta}_0 = (5, 4, 1, -1, -2)^\top$. We present the estimate of β's from CAP and CAP-C over 200 simulations in Figure E.2 in the supplementary material. Our estimate of β_1 is centered around zero with an average of 0.01 under CAP and -0.01 under CAP-C, both much smaller than the corresponding standard errors.

Under the alternative scenario, we set $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta} = \begin{pmatrix} 5 & 4 & 1 & -1 & -2 \\ 0 & -1 & 1 & 0 & 0 \end{pmatrix},$$

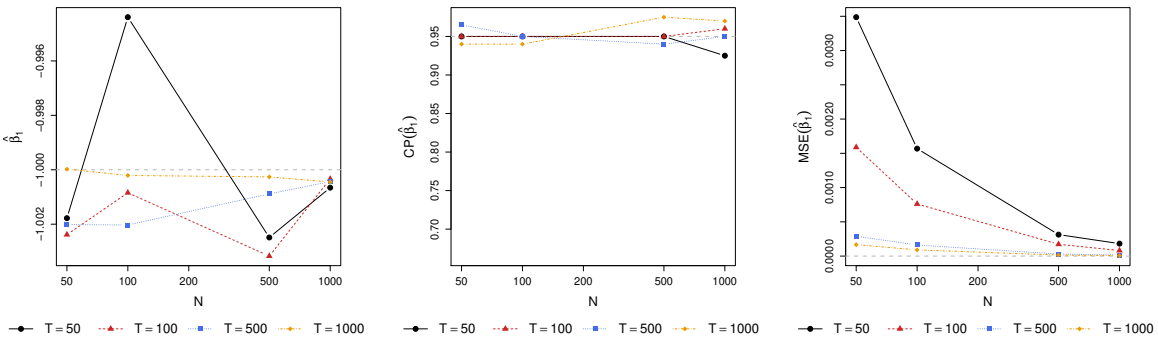
where the first row is for the intercept term (β_0's). Under this setting, the second and third eigencomponents of the covariance matrices follow the log-linear model (1) and the eigencondition (Condition 1) is satisfied. Table 1 presents the estimate of model coefficients β_1 over 200 simulations with $n = 100$ and $T_i = 100$. Since the intercept term is not of our study interest, we will not report the results here. For our proposed methods (CAP and CAP-C), the coverage probability is obtained by both the asymptotic variance in Theorem 1 (CP-A) and 500 bootstrap samples (CP-B); while for PCA and CPCA approaches, only CP-B is reported. As the data is generated under the complete common principal component assumption, the CAP-C approach yields the estimate of β with the lowest bias. The estimated β from CAP (or CAP-OC) for the first direction (the second eigencomponent) is very close to those from CAP-C, and the coverage probability from either the asymptotic variance or bootstrap achieves the designated level ($\alpha = 0.05$). For the second direction (the third eigencomponent), the estimate from CAP has slightly higher bias and the coverage probability is smaller than 0.95. The estimation bias of CAP-OC is higher, due to the orthogonality restriction. The higher bias in the proposed CAP approaches is possibly due to the data manipulation step in Algorithm 2. Both PCA and CPCA do not take into account the covariate information and thus the first two direction estimates are associated with the β components corresponding the largest two eigenvalues, even though the first β component is zero. The estimate of $\boldsymbol{\gamma}$ from CAP, CAP-OC and CAP-C are presented in Table E.1 of Section E in the supplementary material.

To further assess the finite sample performance of our proposed CAP approach, we vary the number of subjects with $n = 50, 100, 500, 1000$ and the number of observations within subject with $T_i = 50, 100, 500, 100$. Figure 2 shows the estimate, coverage probability from the asymptotic variance, and the mean squared error (MSE) of model coefficients of the first two directions. From the figure, as both n and T_i increase, the estimate of β_1 in

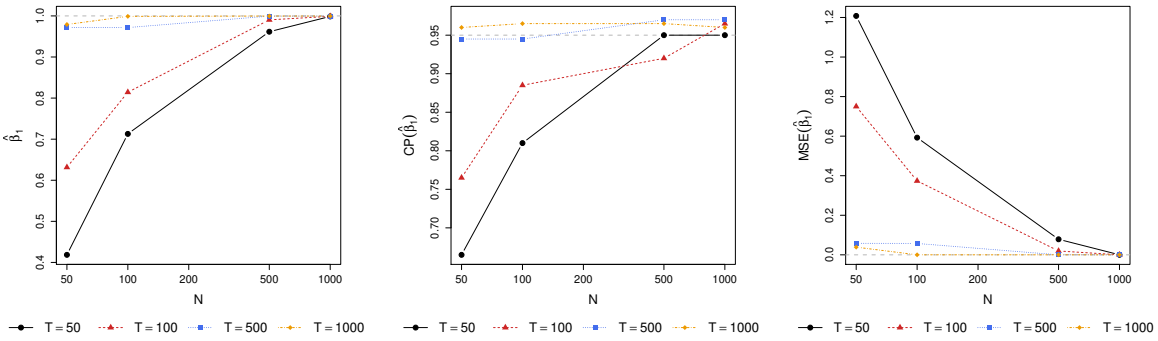
Table 1: Estimate (Est.) of β_1, as well as standard error (SE), coverage probability with asymptotic variance in Theorem 1 (CP-A) and coverage probability from 500 bootstrap samples (CP-B) from different methods under the alternative hypothesis. All values are computed with $n = 100$ and $T_i = 100$ over 200 simulations.

Method	First Direction			Second Direction		
	Est. (SE)	CP-A	CP-B	Est. (SE)	CP-A	CP-B
Truth	-1.00	-	-	1.00	-	-
CAP	-1.00 (0.03)	0.950	0.950	0.81 (0.58)	0.885	0.870
CAP-OC	-1.00 (0.03)	0.950	0.950	0.52 (0.84)	0.730	0.715
CAP-C	-1.00 (0.03)	0.950	0.955	1.00 (0.03)	0.975	0.960
PCA	-0.02 (0.10)	-	0	-0.98 (0.03)	-	0
CPCA	-0.01 (0.11)	-	0	-1.00 (0.03)	-	0

both the first and second identified direction converge to the true value. The coverage probability of the first direction is always close to the designated level, and the coverage probability of the second direction converges to 0.95 as both n and T_i increase. The MSE of both directions converge to zero. The simulation results demonstrate that our proposed method (CAP) can successfully recover the eigencomponents that possess multiplicative heteroscedasticity. Similar results from CAP-C are shown in Figure E.3 of Section E in the supplementary material.



(a) D1: Estimate of β_1. (b) D1: CP of β_1. (c) D1: MSE of $\hat{\beta}_1$.



(d) D2: Estimate of β_1. (e) D2: CP of β_1. (f) D2: MSE of $\hat{\beta}_1$.

Figure 2: Estimate and coverage probability (CP) with asymptotic variance (Theorem 1) of β_1 for the first (D1) and second (D2) projection directions, as well as the mean squared error (MSE) of β estimates under various combination of n and T values using CAP. The gray dashed lines are the target of estimates in (a) and (d), the designated level 0.95 in (b) and (e), and zero in (c) and (f).

5 Resting-state fMRI data example

We apply our proposed method to Human Connectome Project (HCP) resting-state fMRI (rs-fMRI) data. Our dataset includes $n = 118$ healthy young adults (39 aged 22-25 and 79 aged 26-30; 42 female and 76 male) from the most recent S1200 release. The sample size selected here is typical for fMRI studies. The rs-fMRI dataset was preprocessed following the minimal preprocessing pipeline in Glasser et al. (2013). Global signal regression was performed to address whole brain fluctuations typically seen as nuisances (Murphy et al., 2009; Fox et al., 2009). The blood-oxygen-level dependent (BOLD) signals are extracted from $p = 20$ functional brain regions in the default mode network (DMN) (Power et al., 2011) and averaged over voxels within the 5 mm radius. The BOLD time series are temporally correlated, thus we first calculate the effective sample size (ESS) defined by Kass et al. (1998), which infers the equivalent sample size of independent samples,

$$\mathrm{ESS}(p) = \min_{j\in\{1,\ldots,p\}} \left(\frac{T}{1 + 2\sum_{s=1}^{\infty} \mathrm{Cor}(\mathbf{y}_1^{(j)}, \mathbf{y}_{1+s}^{(j)})} \right),$$

where $\mathbf{y}_t^{(j)} = (y_{1t}^{(j)}, \ldots, y_{nt}^{(j)})$ is the data at time t of the jth brain region from all subjects, for $t = 1, 2, \ldots$ and $j = 1, \ldots, p$, and $T = 1200$ is the number of time points. We subsample $\mathrm{ESS}(p) = 660$ time points (demeaned and variance stabilized (Beckmann and Smith, 2004)) for analysis.

It has been shown that there exists sex discrepancy in functional connectivity in the DMN (Gong et al., 2011; Zhang et al., 2018). In this study, the individual demographic information, i.e., age and sex (both as categorical variables), together with their interaction are considered as the explanatory variables. For age, the category 22-25 is the reference level and labeled as Age1 and 26-30 as Age2; for sex, sex = 1 for male and 0 for female.

Four methods are compared in this study, including *(i)* element-wise correlation re-

gression (Wang et al., 2007); *(ii)* common principal component method, i.e., the CPCA method in Section 4; *(iii)* our CAP-C method; and *(iv)* our CAP method. In the simulation study (Section 4), it is shown that the proposed CAP approach without orthogonal constraint overperforms the approach with orthogonal constraint (CAP-OC). In this real data analysis, we employ the CAP approach and include a *post hoc* procedure to examine the orthogonality among the identified projection directions.

For the element-wise correlation regression, each off-diagonal element in the correlation matrix is Fisher z-transformed and multiple testing adjustment is performed following the Benjamini and Hochberg (1995) procedure to control the false discovery rate (FDR). None of the FDR corrected p-values are significant at level 0.05. See Figure F.1 and Figure F.2 in the supplementary material for the raw and FDR corrected p-values.

We present the estimated regression coefficients (together with 95% confidence intervals) of first ten common PCs from the CPCA approach in Figure F.3 in the supplementary material. From the figure, only the model coefficients of the fourth component (CPC4) are significant, indicating that not all of the top PCs are related to either age or sex. Under the same common PCA assumption, CAP-C directly discovers the PCs that are relevant to the covariates. Thus, the first component identified by CAP-C is CPC4. Figure F.4 shows the estimated model coefficients (and 95% confidence intervals from 500 bootstrap samples) of the top seven discovered PCs.

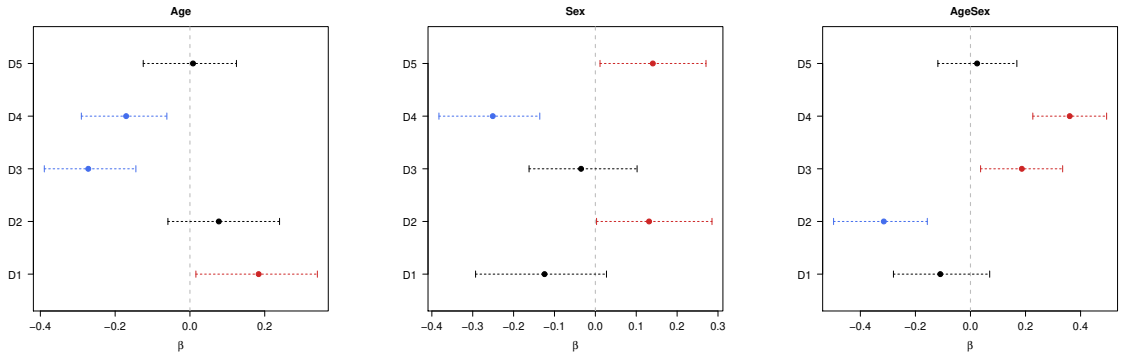
Our CAP approach discovers five projection directions, where the number five is chosen based on the average DfD (see Figure F.5 in Section F of the supplementary material), and the orthogonality of these five directions are verified in Figure F.10. Figure 3 exhibits the model coefficients (and 95% confidence intervals from 500 bootstrap samples) of the five projection directions. From the figures, for each identified projection direction, at least one of the covariates is significant. We use D1 as an example, which presents a significant

age effect. To interpret the loadings, Figure 4a shows the loading profile, and six brain regions have the loading magnitude greater than 0.2 (see Figure 4b in the brain map). This suggests that the connectivity between these brain regions show significant difference in the comparison (see Figure F.9 in the supplementary material for the scatter plot). In in the supplementary material, additional loading plots and brain maps are available in Figures F.7 and F.8, comparisons under different contrasts are shown in Figure F.6. To compare with CAP-C, Table F.1 in the supplementary material displays the similarity (similarity between -1 and 1, and 0 indicates orthogonal) of the projection directions to the common PCs from CAP-C. The two significant PCs identified by CAP-C, CPC4 and CPC18, have similarity greater than 0.6 to the projections D2 and D4 identified by CAP, respectively.

To study the reliability of our proposed methods, we apply the estimated projections to three other scanning sessions of resting-state fMRI data acquired from the same subjects. Figure F.11 in the supplementary material shows the estimated model coefficients and 95% bootstrap confidence intervals and Figure F.12 presents the comparisons under different contrasts. From the figures, the estimate and significance are very similar to the result presented in Figure 3, which validates the existence of difference between age groups and/or sex within these five components (also known as brain subnetworks) of the DMN.

6 Discussion

In this study, we introduce a Covariate Assisted Principal regression model for multiple covariance matrix outcomes. Our approach allows the identification of projection directions that are associated with the explanatory variables or covariates. Under certain regularity conditions, our proposed estimators are asymptotically consistent. Using extensive

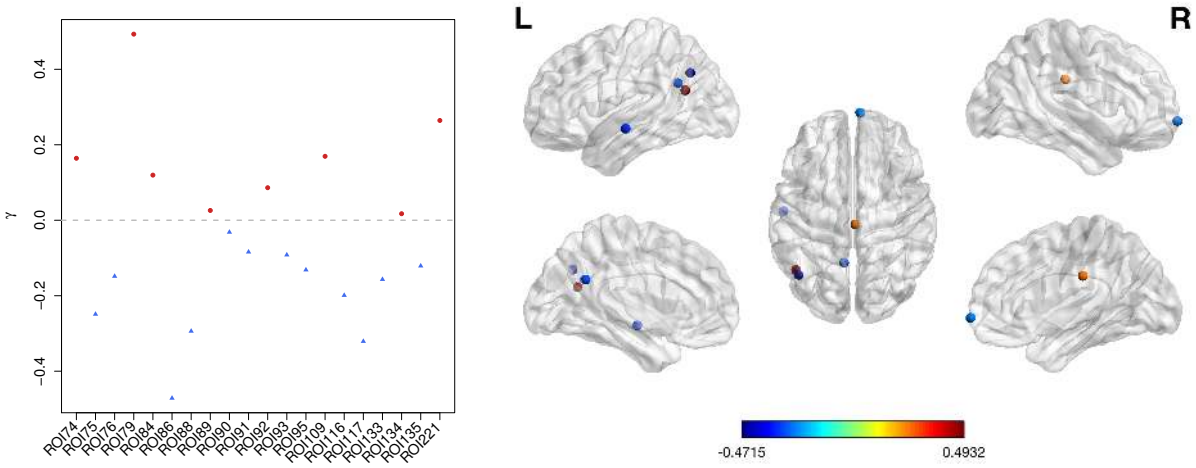


(a) Age. (b) Sex. (c) Age-Sex interaction.

Figure 3: Estimated model coefficients and 95% bootstrap confidence intervals of the six identified projection directions by CAP.

simulation studies, our model shows high estimation accuracy. Applied to resting-state fMRI studies, our method avoids the massive number of hypothesis testing suffered in the element-wise regression approach.

One challenge in modeling covariance matrices directly is having a constraint of positive definiteness. Via projections, the study of a positive definite matrix is decomposed into modeling the eigenvalues in orthogonal spaces. This relaxes the constraint and preserves geometric interpretation. The existing spectral decomposition based methods rely on the assumption that there exists a common diagonalization of the covariance matrices. In practice, this can be unrealistic, especially when p is large. Researchers are often more interested in studying a subset of the components related to the covariates. Though CAP enables identification of a small set of components, the theoretical analysis is challenging without the complete common diagonalization regularity condition in CPCA. One future direction will consider relaxing these assumptions.



(a) The loadings.

(b) Regions with $|\gamma_j|$ above 0.2 in brain map.

Figure 4: The loading profile and brain regions with absolute loading greater than 0.2 in projection direction D1 identified by CAP.

The current framework assumes the dimension of the data, p, is fixed and less than both the number of observations within a subject and the number of subjects. Another future direction is to extend the method to settings of large p, small n.

References

Anderson, T. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *The Annals of Statistics*, pages 135–141.

Beckmann, C. F. and Smith, S. M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE transactions on medical imaging*, 23(2):137–152.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300.

Boik, R. J. (2002). Spectral models for covariance matrices. *Biometrika*, 89(1):159–182.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252.

Carroll, R. J., Ruppert, D., and Holt Jr, R. N. (1982). Some aspects of estimation in heteroscedastic linear models. *Statistical decision theory and related topics, III*, 1:231–241.

Chiu, T. Y., Leonard, T., and Tsui, K.-W. (1996). The matrix-logarithmic covariance model. *Journal of the American Statistical Association*, 91(433):198–210.

Cohen, M., Dalal, S. R., and Tukey, J. W. (1993). Robust, smoothly heterogeneous variance regression. *Applied statistics*, pages 339–353.

Engle, R. F. and Kroner, K. F. (1995). Multivariate simultaneous generalized arch. *Econometric theory*, 11(1):122–150.

Flury, B. (1988). *Common principal components & related multivariate models*. John Wiley & Sons, Inc.

Flury, B. N. (1984). Common principal components in k groups. *Journal of the American Statistical Association*, 79(388):892–898.

Flury, B. N. (1986). Asymptotic theory for common principal component analysis. *The annals of Statistics*, pages 418–430.

Flury, B. N. and Gautschi, W. (1986). An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. *SIAM Journal on Scientific and Statistical Computing*, 7(1):169–184.

Fox, E. B. and Dunson, D. B. (2015). Bayesian nonparametric covariance regression. *Journal of Machine Learning Research*, 16:2501–2542.

Fox, M. D., Zhang, D., Snyder, A. Z., and Raichle, M. E. (2009). The global signal and observed anticorrelated resting state brain networks. *Journal of neurophysiology*, 101(6):3270–3283.

Franks, A. and Hoff, P. (2016). Shared subspace models for multi-group covariance estimation. *arXiv preprint arXiv:1607.03045*.

Friston, K., Frith, C., Liddle, P., and Frackowiak, R. (1993). Functional connectivity: the principal-component analysis of large PET data sets. *Journal of Cerebral Blood Flow & Metabolism*, 13(1):5–14.

Friston, K. J. (2011). Functional and effective connectivity: a review. *Brain connectivity*, 1(1):13–36.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., and Polimeni, J. R. (2013). The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124.

Gong, G., He, Y., and Evans, A. C. (2011). Brain connectivity: gender makes a difference. *The Neuroscientist*, 17(5):575–591.

Hafkemeijer, A., Möller, C., Dopper, E. G., Jiskoot, L. C., Schouten, T. M., van Swieten, J. C., van der Flier, W. M., Vrenken, H., Pijnenburg, Y. A., and Barkhof, F. (2015). Resting state functional connectivity differences between behavioral variant frontotemporal dementia and Alzheimer's disease. *Frontiers in human neuroscience*, 9.

Harvey, A. C. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica: Journal of the Econometric Society*, pages 461–465.

Hoff, P. D. (2009). A hierarchical eigenmodel for pooled covariance estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):971–992.

Hoff, P. D. and Niu, X. (2012). A covariance regression model. *Statistica Sinica*, pages 729–753.

Just, M. A., Cherkassky, V. L., Keller, T. A., Kana, R. K., and Minshew, N. J. (2006). Functional and anatomical cortical underconnectivity in autism: evidence from an fMRI study of an executive function task and corpus callosum morphometry. *Cerebral cortex*, 17(4):951–961.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1998). Markov chain Monte Carlo in practice: a roundtable discussion. *The American Statistician*, 52(2):93–100.

Lewis, C. M., Baldassarre, A., Committeri, G., Romani, G. L., and Corbetta, M. (2009). Learning sculpts the spontaneous activity of the resting human brain. *Proceedings of the National Academy of Sciences*, 106(41):17558–17563.

Luo, C., Li, Q., Lai, Y., Xia, Y., Qin, Y., Liao, W., Li, S., Zhou, D., Yao, D., and Gong, Q. (2011). Altered functional connectivity in default mode network in absence epilepsy: a resting-state fMRI study. *Human brain mapping*, 32(3):438–449.

Mennes, M., Vega Potler, N., Kelly, C., Di Martino, A., Castellanos, F. X., and Milham, M. P. (2012). Resting state functional connectivity correlates of inhibitory control in children with attention-deficit/hyperactivity disorder. *Frontiers in psychiatry*, 2:83.

Murphy, K., Birn, R. M., Handwerker, D. A., Jones, T. B., and Bandettini, P. A. (2009). The impact of global signal regression on resting state correlations: are anti-correlated networks introduced? *Neuroimage*, 44(3):893–905.

Park, B.-y., Kim, J., and Park, H. (2016). Differences in connectivity patterns between child and adolescent attention deficit hyperactivity disorder patients. In *Engineering in Medicine*

and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the, pages 1127–1130. IEEE.

Poldrack, R. A., Mumford, J. A., and Nichols, T. E. (2011). *Handbook of functional MRI data analysis*. Cambridge University Press.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690.

Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., and Schlaggar, B. L. (2011). Functional network organization of the human brain. *Neuron*, 72(4):665–678.

Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 329–358.

Rao, C. R. (1973). Algebra of vectors and matrices. *Linear Statistical Inference and its Applications: Second Editon*, pages 1–78.

Seiler, C. and Holmes, S. (2017). Multivariate heteroscedasticity models for functional brain connectivity. *Frontiers in neuroscience*, 11.

Smyth, G. K. (1989). Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 47–60.

Wang, K., Liang, M., Wang, L., Tian, L., Zhang, X., Li, K., and Jiang, T. (2007). Altered functional connectivity in early Alzheimer's disease: a resting-state fMRI study. *Human brain mapping*, 28(10):967–978.

Zhang, C., Dougherty, C. C., Baum, S. A., White, T., and Michael, A. M. (2018). Functional connectivity predicts gender: Evidence for gender differences in resting brain connectivity. *Human brain mapping*.

Zou, T., Lan, W., Wang, H., and Tsai, C.-L. (2017). Covariance regression analysis. *Journal of the American Statistical Association*, 112(517):266–281.

Supplementary Materials of "Covariate Assisted Principal Regression for Covariance Matrix Outcomes"

A Theory and Proof

A.1 A proposition for Algorithm 1 and proof of Proposition 1

Proposition A.1. *Suppose the vector* $\mathbf{x} \in \mathbb{R}^p$ *is subject to the restriction*

$$\mathbf{x}^\top \mathbf{H} \mathbf{x} = 1,$$

where matrix $\mathbf{H}$ *is positive definite. Then, the stationary points and values of* $\mathbf{x}^\top \mathbf{A} \mathbf{x}$ *are the eigenvectors and values of* $\mathbf{A}$ *with respect to* $\mathbf{H}$.

Proof. The Lagrangian of the optimization problem is

$$\mathcal{L}(\mathbf{X}, \lambda) = \mathbf{X}^\top \mathbf{A} \mathbf{X} - \lambda \left(\mathbf{X}^\top \mathbf{H} \mathbf{X} - 1 \right).$$

Taking partial derivatives gives

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{X}} &= 2\mathbf{A}\mathbf{X} - 2\lambda \mathbf{H}\mathbf{X} = \mathbf{0}, \\ \frac{\partial \mathcal{L}}{\partial \lambda} &= \mathbf{X}^\top \mathbf{H} \mathbf{X} - 1 = 0. \end{aligned}$$

Then

$$\mathbf{A}\mathbf{X} - \lambda \mathbf{H}\mathbf{X} = \mathbf{0}. \tag{A.1}$$

Thus the solution $(\mathbf{X}, \lambda)$ is the eigenvector and eigenvalue of $\mathbf{A}$ with respect to $\mathbf{H}$. The proof of Proposition 1 is straight forward by replacing $\mathbf{H}$ with $\mathbf{I}$. □

To find the eigenvectors and eigenvalues of $\mathbf{A}$ with respect to $\mathbf{H}$, we first assume $\mathbf{x}_0$ is a solution eigenvector that has Euclidean norm 1, i.e, $\mathbf{x}_0^\top \mathbf{x}_0 = 1$. Since $\mathbf{H}$ is positive definite, let $\mathbf{x} = \mathbf{H}^{-1/2}\mathbf{x}_0$, then

$$\mathbf{x}^\top \mathbf{H}\mathbf{x} = \mathbf{x}_0^\top \mathbf{H}^{-1/2}\mathbf{H}\mathbf{H}^{-1/2}\mathbf{x}_0 = \mathbf{x}_0^\top \mathbf{x}_0 = 1,$$

which satisfies the constraint condition. Replace $\mathbf{X}$ with $\mathbf{x} = \mathbf{H}^{-1/2}\mathbf{x}_0$ in (A.1),

$$\mathbf{A}\mathbf{H}^{-1/2}\mathbf{x}_0 - \lambda \mathbf{H}\mathbf{H}^{-1/2}\mathbf{x}_0 = \mathbf{0},$$

$$\Rightarrow \quad \mathbf{H}^{-1/2}\mathbf{A}\mathbf{H}^{-1/2}\mathbf{x}_0 = \lambda \mathbf{x}_0.$$

Therefore, $\mathbf{x}_0$ is the eigenvector of matrix $\mathbf{H}^{-1/2}\mathbf{A}\mathbf{H}^{-1/2}$.

In Algorithm 1, for the $(s+1)$the step,

$$\mathbf{A} = \sum_{i=1}^{n} \exp(-\mathbf{x}_i^\top \boldsymbol{\beta}^{(s)})\mathbf{S}_i, \quad \mathbf{H} = \begin{cases} \mathbf{I}, & \text{if under constraint (C1)} \\ \bar{\Sigma}, & \text{if under constraint (C2)} \end{cases}.$$

We can first find the eigenvectors of $\mathbf{H}^{-1/2}\mathbf{A}\mathbf{H}^{-1/2}$, left multiplied by $\mathbf{H}^{-1/2}$, solve for $\boldsymbol{\beta}$ using formula (5). The update of γ and $\boldsymbol{\beta}$ will be the pair that jointly minimizes the objective function.

A.2 Details of Algorithm 2

In Algorithm 2, with the new data $\tilde{\mathbf{Y}}_i^{(k)}$, we need to solve the following optimization problem

$$\begin{aligned} \underset{\boldsymbol{\gamma},\boldsymbol{\beta}}{\text{minimize}} \quad & \frac{1}{2}\sum_{i=1}^{n} T_i(\mathbf{x}_i^\top \boldsymbol{\beta}) + \frac{1}{2}\boldsymbol{\gamma}^\top \left(\sum_{i=1}^{n} \frac{T_i \tilde{\Sigma}_i^{(k)}}{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})} \right) \boldsymbol{\gamma}, \\ \text{subject to} \quad & \boldsymbol{\gamma}^\top \left(\frac{1}{n}\sum_{i=1}^{n} \tilde{\Sigma}_i^{(k)} \right) \boldsymbol{\gamma} = 1, \\ & \Gamma^{(k-1)\top}\boldsymbol{\gamma} = \mathbf{0}, \end{aligned}$$

where $\tilde{\Sigma}_i^{(k)} = \tilde{\mathbf{Y}}_i^{(k)\top}\tilde{\mathbf{Y}}_i^{(k)}/T_i$. With given (or an initial) $\boldsymbol{\beta}$, let

$$\mathbf{A} = \sum_{i=1}^{n} \frac{T_i\tilde{\Sigma}_i^{(k)}}{\exp(\mathbf{x}_i^\top\boldsymbol{\beta})}, \quad \mathbf{H} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\Sigma}_i^{(k)}, \quad \mathbf{C} = \Gamma^{(k-1)},$$

we first apply the solution in Rao (1964, 1973) to find the stationary points, which are the eigenvectors of $(\mathbf{I}-\mathbf{P})\mathbf{A}$ with respect to $\mathbf{H}$, where $\mathbf{P} = \mathbf{C}(\mathbf{C}^\top\mathbf{H}^{-1}\mathbf{C})^{-}\mathbf{C}^\top\mathbf{H}^{-1}$ is the projection operator onto $\mathcal{M}(\mathbf{C})$ (the linear manifold spanned by $\mathbf{C}$). For each eigenvector, find the solution for $\boldsymbol{\beta}$ using the formula in Algorithm 1. The update of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ will be the pair that jointly minimizes the objective function.

B Proof of Theorem 1

Proof. With true $\boldsymbol{\gamma}$, our proposed estimator of $\boldsymbol{\beta}$ is the maximum likelihood estimator (MLE). Therefore, the asymptotic results of MLE can be applied.

For subject i $(i = 1, \ldots, n)$ observation t $(t = 1, \ldots, T_i)$, the log-likelihood function (with a constant difference) is

$$\ell_{it} = -\frac{1}{2}\mathbf{x}_i^\top\boldsymbol{\beta} - \frac{1}{2}(\boldsymbol{\gamma}^\top\mathbf{y}_{it}\mathbf{y}_{it}^\top\boldsymbol{\gamma})\exp(-\mathbf{x}_i^\top\boldsymbol{\beta}).$$

For the full dataset, let $M_n = \sum_{i=1}^{n} T_i$ and

$$\mathcal{L}_{nT}(\boldsymbol{\beta}) = \frac{1}{M_n}\sum_{i=1}^{n}\sum_{t=1}^{T_i}\ell_{it}.$$

$\hat{\boldsymbol{\beta}}$ is the solution to $\mathcal{L}'_{nT} = 0$. We expand the function at the true parameter $\boldsymbol{\beta}_0$ as

$$0 = \mathcal{L}'_{nT}(\hat{\boldsymbol{\beta}}) = \mathcal{L}'_{nT}(\boldsymbol{\beta}_0) + \mathcal{L}''_{nT}(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \mathcal{R}_{nT},$$

where $\mathcal{R}_{nT}$ is the residual term.

$$\Rightarrow \quad \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = -\left(\mathcal{L}''_{nT}(\boldsymbol{\beta}_0)\right)^{-1}\left[\mathcal{L}'_{nT}(\boldsymbol{\beta}_0) + \mathcal{R}_{nT}\right]$$

$$\ell'_{it}(\boldsymbol{\beta}) = -\frac{1}{2}\mathbf{x}_i + \frac{1}{2}\exp(-\mathbf{x}_i^\top\boldsymbol{\beta})(\boldsymbol{\gamma}^\top\mathbf{y}_{it}\mathbf{y}_{it}^\top\boldsymbol{\gamma})\mathbf{x}_i \quad \Rightarrow \quad \mathbb{E}_{\boldsymbol{\beta}_0}\ell'_{it}(\boldsymbol{\beta}) = \mathbf{0}.$$

$$\ell''_{it}(\boldsymbol{\beta}) = -\frac{1}{2}\exp(-\mathbf{x}_i^\top\boldsymbol{\beta})(\boldsymbol{\gamma}^\top\mathbf{y}_{it}\mathbf{y}_{it}^\top\boldsymbol{\gamma})\mathbf{x}_i\mathbf{x}_i^\top \quad \Rightarrow \quad -\mathbb{E}_{\boldsymbol{\beta}_0}\ell''_{it}(\boldsymbol{\beta}) = \frac{1}{2}\mathbf{x}_i\mathbf{x}_i^\top.$$

Under the assumption that $\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top/n \to \mathbf{Q}$,

$$\sqrt{M_n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, 2\mathbf{Q}^{-1}\right), \quad \text{as } n, T \to \infty,$$

where $T = \min_i T_i$. □

C Proof of Theorem 2

Proof. We propose to estimate $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ by maximizing the likelihood function. Under the complete common principal component assumption, Flury (1986) showed the asymptotic distribution of $\hat{\boldsymbol{\gamma}}$. Together with the conclusion of Theorem 1, the consistency of $\hat{\boldsymbol{\beta}}$ follows. □

D Toy examples

We use three examples to demonstrate the property of the two considered constraints. Assume X_i is generated from a Bernoulli distribution with probability 0.5 to be 1.

D.1 Example I

Let $\boldsymbol{\beta}_1 = (2,3)^\top$ and $\boldsymbol{\beta}_2 = (2,-3)^\top$, and assume $\Sigma_i = \Gamma\Lambda_i\Gamma^\top$, where

$$\Gamma = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad \Lambda_i = \begin{pmatrix} \exp(\mathbf{x}_i^\top\boldsymbol{\beta}_1) & 0 \\ 0 & \exp(\mathbf{x}_i^\top\boldsymbol{\beta}_2) \end{pmatrix},$$

where $\mathbf{x}_i = (1, X_i)^\top$. When $X_i = 1$, $\Sigma_i = \Gamma\Lambda_i^{(1)}\Gamma^\top$ with $\Lambda_i^{(1)} = \text{diag}\{\exp(5), \exp(-1)\}$, and when $X_i = 0$, $\Sigma_i = \Gamma\Lambda_i^{(0)}\Gamma^\top$ with $\Lambda_i^{(0)}\text{diag}\{\exp(2), \exp(2)\}$, where the projection onto the first eigenspace contains larger variation in the data. We generate $\mathbf{y}_{it}$'s from the multivariate normal distribution with mean zero and covariance matrix Σ_i, for $t = 1, \ldots, T_i = 100$ and $i = 1, \ldots, n = 100$.

Then$\Gamma^\top \mathbf{y}_{it}$ follows the multivariate normal distribution with covariance matrix Λ_i. Figures D.1a presents the contour plot of the objective function in 2 with $\boldsymbol{\beta}^* = \boldsymbol{\beta}_1$. Under (C1), from Algorithm 1, the solution for$\boldsymbol{\gamma}$ is the eigenvector corresponding to the minimum eigenvalue of matrix $\sum_{i=1}^{n} \exp(-\mathbf{x}_i^\top \boldsymbol{\beta})\mathbf{S}_i$. Constraint (C2) regulates the shape of the constraint set by the average sample covariance matrix.

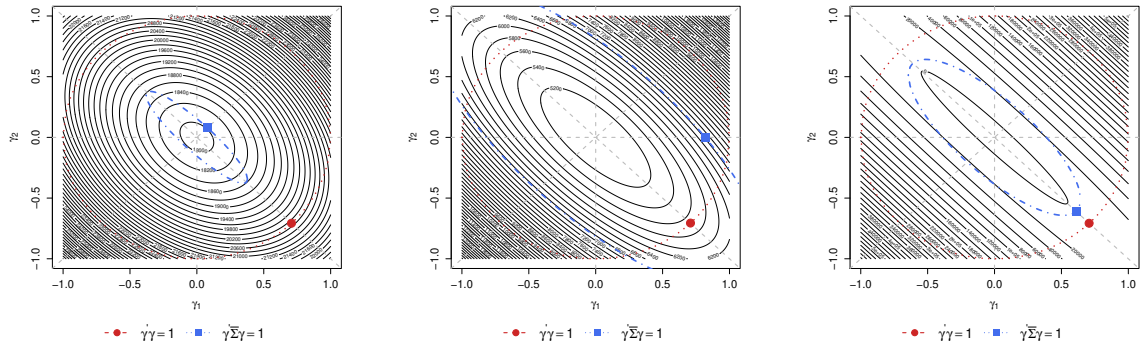
D.2 Example II

Let $\boldsymbol{\beta}_1 = (1, 0)^\top$ and $\boldsymbol{\beta}_2 = (-1, 0)^\top$, which is the null scenario of $\boldsymbol{\beta}$. The rest parameter settings are the same as in Example I. Under this scenario, $\exp(-\mathbf{x}_i^\top \boldsymbol{\beta})$ is a constant, and thus the constraint set under (C2) is parallel to the contour plot of the objective function under the true $\boldsymbol{\beta}$ (see Figure D.1b). Therefore, the estimate of $\boldsymbol{\gamma}$ can be any value in the constraint set.

D.3 Example III

Let $\boldsymbol{\beta} = (1, -3)^\top$, and $\Sigma_i = \Gamma \Lambda_i \Gamma^\top$, where $\Lambda_i = \text{diag}\{\sigma_{i1}^2, \exp(\mathbf{x}_i^\top \boldsymbol{\beta})\}$ and $\log(\sigma_{i1}^2)$ follows a normal distribution with mean two and standard deviation one. The rest parameters are set to be the same as in Example I. In this example, the component with lower variation is relevant to the covariate X. Figure D.1c shows the contour plot under the true $\boldsymbol{\beta}$. Both constraints identify the second component as the estimator of $\boldsymbol{\gamma}$.

Using Examples I to III, we conclude that under constraint (C1), the proposed method yields the estimate of the component with the lowest variation in the data; while constraint (C2) identifies the component that satisfies the model assumption (1).



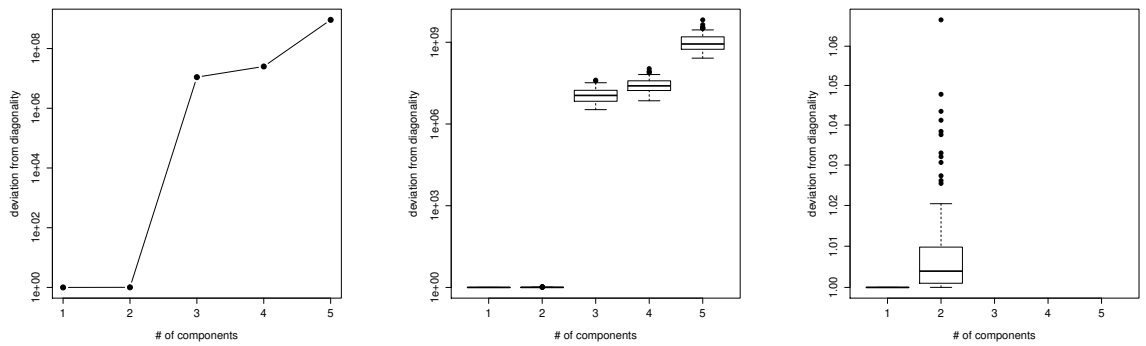
(a) Example I with $\boldsymbol{\beta}^* = \boldsymbol{\beta}_1$. (b) Example II with $\boldsymbol{\beta}^* = \boldsymbol{\beta}_1$. (c) Example III with $\boldsymbol{\beta}^* = \boldsymbol{\beta}$.

Figure D.1: The contour plot of the negative log-likelihood function in (a) Example I, (b) Example II and (c) Example III. The blue curve and point are the constraint (C2) function and the estimate, respectively; and the red are under constraint (C1).

E Additional Simulation Results

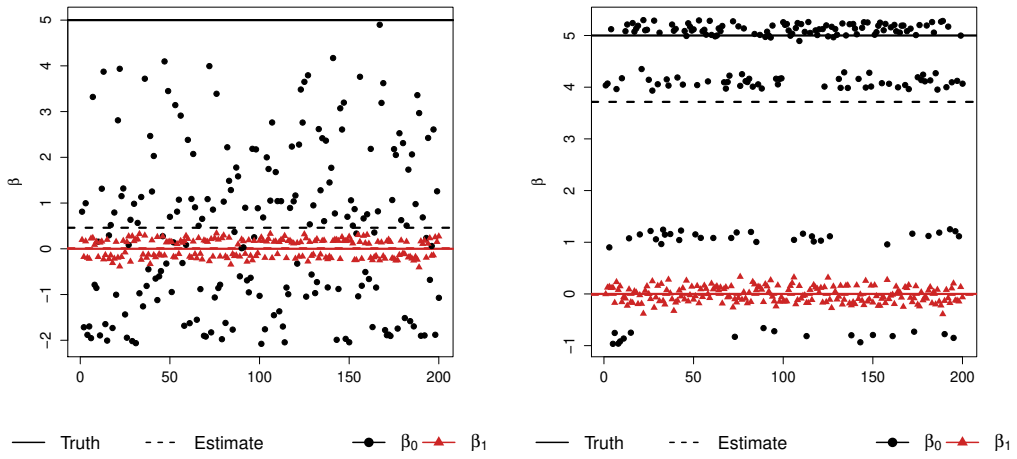
We first use a simulated example to demonstrate the performance of the "deviation from diagonality" metric defined in (8). The data is generated following the alternative scenario in Section 4. Figure E.1a shows the average DfD and Figures E.1b and E.1c are the boxplot of individual DfD, where the γ's are estimated using our proposed CAP method. From the figures, for all samples, when moving to the third component, the DfD value jumps to over 10^6. Thus, two is the proper number of components to be chosen, which is the same as the truth. Therefore, the proposed average DfD is an appropriate metric to chose the number of projection directions.

Under the null case, we present the estimate of β's from CAP and CAP-C over 200 simulations in Figure E.2. As demonstrated in the toy example II in Section D, under constraint (C2), our method could not identify the principal direction of projection, and thus the estimate of β_0 from CAP and CAP-C varies according to the estimated $\boldsymbol{\gamma}$. However, the estimate of β_1 is centered around zero with an average of 0.01 (SE: 0.20) under CAP and -0.01 (SE: 0.15) under CAP-C.



(a) Average DfD. (b) Individual DfD. (c) Individual DfD (zoomed in).

Figure E.1: Average and individual "deviation from diagonality" of a simulated example.



(a) CAP (b) CAP-C

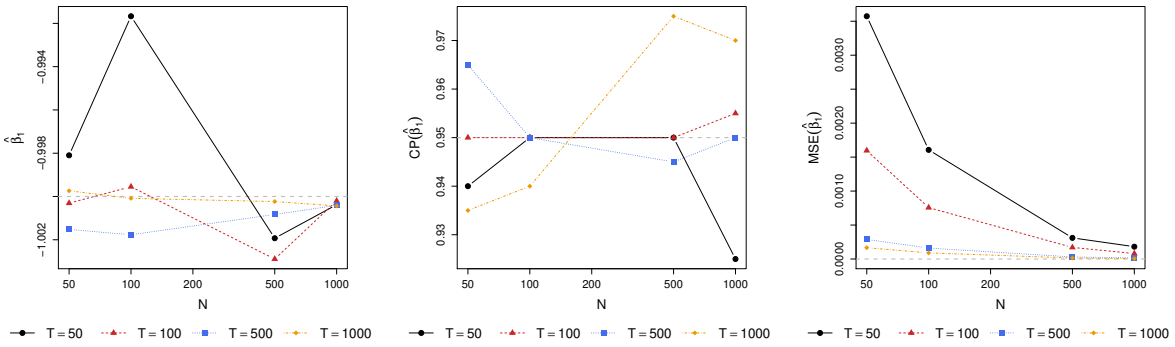
Figure E.2: Estimate of β_0 and β_1 in the 200 simulations from (a) CAP and (b) CAP-C methods with $n = 100$ and $T_i = 100$ under the null case.

Table E.1 presents the estimate of $\boldsymbol{\gamma}$ using CAP and CAP-C methods with $n = 100$ and $T_i = 100$. Both methods yield correct identification of the two principal directions. CAP-C attains lower bias and variation, which is optimal under the complete common principal component assumption.

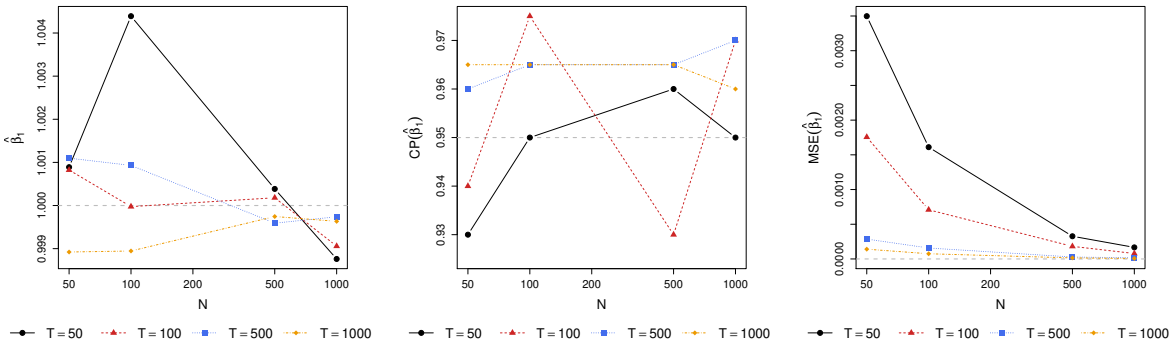
Table E.1: Estimate (standard error) of $\boldsymbol{\gamma}$ under the alternative scenario with $n = 100$ and $T_i = 100$.

	γ	Truth	CAP	CAP-OC	CAP-C
First Direction	γ_1	0.45	0.42 (0.125)	0.42 (0.125)	0.45 (0.002)
	γ_2	-0.86	-0.82 (0.057)	-0.82 (0.057)	-0.86 (0.002)
	γ_3	0.14	0.13 (0.050)	0.13 (0.050)	0.14 (0.004)
	γ_4	0.14	0.13 (0.161)	0.13 (0.161)	0.14 (0.002)
	γ_5	0.14	0.14 (0.210)	0.14 (0.210)	0.14 (0.002)
Second Direction	γ_1	0.45	0.43 (0.085)	0.43 (0.111)	0.45 (0.002)
	γ_2	0.14	0.15 (0.110)	0.13 (0.157)	0.14 (0.003)
	γ_3	-0.86	-0.76 (0.282)	-0.61 (0.423)	-0.86 (0.001)
	γ_4	0.14	0.13 (0.087)	0.11 (0.171)	0.14 (0.003)
	γ_5	0.14	0.06 (0.288)	-0.06 (0.408)	0.14 (0.002)

Figure E.3 shows the estimate of β using CAP-C as both n and T_i increases. As CAP-C correctly identifies the two components that satisfy the model assumption (1), the estimate of β is close to the true value and the coverage probability reaches the designated level under all combinations of n and T_i values.

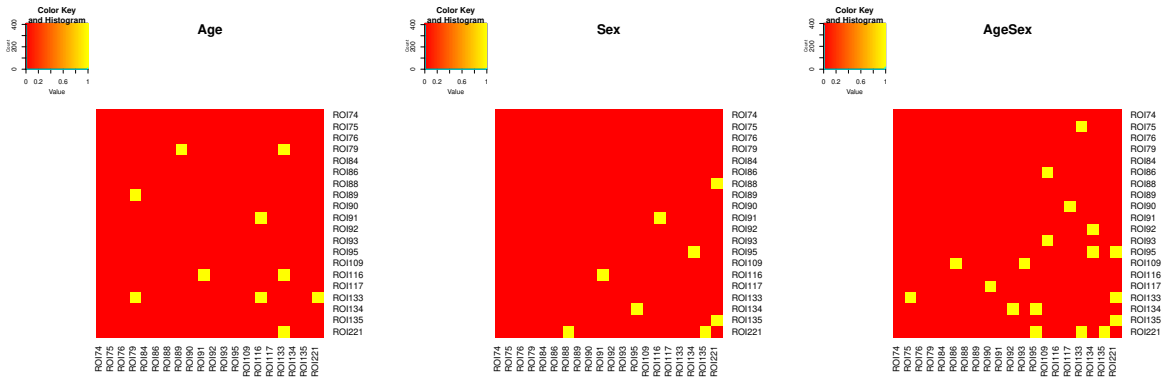


(a) D1: Estimate of β_1. (b) D1: Coverage probability of β_1. (c) D1: MSE of $\hat{\beta}_1$.



(d) D2: Estimate of β_1. (e) D2: Coverage probability of β_1. (f) D2: MSE of $\hat{\beta}_1$.

Figure E.3: Estimate and coverage probability (CP) with asymptotic variance (Theorem 1) of β_1 for the first (D1) and second (D2) projecting direction, as well as the mean squared error (MSE) of β estimates under various combination of n and T values using CAP-C. The gray dashed line in (a) and (d) are the target of estimates and zero in the rest.



(a) Age. (b) Sex. (c) Age-Sex interaction.

Figure F.1: Significance of model coefficients with original p-value at level of 0.05 in the element-wise correlation regression. The yellow elements are significant, and the red are not.

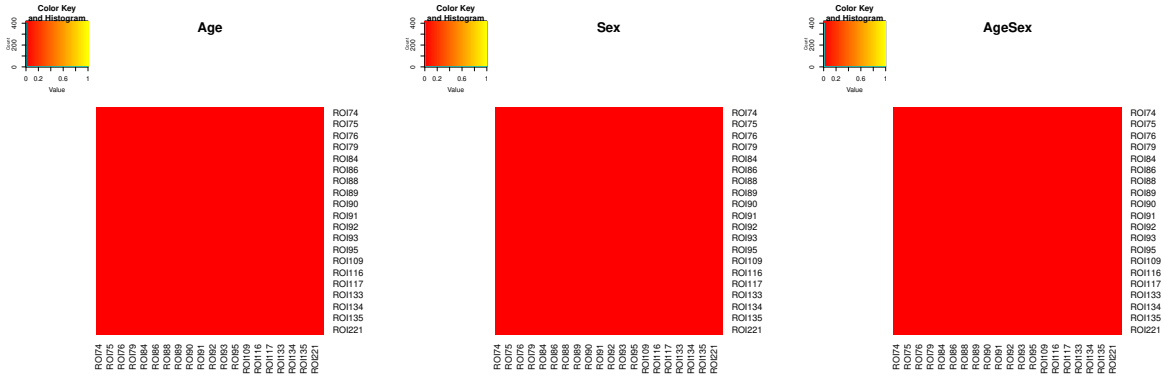
F Additional Real Data Analysis Results

F.1 The element-wise regression approach

Figure F.1 shows the significance of model coefficients with original p-value less than 0.05 in the element-wise regression analysis. Figure F.2 shows the significance after multiple testing correction, where all become insignificant.

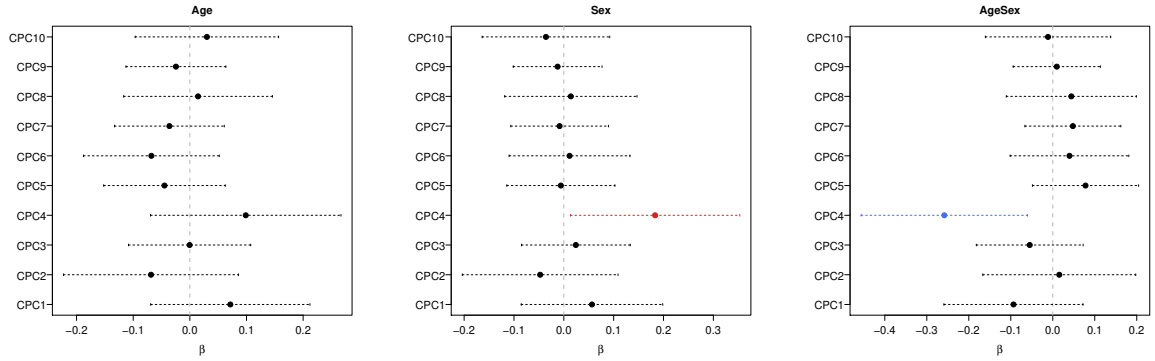
F.2 The CPCA approach

We present the estimated model coefficients (together with 95% confidence interval from the regression model) of the first ten common PCs from the CPCA approach in Figure F.3. From the figure, the model coefficients of CPC5, CPC6 and CPC7 are not significant, indicating that brain connectivity within the corresponding brain network does not show any difference when comparing age and sex groups. The CAP-C method builds on the common diagonalization assumption as in the CPCA approach, but targets on the PCs that satisfy the log-linear model assumption.



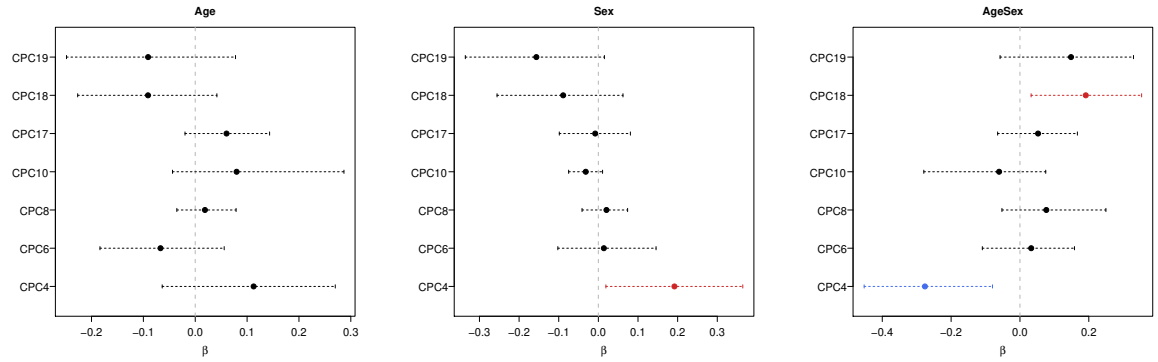
(a) Age. (b) Sex. (c) Age-Sex interaction.

Figure F.2: Significance of model coefficients with adjusted p-value at level of 0.05 in the element-wise correlation regression. The yellow elements are significant, and the red are not.



(a) Age. (b) Sex. (c) Age-Sex interaction.

Figure F.3: Estimated model coefficients and 95% confidence interval of the first ten common PCs in the CPCA approach.



(a) Age. (b) Sex. (c) Age-Sex interaction.

Figure F.4: Estimated model coefficients and 95% bootstrap confidence interval of the three identified common PCs from the CAP-C approach.

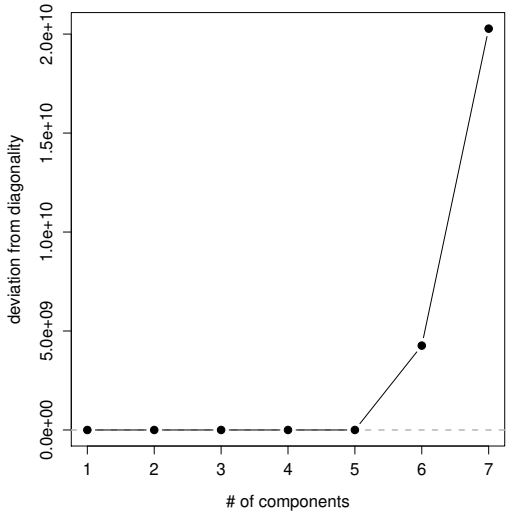
F.3 The CAP-C approach

Figure F.4 shows the estimated model coefficients (and 95% confidence interval from 500 bootstrap samples) of the three discovered PCs, which also satisfy the eigenvalue condition (Condition 1). Though CPC3 has significant coefficient in sex, the corresponding eigenvalue condition is violated and thus is not identified by the CAP-C approach.

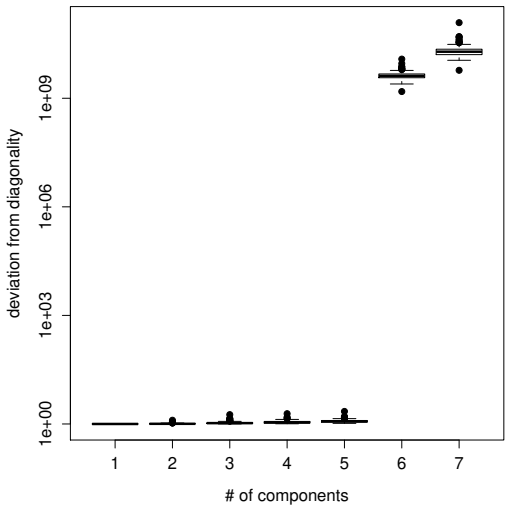
F.4 The CAP approach

Figure F.5 presents the average and individual "deviation from diagonality" of the first seven projection directions in the real data analysis. We observe a sudden jump on the sixth direction, therefore we choose the first five components.

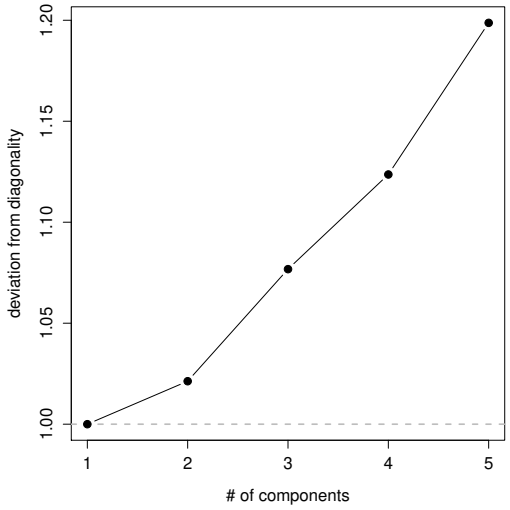
Figure F.7 presents the loadings of the five projection directions from the CAP approach, and Figure F.8 is the visualization of the loadings in the brain map. Figure F.9 shows the scatter plot of the model outcome $\log(\boldsymbol{\gamma}^\top \hat{\Sigma}_i \boldsymbol{\gamma})$ by age and sex group for the five projection directions from the CAP approach. From the figure, we observe the interaction effect in D2, D4 and D5.



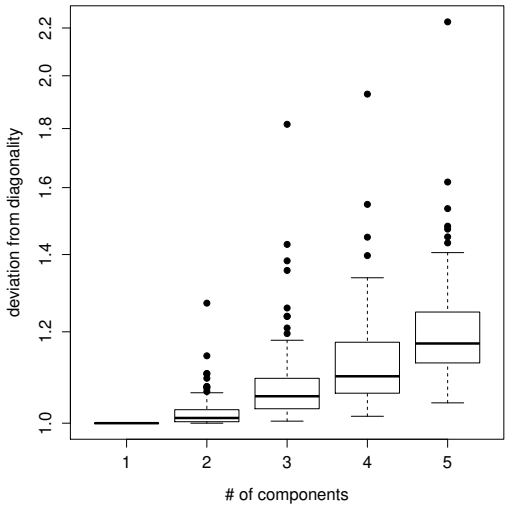
(a) Average "deviation from diagonality".



(b) Individual "deviation from diagonality".

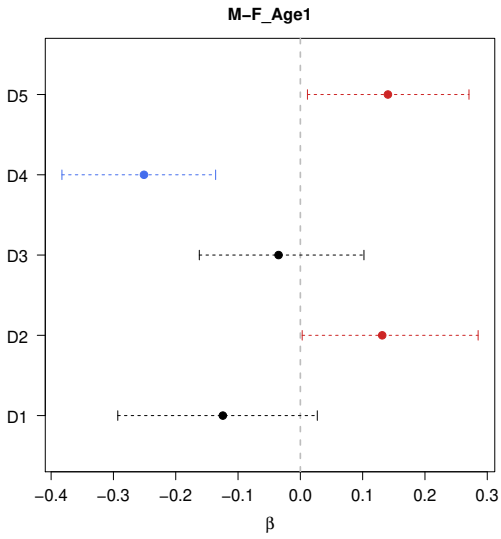


(c) Average "deviation from diagonality".

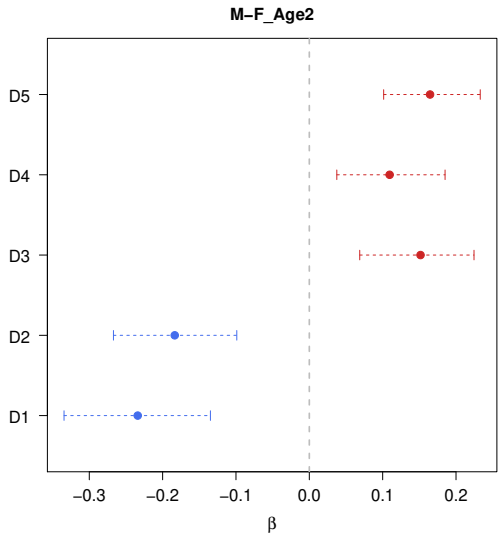


(d) Individual "deviation from diagonality".

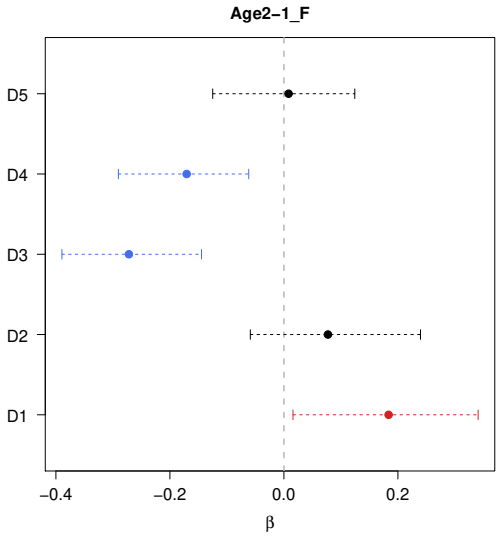
Figure F.5: The average and individual "deviation from diagonality" of the first seven ((a)-(b)) and first five ((c)-(d)) projection directions in the real data analysis.



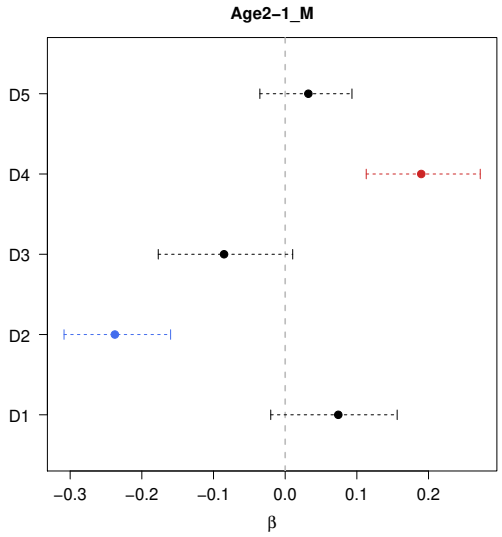
(a) Age 22-25: Male vs. Female.



(b) Age 26-30: Male vs. Female.



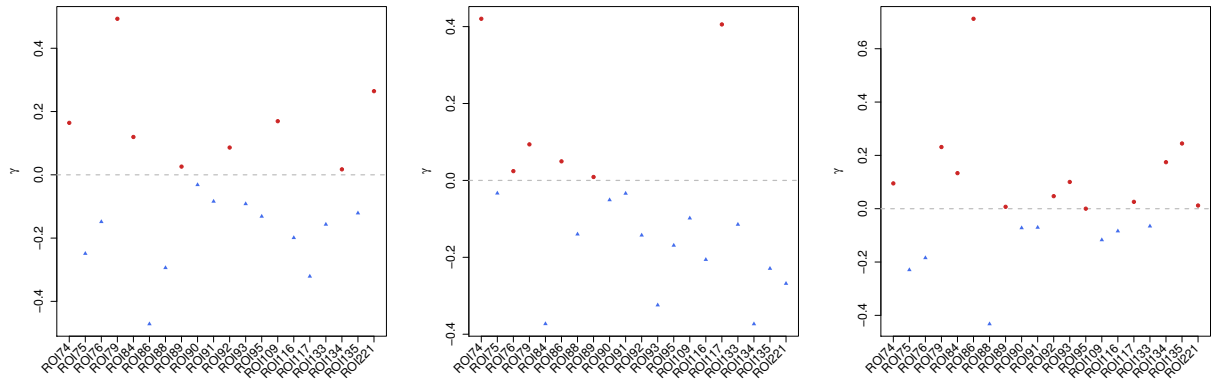
(c) Female: Age 22-25 vs. Age 26-30.



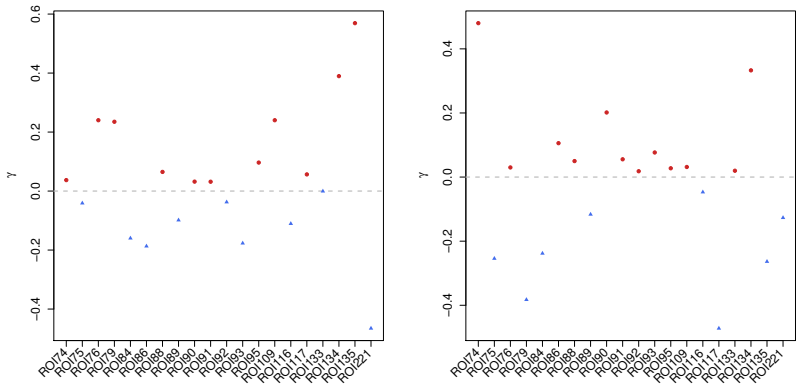
(d) Male: Age 22-25 vs. Age 26-30.

Figure F.6: Pair-wise comparison of the five identified projection directions from the CAP approach. The confidence interval is obtained from 500 bootstrap sample.

Table F.1 displays the similarity (similarity between -1 and 1, and 0 indicates orthogonal) of the projecting directions to the PCs from CAP-C. The proposed CAP approach recovers the three PCs with high similarity and detects three additional. Using the definition in Krzanowski (1979),

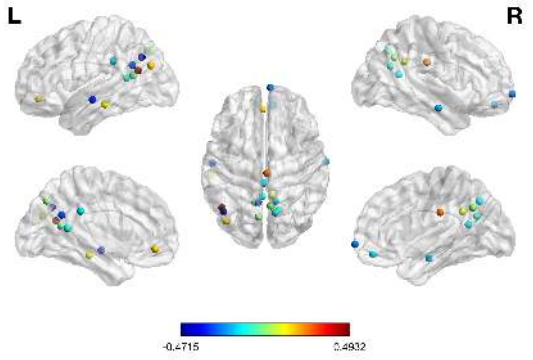


(a) D1. (b) D2. (c) D3.

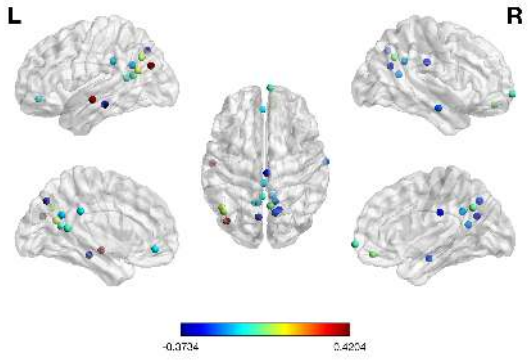


(d) D4. (e) D5.

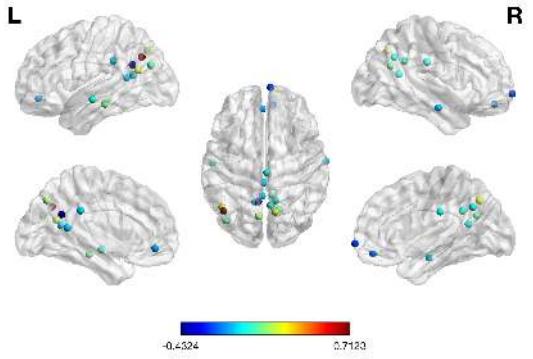
Figure F.7: The loadings of the five projection directions from the CAP approach.



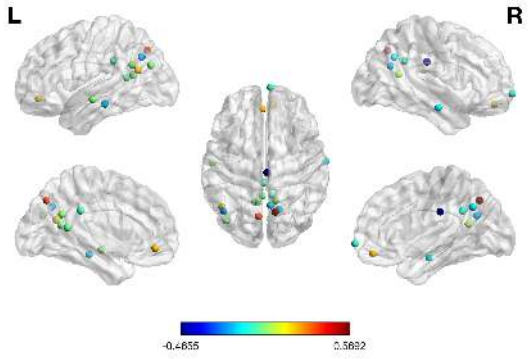
(a) D1.



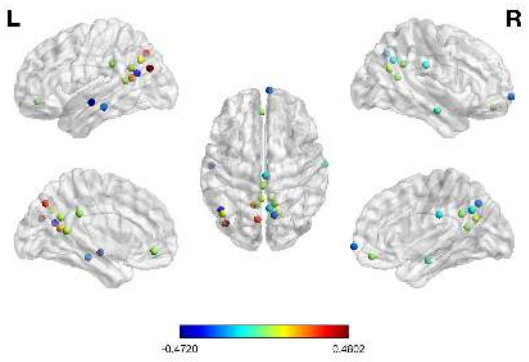
(b) D2.



(c) D3.

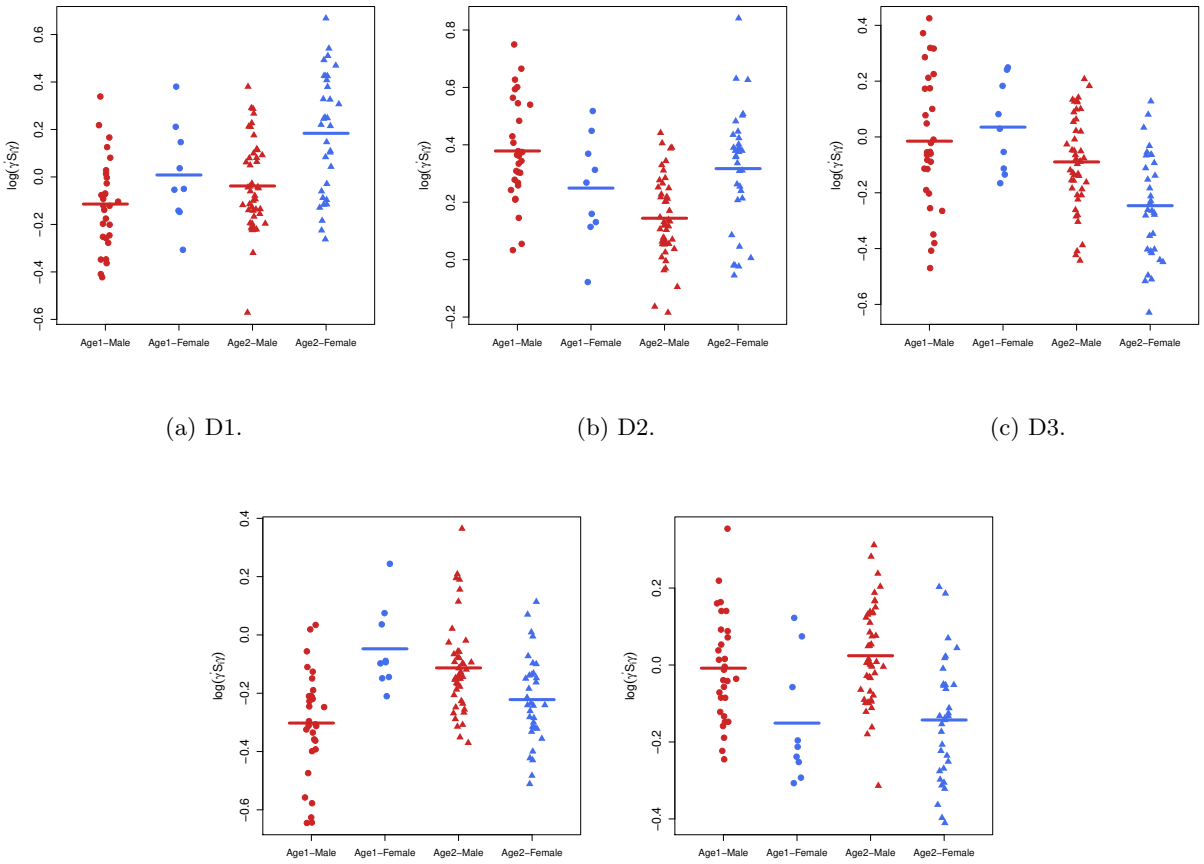


(d) D4.



(e) D5.

Figure F.8: The loading map of the five projection directions from the CAP approach. The color legend indicates the value of $\boldsymbol{\gamma}$.



(a) D1. (b) D2. (c) D3.

(d) D4. (e) D5.

Figure F.9: Scatter plot of the outcome in the log-linear model ($\log(\boldsymbol{\gamma}^\top \hat{\Sigma}_i \boldsymbol{\gamma})$) by age and sex groups for the five projection directions from the CAP approach.

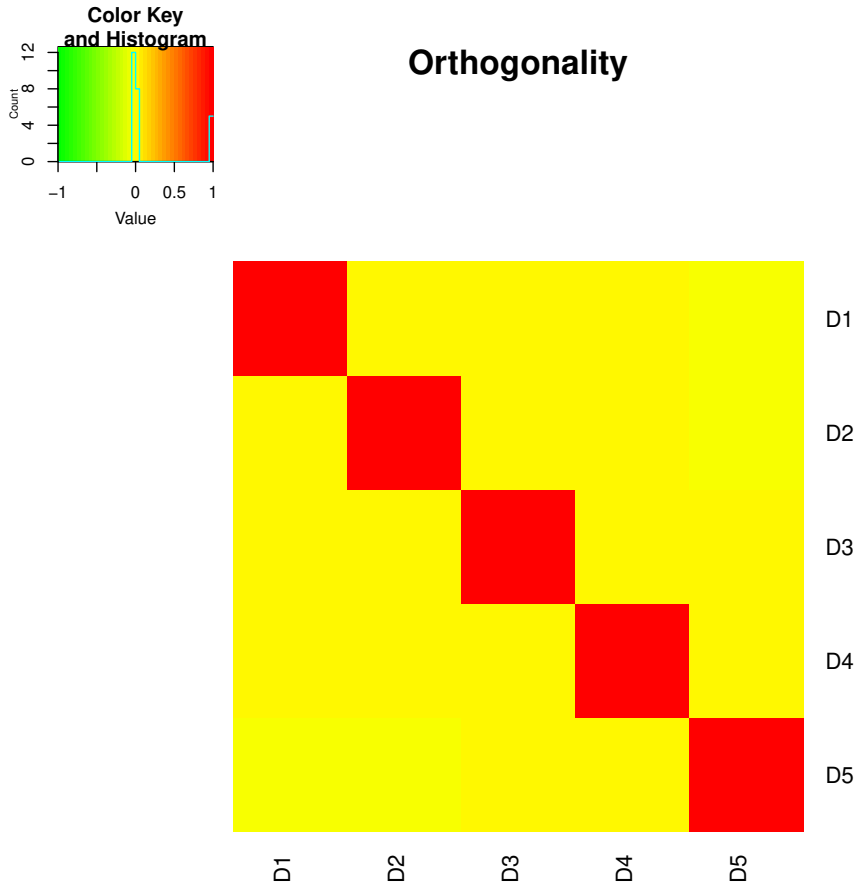


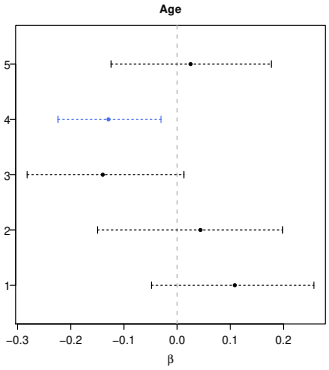
Figure F.10: Orthogonality of the five identified projection directions from CAP.

Table F.1: Similarity between the five projecting directions from CAP and the seven PCs from CAP-C method.

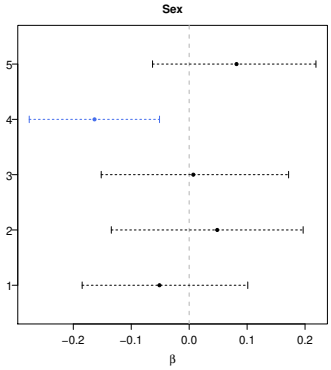
		CAP-C						
		CPC4	CPC6	CPC8	CPC10	CPC17	CPC18	CPC19
CAP	D1	-0.115	-0.065	0.020	0.140	0.024	0.202	0.165
	D2	**0.638**	0.144	0.053	-0.051	-0.086	0.139	-0.078
	D3	-0.193	-0.311	-0.007	0.056	-0.067	-0.118	-0.096
	D4	-0.072	0.051	0.208	0.234	-0.329	**-0.669**	-0.339
	D5	-0.214	0.432	0.192	-0.016	0.009	-0.153	0.349

the similarity between the two spaces discovered by CAP and CAP-C is 0.386, indicating that the space spanned by the seven identified PCs from CAP-C is different from the one spanned by the five components discovered by CAP.

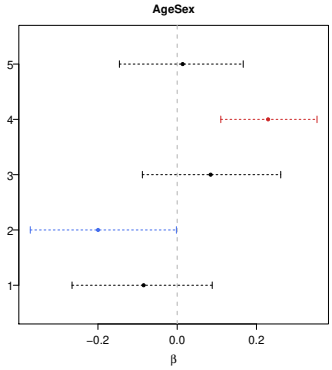
To study the reliability of our proposed method, we apply the same linear projection to the rest three sessions of resting-state fMRI data acquired from the same subjects in the HCP study. Figure F.11 shows the estimated model coefficients and 95% bootstrap confidence interval. From the figure, the estimate and significance are very similar to the result presented in Figure 3 of Section 5, which postulates the existence of difference between age groups and/or sex within these five subnetworks of the DMN.



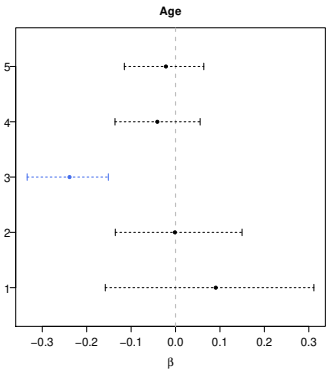
(a) REST1 RL: Age.



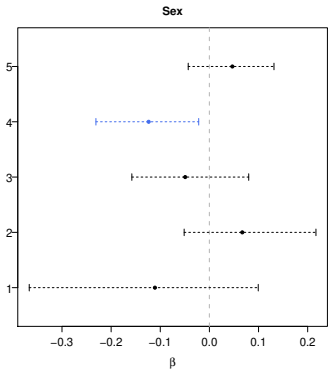
(b) REST1 RL: Sex.



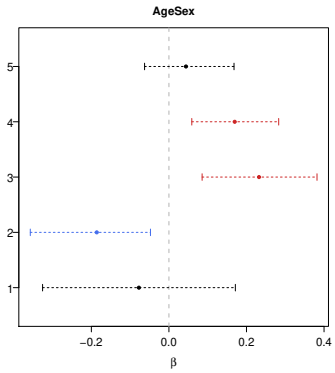
(c) REST1 RL: Age-Sex interaction.



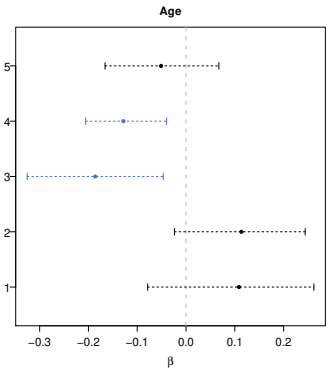
(d) REST2 LR: Age.



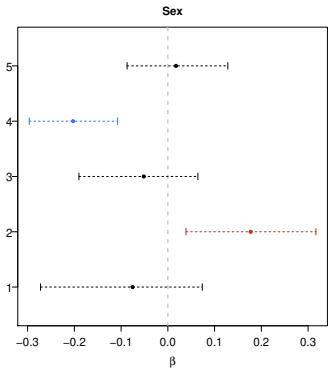
(e) REST2 LR: Sex.



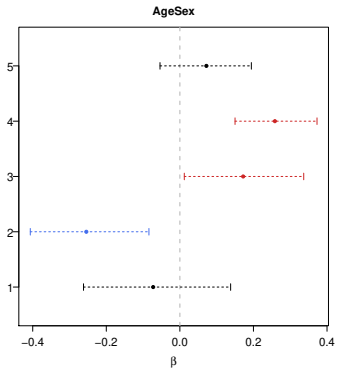
(f) REST2 LR: Age-Sex interaction.



(g) REST2 RL: Age.

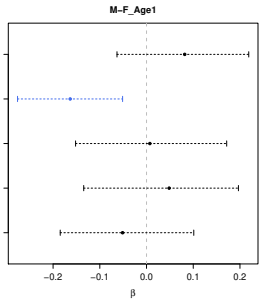


(h) REST2 RL: Sex.

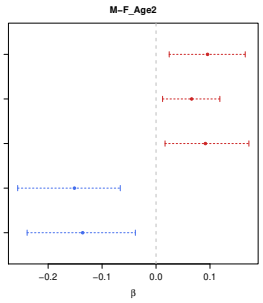


(i) REST2 RL: Age-Sex interaction.

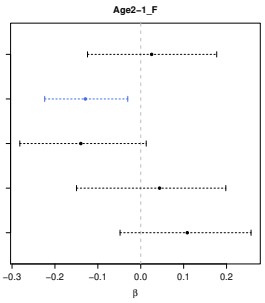
Figure F.11: Estimated model coefficients and 95% bootstrap confidence interval of the five identified projection directions from the CAP approach in Section 5 tested on the rest three sessions of resting-state data collected from the same subjects in the HCP study.



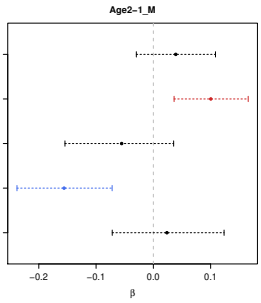
(a) REST1 RL



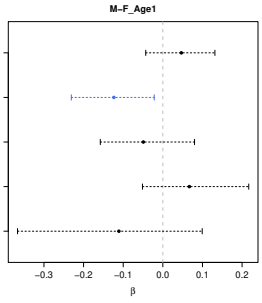
(b) REST1 RL



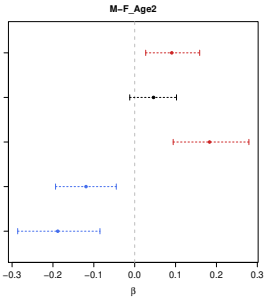
(c) REST1 RL



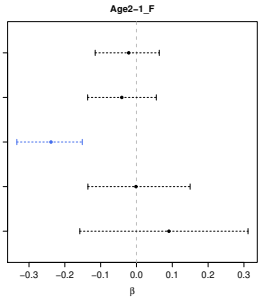
(d) REST1 RL



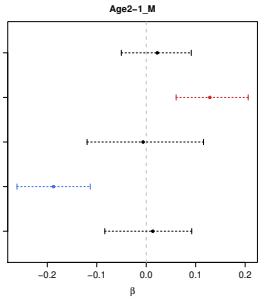
(e) REST2 LR



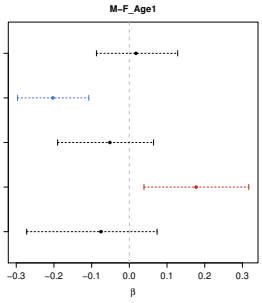
(f) REST2 LR



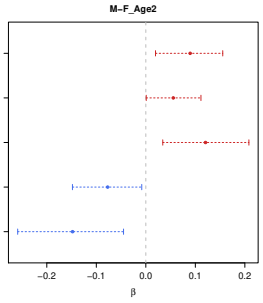
(g) REST2 LR



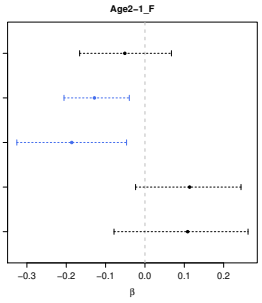
(h) REST2 LR



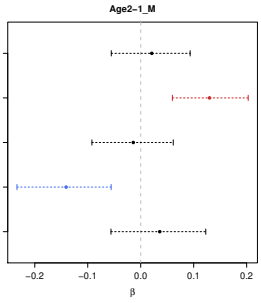
(i) REST2 RL



(j) REST2 RL



(k) REST2 RL



(l) REST2 RL

Figure F.12: Pair-wise comparison of the five identified projection directions from the CAP approach in Section 5 tested on the rest three sessions of resting-state data collected from the same subjects in the HCP study.

References

Flury, B. N. (1986). Asymptotic theory for common principal component analysis. *The annals of Statistics*, pages 418–430.

Krzanowski, W. (1979). Between-groups comparison of principal components. *Journal of the American Statistical Association*, 74(367):703–707.

Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 329–358.

Rao, C. R. (1973). Algebra of vectors and matrices. *Linear Statistical Inference and its Applications: Second Editon*, pages 1–78.