Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Raphiel J. Murden Date

Topics in Data Integration Methods for Neuroimaging and Generalized Additive Mixed Models for Ambulatory Blood Pressure Curves and Psychosocial Stressors

By

Raphiel J. Murden
Doctor of Philosophy

Biostatistics and Bionformatics

Benjamin Risk, Ph.D.
Advisor

Ying Guo, Ph.D.
Committee Member

Deqiang Qiu, Ph.D.
Committee Member

Lance Waller, Ph.D.
Committee Member

Accepted:

Lisa Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Topics in Data Integration Methods for Neuroimaging and Generalized Additive Mixed Models for Ambulatory Blood Pressure Curves and Psychosocial Stressors

By

Raphiel J. Murden
B.A., Morehouse College, GA, 2008
M.A., Washington University, MO, 2011
M.Sc., Emory University, GA, 2018

Advisor: Benjamin Risk, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics and Bionformatics
2021

Abstract

Topics in Data Integration Methods for Neuroimaging and Generalized Additive Mixed Models for Ambulatory Blood Pressure Curves and Psychosocial Stressors
By Raphiel J. Murden

Data integration methods, e.g., Joint and Individual Variation Explained (JIVE), simultaneously explore and analyze similarities between two or more sets of measures captured on the subjects. JIVE estimates shared and unique subspaces, which can be challenging to interpret. Chapter 1 expands upon insights into AJIVE as a canonical correlation analysis of principal component scores. This reformulation, which we call CJIVE, provides an ordering of joint components, uses a computationally efficient permutation test for the number of joint components, and can predict subject scores for out-of-sample observations. Extensive simulations show that AJIVE and CJIVE tend to select the joint rank correctly when true total signal ranks are provided. Using JIVE to integrate functional and structural connectivity from the Human Connectome Project, we find that joint scores from the first of two components are associated with fluid intelligence.

CJIVE only improves interpretation for two datasets. Furthermore, it remains unclear how to interpret JIVE decomposition for a single subject. Chapter 2 proposes Probablistic JIVE (ProJIVE), a model-based method for conducting JIVE analysis. ProJIVE provides a subject-level interpretation of the JIVE framework by modeling subject scores as random effects. Simulation studies show that ProJIVE estimates scores and loadings as well or better than existing methods. We applied ProJIVE to brain morphometry and cognitive/behavioral measures from the Alzheimer's Disease Neuroimaging Initiative (ADNI), which revealed associations between subject scores and Alzheimer's diagnoses. Variable loadings show that measurements of cortical and subcortical volume are strongly related to cognition measures.

Chapter 3 examines the relationship between household financial responsibility and ambulatory blood pressure (ABP) among black women in metro Atlanta. Previous studies of ABP use either a summary measure or inflexible parametric models. However, these approaches may result in the loss of substantial variability or unnecessarily constrain profile shape. Furthermore, ABP profiles are non-linear in time. We use generalized additive mixed models (GAMMs) to estimate ABP profiles for participants who are primarily responsible for earning household finances versus those who are not. GAMMs enable the assessment of periods during which the groups differ significantly, which may lead to interventions to help prevent adverse cardiovascular events.

Topics in Data Integration Methods for Neuroimaging and Generalized Additive Mixed Models for Ambulatory Blood Pressure Curves and Psychosocial Stressors

By

Raphiel J. Murden
B.A., Morehouse College, GA, 2008
M.A., Washington University, MO, 2011
M.Sc., Emory University, GA, 2018

Advisor: Benjamin Risk, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics and Bionformatics
2021

Acknowledgments

Getting to this phase of my life has taken me down a long, winding road of successes and failures. Without the amazing have supported me in the past and continue to support me now there is no way I would be here. I want to make sure I express my deep gratitude to many of those people.

To my dissertation advisor, mentor, and now colleague, Benjamin B. Risk: your patience, kindness, tenacity, and persistence have been a driving force in my academic career since we met. Your curiosity and intellect inspire me. It is a great honor to be your first PhD student, for which I will forever remain grateful. Similarly, to my other committee members, Drs. Ying Guo, Lance Waller, and Deqiang Qiu: your insight, knowledge and ability to provide thoughtful feedback have pushed me to new heights as a statistician. The four of you have helped me navigate some tough times and I deeply appreciate your guidance.

I also want to acknowledge the other faculty and staff I have worked with while here at Emory University. Drs. Renee Moore, Amita Manatunga, Drenna Waldrop (School of Nursing), and Tene Lewis have each made significant contributions to my life and career by mentoring me through various phases of my time as a PhD student. From life and career to writing and interpreting statistical analyses, each of you have been invaluable in helping me arrive to this moment. To the faculty from whom I have taken classes, thank you for your hard work and constructive feedback. I have learned so much from the statistics and other courses here at Emory and very much look forward to continuing to learn from you now as a colleague. While he has only been here less than a year's time, I also want to particularly acknowledge Dr. Robert Krafty. You have encouraged me much more than you know. I appreciate your constant support. Similarly, Mary Abosi, Angela Guinyard, and Joy Hearn have been instrumental in providing support that I have needed to complete this journey. Thank you!

To my mother, father, step-father, grandmothers, and all of my ancestors who have gone on: there is no way I would have ever seen Emory or gotten out of Memphis without your encouragement, love, and support. I often hear people describe me as gracious, thoughtful, and considerate. If any of those things are true, I learned them from modeling the examples that you have provided. I love you deeply and that will never change. Thank you for helping me become the person I am today and the person I will become.

I have the best friends a human could ask for! Marcel, Anthony, Adrian, Andrea and Josh, to name a few. I hope that each of you knows that without you I would be a very different individual. If my light shines at all it is because I reflect the amazing light that my friends shine on me. From study support to telling me when to stop work, each of you have made massive contributions to my life that I will continue trying to repay as long as I live. Thank you!

Last, but certainly not least, thank you to the rest of the Rollins and Biostatistics staff, faculty, and students who I have not named directly. From facilities and waste management to the Master's students I've taught and faculty who I may have only chatted with. I always tell people about how grateful I am to be at a wonderful place like Rollins. We have some of the best people in the world here. Your feedback, support, and conversations have nearly always made me feel warm and welcome. I am grateful to each of you.

Contents

List of Figures

List of Tables

Chapter 1

Canonical JIVE

1.1 Introduction

Modern biomedical and scientific studies often collect multiple datasets in which the number of variables may greatly exceed the number of participants. This phenomenon is especially prevalent in neuroimaging studies, where multiple neuroimaging data types, referred to as modalities, as well as behavioral and demographic data, are often collected [17, 32]. The importance of such multi-dataset studies underscores the urgent need for quantitative methods capable of simultaneous analysis of multi-block datasets, i.e., data integration or multi-view data analysis.

A fundamental goal in neuroimaging is understanding the similarities between structural connectivity (SC) and functional connectivity (FC), where FC can be quantified by cross correlations between brain region time series revealed through functional magnetic resonance imaging (fMRI) and SC by measures of anatomical connections revealed using diffusion-weighted MRI (dMRI) [19]. Studies have reported that brain regions with strong SC demonstrate more reliable functional connections [19, 23], and incorporating SC information leads to more reproducible FC network estimation [18]. However, additional research is needed to elucidate the information

shared between measures of connectivity and the information unique to structural or functional connectivity. Increasing attention has been paid to data integration and data fusion methods [51], which may provide insight into shared structure without imposing a priori spatial constraints.

Statistical approaches to data integration, which seeks to find shared structure across multiple datasets collected on a common set of subjects, date back to the 1930s with canonical correlation analysis (CCA) [20]. Smith et al. [48] used principal component analysis (PCA) and CCA to integrate functional MRI (fMRI) and behavioral data from the Human Connectome Project (HCP). Recently, novel methods that assess the shared structure between datasets have arisen [57, 25], including several which also explore structure unique to each dataset [27, 8, 15, 46, 66].

Joint and Individual Variation Explained (JIVE) has been used in studies to integrate genetic data [36], behavioral and brain imaging data [63], and other applications [27]. Common and orthogonal basis extraction (COBE), which is closely related to JIVE [66], was recently applied to multi-subject resting-state correlation matrices where individual structure was used in connectome fingerprinting [22]. Throughout the remainder of this chapter, we will refer to the JIVE implementation in Lock et al. [27] and the follow-up paper in O'Connell and Lock [36] as R.JIVE. An alternative algorithm and rank-estimation routine for JIVE were recently proposed in Angle-based JIVE (AJIVE) [8]. AJIVE uses matrix perturbation theory (Wedin, 1972) to determine when two similar directions of variation represent noisy estimates of the same direction and proposed a novel non-iterative algorithm that can decrease computational costs. Although there are a number of promising methods for analyzing joint variation, in this paper we focus on R.JIVE and AJIVE, as they both have R-packages and have been applied in the neuroimaging literature [63, 65].

Despite the advancement in statistical methodology, there are limitations that may limit its widespread application. JIVE is formulated as a subspace decomposition,

and the results can be difficult to interpret. For instance, singular value decomposition of joint matrices results in subject scores that differ across datasets. Moreover, the components of the estimated joint subspace have no clear ordering. A related problem is that there is currently no clear method for applying results to new study participants/patients. If JIVE is used for biomarker development, as in Sandri et al. [43], we may want to estimate a subject score for a new patient, which can then be used to classify her or his risk. Additionally, simulation studies examining the accuracy of the rank selection procedures and estimated components are needed to provide guidance to scientific applications.

Our contributions are the following.

- We propose Canonical JIVE (i.e. CJIVE), an adaptation to AJIVE, which improves interpretation of subspaces obtained via JIVE analysis.
- CJIVE also allows prediction of joint scores in new subjects.
- We conduct simulation studies that address important gaps in our understanding of AJIVE versus R.JIVE.
- We apply JIVE to the integration of functional and structural connectivity using a state-of-the-art pipeline applied to 998 subjects from the Human Connectome Project. CJIVE reveals new insights into the shared variation, in particular revealing relationships that go beyond conventional spatial priors.

Section 1.2 describes the statistical methodology employed in AJIVE, R.JIVE, and sCCA, and introduces CJIVE. Section 1.3 conducts simulation studies. Section 1.4 analyzes the HCP data. We discuss our findings and recommendations in Section 1.5.

1.2 Statistical Methodology

1.2.1 JIVE Decomposition

Consider a collection of K data blocks/matrices, $\{\mathbf{X}_k \in \mathbb{R}^{n \times p_k} : k = 1, \ldots, K\}$, where n is the number of subjects and p_k the number of features or variables in the k^{th} dataset. Each data block can be written as $\mathbf{X}_k = \mathbf{G}_k + \mathbf{E}_k$, where $\mathbf{G}_k$ represents the rank-reduced signal (with rank $r_k << \min(n, p_k)$) and $\mathbf{E}_k$ represents full-rank isotropic noise. The JIVE model assumes that each $\mathbf{G}_k$ can be decomposed into a subspace of $\mathbb{R}^n$ that is common across $\mathbf{X}_k$ (the joint subspace) and a subspace that is unique to the k^{th} dataset and orthogonal to the joint subspace (the individual subspaces) [8]. In our presentation, we expand on one of three ways to represent the joint subspace, called the "common normalized score" representation in [8]. We emphasize this representation because it results in a correspondence between the joint components of each dataset, whereas the other representations are arguably less interpretable. The common basis, $\mathbf{Z} \in \mathbb{R}^{n \times r_J}$, is derived from joint analysis of all data blocks, and the other, $\mathbf{B}_k \in \mathbb{R}^{n \times r_{Ik}}$ from the part that remains after joint analysis, where $r_{Ik} = r_k - r_J$. Let $\mathbf{I}_d$ be the $d \times d$ identity matrix and $\mathbf{0}$ a matrix of zeros. Then the JIVE model corresponds to the matrix decomposition

$$\mathbf{X}_k = \mathbf{G}_k + \mathbf{E}_k, \tag{1.1}$$

$$\mathbf{G}_k = \mathbf{Z}\mathbf{W}_{Jk} + \mathbf{B}_k\mathbf{W}_{Ik},$$

$$\mathbf{B}_k^\top \mathbf{Z} = \mathbf{0},\ \mathbf{Z}^\top \mathbf{Z} = \mathbf{I}_{r_J},\ \mathbf{B}_k^\top \mathbf{B}_k = \mathbf{I}_{r_{Ik}}. \tag{1.2}$$

We call $\mathbf{Z}$ joint subject scores and $\mathbf{W}_{Jk}$ joint variable loadings. We also define the joint and individual signal matrices of the k^{th} data block as $\mathbf{J}_k = \mathbf{Z}\mathbf{W}_{Jk}$ and $\mathbf{A}_k = \mathbf{B}_k\mathbf{W}_{Ik}$, respectively, with $\mathbf{G}_k = \mathbf{J}_k + \mathbf{A}_k$. While the model holds for any integer $K > 1$, this study focuses on the case $K = 2$.

In this representation, we do not enforce orthogonality between $\mathbf{B}_k$ and $\mathbf{B}_{k'}$. Later, we propose a permutation test for the joint rank, r_J, that determines when the correlation between signal is sufficiently large to be deemed joint, but allows insignificant correlation between individual subject scores. Our proposed approach will also result in an intuitive ordering of components by the strength of evidence that they are joint. Also note that in (1.1), the rows of the loadings matrices $\mathbf{W}_{Jk}$ are not orthogonal.

For the HCP network data that we examine in section 1.4, we can translate each row of the score $(\mathbf{Z}, \mathbf{B}_k)$ matrix into a low-dimensional vector summary of a participant's k^{th} network data (e.g., FC). The joint scores $\mathbf{Z}$ surmise information that is common across modalities, while $\mathbf{B}_k$ comprises information unique to an individual modality. The l^{th} row of the loading matrix $\mathbf{W}_{Jk}$ exhibits the magnitude with which network edges contribute to the l^{th} column of the summary scores in $\mathbf{Z}$. In section 1.4.4 we examine variable loadings to develop insight into latent structures which are common within both modalities and those which are unique to each. For instance, section 1.4.3 shows that CJIVE joint scores are more strongly associated with a measure of fluid intelligence than individual scores.

R.JIVE Estimation

R.JIVE uses an iterative algorithm that simultaneously estimates signal matrices as well as their ranks. Assume each dataset is column-centered and scaled by its Frobenius norm. Each iteration involves two steps: 1) estimating joint and individual signal ranks, and 2) estimating signal matrices using the ranks from step 1. The first iteration assumes that the joint signal matrix has the same column space as a matrix formed by concatenating the data matrices. This procedure is iterated until convergence; details are in the Web Appendix A.1.1. Two methods for choosing the joint rank were proposed: a permutation test and Bayesian Information Criterion (BIC). In the default R.JIVE implementation, the individual subspaces are orthogonal

O'Connell and Lock [36].

AJIVE Estimation

In AJIVE, the joint rank r_J is determined using principal-angle analysis (PAA) and requires user-specified signal ranks $r_1 = r_J + r_{I1}$ and $r_2 = r_J + r_{I2}$. The main idea is to investigate when basis vectors in the signal subspaces should be considered 'noisy' estimates of the same direction. This problem can be translated into finding the singular values of the concatenated signal bases that exceed a given threshold.

For the remainder of this paper, we standardize the columns of $\mathbf{X}_1$ and $\mathbf{X}_2$ to have mean zero and variances equal to one, as commonly done in PCA.

First, the user specifies the ranks used in PCA of $\mathbf{X}_1$ and $\mathbf{X}_2$. Let $\widetilde{\mathbf{U}}_1$ and $\widetilde{\mathbf{U}}_2$ denote the r_1 and r_2 left singular vectors of $\mathbf{X}_1$ and $\mathbf{X}_2$. Define $\mathbf{C} = [\widetilde{\mathbf{U}}_1, \widetilde{\mathbf{U}}_2]$. Let $\mathbf{U_C}$ denote the left singular vectors of $\mathbf{C}$. Feng et al. [8] develop two bounds to determine whether the jth column of $\mathbf{U_C}$ represents a joint direction of variance. These bounds are based on the principal angles between $\mathbf{U}_1$ and $\mathbf{U}_2$, which can be extracted from the singular values of $\mathbf{C}$. The first bound is based on Wedin's theorem and several corollaries thereof, as discussed in the Web Appendix A.1.2. The second is a random direction threshold based on the principal angles between simulated noise subspaces.

1.2.2 Using CCA to Interpret JIVE: CJIVE

Equivalence of estimators

We review CCA and describe how it relates to the AJIVE algorithm. Given standardized data matrices $\mathbf{X}_1$, $\mathbf{X}_2$, and a number of joint components r_J, CCA aims to solve

$$\begin{gathered} \underset{\omega_{1j}\in\mathbb{R}^{p_1},\omega_{2j}\in\mathbb{R}^{p_2}}{\arg\max}\ \omega_{1j}^\top \mathbf{X}_1^\top \mathbf{X}_2 \omega_{2j},\ j = 1 \dots r_J, \\ \text{subject to } ||\omega_{kj}|| = 1 \text{ and } \omega_{kj}^\top \omega_{kj'} = 0,\ k = 1, 2,\ j \neq j'. \end{gathered} \tag{1.3}$$

The solutions to (1.3), which we denote as $\hat{\omega}_{1j}$ and $\hat{\omega}_{2j}$, are given by the left and right singular vectors of $\mathbf{X}_1^\top \mathbf{X}_2$, which are unique up to a change in sign [20]. Additionally, $\rho_j = \frac{1}{n}\hat{\omega}_{1j}\mathbf{X}_1^\top \mathbf{X}_2 \hat{\omega}_{2j}$ is the jth canonical correlation.

Classic CCA can not be applied to $p_k > n$. Sparse CCA is one alternative [57], and it turns out JIVE is a reduced-rank alternative. [8] show that the j^{th} joint subject score from AJIVE is equivalent to the average of the j^{th} canonical variables of the CCA of the scores from the separate PCAs, up to scaling. Our theorem, below, formalizes their finding. A proof is provided in the Web Appendix A.3.

Theorem 1.2.1. *Let the columns of $\widetilde{\mathbf{U}}_1$ and $\widetilde{\mathbf{U}}_2$ represent orthonormal bases for the signal matrices $\widehat{\mathbf{G}}_1$ and $\widehat{\mathbf{G}}_2$. Let $\hat{\mathbf{z}}_j$ be the j^{th} joint subject score from AJIVE analysis. Let $\hat{\omega}_{1j} \in \mathbb{R}^{r_1}$ and $\hat{\omega}_{2j} \in \mathbb{R}^{r_2}$ represent the canonical vectors from the CCA of $\widetilde{\mathbf{U}}_1^\top \widetilde{\mathbf{U}}_2$. Let σ_{Cj} denote the jth singular value of $[\widetilde{\mathbf{U}}_1, \widetilde{\mathbf{U}}_2]$. Then*

$$\hat{\mathbf{z}}_j = \frac{1}{\sqrt{2}\sigma_{Cj}}(\widetilde{\mathbf{U}}_1\hat{\omega}_{1j} + \widetilde{\mathbf{U}}_2\hat{\omega}_{2j}).$$

Additionally, the canonical correlation $\rho_j = \sigma_{Cj}^2 - 1$.

In summary, the j^{th} joint scores from AJIVE are equivalent to a scaled average of the j^{th} canonical variables of the principal component scores. This perspective is illustrated in Figure 1.1, and we define CJIVE (CCA JIVE) in the next section.

CJIVE: ordering, permutation test, and unique components

The CCA perspective on the signal subspaces provides a useful way to interpret the joint components. We view the canonical correlations defined in Theorem 1.2.1 as a measure of the strength of the corresponding joint component, which provides an ordering.

This motivates the use of a permutation test of the canonical correlations of the PCs. For $b = 1, \ldots, n_{perms}$, let $\widetilde{\mathbf{U}}_2^{(b)}$ represent a copy of $\widetilde{\mathbf{U}}_2$ with the rows permuted

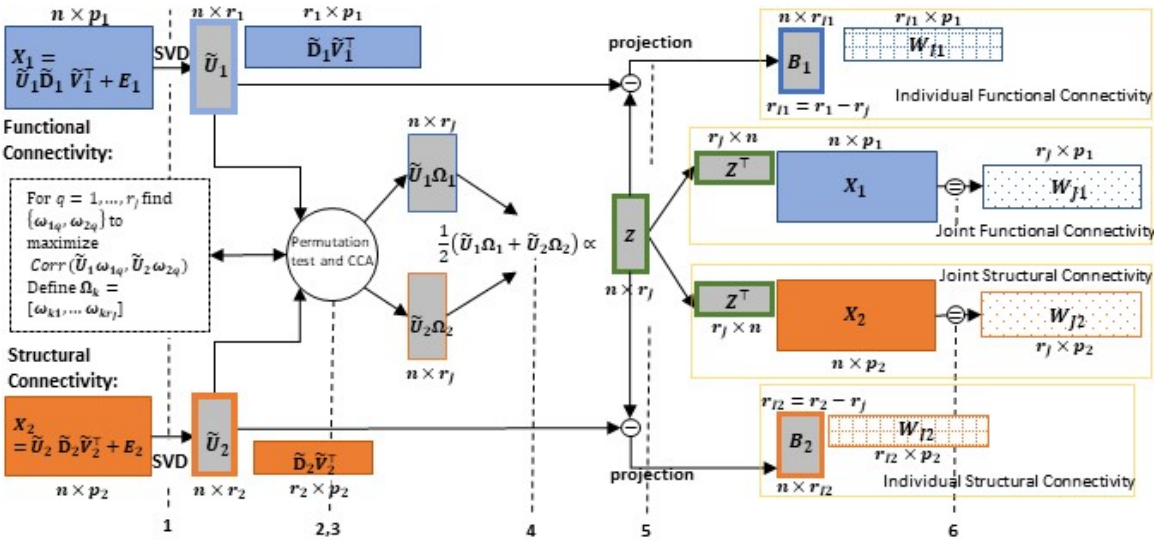


Figure 1.1: Schematic of the CJIVE decomposition for obtaining joint subject scores and loadings. Quantities specific to $\mathbf{X}_1$ are shown in blue; those specific to $\mathbf{X}_2$, orange. Gray boxes illustrate scores, with a green outline for joint scores. Checked and dotted boxes represent loadings. Steps are outlined in section 1.2.2.

so that they no longer represent the same ordering of participants as in $\widetilde{\mathbf{U}}_1$. We then obtain the null distribution of the canonical correlations from the max of the singular values of $\widetilde{\mathbf{U}}_1^\top \widetilde{\mathbf{U}}_2^{(b)}$, $b = 1, \ldots, n_{perms}$. For each component, we calculate a p-value as the proportion of maximal null correlations which exceed that component's canonical correlation. We then calculate r_J for a specified α-level. Once we have estimated r_J via the permutation test, we calculate joint scores using the results of Theorem 1.2.1 and estimate the signal matrices $\mathbf{J}_k$ and $\mathbf{A}_k$ using the same procedure in AJIVE.

Here,we summarize the CJIVE procedure depicted in Figure 1.1.

CJIVE Procedure

1. For $k = 1, 2$ conduct PCA of $\mathbf{X}_k$ and determine total rank r_k by examining the scree plot. Obtain PC scores $\widetilde{\mathbf{U}}_k$.

2. Calculate canonical correlations: $\rho_j,\ j = 1 \ldots min(r_1, r_2)$, as in Theorem 1.2.1; use these to order joint components.

3. Use a permutation test to determine which canonical correlations are significant.

4. Calculate joint scores $\mathbf{Z}$ as in Theorem 1.2.1.

5. Project data onto the orthogonal complement of the joint subspace to obtain individual signal matrices $\mathbf{A}_K = \mathbf{B}_k \mathbf{W}_{Ik}$.

6. Calculate loadings for joint structure and visualize.

7. Examine the scores for structure, e.g., associations with exogenous variables.

8. Calculate the variance explained for both joint and individual components, which provide insight into the importance of joint and individual sources of variation.

CJIVE provides a unique decomposition of $\hat{\mathbf{J}}_1$ and $\hat{\mathbf{J}}_2$ (up to sign) when the canonical correlations differ across components, as expected to occur in data. In the JIVE model given by (1.1), it is assumed that the subject score subspaces are equivalent. Then, the components are not unique. To see this, let $\mathbf{Z} \in \mathbb{R}^{n \times r_J}$ denote the joint scores. Let $\mathbf{O}$ denote any orthogonal matrix of dimensions $r_J \times r_J$. Then $\mathbf{J}_1 = \mathbf{ZOO}^\top \mathbf{W}_{J1}$ and $\mathbf{J}_2 = \mathbf{ZOO}^\top \mathbf{W}_{J2}$. Consequently, the basis $\mathbf{ZO}$ also provides a set of joint subject scores.

Although we focus on the case of $K = 2$ datasets here, CJIVE also generalizes to $K > 2$, where the size of the singular values denotes the amount of information shared between the datasets. For $K > 2$, the interpretation becomes more nuanced. For example, a joint component with canonical correlation equal to one between two datasets but zero with a third dataset (i.e. partially shared structure), could be a larger source of joint variation than a component which is weakly correlated across the three datasets. Partially shared structure is modeled in [15]. We note that examining the variance explained by each component provides insight into partially shared structure, but additional investigation lies beyond the scope of this study.

Predicting joint scores in new participants

An important problem is how to apply the results from JIVE analysis to a new participant. For example, if JIVE is used for biomarker development, we may want to estimate a subject score for a patient, which can then be used to classify their risk.

One straightforward way of using JIVE to predict new joint scores is to regress each new pair of observations onto the generalized inverse of joint loadings to obtain block-specific joint scores and then compute their average. Let $\widehat{\mathbf{W}}_{Jk}$, $k = 1, 2$, represent joint loadings from applying JIVE on the data blocks $\mathbf{X}_1$ and $\mathbf{X}_2$. Let $\mathbf{x}_i \in \mathbb{R}^{p_1}$ and $\mathbf{y}_i \in \mathbb{R}^{p_2}$ be data for a new participant. Then define predicted joint scores as

$$\hat{\mathbf{z}}_i^\top = (\mathbf{x}_{i1}^\top \widehat{\mathbf{W}}_{J1}^- + \mathbf{x}_{i2}^\top \widehat{\mathbf{W}}_{J2}^-)/2,$$

where $\widehat{\mathbf{W}}_{Jk}^-$ represents the g-inverse of $\widehat{\mathbf{W}}_{Jk}$.

An alternative approach is based on the canonical variables given in Theorem 1.2.1. First, we predict the PC scores for a new subject; second, we estimate the canonical variables of the PC scores from each dataset; third, we sum the canonical variables and normalize to length one. Recall the rank r_1 and r_2 approximations to $\mathbf{X}_1$ and $\mathbf{X}_2$: $\widehat{\mathbf{G}}_1 = \widetilde{\mathbf{U}}_1\widetilde{\mathbf{D}}_1\widetilde{\mathbf{V}}_1^\top$, $\widehat{\mathbf{G}}_2 = \widetilde{\mathbf{U}}_2\widetilde{\mathbf{D}}_2\widetilde{\mathbf{V}}_2^\top$. Using CCA on $\widetilde{\mathbf{U}}_1$ and $\widetilde{\mathbf{U}}_2$ yields a matrix of canonical vectors: $\widehat{\mathbf{\Omega}}_1 = [\hat{\omega}_{1j}, \ldots, \hat{\omega}_{1r_J}]$ and $\widehat{\mathbf{\Omega}}_2 = [\hat{\omega}_{2j}, \ldots, \hat{\omega}_{2r_J}]$. The predicted estimate for each canonical variable is given by $\hat{\mathbf{c}}_{1i} = \mathbf{x}_i^\top \widetilde{\mathbf{V}}_1 \widetilde{\mathbf{D}}_1^{-1} \widehat{\mathbf{\Omega}}_1$ and $\hat{\mathbf{c}}_{2i} = \mathbf{y}_i^\top \widetilde{\mathbf{V}}_2 \widetilde{\mathbf{D}}_2^{-1} \widehat{\mathbf{\Omega}}_{2j}$. Then the j^{th} joint score is

$$\hat{z}_{ij} = \frac{\hat{c}_{1ij} + \hat{c}_{2ij}}{\sqrt{2(1+\rho_j)}},$$

for $j = 1, \ldots, r_J$. We apply and evaluate the proposed method in both the simulation study of section 1.3 and analysis of the HCP data, section 1.4.

1.3 Simulation study

1.3.1 Simulations comparing JIVE methods

We conduct simulation studies to address the following gaps in the current understanding of the performance of R.JIVE and AJIVE: 1) accuracy when the joint signal strength is low versus high; 2) rank selection when the number of joint components is greater than 1; and 3) the impact of the initial signal rank selection on joint rank selection. We use a full factorial design with the following factors:

1. The number of features in $\mathbf{X}_2$: with levels (a) $p_2 = 200$ and (b) $p_2 = 10000$,
2. Joint Variation Explained in $\mathbf{X}_1$: with levels (a) $R^2_{J1} = 0.05$ and (b) $R^2_{J1} = 0.5$,
3. Joint Variation Explained in $\mathbf{X}_2$: with levels (a) $R^2_{J2} = 0.05$ and (b) $R^2_{J2} = 0.5$.

The joint rank was 3 in all settings. The entries of the error matrices $\mathbf{E}_1$ and $\mathbf{E}_2$ were randomly drawn from a standard Gaussian distribution. The number of features in $\mathbf{X}_1$ and the individual variation explained for both data blocks were held constant at $p_1 = 200$ and $R^2_{I1} = R^2_{I2} = 0.25$, respectively.

Experimental factor (b) (i.e., p_2) allows us to assess the impact of p_k on the accuracy of r_J estimates. Factors three and four (i.e., R^2_{J1} and R^2_{J2}) allow us to examine the impact of the joint signal's magnitude within each dataset.

The joint and individual signals, defined as $\mathbf{J}_k = \mathbf{Z}\mathbf{W}_{Jk}$ and $\mathbf{A}_k = \mathbf{B}_k\mathbf{W}_{Ik}$, were constructed by generating score matrices ($\mathbf{Z}$, $\mathbf{B}_k$) and loading matrices ($\mathbf{W}_{Jk}, \mathbf{W}_{Ik}$) in the following manner. For each simulation, the subject score matrix $[\mathbf{Z}, \mathbf{B}_1, \mathbf{B}_2]$ was drawn from a Bernoulli distribution, with probability 0.2 for $\mathbf{Z}$ and 0.4 for $\mathbf{B}_k$. The use of two values is similar to the toy examples from [8], which used ± 1. Next, we defined diagonal matrices $\mathbf{L}_{Jk} = \text{diag}(r_J, \ldots, 1)$ and $\mathbf{L}_{Ik} = \text{diag}(r_{Ik}, \ldots, 1)$. Then we defined $\mathbf{M}_{Jk}$ and $\mathbf{M}_{Ik}$ with entries from independent standard multivariate Gaussian distributions. Then we initially set $\mathbf{W}_{Jk} = \mathbf{L}_{Jk}\mathbf{M}_{Jk}$, $\mathbf{W}_{Ik} = \mathbf{L}_{Ik}\mathbf{M}_{Ik}$. Note that this

set-up results in approximately orthogonal $\mathbf{A}_1$ and $\mathbf{A}_2$. In R.JIVE, we use the option enforcing this orthogonality. This set-up favors the rank-selection procedure in AJIVE since principal angles between $\mathbf{A}_1$ and $\mathbf{A}_2$ are large and corresponding singular values are unlikely to exceed the Wedin and random bounds described in 1.2.1.

In order to achieve the desired values of R^2_{Jk} and R^2_{Ik}, we rescale the joint and individual matrices such that $\mathbf{X}_k = d_k\mathbf{J}_K + c_k\mathbf{A}_k + \mathbf{E}_k$ for appropriate constants c_k and d_k. R^2_{Jk} and R^2_{Ik} can be expressed as equations which are quadratic in c_k and d_k, as described in Web Appendix B. We approximated solutions for c_k and d_k numerically.

The chordal subspace norm is a distance metric for linear subspaces that has been generalized to matrices, say $\mathbf{F}_1, \mathbf{F}_2$, of possibly different ranks [62] and can be calculated as

$$\delta(\mathbf{F}_1, \mathbf{F}_2) = \sqrt{\sum_{m=1}^{q} \sin^2 \theta_m}, \tag{1.4}$$

where $q = \min_k(\text{rank}(\mathbf{F}_k))$ and θ_m are the principal angles between the column space of $\mathbf{F}_1$ and $\mathbf{F}_2$. We use this metric in our simulation studies to describe the accuracy of JIVE estimates. Note when the column space of $\mathbf{F}_1$ is contained in the column space of $\mathbf{F}_2$, $\delta(\mathbf{F}_1, \mathbf{F}_2) = 0$. Therefore comparing results from different methods requires examination of rank estimates and subspace estimates.

We performed 100 simulations using three methods: (1) R.JIVE, with its permutation based algorithm for choosing ranks; (2) AJIVE-Oracle, where we used the true number of components r_k (joint rank + individual rank) as input; and (3) AJIVE-Over, where the total number of components was chosen to retain 95% of the variance. We also defined CJIVE-Oracle and CJIVE-Over using the same approach for total signal ranks and selecting the joint rank using our permutation test with $n_{perms} = 500$ and $\alpha = 0.05$.

Results from different methodologies are not all directly comparable. Note that AJIVE and CJIVE both return a single joint subject score per observation for each

component found in the joint subspace. On the other hand, R.JIVE returns signal matrices, from which we can derive block-specific joint scores. To compare R.JIVE results to those from AJIVE/CJIVE, we apply SVD to each joint signal matrix estimated via R.JIVE using the selected joint rank, concatenate left singular values, apply SVD and then use the first $\hat{r}_J$ left singular vectors as estimates of subject scores.

To investigate the prediction methods outlined in section 1.2.2, each pair of replicate datasets was randomly divided into a pair of training datasets and a pair of test datasets, both with sample sizes $n/2 = 100$. AJIVE-Oracle and CJIVE-Oracle were applied on the pair of training datasets. Subject scores were predicted for "new subjects", represented by the test datasets. We then assessed performance by calculating the Pearson correlation coefficient between predicted joint scores for the test datasets and true joint scores for the same datasets for each of the r_J joint score components.

1.3.2 Simulation Results

Figures 1.2(a) and (b) show that CJIVE-Oracle and AJIVE-Oracle chose the correct joint rank in nearly 100% of simulations in all settings except for the low-signal lower-dimensional case, in which AJIVE-Oracle selected the correct rank more frequently than CJIVE-Oracle (approximately 85% versus 75% of the simulations). AJIVE-Over and CJIVE-Over both routinely underestimated the number of joint components in all scenarios except the high joint variation with lower dimensional matrices ($R^2_{J1} = R^2_{J2} = 0.5,\ p_1 = p_2 = 200$). When an estimate of r_k is very large, the correlation between permuted datasets can be very large, such that zero joint components are significant. The joint rank estimated in R.JIVE tends equal 2 when the joint signal in both datasets is relatively large (bottom-right panels in both sub-figures of Figure 1.2: $R^2_{J1} = R^2_{J2} = 0.5$), while it is mostly 0 or 1 in the other scenarios.

Figure 1.3 shows that the chordal distances between true score subspaces and

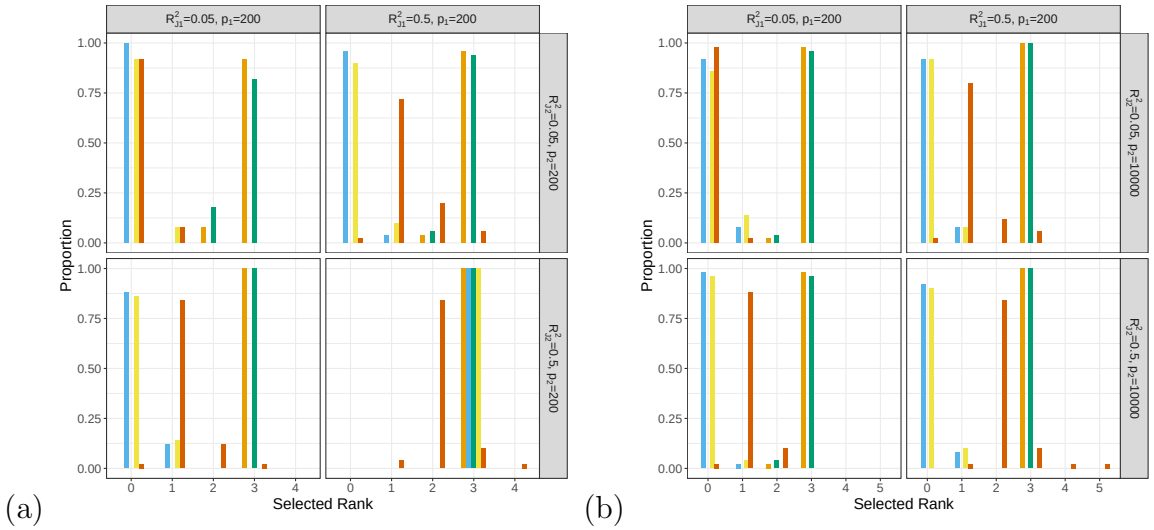


Figure 1.2: Results of simulation studies: (a) $p_2 = 200$, (b) $p_2 = 10000$. Sub-figure show proportions of values chosen as rank of the joint signal space for each method and combination of simulation settings. True rank equals 3 in all simulations. Color key: Orange = AJIVE-Oracle. Light Blue = AJIVE-Over. Green = CJIVE-Oracle. Yellow = CJIVE-Over. Red = R.JIVE.

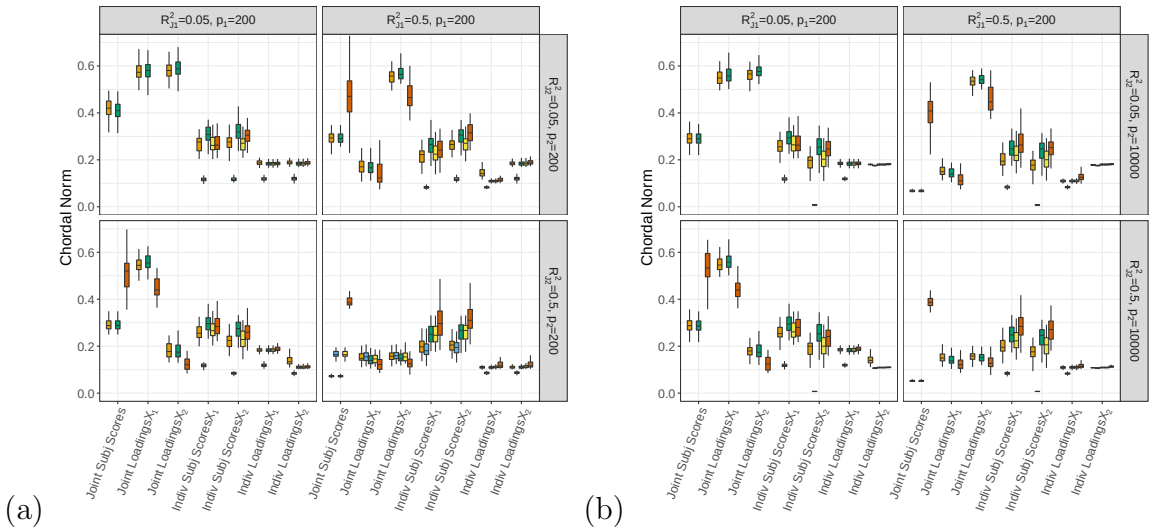
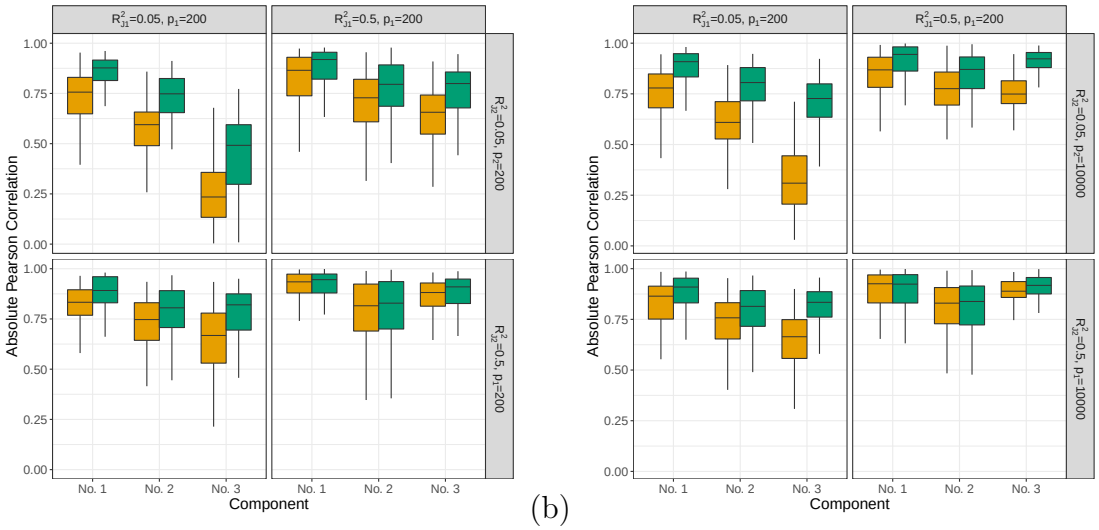


Figure 1.3: Results of simulation studies: (a) $p_2 = 200$, (b) $p_2 = 10000$. Each sub-figure exhibits boxplots of chordal norms for each of the post-JIVE measurements described in section 1.2.1. Methods that are not shown had median chordal norms of 1. Color key: Orange = AJIVE-Oracle. Light Blue = AJIVE-Over. Green = CJIVE-Oracle. Yellow = CJIVE-Over. Red = R.JIVE.

their estimates. CJIVE-Oracle and AJIVE-Oracle score subspaces trended less than the same distances for R.JIVE, CJIVE-Over and AJIVE-Over in all settings. In AJIVE-Over, all detected signal components are allocated to the individual subspace since the estimated joint rank is almost always 0. Since the true individual signal components are likely to lie mostly within the over-estimated individual signal subspaces, their chordal distance is small. Similarly, although the chordal distances for loading subspaces from R.JIVE trended less than those from AJIVE-Oracle, the lack of accurate joint rank estimates from R.JIVE may indicate that estimated subspaces partially lie within true subspaces.

To summarize, we find that CJIVE-Oracle and AJIVE-Oracle choose the joint rank correctly in most simulations. For both CJIVE-Over and AJIVE-Over, including too many initial signal components results in a noise-contaminated signal for each data matrix, which increases the chance of finding angles between noise components that are near zero and thus results in too few joint components or none at all.



(a) (b)

Figure 1.4: Results of simulation studies: (a) $p_2 = 200$, (b) $p_2 = 10000$. Boxplots of absolute Pearson correlations between predicted joint scores and true joint scores in simulation study for settings where $r_J = 3$. Color key: Orange = AJIVE-Oracle. Green = CJIVE-Oracle

Lastly, Figure 1.4 exhibits that out-of-sample subject scores can be predicted more

accurately across joint components using the CJIVE method when compared to the straightforward method using results of AJIVE. The Pearson correlation coefficients tend to be close to 1, on average, for the first joint component of subject scores across all simulation settings with. Neither method does very well at predicting joint scores for the third component when the joint signal is small, i.e. $R^2_{J1} = R^2_{J2} = 0.05$.. Recall data were simulated so that the proportion of variance attributable to the j^{th} joint component in $\mathbf{X}_k$, $k = 1, 2$, $j = 1, 2, 3$ is given by $R^2_{Jk}\left(\frac{3-(j-1)}{6}\right)$. Therefore components are ordered (from highest to lowest) by the proportion of joint variation that they contribute, which may contribute to the poor prediction at higher components.

1.4 Joint Analysis of Structural and Functional Connectivity in HCP Data

1.4.1 Human Connectome Project

Our data application uses measures of FC and SC from $n = 998$ study participants (532 females) in the young adult Human Connectome Project (HCP). Web Appendix Table 4 provides demographics. We applied R.JIVE, AJIVE, CJIVE, and sCCA to examine multivariate relationships across brain networks as measured by Fisher z-transformed correlations from rs-fMRI (FC) and log-transformed streamline counts from dMRI (SC).

Data preprocessing

HCP rs-fMRI data comprise two left-right phase encoded and two right-left phase encoded 15-minute eyes-open rs-fMRI runs [17]. Each run used 2-mm isotropic voxels with 0.72s repetition time. For each run, we calculated the average time series for

each of the 68 cortical regions of interest (ROIs) from Desikan et al. [3] plus the 19 subcortical gray-matter ROIs from Glasser et al. [17]. For each participant and pair of ROIs, the Pearson correlation was calculated, Fisher z-transformed, and then averaged across the four runs. The lower diagonal of each subject's connectivity matrix was vectorized, resulting in $p_1 = 3{,}741$.

For each HCP participant, three left-right and three right-left phase-encoded runs of dMRI from three shells of $b = 1000, 2000$ and 3000 s/mm^2 with 90 directions and 6 b_0 acquisitions interspersed throughout were acquired [17]. Whole-brain tractography for each participant was conducted using probabilistic tractography as detailed in Zhang et al. [64]. On average, around 10^5 voxels occurring along the white matter/gray matter interface were identified as seeding regions for each participant. Sixteen streamlines were initiated for each seeding voxel, resulting in approximately 10^6 streamlines for each participant. Nodes of the SC networks were defined from the same ROIs as the rs-fMRI. Edges were represented by the number of viable streamlines between ROIs, with viability determined by three procedures: (1) each gray matter ROI is dilated to include a small portion of white matter region; (2) streamlines connecting multiple ROIs were cut into pieces such that no streamlines pass through ROIs; and (3) apparent outliers were removed. Finally, edges where at least 99% of subjects had zero streamlines were removed, and the remaining streamline counts were log transformed. There were $p_2 = 3{,}330$ edges in the resultant SC data matrix.

1.4.2 Dimension Selection and Joint and Individual Variation Explained

Three methods were employed to choose total signal ranks: 1) visually determining the elbow in eigenvalue scree plots, 2) 95% variance, and 3) R.JIVE permutation tests. Joint ranks were also chosen using three methods: 1) AJIVE, 2) the permutation test

Table 1.1: Joint, Total Signal Ranks Chosen and Joint, Individual Variation Explained in the dMRI data (Streamline Counts) and functional connectivity data (Pearson correlations) from the HCP.

Method		**Chosen Rank**			**Variation Explained**			
Jnt	Ttl	Jnt	Ttl FC	Ttl SC	Jnt FC	Ind FC	Jnt SC	Ind SC
AJIVE	Scree plot	2	7	10	0.113	0.499	0.032	0.216
	95% Var.	1	225	683	0.005	0.945	0.002	0.948
CJIVE	Scree plot	2	7	10	0.113	0.499	0.032	0.216
	95%	0	225	683	0	0.950	0	0.950
R.JIVE	Perm.	1	54	98	0.042	0.794	0.012	0.507
	Scree Plot	2	7	10	0.074	0.569	0.012	0.224

in R.JIVE and 3) CJIVE. In sCCA, permutations tests resulted in sparsity parameters equal to 0.1 using the *PMA* R package [57].

The total ranks estimated from the three methods are in Table (1). Both AJIVE and CJIVE with the scree-plot method estimated 2 joint components, which implies that results from these methods are equivalent. Similar to results of our simulation study, AJIVE estimated 1 and CJIVE estimated 0 when each was combined with the 95% variation method. R.JIVE estimated 1 joint component. Guided by these results, we also estimated two pairs of canonical variables with sCCA. Lastly, CJIVE-Scree plot ranks were used as input for R.JIVE and vice versa.

The canonical correlations were $\rho_1 = 0.31$ and $\rho_2 = 0.21$ using 1000 permutations. The proportion of variation attributable to joint component 1 was 0.094 in FC and 0.017 in SC (Table 1). For component 2, the values were 0.018 and 0.015, respectively.

1.4.3 Subject scores

Note subject scores are equal in AJIVE-Scree plot and CJIVE-Scree plot because they selected the same ranks; hereafter, we refer to these results as CJIVE-Scree plot. In order to compare results from sCCA to CJIVE, we averaged canonical variables across datasets to obtain a single subject score vector for each joint component. Next, joint subject scores from CJIVE, R.JIVE, and sCCA, and individual scores from CJIVE

Table 1.2: Multiple regression of gF onto joint subject scores estimated with sCCA, AJIVE, R.JIVE, sCCA. Here, AJIVE and CJIVE are equivalent as both methods selected two joint components.

	Partial Correlation Coefficients			
	Joint Signal	Indiv FC Signal	Indiv SC Signal	Total Signal
CJIVE-Scree plot	0.251	0.091	0.080	0.278
R.JIVE	0.210	0.364	0.369	0.559
sCCA	0.200	–	–	0.200

and R.JIVE were examined for associations with fluid intelligence (gF), measured in the HCP as the number of correct responses to the Penn Progressive Matrices Test.

Among the joint scores, CJIVE-Scree plot resulted in the highest partial correlation coefficient. Partial correlation coefficients for individual scores and total scores (joint + individual) were highest in R.JIVE (Table 2). However, R.JIVE contained a total of 151 components while CJIVE included 15 components. Moreover, in all three methods, only the first joint component and no individual components were significantly associated with gF (CJIVE: $p < 10^{-13}$, Bonferroni-$\alpha \approx 0.003$; R.JIVE: $p < 10^{-9}, \alpha \approx 0.0003$; sparse-CCA: $p < 10^{-4}, \alpha = 0.025$).

1.4.4 Variable Loadings

Since edges from FC and SC networks comprise the features in our input data blocks, loadings are imposed onto symmetric matrices. The sign indeterminacy of the joint loadings for each component was chosen to result in positive skewness. In Figure 1.5a, we see that there were strong positive loadings throughout the FC. Overall, there was not clear spatial correspondence between FC and SC, and the correlation between loadings was -0.04. Instead, overall higher FC was associated with higher SC in many regions, particularly subcortical and frontal, with SC loadings in the opposite direction in certain connections between occipital, parietal, temporal, and subcortical.

Taking the L1 norm of each row within each loading matrix reduces the num-

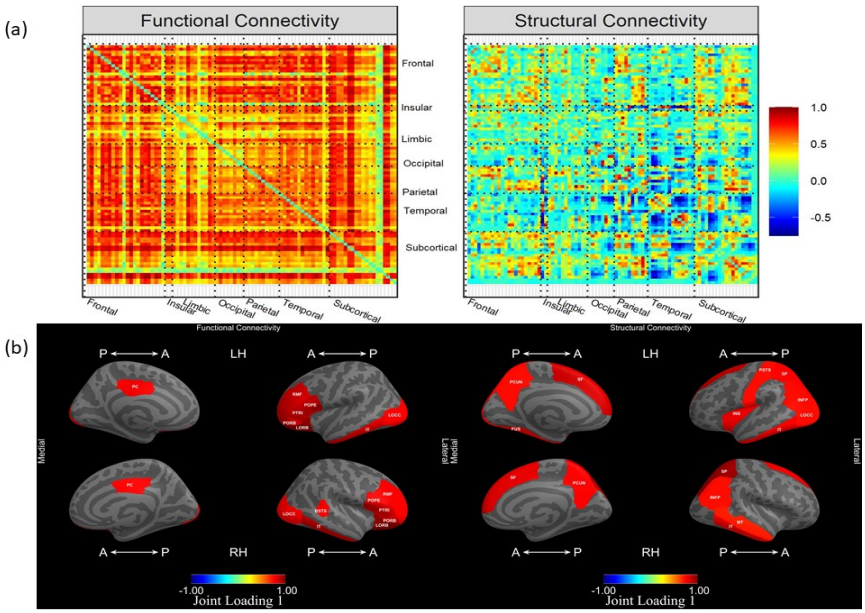


Figure 1.5: (a) Variable loadings for the first component of the joint signal space estimated by CJIVE and displayed on heatmaps. Sub-figure (b) displays the top 25^{th} percent of L1 norms of the variable loadings related to each cortical ROI for joint component 1. L1 norm for an ROI equals the sum of the absolute values of the rows of (a), excluding subcortical regions.

ber of features to the number of nodes, which provides a more detailed examination of the patterns. In this analysis, we are particularly interested L1 norms that are large in both the left and right hemispheres, which suggests the loadings are capturing meaningful biological structure. In the FC loadings, Figure 1.5b shows that the most prominent cortical regions in the first joint component correspond to ROIs from the frontal, occipital and temporal lobes, with extensive left-right hemispheric correspondence. In the SC loadings, we again see left-right hemispheric correspondence, this time in the parietal and temporal lobes, as well as regions that did not exhibit hemispheric correspondence. L1 norms of subcortical regions (not shown) were large in the left and right accumbens, left caudate, and left putamen in both modalities. Additionally, the right putamen and right caudate were prominent in FC, while both left and right hippocampus were prominent in SC.

1.4.5 Reproducibility and prediction of new subjects

Subjects from the HCP data were split into two sets to examine the reproducibility of our results. We will refer to the first sub-sample as 'sample A' and the second as 'sample B'. CJIVE with total signal ranks from scree plots (see Table 1) found $r_J = 1$ for both samples, while AJIVE found $r_J = 2$ for sample A and $r_J = 1$ for sample B. The correlations between the joint loadings from sample A and B were equal to 0.61 for FC and 0.65 for SC. When a second joint component was estimated, as in analysis of the full sample, the correlation of the FC loadings was 0.29 and the SC loadings was 0.38.

CJIVE canonical vectors from sample A were used to predict joint subject scores for sample B. We then compared the predicted joint scores to those from the CJIVE analysis of sample B (Figure 1.6). Pearson correlations between sample A subject scores and predicted sample B subject scores were 0.68 and 0.20, for components 1 and 2, respectively. We used a permutation test to examine whether the observed correlations were significantly different from 0. P-values were 0 for both components using 10,000 permutations. Similar results were achieved when CJIVE canonical scores from sample B data were used to predict subject scores for sample A. In our simulation study, the joint subspace components with higher indices had reduced predictive power (Figure 1.4).

1.5 Discussion

We propose CJIVE, an adaptation to AJIVE which improves interpretation: 1) the joint scores are an average of the canonical variables of the principal component scores of each dataset; 2) joint scores are ordered by canonical correlations; 3) p-values from permutation tests indicate the significance of each joint component; 4) the proportion of variance explained for each of the joint and individual components complements

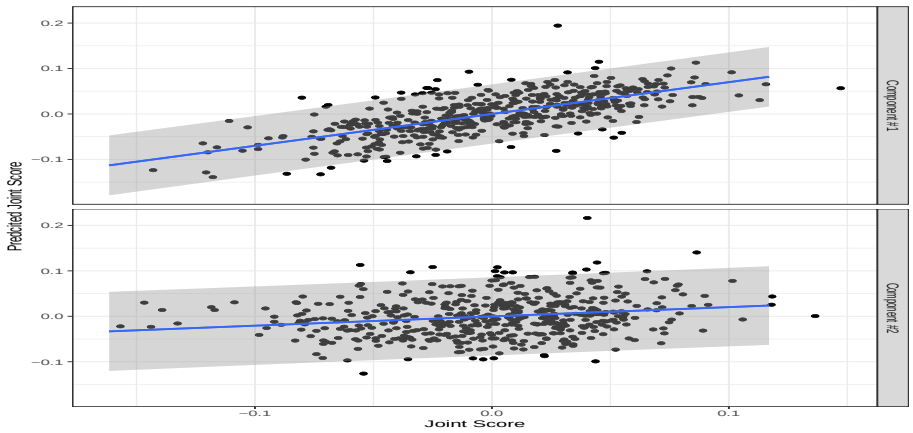


Figure 1.6: Joint subject scores predicted for sample B using CJIVE of sample A versus joint subject scores estimated from the full CJIVE analysis of sample B with prediction interval. Note only one joint component was selected in sample A.

this information. Simulation study results indicate that when total signal ranks are accurately estimated, AJIVE/CJIVE generally choose a more appropriate number of joint components and provide more accurate estimates of the subspaces of interest compared to R.JIVE.

We apply CJIVE to obtain novel insight into the relationship between structural and functional connectivity. Interestingly, we did not find much spatial correspondence between prominent communities in FC and those in SC. However, the biological relevance of subject scores was revealed by their association with gF, and reproducibility through the data splitting and prediction of the joint scores. Recent studies suggest that the correlation between the weighted edges in FC and SC is roughly 0.20 [26], which is much lower than a landmark study that contained just five subjects [19]. In the current analyses, correlation between mean FC and mean SC was 0.22, with canonical correlations of 0.31 and 0.21. Note these approaches treat the edge as the unit of observation, averaged across subjects, and the correlations are not comparable to the variation explained in Table 1.1. Some models assume that higher SC for a given edge leads to higher FC [18]. CJIVE allows the extraction of patterns of covariation to provide novel insight not provided by spatial assumptions.

We found that CJIVE joint scores were more strongly related to gF than joint scores from R.JIVE or sCCA. The overall correlation from R.JIVE was higher than CJIVE (0.56 versus 0.28), but used many more components (151 versus 15), and individual components were not significant in R.JIVE or CJIVE. When examining gF and all pair-wise correlations (i.e., FC only) in the Web Explorer "HCP820-MegaTrawl", no edges survive corrections for multiple comparisons, and using the elastic net, $r = 0.21$. Initial studies with a subsample of the HCP rs-fMRI subjects found correlations between predicted and observed gF ranging from $r = 0.4$ to $r = 0.5$ [48, 10]. Brain wide association studies with hundreds of subjects may have inflated effect sizes relative to larger cohorts [30]. Moreover, previous studies did not examine the relationship between gF, FC and SC. Interestingly, CJIVE individual scores were not related to gF. This may suggest that FC and SC are simultaneously associated with gF in a manner that neither is independently. This result combined with the ability to predict out-of-sample subject scores via CJIVE suggests that results from JIVE methods may be a promising direction for biomarker development.

In practice, choosing the total signal rank remains a challenge. In simulations, the total signal rank chosen for a data block via R.JIVE permutation tests varied with the level of joint signal and the number of features within that block (Web Figure 1), and the number of components was relatively large in the real data. Additionally, scree plots of simulated data provide a much clearer distinction between eigenvalues that correspond to signal and those lying outside the signal subspace when compared to scree plots of real data. Most pertinent to our analyses is the result that both the CJIVE and AJIVE methods for estimating the joint rank are sensitive to estimates of the total signal ranks. If r_k approaches n, the maximum correlation between permuted datasets is very high, which leads to the estimation of zero joint components. In fact, when $r_k = n$, the correlation between permuted datasets equals one, and hence zero components are selected by CJIVE. The same issue occurs in AJIVE. Erring on the

side of a smaller total signal rank tends to result in more powerful tests of joint rank.

Further research is needed to explore connections between CJIVE and AJIVE estimates for more than two datasets. Multiset CCA (mCCA) [25] extends CCA to multiple datasets by maximizing the sum of pairwise correlations. A CJIVE variant on mCCA may provide novel insights into individual structure. A related issue is that for more than two datasets, joint signal may be shared by a subset of datasets [15]. When combining more than two datasets, future research should examine optimal ways of combining the canonical variables of the PC scores.

Chapter 2

Probabilistic JIVE

2.1 Introduction

Data integration encompasses a large framework of statistical and other methods designed to simultaneously explore and analyze multiple collections of features collected from the same observational units e.g., patients. Such collections of data have become nearly universal in many large- and moderate-size neuroimaging studies [32, 44, 39]. Statistical approaches to data integration date as far back as the 1930s, with the introduction of Canonical Correlation Analysis (CCA) [20]. While CCA focuses on exploring structure that is shared (i.e., joint) within two datasets, more recently methods, such as JIVE (Joint and Individual Variance Explained) [27] add to the data integration framework by teasing out structure that is unique (.i.e., individual) within a collection of datasets.

The current manuscript develops a probabilistic, model-based method to implement JIVE, which integrates differing modes of data on a common set of subjects in order to find low-rank approximations of the datasets' joint variability as well as low rank approximations of the variability unique to each dataset. Decomposition is generally achieved in two steps, both of which involve singular value decompositions

(SVD). JIVE has been used to explore the relationship between microRNA and gene expression in patients with a fatal form of malignant brain tumors in Lock et al. [27]. This approach was also employed to integrate behavioral and imaging data from the Human Connectome Project [63]. Several data integration methods have been developed in recent years. R.JIVE [36] is an iterative method that finds initial estimates of both the joint and individual variability, and iteratively uses each to re-estimate the other until convergence. AJIVE [8] uses matrix perturbation theory [56] to develop a non-iterative method for JIVE analysis. More recently, Canonical JIVE, or CJIVE, interprets joint scores (subject-level summaries of shared information) as a linear combination of canonical variables that arise from canonical correlation analysis of PC scores (Chapter 1). Other data integration methods published in recent years include robust-JIVE, which utilizes an L-1 norm minimization technique [42]; Common and Orthogonal Basis Extraction, which is similar to JIVE [66]; Structural Learning and Integrative Decomposition, which allows for partially shared structure [16]; Decomposition-based Canonical Correlation Analysis, a CCA method that decomposes data based on the $\mathcal{L}^2$ space of random variables rather than Euclidean space [47]; and Simulatneous Non-Gaussian Component Analysis, derived specifically for neuroimaging data.

While much progress has been made in data integration studies, the previous methods do not propose a statistical model. Interpreting their results can be quite challenging, whereas a likelihood-based approach may improve efficiency and interpretability. JIVE was formulated as a framework for decomposing data subspaces. However, data subspaces do not easily translate into fields of application where JIVE might be employed. For example, it is unclear what a JIVE decomposition means for a single observation, whereas a model-based framework posits the observation is a realization of a random variable. Moreover, the quantities of interest often discussed in post-analysis (i.e., subject scores and variable loadings) must be derived from JIVE

results and may require additional interpretation.

Principal Components Analysis (PCA) is closely related to JIVE: both methods involve SVD of datasets and define subject scores and variable loadings as left and right singular vectors, respectively, when the datasets have the form $n \times p$. Probabilistic PCA [52] formulates a statistical model for PCA in which subject scores are normally distributed random effects and variable loadings are fixed parameters. We generalize the PPCA framework to two datasets and develop a probabilistic approach to JIVE decomposition. We call our proposed method Probabilistic JIVE or ProJIVE. ProJIVE models sources of joint variation as subject random effects shared between datasets and sources of individual variation as subject random effects unique to each dataset. Such an MLE-based approach increases interpretability by directly modelling quantities of interest and may be more efficient than a two-step decomposition process that could introduce error at each step.

In this chapter, we propose a likelihood-based JIVE decomposition that may be more statistically efficient than SVD-based approaches. We evaluate the effectiveness of ProJIVE via simulation studies and an analysis of data obtained from The Alzheimer's Disease Neuroimaging Initiative (ADNI). Our simulation study shows that ProJIVE estimates subspaces at least as accurately as R.JIVE and AJIVE, whether or not our model assumptions requiring Gaussianity hold. In applying ProJIVE to the ADNI data, we combine measures of brain morphometry (i.e., thickness, surface area, and volume), and estimate their joint sources of variation. Results demonstrate the utility of ProJIVE: joint subject scores strongly associate with exogenous variables such as AD diagnosis and the presence of genetic marker apolipoprotein E4 (ApoE4).

In section 2.2, we describe the ProJIVE model and derive an EM algorithm to estimate its parameters. Section 2.3 conducts a simulation study comparing ProJIVE to existing JIVE methods. Our analysis of ADNI data in section 2.4 exhibits the

utility of ProJIVE in a real data setting. Finally, section 2.5 discusses our findings.

2.2 Methods

2.2.1 The original JIVE decomposition

Suppose features arising from the same set of n observational units (e.g., subjects) are collected in different datasets, $\mathbf{X}_k$ for $k = 1, \ldots, K$, such that each datablock (i.e., dataset) has dimension $n \times p_k$. For simplicity, we focus on the the case $K = 2$ here. The intuition underlying JIVE states that each data matrix $\mathbf{X}_k$ can be additively decomposed into a joint signal, $\mathbf{J}_k$, and individual signal, $\mathbf{A}_k$, and noise $\mathbf{E}_k$ where $\mathrm{E}(\mathbf{E}_k) = \mathbf{0}_{n \times p_k}$ and entries are mutually independent. We use the notation $C(\mathbf{G})$ to denote the vector subspace spanned by the eigenvectors of $\mathbf{G}$. The JIVE framework assumes that joint and individual signals lie in orthogonal vector subspaces (i.e. $C(\mathbf{J}_k) \perp C(\mathbf{A}_k)$ for $k = 1, 2$); and that both joint signals lie in the same vector subspace (i.e. for $C(\mathbf{J}_1) = C(\mathbf{J}_2)$). The vector subspaces of the individual signal matrices can either be mutually orthogonal or assumed to have null intersection. The formal model is given as (2.1):

$$\begin{gathered} \mathbf{X}_k = \mathbf{J}_k + \mathbf{A}_k + \mathbf{E}_k, \text{ subject to} \\ C(\mathbf{J}_k) = C(\mathbf{J}_{k'}), \;\; C(\mathbf{A}_k) \cap C(\mathbf{A}_{k'}) = 0, \text{ and } \mathbf{J}_k \mathbf{I}_k^\top = 0 \end{gathered} \tag{2.1}$$

2.2.2 Probabilistic JIVE

We propose a probabilistic model by first writing the k^{th} dataset as $\mathbf{X}_k = [\mathbf{x}_{1k}^\top, \ldots, \mathbf{x}_{nk}^\top]^\top$, where n is the number of subjects/observations in the study and each vector $\mathbf{x}_{ik} \in \mathbb{R}^{p_k}$

contains the p_k features captured on subject i in datablock k. Without loss of generality, assume $\mathbb{E}(\mathbf{x}_{i1}^\top, \mathbf{x}_{i2}^\top)^\top = \mathbf{0}$. If the data are not mean $\mathbf{0}$, demean the columns of each dataset prior to analysis. JIVE applications usually focus on factors of the joint and individual signal matrices instead of the matrices themselves. To that end, let $\mathbf{J}_k = \mathbf{Z}\mathbf{W}_{JK}$ and $\mathbf{A}_k = \mathbf{B}_k\mathbf{W}_{IK}$. We call the quantities $\mathbf{Z} = (\mathbf{z}_1^\top, \ldots, \mathbf{z}_n^\top)^\top$ and $\mathbf{B}_k = (\mathbf{b}_{1k}^\top, \ldots, \mathbf{b}_{nk}^\top)^\top$ subject scores, which, respectively, represent subject-specific summaries of the joint variance, and variance unique to the k^{th} block (i.e. individual variance). Loading matrices $\mathbf{W}_{Jk}, \mathbf{W}_{Ik}$ represent variable-specific summaries of joint and individual variance, respectively, for the k^{th} data-block. Our model is given by

$$\mathbf{x}_{ik} = \mathbf{W}_{Jk}\mathbf{z}_i + \mathbf{W}_{Ik}\mathbf{b}_{ik} + \epsilon_{ik},$$
$$\text{where } (\mathbf{z}_i^\top, \mathbf{b}_{i1}^\top, \mathbf{b}_{i2}^\top)^\top \overset{iid}{\sim} N(\mathbf{0}, \mathbf{I}) \text{ and } \epsilon_{ik} \sim N(\mathbf{0}, \sigma_k^2\mathbf{I}) \text{ for } k = 1, 2. \tag{2.2}$$

The dimension of each $\mathbf{z}_i$, r_J, describes the number of latent components giving rise to the joint variability within each dataset. Similarly, the number of individual components, r_{Ik}, equals the dimension of $\mathbf{b}_{ik}$. Since the joint and individual scores are assumed independent and the latent components are unobserved, careful consideration must be given to choosing values for r_J and r_{Ik}. Hereafter, we refer to r_J and r_{Ik} as joint and individual rank, respectively, and their sum $r_J + r_{Ik}$ as the signal rank.

In many applications, joint subject-scores, $\{\mathbf{z}_i : i = 1, \ldots, n\}$, or individual subject-scores $\{\mathbf{b}_{i1}, \mathbf{b}_{i2} : i = 1, \ldots, n\}$ are of particular interest as they can be used a potential biomarkers, prodromes, or discriminating factors among subgroups, by borrowing information from both datasets, as in Lock et al. [27]. On the other hand, examination of the joint and individual variable loadings $\{\mathbf{W}_{Jk}, \mathbf{W}_{Ik} : k = i, 2\}$ can lead to discovery of multivariate relationships between features of each dataset or unique patterns within datasets, as in Yu et al. [63] and Kashyap et al. [21].

The distribution of the data, conditioned on the latent variables, is given by

$$\mathbf{x}_{ik}|\mathbf{z}_i, \mathbf{b}_{ik} \sim N(\mathbf{W}_{Jk}\mathbf{z}_i + \mathbf{W}_{Ik}\mathbf{b}_{ik}, \sigma_k^2\mathbf{I}_{p_k}). \tag{2.3}$$

By stacking the two data vectors from a subject, we rewrite (2.2) as

$$\begin{pmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \end{pmatrix} = \begin{pmatrix} \mathbf{W}_{J1} \\ \mathbf{W}_{J2} \end{pmatrix} \mathbf{z}_i + \begin{pmatrix} \mathbf{W}_{I1} \\ \mathbf{0} \end{pmatrix} \mathbf{b}_{i1} + \begin{pmatrix} \mathbf{0} \\ \mathbf{W}_{I2} \end{pmatrix} \mathbf{b}_{i2} + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix}.$$

Then, the covariance of the JIVE model is

$$\text{Cov} \begin{pmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \end{pmatrix} = \begin{pmatrix} \mathbf{W}_{J1}\mathbf{W}_{J1}^\top + \mathbf{W}_{I1}\mathbf{W}_{I1}^\top + \sigma_1^2 I & \mathbf{W}_{J1}\mathbf{W}_{J2}^\top \\ \mathbf{W}_{J2}\mathbf{W}_{J1}^\top & \mathbf{W}_{J2}\mathbf{W}_{J2}^\top + \mathbf{W}_{I2}\mathbf{W}_{I2}^\top + \sigma_2^2 I \end{pmatrix}. \tag{2.4}$$

As the latent variables and errors follow marginal Gaussian distributions, we also obtain a marginal Gaussian distribution for the data captured on subject i:

$$\begin{pmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \end{pmatrix} \sim MVN(\mathbf{0}, \mathbf{C}), \tag{2.5}$$

where $\mathbf{C}$ is given by equation (2.4) above. The corresponding log-likelihood of the data is

$$\mathcal{L} = -\frac{n}{2}\{(p_1 + p_2)\log(2\pi) + \log(|\mathbf{C}|) + tr(\mathbf{C}^{-1}\mathbf{S})\} \tag{2.6}$$

where $\mathbf{S} = \frac{1}{n}\Sigma_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$.

2.2.3 Model identifiability

Here we discuss identifiability of the parameters in our model. Specifically, the log-likelihood is completely determined by the data and loading matrices, which are

unique up to orthogonal rotations.

Lemma: Suppose K data blocks, $\mathbf{X}_k \in \mathbb{R}^{n \times p_k}$, all satisfy the model (2.2), and the set of matrices $\{\mathbf{W}_{Ik}, \mathbf{W}_{Jk} : k = 1, \ldots, K\}$ maximize the likelihood function, $\mathcal{L}$, given in equation (2.6). Consider some orthogonal matrices $\mathbf{O}_J \in \mathbb{R}^{r_J \times r_J}$ and $\mathbf{O}_k \in \mathbb{R}^{r_{Ik} \times r_{Ik}}$ and define $\mathbf{W}^*_{Jk} = \mathbf{W}_{Jk}\mathbf{O}_J$ and $\mathbf{W}^*_{Ik} = \mathbf{W}_{Ik}\mathbf{O}_k$. Then $\mathcal{L}(\mathbf{W}) = \mathcal{L}(\mathbf{W}^*)$, where $\mathbf{W}$ and $\mathbf{W}^*$ are defined in 2.7.

Proof: The ProJIVE model relies on an underlying Gaussian distribution with mean 0. Thus, the log-likelihood, is completely determined by the data covariance matrix $\mathbf{C}$ and the data matrices themselves. Now, for any $\mathbf{C} \neq \tilde{\mathbf{C}}$, we have $\mathcal{L}(\mathbf{C}) = \mathcal{L}(\tilde{\mathbf{C}}) \iff |\mathbf{C}| = |\tilde{\mathbf{C}}|$ and $\text{tr}(\mathbf{C}^{-1}\mathbf{S}) = \text{tr}(\tilde{\mathbf{C}}^{-1}\mathbf{S})$. This produces an equivalence class of covariance matrices which return identical values for the log-likelihood.

By their definitions, $\mathbf{O}_J^\top \mathbf{O}_J = \mathbf{O}_J \mathbf{O}_J^\top = \mathbf{I}_{r_J}$ and $\mathbf{O}_k^\top \mathbf{O}_k = \mathbf{O}_k \mathbf{O}_k^\top = \mathbf{I}_{r_{Ik}}$, which implies $\mathbf{W}^*_{Jk}{\mathbf{W}^*_{Jk}}^\top = \mathbf{W}_{Jk}\mathbf{W}_{Jk}^\top$, $\mathbf{W}^*_{Ik}{\mathbf{W}^*_{Ik}}^\top = \mathbf{W}_{Ik}\mathbf{W}_{Ik}^\top$, and $\mathbf{W}^*_{J1}{\mathbf{W}^*_{J2}}^\top = \mathbf{W}_{J1}\mathbf{W}_{J2}^\top$. Then for any $i = 1, \ldots, n$ and $k = 1, 2$

$$\begin{aligned}\text{Cov}\left(\mathbf{x}_{ik}\right) &= \mathbf{W}_{Jk}\mathbf{W}_{Jk}^\top + \mathbf{W}_{Ik}\mathbf{W}_{Ik} + \sigma_k^2 I \\ &= \mathbf{W}^*_{Jk}{\mathbf{W}^*_{Jk}}^\top + \mathbf{W}^*_{Ik}{\mathbf{W}^*_{Ik}}^\top + \sigma_k^2 I.\end{aligned}$$

Let $\mathbf{W}$ and $\mathbf{W}^*$ be defined as

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_{J1} & \mathbf{W}_{I1} & 0 \\ \mathbf{W}_{J2} & 0 & \mathbf{W}_{I2} \end{pmatrix} \quad \mathbf{W}^* = \begin{pmatrix} \mathbf{W}^*_{J1} & \mathbf{W}^*_{I1} & 0 \\ \mathbf{W}^*_{J2} & 0 & \mathbf{W}^*_{I2} \end{pmatrix} \tag{2.7}$$

Then $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \mathbf{D} = \mathbf{W}^*{\mathbf{W}^*}^\top + \mathbf{D}$, where $\mathbf{D} = \begin{pmatrix} \sigma_1^2\mathbf{I}_{p_1} & \mathbf{0} \\ \mathbf{0} & \sigma_2^2\mathbf{I}_{p_2} \end{pmatrix}$. Thus, $\mathcal{L}(\mathbf{W}) = \mathcal{L}(\mathbf{W}^*)$, which defines an equivalence class of loadings matrices via right-multiplication by an orthogonal matrix. When the joint and individual signals each consist of only one component (i.e., $r_J = r_{Ik} = 1$), variable loadings are identifiable

up to sign changes.

2.2.4 Expectation-Maximization Algorithm for ProJIVE

We now develop an Expectation-Maximization, or EM, algorithm to estimate the variable loadings and error variances. Subject scores will be estimated using their BLUPs (Best Linear Unbiased Predictors), given by $E[\theta_i|\mathbf{x}_i]$.

Consider the latent subject sores, $\{\theta_i = (\mathbf{z}_i^\top, \mathbf{b}_{i1}^\top, \mathbf{b}_{i2}^\top)^\top : i = 1, \ldots, n\}$, as "missing" data, so that the "complete" data include the latent scores and the observed variables $\{\mathbf{x}_i = (\mathbf{x}_{i1}^\top, \mathbf{x}_{i2}^\top)^\top : i = 1, \ldots, n\}$. Using notation from (2.7), equation (2.2) is equivalent to

$$\mathbf{x}_i = \mathbf{W}\theta_i + \mathbf{E}_i. \tag{2.8}$$

Let $p = \sum_{k=1}^K p_k$ represent the total number of variables/features in both datasets and $r = r_J + \sum_{k=1}^K r_{Ik}$ the total number of latent components. Then, we can write the complete-data likelihood as

$$\begin{aligned}\mathcal{L}_C(\mathbf{W}, \mathbf{D}) = &-\frac{n}{2}\left((p+r)\log(2\pi) + \log(|\det(\mathbf{D})|)\right) \\ &-\frac{1}{2}\sum_{i=1}^{n}\left\{(\mathbf{x}_i - \mathbf{W}\theta_i)^\top \mathbf{D}^{-1}(\mathbf{x}_i - \mathbf{W}\theta_i) + \theta_i^\top\theta_i\right\}.\end{aligned} \tag{2.9}$$

Recall from section 2.2.3, $\mathrm{Cov}\,(\mathbf{x}_i) = \mathbf{C} = \mathbf{W}\mathbf{W}^\top + \mathbf{D}$. Since the complete data likelihood can be written as a product of Gaussian likelihoods, we also obtain a multivariate normal distribution for the complete data vector $(\theta_i^\top, \mathbf{x}_i^\top)^\top \sim N(\mathbf{0}, \mathbf{\Sigma})$ where,

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{I}_r & \mathbf{W}^\top \\ \mathbf{W} & \mathbf{C} \end{pmatrix}.$$

Thus, the mean and covariance of the conditional latent scores are

$$\begin{aligned}
\mathbb{E}(\theta_i|\mathbf{x}_i) &= \mathbf{W}^\top\mathbf{C}^{-1}\mathbf{x}_i, \\
\mathrm{Cov}\,(\theta_i|\mathbf{x}_i) &= \mathbf{I}_r - \mathbf{W}^\top\mathbf{C}^{-1}\mathbf{W}, \\
\mathbb{E}(\theta_i\theta_i^\top|\mathbf{x}_i) &= \mathbf{I}_r - \mathbf{W}^\top\mathbf{C}^{-1}\mathbf{W} + \mathbf{W}^\top\mathbf{C}^{-1}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{C}^{-1}\mathbf{W}.
\end{aligned} \tag{2.10}$$

Define loadings and scores for block k as $\mathbf{W}_k = (\mathbf{W}_{Jk}, \mathbf{W}_{Ik})$ and $\theta_{ik} = (\mathbf{z}_i, \mathbf{b}_{ik})$, and "selection matrices" $\mathbf{L}_k = (\mathbf{0}_{p_k\times p_1} \ldots \mathbf{I}_{p_k\times p_k} \ldots \mathbf{0}_{p_k\times p_K})$ and $\mathbf{M}_k = \begin{pmatrix} \mathbf{I}_{r_J\times r_J} \ldots \mathbf{0} \ldots \mathbf{0} \\ \mathbf{0} \ldots \mathbf{I}_{r_{Ik}\times r_{Ik}} \ldots \mathbf{0} \end{pmatrix}$. With this parameterization, $\mathbf{x}_{ik} = \mathbf{L}_k\mathbf{x}_i$, $\theta_{ik} = \mathbf{M}_k\theta_i$, and $\mathbf{L}_k\mathbf{W} = \mathbf{W}_k\mathbf{M}_k$, so that the conditional expectation of the log-likelihood is

$$\begin{aligned}
\mathbb{E}\{\mathcal{L}_C\} = &-\frac{n}{2}\left((p+r)\log(2\pi) + \sum_{k=1}^{K}\frac{p_k}{2}\log\sigma_k^2\right) \\
&-\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\sigma_k^{-2}\left[\mathbf{x}_{ik}^\top\mathbf{x}_{ik} + \mathbf{W}_k^\top\mathbf{W}_k\mathbb{E}(\theta_{ik}\theta_{ik}^\top|\mathbf{x}_i) - 2\mathbf{x}_i^\top\mathbf{W}_k\mathbb{E}(\theta_{ik}|\mathbf{x}_i)\right] + \mathbb{E}(\theta_{ik}^\top\theta_{ik}|\mathbf{x}_i).
\end{aligned} \tag{2.11}$$

Note that the first and second conditional moments of the scores take the form

$$\begin{aligned}
\mathbb{E}(\theta_{ik}|\mathbf{x}_i) &= \mathbf{M}_k\mathbb{E}(\theta_i|\mathbf{x}_i) = \mathbf{M}_k\mathbf{W}^\top\mathbf{C}^{-1}\mathbf{x}_i, \\
\mathrm{Cov}\,(\theta_{ik}|\mathbf{x}_i) &= \mathbf{M}_k\mathrm{Cov}\,(\theta_i|\mathbf{x}_i)\mathbf{M}_k^\top = \mathbf{M}_k(\mathbf{I}_r - \mathbf{W}^\top\mathbf{C}^{-1}\mathbf{W})\mathbf{M}_k^\top, \\
\mathbb{E}(\theta_{ik}\theta_{ik}^\top|\mathbf{x}_i) &= \mathbf{M}_k(\mathbf{I}_r - \mathbf{W}^\top\mathbf{C}^{-1}\mathbf{W} + \mathbf{W}^\top\mathbf{C}^{-1}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{C}^{-1}\mathbf{W})\mathbf{M}_k^\top.
\end{aligned} \tag{2.12}$$

Differentiating the conditional expected log-likelihood with respect to the param-

eters $\mathbf{W}_k, \sigma_k^2$, then setting each equal to 0 yields closed form solutions given by

$$\begin{aligned}\widetilde{\mathbf{W}}_k =& \left(\sum_i \mathbf{x}_i \mathbb{E}(\theta_i^\top | \mathbf{x}_i)\right)\left(\sum_i \mathbb{E}(\theta_{ik}\theta_i^\top | \mathbf{x}_i)\right)^{-1}, \\ \widetilde{\sigma}_k^2 =& \frac{1}{np_k}\sum_i \mathrm{tr}\{\mathbf{L}_k\mathbf{L}_k^\top \mathbf{x}_i\mathbf{x}_i^\top + \mathbf{W}\mathbb{E}(\theta_{ik}\theta_{ik}^\top|\mathbf{x}_i)\mathbf{W}^\top - 2\mathbf{W}\mathbb{E}(\theta_{ik}|\mathbf{x}_i)\mathbf{x}_i^\top\}.\end{aligned} \tag{2.13}$$

We initiate the EM algorithm with CJIVE estimates, described in chapter 1.

2.3 Simulation study

2.3.1 Simulations comparing JIVE methods

Simulation studies assessed the utility of ProJIVE by comparing its results to those obtained via AJIVE and R.JIVE. The simulation studies examine two issues: 1) accuracy of estimates when the joint signal strength is low versus high, and 2) robustness against model misspecification. We use a full factorial design with the following factors:

1. The number of features in $\mathbf{X}_2$: with levels (a) $p_2 = 20$ and (b) $p_2 = 200$,
2. Joint Variation Explained in $\mathbf{X}_1$: with levels (a) $R^2_{J1} = 0.05$ and (b) $R^2_{J1} = 0.5$,
3. Joint Variation Explained in $\mathbf{X}_2$: with levels (a) $R^2_{J2} = 0.05$ and (b) $R^2_{J2} = 0.5$,
4. Data generating distributions: with levels (a) Gaussian scores and loadings and (b) mixture of Gaussian joint scores and Rademacher loadings (joint and individual).

The joint rank was 3 and individual ranks were 2 in all settings. A simulation study with joint rank equal to 1 showed results similar to those described here. The sample size, number of features in $\mathbf{X}_1$, and proportions of individual variation explained for both data blocks were held constant at $p_1 = 20$ and $R^2_{I1} = R^2_{I2} = 0.25$, respectively.

Lastly, entries of the error matrices $\mathbf{E}_1$ and $\mathbf{E}_2$ were randomly drawn from a standard Gaussian distribution.

Simulated data were generated in a manner similar to that described in 1.3.1 for setting (a) of experimental factor 4, i.e. with subject scores and variable loadings from Gaussian distributions. For setting (b) of experimental factor 4, joint subject scores are drawn from a mixture of Gaussian distributions with unit variance: 20% with mean -4, 50% mean 0, and 30% mean 4.

For each combination of settings, we performed 100 simulations using three methods of JIVE analysis: ProJIVE, AJIVE [8], and R.JIVE [27]. True signal ranks were used as input for each method, since rank selection was not a target for evaluation in the current study. The chordal norm between true and estimated parameters (equation 1.4) evaluated the accuracy of estimated of each method.

2.3.2 Simulation Results

Figure 2.1 shows the chordal distances between true score/loading subspaces and their estimates when simulated data conform to model assumptions, i.e. setting (a) of the 4^{th} experimental factor. For these simulations, ProJIVE score subspaces distance from the true score subspaces trended less than the same distances for R.JIVE and AJIVE, especially in settings with $p_2 = 200$. Variable loadings estimated via ProJIVE were more accurate than those form AJIVE and R.JIVE in the low joint variation settings ($i.e., R^2_{J1} = R^2_{J2} = 0.05$). Figure 2.2 shows chordal distances for simulations in setting (b) of the 4^{th} experimental factor. These results provide evidence that ProJIVE is robust against failure to satisfy the Gaussianity assumptions in equation (2.3). Although the chordal distances between estimated and true loadings are more variable in these settings, ProJIVE estimates are closer to the truth, on average, than those from other methods when the joint variation in at least one data block is relatively small, i.e. $R^2_{Jk} = 0.05$ for at least one$k = 1, 2$.

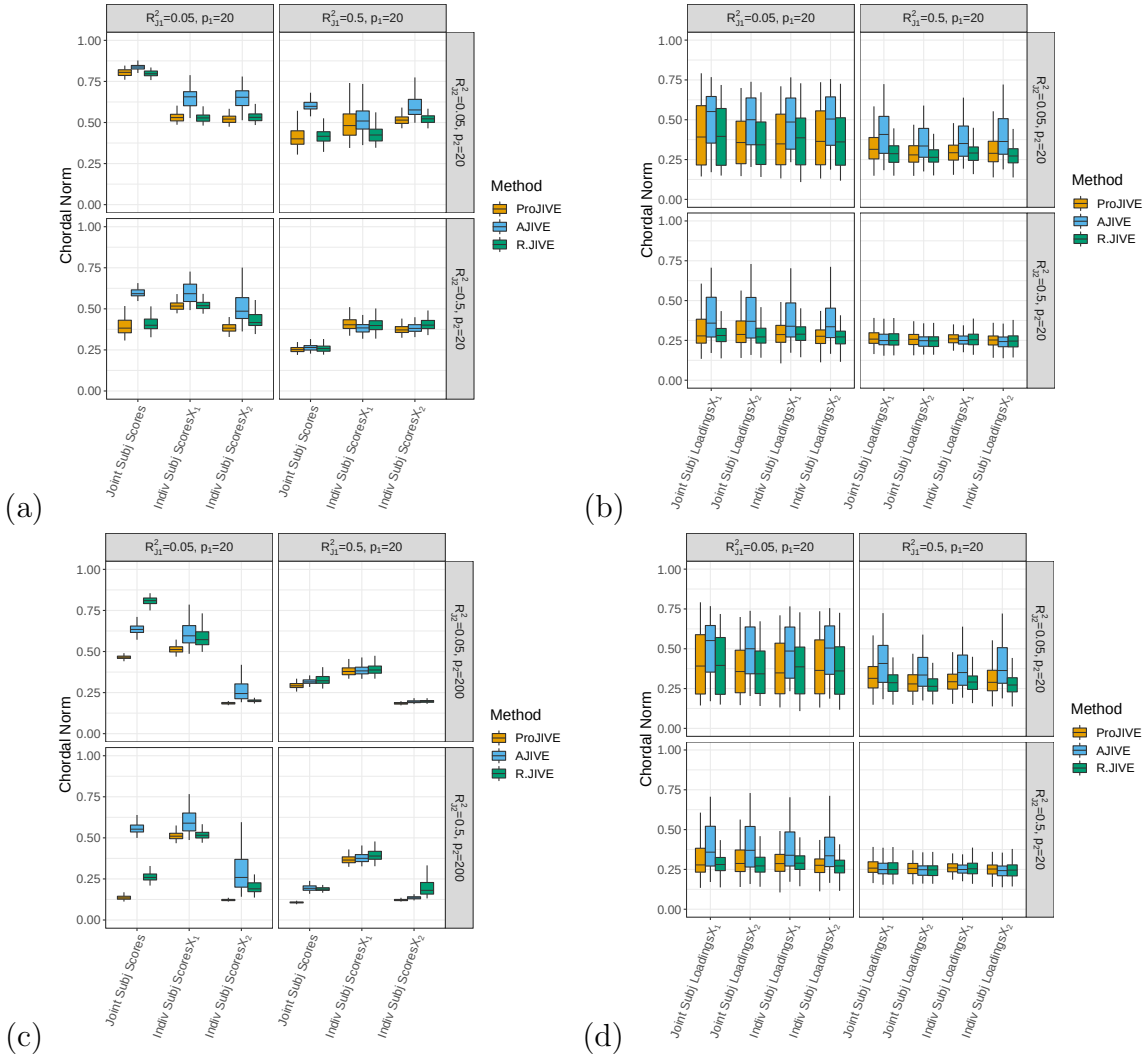


Figure 2.1: Results of simulation studies with data generated from the ProJIVE model using subject scores and variable loadings generated from standard Gaussian distributions: (a and b) $p_2 = 20$, (c and d) $p_2 = 200$. Each sub-figure exhibits boxplots of chordal norms for subject scores (a and c) and variable loadings (b and d). Color key: Orange = ProJIVE. Light Blue = AJIVE. Green = R.JIVE.

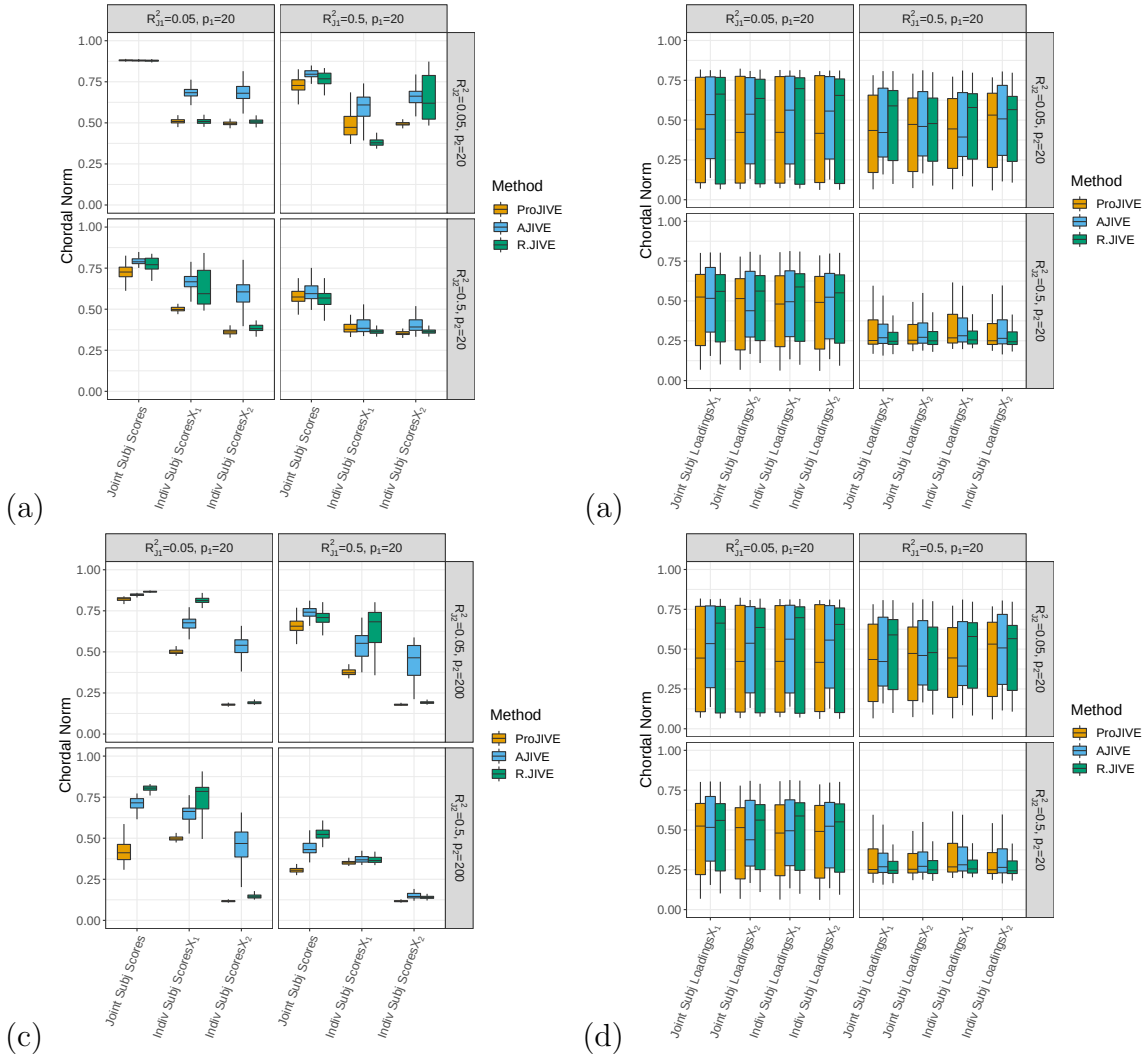


Figure 2.2: Results of simulation studies with data generated from the ProJIVE model using joint subject scores generated from a mixture of Gaussian and individual subject scores from standard Gaussian distributions. Variable loadings (joint and individual) were generated from Rademacher loadings. Sub-figures (a and b) show results when $p_2 = 20$; in (c and d) $p_2 = 200$. Each sub-figure exhibits boxplots of chordal norms for subject scores (a and c) and variable loadings (b and d). Color key: Orange = ProJIVE. Light Blue = AJIVE. Green = R.JIVE.

To summarize, we find that ProJIVE performs at least as well at estimating subject scores when compared to other methods. Estimation of both subject scores and variable loadings was markedly than better AJIVE in nearly all settings. When the number of variables in the second data block exceeded that of the first, joint subject scores from ProJIVE were more accurate in settings with mixed proportions of joint variation.

2.4 Joint Analysis of Brain Morphometry and Cognition in ADNI Data

Our data application examines shared variability in cognitive/behavioral measures and measures of brain structural integrity. Both sets of measures were residualized after regressing out age and sex to avoid confounding. We applied ProJIVE, AJIVE, and R.JIVE using total ranks. Joint ranks were chosen via the permutation test outlined in section 1.2.1.

Specifically, the data used in preparation of this chapter were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). The ADNI is an ongoing, longitudinal study that uses neuroimaging techniques, cognitive assessments, and other biomarkers (e.g., the number of ApoE4 singular nucleotide polymorphisms or SNPs) to better understand the natural history of Alzheimer's disease (AD) and improve the way that AD is diagnosed. After the baseline visit, ADNI participants have follow-up visits every three months during the first year, every six months during the second year, and annually after that. [32]

2.4.1 TADPOLE Challenge

The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge (`https://tadpole.grand-challenge.org/Home/`) was an open competition

to predict the onset of AD in the next phase of ADNI recruitment and retention, ADNI-3. We exhibit the utility of ProJIVE with an application to data from $n = 587$ older adults participating in ADNI, obtained via TADPOLE. Although there are data available for more than $n = 1600$ in the TADPOLE datasets, the data analyzed herein comprise participants for whom the full battery of cognitive/behavioral assessments was available. The battery includes 22 scales and sub-scales, given to those in the ADNI-GO and ADNI-2 phases of the study. We limit our analysis to observations taken at the participants' 6-month follow-up visit to minimize missingness and maximize sample size.

Both ADNI-2 and ADNI-GO limit newly recruited participants to those between 55-90 years of age (inclusive) and require that each has an English- or Spanish-speaking study partner to provide independent evaluation of patient functioning. ADNI-GO inclusion criteria required that participants who rolled over from ADNI-1 were either diagnosed as cognitively normal (CN) or having mild cognitive impairment (MCI) at baseline. Newly recruited participants in ADNI-GO were all diagnosed with early mild cognitive impairment (EMCI). ADNI-2 participants were either rolled-over from ADNI-1/ADNI-GO or newly recruited participants diagnosed as CN, EMCI, LMCI (late mild cognitive impairment), or AD. For the remainder of this manuscript, we combine EMCI and LMCI participants as MCI and only use diagnoses from the 6-month follow-up visit, as participants were diagnosed again at each follow-up visit. Table 2.1 provides summary statistics for age, gender, and ApoE4 SNP counts stratified by diagnosis.

2.4.2 Dimension Reduction, Preprocessing, and Summary

Cognition

Cognitive and behavioral measures, hereafter cognitive measures or the cognition dataset, included the following assessments:

Table 2.1: Summary statistics for selected covariates of participants in ADNI-GO and ADNI2.

	AD (N=88)	MCI (N=340)	CN (N=159)	Total (N=587)	p value
	Mean (S.D.) or N (%)				
Age					0.006
	74.0 (7.92)	71.5 (7.57)	72.8 (5.85)	72.2 (7.25)	
Gender					0.006
Female	28 (31.8%)	150 (44.1%)	84 (52.8%)	262 (44.6%)	
ApoE4					<0.001
0	21 (23.9%)	178 (52.4%)	111 (69.8%)	310 (52.8%)	
1	45 (51.1%)	126 (37.1%)	46 (28.9%)	217 (37.0%)	
2	22 (25.0%)	36 (10.6%)	2 (1.3%)	60 (10.2%)	

- Clinical Dementia Rating - Sum of Boxes (CDR-SB) [31]
- Alzheimer's Disease Assessment Scale - Cognition (ADAS) [40]
  - The 11-item and 13-item scores used as separate variables
- The Mini-Mental State Exam (MMSE) [11]
- Rey's Auditory Verbal Learning Test (RAVLT) [37, 38]
  - Forgetting, Immediate, and Learning sub-scales used as separate variables
- Montreal Cognitive Assessment (MOCA) [33]
- Everyday Cognition (ECOG) [6]
  - 7 pairs of sub-scales used: each pair includes a response from participant (PT) and their study partner (SP)

Summary statistics for each cognition measure used in JIVE analyses are shown in table 4. All cognitive assessments were associated with diagnosis. However, we note that some cognitive measures were used to inform diagnoses (e.g. ADAS13, MOCA, and MMSE) and, thus, are expected to correspond strongly with diagnosis.

The scree plot in figure 2.3 shows rank choices based on three methods: 1) choosing the 'elbow' of the scree plot (rank=5), 2) the number of eigenvalues that account for at

least 90% of total variability (rank=9), and 3) the number of eigenvalues accounting for at least 95% of total variation (rank = 13). Our final analysis uses total rank of $r_C = 5$.

Brain Morphometry

Brain morphometry is used here as a catch-all to describe measures of volume, thickness, and surface area of regions of interest (ROIs) within the brain. Cortical ROIs in the TADPOLE dataset largely reflect those described in Desikan et al. [3], which calls for 34 gray matter ROIs within each hemisphere. The morphometry contains measures of cortical thickness (CT) and cortical surface area (SA) for each of ROI. Inter-cranial volume is also included as a measure of cortical volume (CVol) for a total of 69 measures. The volumes of subcortical ROIs (both gray matter and white matter regions/structures) are labelled as white matter volumes (WMVol) in the TADPOLE Challenge dataset. These include 17 subcortical gray matter structures, 16 of which form eight hemispheric left-right pairs. The unpaired subcortical gray matter ROI is the brainstem. There are 13 remaining WMVol measures which include the ventricles, CSF, optic chiasm, corpus callosum, and others. In total, 245 measures of brian morphometry are included. A complete list along with descriptive statistics are included in Appendix 2.

Morphometry measures in TADPOLE were preprocessed using the cross-sectional Freesurfer pipeline, which is documented and freely available online (`http://surfer.nmr.mgh.harvard.edu/`). MRI scans were skull-stripped, corrected for B1 field bias, and segmented into gray matter, white matter, and CSF. Next, image reconstruction ensured correspondence to cortical surface models, i.e. gray-white matter boundary and pial surfaces. Finally, cortical and subcortical regions were labelled and registered to a standardized template via nonlinear transformations.

As with the cognition dataset, three methods were applied for choosing the total

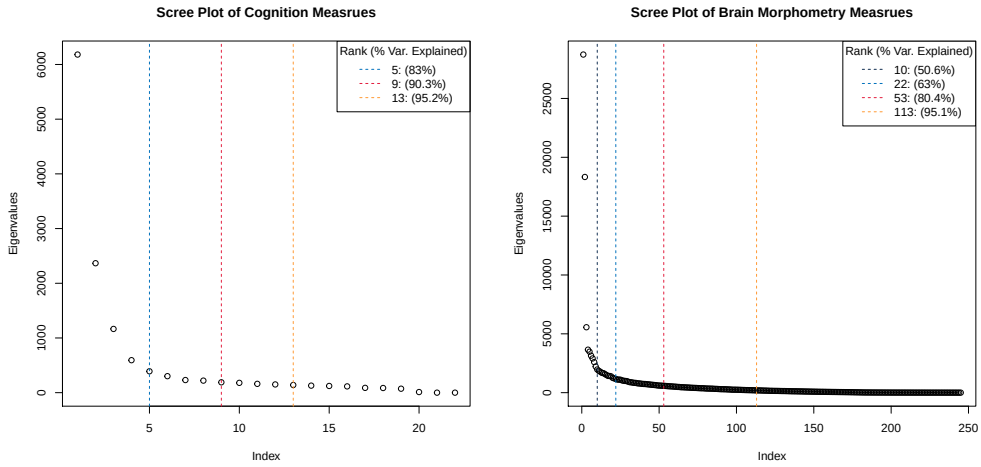


Figure 2.3: Scree plots shows which values were chosen as total signal ranks. Choices depended the 'elbow' of the scree plot or by accounting for at least 80%, 90% ,or 95% of total variance.

signal rank. With brain morphometry, two values were proposed as the 'elbow' of the scree plot (rank = 10 or 22). The other two ranks were chosen to account for at least 80% (rank = 53) and at least 95% (rank = 113) of total variability. Figure 2.3 shows the scree plot, each chosen rank and their corresponding percentage of total variation. Our final analysis uses total rank of $r_B = 10$.

2.4.3 Joint Subspace

Our permutation test for joint rank found $r_J = 1$ component in the subspace shared by cognition and brain morphometry measures. Recall that section 2.2.3 shows model parameters are identifiable up to orthogonal orthogonal transformation. For $r_J = 1$, this result reduces to identifiability up to a sign flip for both joint subject scores and joint variable loadings.

Subject Scores

We examined joint subject scores for their associations with 6-month diagnosis using a multinomial logistic regression model. Z-transformed regression coefficients and p-

values were $z_{MCI} = \beta_{MCI}/SE(\beta_{MCI}) = 6.45$, $p < 1^{-9}$, and $z_{AD} = \beta_{AD}/SE(\beta_{AD}) = 11.41$, $p < 1^{-10}$ for MCI vs CN and AD vs CN, respectively. The residual deviance was 892.02, with $edf = 4$ effective degrees of freedom. These results indicate that the joint subject scores capture a summary of the subspace shared between cognition and brain morphometry and that subspace strongly associates with diagnoses. The graph in figure 2.4 illustrates this point by showing that the spread of scores within each diagnosis category has a distinctive center and inter-quartile range. The association between joint subject scores and ApoE4 SNP counts was also statistically significant with normalized coefficients $z_{ApoE4=1} = 5.53$, $p < 1^{-7}$, and $z_{ApoE4=2} = 4.54$, $p < 1^{-5}$, for 1 vs 0 and 2 vs 0 SNP counts, respectively. The residual deviance was 1060.07, $edf = 4$.

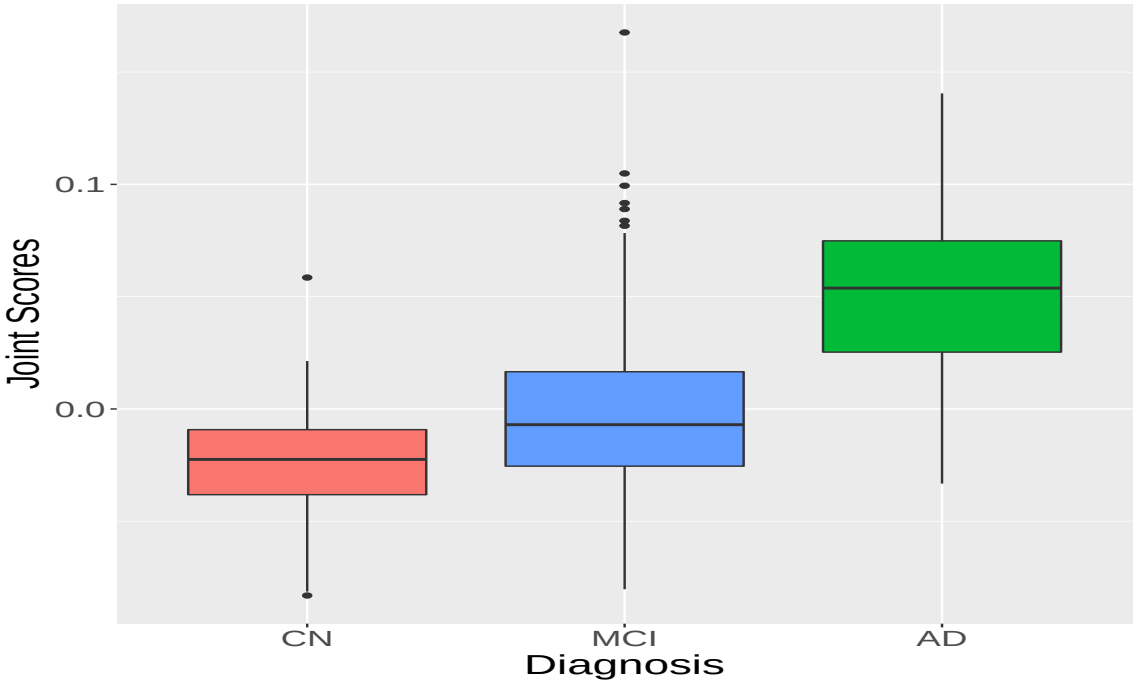


Figure 2.4: Joint subject scores estimated via ProJIVE show separation by diagnosis at 6-month follow-up.

Variable Loadings

Joint loadings for a dataset exhibit the extent to which each measure in that dataset contributes to the shared subspace. To aid interpretation, we normalized joint loadings to the interval $[-1, 1]$, then sign-corrected to result in positive skewness. Of particular interest are measures with normalized loadings near the interval boundaries. Therefore, we focus on the top 10 absolute cognition loadings and the 90^{th} percentile of absolute brain loadings. Figure 2.5 shows that ADAS and MMSE measures were most prominent among cognition loadings. Measures of cortical thickness (CT) and (WMVol) were prominent in morphometry. The 90^{th} percentile of loadings occur mostly within left-right hemispheric pairs, within each type of morphometry measurement. Of the five CVol measures present, however, three appear in the left hemisphere only. No SA measurements were present in the 90^{th} percentile.

The signs of the joint cognition loadings were consistent with the interpretation of the related measures within each data block. For example, the scoring schema for MOCA/MMSE have opposite interpretations for diagnosing AD when compared to the scoring schema for ADAS/CDRSB. Brain loadings were also consistent with associations between certain ROIs and AD in the literature. Both enlargement of ventricles and atrophy of gray matter structures (e.g., hippocampus, amygdala) have been shown to associate with increased risk of AD. [34, 13] Note that ADAS and CDRSB load in opposite directions of MOCA and MMSE. Similarly, only the ventricles (which are not gray matter structures) load in the opposite direction of the remaining 90^{th} of joint loadings.

2.5 Discussion

We propose ProJIVE, a method for conducting JIVE analysis that builds on the CJIVE methodology presented in Chapter 1 and the well-known PPCA methodology

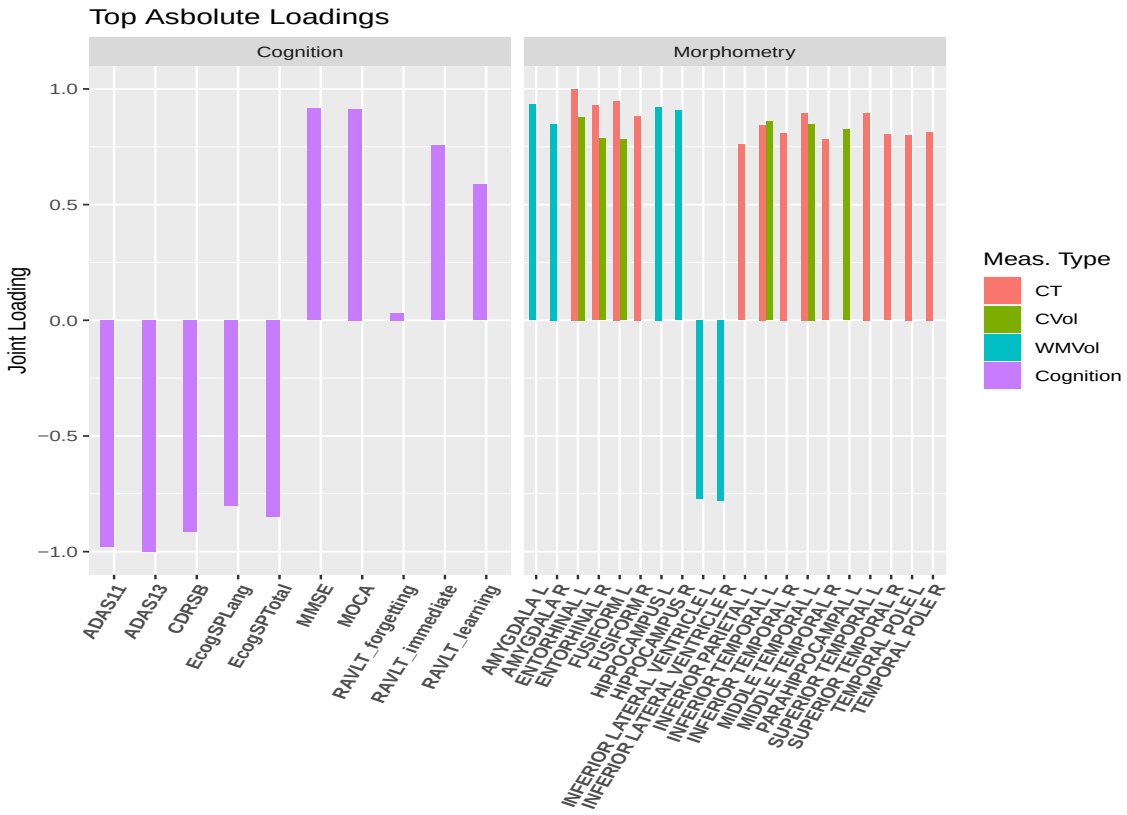


Figure 2.5: Ten most extreme joint cognition loadings and 90^{th} percentile of absolute joint brain loadings estimated via ProJIVE.

[52] to provide a probabilistic model for the JIVE framework. Our proposed model is conceptually consistent with JIVE and improves interpretation by explicitly modelling parameters of interest. Results from simulations studies indicate that ProJIVE improves estimation of joint subject scores when compared to AJIVE and R.JIVE and estimation of variable loadings when compared to AJIVE.

We apply ProJIVE to examine shared information captured via measures of cognition and brain structure in a cohort of older adults participating in the ADNI. Results of our analysis demonstrate ProJIVE's utility as a data reduction method, which uncovers multivariate relationships across datasets. Biological relevance was revealed by the strong association between diagnoses and joint subject scores. Extreme variable loadings show patterns of variation that are consistent with association found in the literature between measures of cognition, brain morphometry, and Alzheimer's disease.

Our proposed method is limited in its increased computational costs due to application of the EM algorithm to estimate model parameters. ProJIVE run-times varied form twice to one-hundred times as long as R.JIVE and AJIVE run-times. Additionally, ProJIVE does not currently consider rank selection. In the future, we consider using the Bayesian Information Criterion (BIC) to inform rank choice, modifying the model to allow for structure that is shared between multiple but not all datasets in settings where $K > 2$, and using EM to incorporate variables with missingness. We also plan to address inference and prediction, two aspects that are not addressed in the current methodology.

Chapter 3

Generalized Additive Models of Ambulatory Blood Pressure Profiles

While the previous two chapters focused on data integration methods, chapter 3 examines the use of another set of multivariate methods, namely generalized additive models, or GAMs. Here we focus on the use of GAMs to model differences in ambulatory blood pressure profiles from a cohort of young to middle-aged Black women living in the metropolitan area surrounding Atlanta, GA, based on psychosocial exposures.

3.1 Introduction

The U.S. has seen steady declines in age-specific cardiovascular disease (CVD) during previous decades. Decreased smoking rates, the treatment and control of elevated blood pressure, and an increased focus on preventative healthcare are leading factors in the decline [12, 54]. Recently, however, rates of CVD mortality have begun to plateau and even increase in some sub-populations. Black Americans (especially those with hypertension) and adults aged 35-64 years are both populations at increased

risk for CVD and related adverse events such as stroke, myocardial infarction, heart disease, and death [61]. Simultaneously, CVD remains the leading cause of death among women in the U.S. This burden disproportionately impacts Black women and appears to persist after accounting for traditional risk factors such as diet, waist circumference, and smoking [7]. These factors prompt additional research into the sources of such disparities and strategies to alleviate them.

The impacts of psychological stress on blood pressure and other biological mechanisms involved in CVD have been documented in laboratory settings and mechanistic studies in real-life settings [24]. Specifically, stress has shown to be moderately associated with CVD risk in, mostly White, study populations. However, studies that examine these associations within Black women are sparse [7]. The Mechanisms Underlying Stress and Emotions (MUSE) in African-American Women's Health Study aims to address gaps in the current understanding of relationships between CVD risk, psychosocial stressors, and sociodemographics. in Black women. The study follows a cohort of young- to middle-aged black women living in the metropolitan area in and around Atlanta, GA.

MUSE brings a comprehensive, state-of-the-art approach to examining CVD risk using ambulatory blood pressure measurements (ABPM), carotid intima-media thickness, and pulse wave velocity. ABPM has been recommended over blood pressure (BP) measurements taken in a clinical setting for at least two reasons. First, the so-called "white coat" effect can lead to elevated BP in people with anxiety about doctors' visits. Second, the ability to capture nighttime BP, which can be indicative of increased CVD risk and other adverse outcomes [35]. In the MUSE Study, participants are asked to wear an ABPM device for 48 hours. During which time, the device will record BP and heart rate (HR) measurements every 30–minutes from 8 am - 6pm, and every 60-minutes. Our objective is to examine differences in these profiles as functions of exposure to psychosocial stressors.

Some previous analyses of ABP have used summary measures such as the mean/median BP, 24-hour or daytime/nighttime, as the outcome in statistical models [55, 9, 1]. Others have used trigonometric models [14, 28]; generalized linear mixed models [45]; linear mixed models with orthonormal polynomials [5]; and smoothing techniques [49, 4] to analyze the longitudinal profiles arising from ABPM. Streitberg et al. [50] concluded that Fourier smoothing or spline smoothing, combined with a robust estimation, method were preferable for analysis of ABP when compared to summary measures.

To that end, we use Generalized Additive Mixed Models (GAMMs) [59, 58] with penalized cyclic cubic splines to develop a statistical model of ABPM that compares two groups within the MUSE cohort, while controlling for important covariates, circadian rhythm, and within-person variability. Penalized splines enable one to balance between model fit and smoothness. Other methods essentially model ABP as curves, which constrains the shape of the estimated profiles. Additionally, using GAMMs to model ABP profiles allows for inference on particular times of day during which ABP differs between groups. This offers additional insight into the physiological impacts of psychosocial stressors not available from conventional linear mixed models.

3.2 Ambulatory Blood Pressure in MUSE

The MUSE study follows a cohort of Black women ages 30-45 who live in the metro area of Atlanta, GA. The study examines associations between biological indicators of cardiovascular disease, socio-demographic characteristics, and exposures to social and psychological stressors. Potential recruits were excluded if they had a history of cardiovascular disease, diabetes, were pregnant or lactating, diagnosed with any chronic illness known to influence atherosclerosis (e.g., autoimmune or chronic inflammatory diseases such as HIV/AIDS, lupus, rheumatoid arthritis, renal disease, liver

disease), currently receiving treatment for psychiatric disorders, currently using illicit drugs (i.e., marijuana, cocaine), or dealt with alcohol abuse. Women who reported working overnight shifts were also excluded because of the known impact of shift work on diurnal rhythms and ambulatory blood pressure. Of the 422 MUSE participants, we examine the profiles of $n = 408$ with non-missing ABP and covariate data.

MUSE provided participants with a wearable electronic device that measures ABP and heart rate (HR) at pre-determined intervals. They were asked to wear the device for 48 hours and received financial compensation if at least 70% of the intended measurements were attained. The cuff was scheduled to take ABP/HR readings once every 30-minutes during "waking" or "daytime" (DT) hours (8 am - 9:59 pm) and once per hour during "nighttime" (NT) hours (10 pm - 7:59 am).

We dichotomized participants into two groups for the present analyses based on whether they held primary financial responsibility for their household, i.e., breadwinners (BWs) and non-breadwinners (non-BWs). Financial responsibility was determined by the survey question: "Are you the primary breadwinner in your household?" We consider participants who answer affirmatively to be financially responsible. BW-status is a primary explanatory variable for the remainder of this manuscript.

3.3 Statistical Methods

We utilized the GAMM framework to estimate average 24-hour ABPM profiles for each group and examined the profiles for time intervals over which they differed significantly between groups. MUSE recorded ABP for each participant over a 48-hour period. However, we consider a model of 24-hour ABP here. Our models assign $t = 0$ to 12 pm on the first day that ABPM readings were obtained. The time of subsequent readings are assigned values $t \in (0, 1]$. The resulting models treat each day as a replicate and therefore borrow information across both days to use in estimating

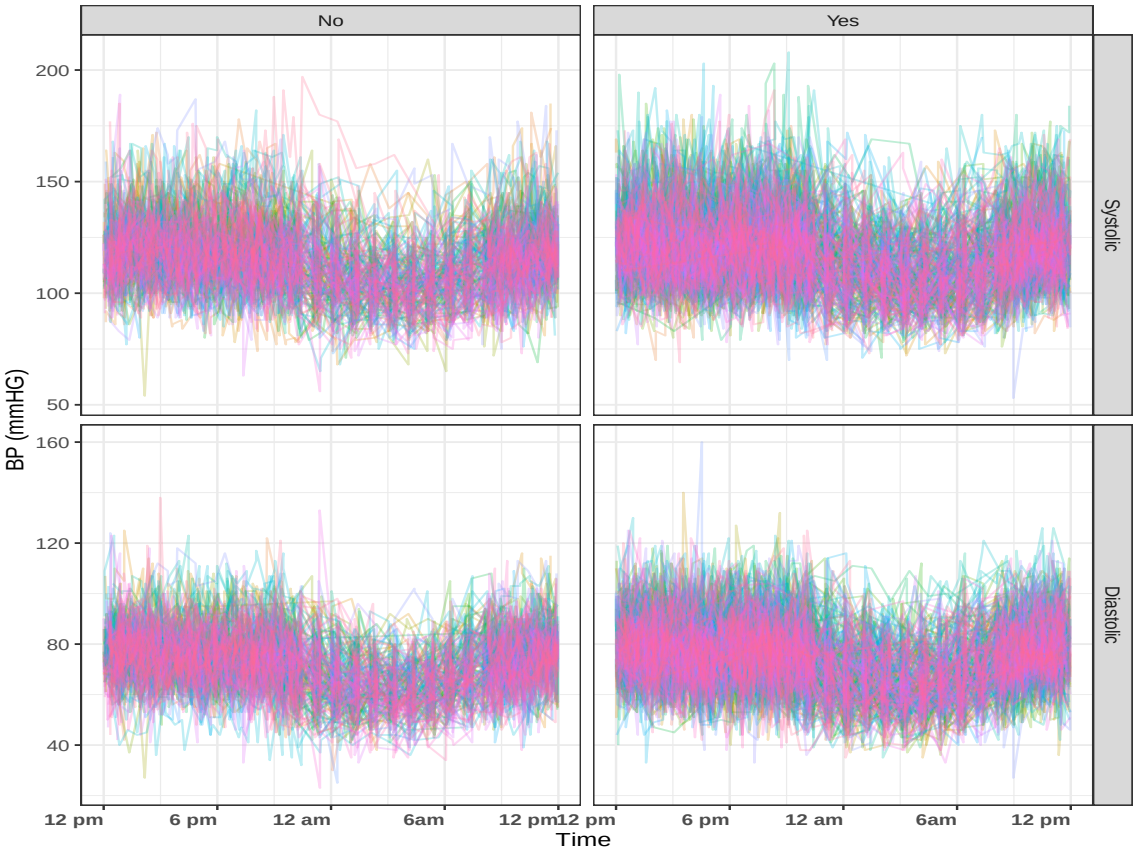


Figure 3.1: 48-hour ABPM data overlaid onto a 24-hour period for $n = 408$ participants in the MUSE study, stratified by BW status.

ABP profiles. The data are structured to preserve the order in which they were recorded. The number of ABP readings per participant ranged $[6, 88]$, with median 73 and interquartile range $(66, 76)$. We examined results using all available data and compared those to a dataset restricted to participants with at least 70% of intended readings (Appendix 3). Graphical summaries of the intervals between readings are also shown in Appendix 3. We consider differences in systolic BP (SBP) independent from differences in diastolic BP (DBP).

Before examining associations between ABP and exposures, we build a 'time model' to ensure a good fit to the overall mean SBP and DBP profiles. Additional details about model selection and diagnostics are described in Appendix 3. The time

model is given by

$$BP_{ij} = \beta_0 + f(t_{ij}) + b_{ij} + \epsilon_{ij}, \tag{3.1}$$

where BP_{ij} is the i^{th} individual's i^{th} ABP reading, which occurs at time point $t_{ij} \in [0, 24)$; $f(\cdot) = c_1\phi_1(\cdot) + \cdots + c_k\phi_k(\cdot)$ is a smooth function of time constructed using penalized cyclic cubic splines of rank $k = 58$; $\phi_1, \ldots, \phi_k$ and $c_1, \ldots, c_k$ are the basis functions and coefficients, respectively, which build $f(\cdot)$; $b_{ij} \sim N(0, \sigma_b^2)$ is a subject-specific random effect with autocorrelation-1 structure, e.g. $\text{Corr}(b_{ij}, b_{ij'}) = \rho^{|j-j'|}$; and $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$.

The components of f are determined by minimizing the objective function $\sum_{i,j}\{BP_{ij} - f(t_{ij})\} + \lambda \int f''(x)^2 dx$. The tuning parameter λ determines balance between the smoothness of the estimated function $f(\cdot)$ and its fit to the observed data. Larger values of λ lead to smoother estimates and smaller values result in a function which fits the data more closely. We used the restricted maximum likelihood method (REML) to estimate λ. This method treats the basis functions as random effects to select λ. Basis coefficients and remaining parametric coefficients are then estimated using penalized iteratively re-weighted least squares (PIRLS).

For a cyclic cubic spline, the basis functions have the form

$$\phi_j(t) = 1 + (h_j c_j)^{-1}\Bigg\{(c_{j+1} - c_j)t + \frac{(t_{j+1} - t)^3 - h_j^2(t_{j+1} - t)}{6}\mathbf{F}_j\mathbf{c} + \frac{(t - t_j)^3 - h_j^2(t - t_j)}{6}\mathbf{F}_{j+1}\mathbf{c}\Bigg\},$$

where $t_1, \ldots, t_k$ are 'knots' chosen to span the interval $[0, 1]$, $h_j = t_{j+1} - t_j$, $\mathbf{c} = (c_1, \ldots, c_k)$, and $\mathbf{F}_j$ is the j^{th} column of $\mathbf{F} = [\mathbf{0},\ \mathbf{D}^\top(\mathbf{B}^{-1})^\top,\ \mathbf{0}]^\top$. Square matrices

$\mathbf{B}$ and $\mathbf{D}$ have non-zero elements

$$\mathbf{B}_{i-1,i} = \mathbf{B}_{i,i-1} = h_{i-1}/6, \;\; \mathbf{B}_{i,i} = (h_{i-1} + h_i)/3,$$

$$\mathbf{D}_{i-1,i} = \mathbf{D}_{i,i-1} = 1/h_{i-1}, \;\; \mathbf{D}_{i,i} = -1/h_{i-1} - 1/h_i, \text{ for } i = 2, \ldots, k-1,$$

$$\mathbf{B}_{1,k-1} = \mathbf{B}_{k-1,1} = h_{k-1}/6, \;\; \mathbf{B}_{1,1} = (h_{k-1} + h_1)/3,$$

$$\mathbf{D}_{1,k-1} = \mathbf{D}_{k-1,1} = 1/h_{k-1}, \text{ and } \mathbf{D}_{1,1} = -1/h_1 - 1/h_{k-1}.$$

The exposure model incorporates our primary predictor variable, i.e., BW-status Hereafter, BW_i is an indicator variable with $BW_i = 1$ for BWs and $BW_i = 0$ for non-BWs and β_l are unknown parameters to be estimated. We examine differences between these groups by modelling the average ABP profile for each separately,indicated by the functions $f_{BW}(\cdot)$ and $f_{non-BW}(\cdot)$. The model is given by

$$BP_{ij} = \beta_0 + \beta_1 BW_i + BW_i f_{BW}(t_{ij}) + (1 - BW_i) f_{non-BW}(t_{ij}) + b_{ij} + \epsilon_{ij}. \qquad (3.2)$$

Next we refine our analyses by adjusting for several time-invariant covariates such as age, body mass index (BMI), depressive symptoms as measured by the Beck Depression Inventory (BDI) score [2], cigarette smoke, and others. For each pair of models, we computed simulation-based 95% simultaneous confidence bands (CBs) for each estimated average ABP curve. CBs are shown as the shaded areas in figures 3.2 -3.4. Simultaneous confidence bands are produced by sampling from the posterior distribution of the GAMMs using 10,000 simulations.

Descriptive statistics were calculated for each covariate used in the fully adjusted model and for mean daytime (DT) and nighttime (NT) values of ABP (table 3.1). We examined each of these for associations with BW-status, using chi-square tests for categorical covariates and two sample t-tests for continuous variables. All statistical analyses were performed and plots created using RStudio 1.3 [41]. The type-I error rate $\alpha = 0.05$ was used to indicate statistical significance. GAMM models were fit

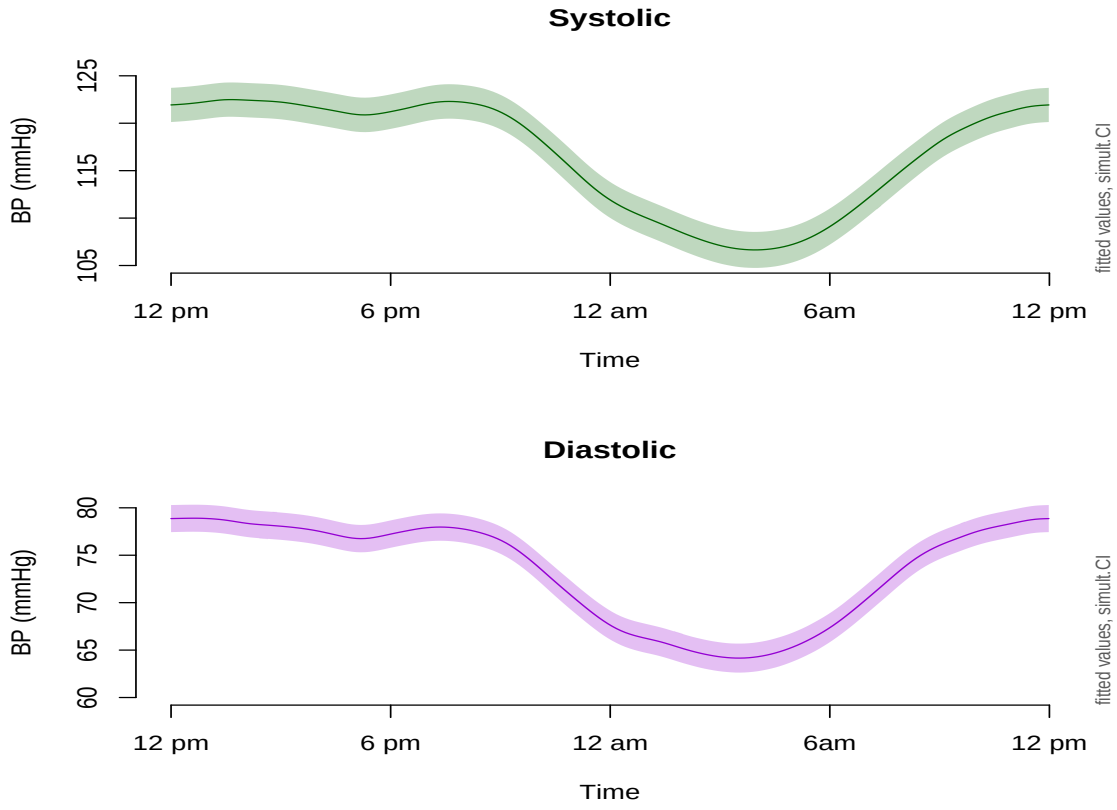


Figure 3.2: Fitted ABP profiles from the 'Stage 1' time model.

using the *mgcv* package [60] and visualization performed with the package *itsadug* [53].

3.4 Results

Table 3.1 shows that average DT and NT ABP were significantly associated with BW-status. ABP for BWs was approximately 3.5/2 mm Hg (systolic/diastolic) higher than non-BWs, on average during both daytime and nighttime hours. Income, partner status, and family size were also significantly associated with BW-status.

Figure 3.2 shows the fitted values from our time model, and importantly, that the time model captures the non-linear nature of the ABP profiles. The dip in the profiles reflects the diurnal patterns seen in the raw profiles (figure 3.1).

Table 3.1: Summary statistics for selected variables, stratified by BW-status, show a statistically significant association between DT/NT ABP and BW-status

	Breadwinner Status			
	No (N=154)	Yes (N=254)	Total (N=408)	p value
	Mean (SD) or Count (%)			
DT SBP				0.005
	119.3 (11.3)	122.8 (12.7)	121.5 (12.3)	
NT SBP				0.005
	109.3 (11.3)	112.7 (11.9)	111.4 (11.7)	
DT DBP				0.020
	76.3 (7.8)	78.4 (9.4)	77.6 (8.8)	
NT DBP				0.006
	67.1 (7.9)	69.5 (8.9)	68.6 (8.6)	
BMI				0.746
	32.5 (8.1)	32.7 (8.0)	32.6 (8.1)	
Age				0.153
	37.5 (4.4)	38.1 (4.1)	37.9 (4.3)	
BDI Score				0.744
	6.1 (7.0)	5.8 (6.8)	5.9 (6.9)	
Family Size				<0.001
	4.3 (1.8)	3.1 (1.6)	3.6 (1.8)	
Current Smoking				0.774
No	139 (90.3%)	227 (89.4%)	366 (89.7%)	
Yes	15 (9.7%)	27 (10.6%)	42 (10.3%)	
Anti.HTN				0.100
No	134 (87.0%)	205 (80.7%)	339 (83.1%)	
Yes	20 (13.0%)	49 (19.3%)	69 (16.9%)	
Partner Status				<0.001
No partner	54 (35.1%)	203 (79.9%)	257 (63.0%)	
Partnered	100 (64.9%)	51 (20.1%)	151 (37.0%)	
Education				0.076
H.S. or less	49 (31.8%)	78 (30.7%)	127 (31.1%)	
Some college	41 (26.6%)	46 (18.1%)	87 (21.3%)	
College or higher	64 (41.6%)	130 (51.2%)	194 (47.5%)	
Income				<0.001
<35k	20 (13.0%)	79 (31.1%)	99 (24.3%)	
35k-50k	25 (16.2%)	60 (23.6%)	85 (20.8%)	
50k-75k	34 (22.1%)	57 (22.4%)	91 (22.3%)	
>75k	69 (44.8%)	56 (22.0%)	125 (30.6%)	
Refused or Don't Know	6 (3.9%)	2 (0.8%)	8 (2.0%)	

In figure 3.3 the top panels exhibit estimated curves from the exposure model. Consistent with the differences in means (table 3.1), the fitted GAMM results show that BWs have higher systolic and diastolic ABP, on average, compared to non-BWs. The bottom row exhibits the differences in ABP profiles over time. Dotted vertical lines and colors overlaid on the x-axes delineate time intervals during which BWs' ABP was significantly higher than non-BWs. Intervals during which systolic ABP differed significantly between groups in the exposure model (figure 3.3) were much wider than those from the covariate-adjusted model (figure 3.4).

The exposure model shows that, marginally, BWs on average had significantly higher systolic ABP approximately 8:15 am - 12 pm and 7:30 pm - 11 pm. Diastolic ABP was higher for BWs during the hours of approximately 8:45 am - 10:45 am and for a short time around 10 pm. After adjusting for covariates, significant fixed-effect predictors of SBP were BW-status ($\beta(SE) = 3.0(1.33), p = 0.024$), BMI ($\beta(SE) = 0.18(0.07), p = 0.018$), Family Size ($\beta(SE) = -0.81(0.35), p = 0.020$), and use of anti-hypertensive medication ($\beta(SE) = 8.5(1.58), p < 0.001$). Time intervals of significantly higher systolic ABP occurred around 9 am - 11 am and during a short period around 10 pm. Significant fixed-effect predictors of DBP were BW-status ($\beta(SE) = 2.5(0.98), p = 0.010$), Partner Status ($\beta(SE) = 2.2(1.08), p = 0.039$), and use of anti-hypertensive medication ($\beta(SE) = 6.8(1.17), p < 0.001$). Time intervals for higher DBP in breadwinners were about the same as those for SBP.

We used the same models to analyze data from participants who completed at least 70% of the intended ABP readings. This restricted the sample to observations from $n = 365$ women. There was no statistically significant difference in the proportion of BWs versus non-BWs with fewer than 70% of the intended readings. Findings in this group were similar to those in the overall sample for ABP profiles (Appendix 3.3).

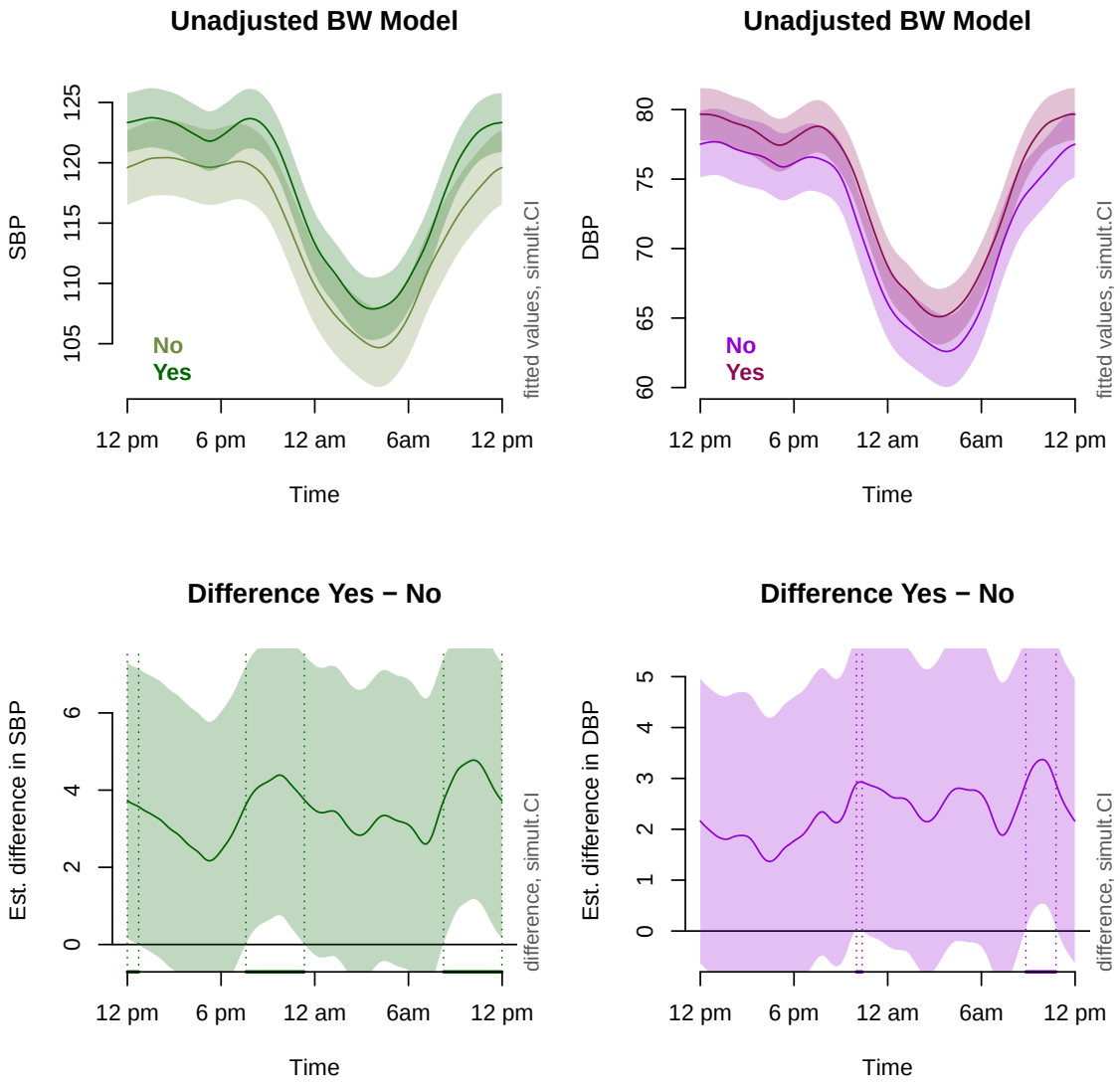


Figure 3.3: Fitted ABP profiles from the exposure model (top row) exhibit the estimated average ABP profiles for BWs vs non-BWs. Estimated difference curves with simultaneous confidence bands (bottom row) show time intervals during which average ABP is significantly different for the two groups.

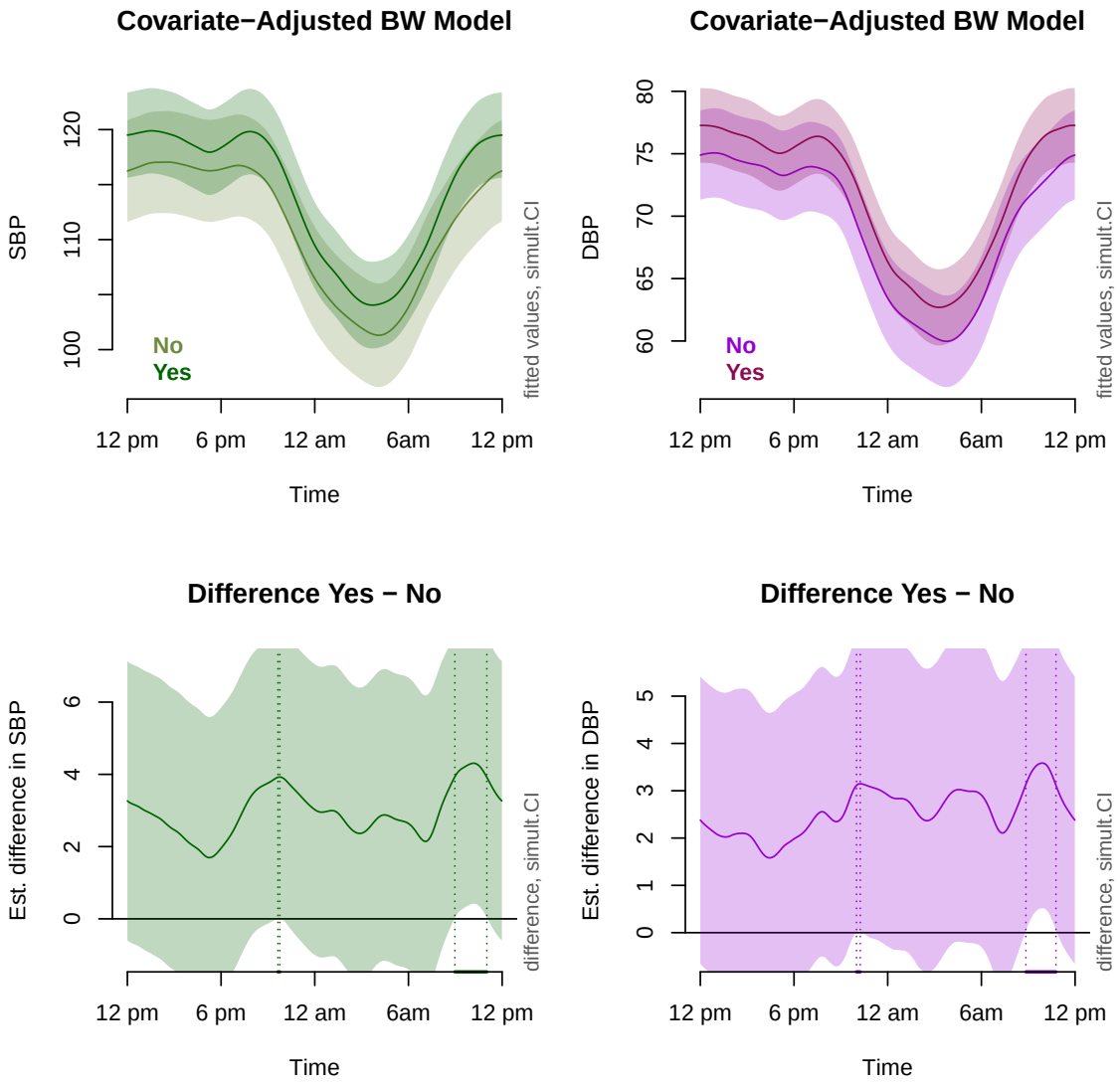


Figure 3.4: Fitted ABP profiles from the 'Stage 3' (covariate-adjusted) model (top row) exhibit the estimated average ABP profiles for BWs vs non-BWs. Estimated difference curves with simultaneous confidence bands (bottom row) show time intervals during which average ABP is significantly different for the two groups.

3.5 Discussion

We proposed a model of mean ABP profiles for BWs versus non-BWs using recent statistical methodology to capture curvilinear relationships in a regression setting. Previous statistical analyses of ABP have used summary measures, which precludes detecting differences within time intervals, or used shape-constrained parametric methods, which may not accurately capture the diurnal patterns of ABP. The proposed GAMMs provide additional nuance when compared to the use of summary measures. It allows for detection of time intervals during which average ABP differed significantly. While other methods can also achieve this goal, many do not allow the flexibility of GAMMs as they constrain the profile shape to be piece-wise linear, polynomial, or sinusoidal. The penalized spline approach used in GAMM avoids this issue.

Results confirm that the differences in ABP profiles are mostly time-invariant, as indicated by the differences in means. However, results from analyses of mean do not capture the additional difference that occurs specifically during late morning hours. Specifically, the largest differences between BWs and non-BWs occurred during morning hours. Mornings surges in BP have been identified as a potential risk factor for adverse cardiovascular events which occur in the morning [28]. Results suggest that interventions which aim to dampen mornings surge might be helpful for Black women BWs.

Our study examines relationships between BW-status and average ABP profiles but does not model individual ABP profiles. BP is quite variable from person to person and within each person. Including only subject-specific random intercepts, not subject-specific smooths of ABP profiles, misses a substantial source of variability. Results presented here apply to the averages BWs and non-BWs. More modelling is needed to address the expected difference between individuals.

Appendix 1: Canonical JIVE

This appendix provides additional details and supplementary information about interpreting AJIVE analysis by using canonical correlation analysis on the estimated signal matrices (i.e. CJIVE). In Section A, we describe the AJIVE and R.JIVE algorithms for model estimation and the equivalence between AJIVE joint scores and CJIVE joint scores. In Section B, additional information about and results from the simulation study described in the main article are provided. Section C presents additional information related to the JIVE analysis of HCP data in the main article.

CJIVE Appendix 1.1: Statistical Methodology

Here we outline the algorithms of R.JIVE and AJIVE. Cursory descriptions of each method are given in the main article. R.JIVE employs permutation tests to estimate the joint and individual signal ranks within each of the datasets analyzed. Principal-angle analysis (PAA) is used in the AJIVE method to determine joint rank. Scree plots are recommended for choosing total rank in AJIVE, although other options are explored in the main article.

CJIVE Appendix 1.1.1: JIVE Methods

For a collection of K data matrices JIVE decomposes each matrix $\mathbf{X}_k$, $k = 1, \ldots, K$ into a joint signal, $\mathbf{J}_k$, individual signal, $\mathbf{A}_k$, and additive noise $\mathbf{E}_k$. Let $\mathbf{X}_k \in \mathbb{R}^{n \times p_k}$,

where n is the number of subjects and p_k the number of features or variables. Let $C(\mathbf{G})$ define the column space of a matrix $\mathbf{A}$, i.e., $C(\mathbf{G}) = \{\mathbf{v} \in \mathbb{R}^n : \exists \mathbf{t} \in \mathbb{R}^p \text{ such that } \mathbf{v} = \mathbf{G}\mathbf{t}\}$.

$$\begin{aligned} \mathbf{X}_k &= \mathbf{J}_k + \mathbf{A}_k + \mathbf{E}_k, \quad \text{where} \\ C(\mathbf{J}_k) &= C(\mathbf{J}_{k'}), \text{ for all } k, k' \subset 1, \ldots, K, \\ C(\mathbf{J}_k) &\perp C(\mathbf{A}_k) \text{ for } k = 1, \ldots, K, \\ \mathbb{E}(\mathbf{E}_k) &= \mathbf{0}_{n \times p_k}. \end{aligned} \tag{3}$$

Let r_k denote the signal rank of the kth dataset, r_J denote the joint rank, and r_{Ik} denote the rank of the individual subspace. Because we assume that the joint and individual signals are orthogonal, $r_k = r_J + r_{Ik}$. In this study, we focus on $K = 2$.

CJIVE Appendix 1.1.2: R.JIVE Estimation

In addition to the assumptions in (3), R.JIVE requires the error matrices $\mathbf{E}_k$ to have independent entries. In Lock et al. [27], two automated methods were proposed for choosing the joint rank. One of these is a permutation test; the other, a strategy that utilizes Bayesian Information Criterion (BIC). In the case that either of the automated rank choice methods is used, R.JIVE will also compute the mean squared error (MSE) between consecutive estimates of the total signal matrices $\mathbf{G}_k = \mathbf{J}_k + \mathbf{A}_k$ in an iterative process which simultaneously chooses ranks and estimates signals.

Algorithm:

1. Calculate the centered/scaled matrices $\mathbf{X}_{1,cs}$, $\mathbf{X}_{2,cs}$:

$$\mathbf{X}_{k,cs} = ||\mathbf{X}_k||_F^{-1}(\mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top)\mathbf{X}_k, \quad k = 1, 2$$

where $||\cdot||_F$ denotes the Frobenius norm and $\mathbf{1}_n$ is an $n \times 1$ vector of ones.

2. Estimate ranks r_J, r_1 and r_2 with n_{perm} permutations and significance level $\alpha \in (0,1)$ (e.g., $n_{perm} = 100$ and $\alpha = 0.05$):

   (a) Estimate r_J:

      i. Let λ_j be the j^{th} singular value of $\mathbf{X} = [\mathbf{X}_{1,cs}, \mathbf{X}_{2,cs}]$ for $j = 1, \ldots, \min(n, p_1 + p_2)$

      ii. Permute the rows within each $\mathbf{X}_{k,cs}$ and calculate the singular values of the resultant concatenated matrix. Repeat n_{perm} times.

      iii. Let λ_j^{perm} be the $100(1-\alpha)$ percentile calculated from the n_{perm} samples of the j^{th} singular value from permuted data.

      iv. Choose r_J to be the largest integer such that $\forall j \leq r_J, \lambda_j > \lambda_j^{perm}$.

   (b) Estimate r_1 and r_2:

      i. Let λ_{j_k} be the j^{th} singular value of $\mathbf{X}_k$ for $k = 1, 2$ and $j_k = 1, \ldots, \min(n, p_k)$

      ii. Permute the rows separately within each column of $\mathbf{X}_k$ and calculate the singular values of the resultant matrix. Repeat n_{perm} times.

      iii. Let $\lambda_{j_k}^{perm}$ be the $100(1-\alpha)$ percentile among the j_k^{th} singular values after permutation.

      iv. Choose r_k to be the largest integer such that $\forall j_k \leq r_k, \lambda_{j_k} > \lambda_{j_k}^{perm}$

3. Use the estimates of $r_J, r_1,$ and r_2 found with the permutation tests above to estimate the signal matrices $\mathbf{J} = [\mathbf{J}_1, \mathbf{J}_2]$ and $\mathbf{G} = [\mathbf{A}_1, \mathbf{A}_2]$. Loop until $\widehat{MSE} = \sum_{k=1}^{n} \sum_{l=1}^{p_1+p_2} \frac{(\hat{\mathbf{G}}^{new}[k,l] - \hat{\mathbf{G}}^{old}[k,l])^2}{n(p_1+p_2)}$ is less than a given threshold:

   (a) Initialize $\mathbf{X}_J = [\mathbf{X}_{1,cs}, \mathbf{X}_{2,cs}]$ and estimate $\hat{\mathbf{J}}$ as a rank r_J SVD of $\mathbf{X}_J$.

$$\hat{\mathbf{J}} = \mathbf{U}_{r_J} \mathbf{D}_{r_J} \mathbf{V}_{r_J}^{\top} = [\hat{\mathbf{J}}_1, \hat{\mathbf{J}}_2],$$

   (b) For $k = 1, 2$, set $\hat{\mathbf{X}}_k = \mathbf{X}_{k,cs} - \hat{\mathbf{J}}_k$ and estimate $\hat{\mathbf{A}}_k$ with a rank r_k SVD of

$(\mathbf{I}_n - \mathbf{U}_{r_J}\mathbf{U}_{r_J}^\top)\hat{\mathbf{X}}_k$

$$\hat{\mathbf{A}}_k = \mathbf{U}_{r_{Ik}}\mathbf{D}_{r_{Ik}}\mathbf{V}_{r_{Ik}}^\top.$$

(c) Set $\hat{\mathbf{X}}_J = [\mathbf{X}_{1,cs}, \mathbf{X}_{2,cs}] - [\hat{\mathbf{A}}_1, \hat{\mathbf{A}}_2]$

(d) Use procedure (i) in step (b) above to re-estimate r_J using the concatenated matrix $\hat{\mathbf{X}}_J$ from step iii. Similarly, re-estimate r_1 and r_2 by procedure (ii) in step (b) above using the the matrices $\mathbf{X}_{k,cs} - \hat{\mathbf{J}}_k, k = 1, 2$.

(e) Repeat (c) loop with $\mathbf{X}_J = \hat{\mathbf{X}}_J$

CJIVE Appendix 1.1.3: AJIVE Estimation

AJIVE imposes additional constraints on the model given in equation (3). The error matrices $\mathbf{E}_k$ both follow an isotropic error model, which implies that the energy of projection is invariant to the direction in both row and column spaces. The standard multivariate Gaussian distribution and the multivariate student t-distribution with shape matrix equal to the identity matrix are both examples of isotropic models ([8]). Furthermore, individual signals' vector subspaces have null intersection: $\mathbf{A}_1 \cap \mathbf{A}_2' = \mathbf{0}$.

The number of joint components r_J (i.e. the joint rank) is determined using principal-angle analysis (PAA). Define $\mathbf{U}_k$ as the orthonormal left singular vectors of matrices $\mathbf{G}_k$, $k = 1, 2$. Then we can write an SVD of their inner product as $\mathbf{U}_1^\top \mathbf{U}_2 = \mathbf{U}\cos(\mathbf{\Theta})\mathbf{V}^\top$, where $\mathbf{\Theta} = (\theta_1, \ldots, \theta_q)$ is the vector of principal angles between $C(\mathbf{U}_1)$ and $C(\mathbf{U}_2)$, and $\mathbf{U}, \mathbf{V}$ are the left- and right- singular vectors, respectively of the inner product $\mathbf{U}_1^\top \mathbf{U}_2$. The following theorem and subsequent lemma are used to develop a bound on the angle between two common subspaces perturbed by isotropic error, which is complimented by a random direction bound.

Theorem .0.1 ([56]). *For $k = 1 \ldots K$, let θ_k be the largest principal angle between the subspace spanned by $\mathbf{G}$ and that spanned by $\widetilde{\mathbf{G}}_k = \mathbf{G}_k + \mathbf{E}_k$ and denote the SVD*

of $\widetilde{\mathbf{G}}_k = \widetilde{\mathbf{U}}_k \widetilde{\mathbf{D}}_k \widetilde{\mathbf{V}}_k^\top$. *Then*

$$\sin(\theta_k) \leq \frac{\max(||\mathbf{E}_k \widetilde{\mathbf{V}}_k||, ||\mathbf{E}_k^\top \widetilde{\mathbf{U}}_k||)}{\sigma_{\min}(\widetilde{\mathbf{D}}_k)},$$

where $\sigma_{\min}(\widetilde{\mathbf{D}}_k)$ *is the minimal non-zero singular value of* $\widetilde{\mathbf{G}}_k$

Lemma, Feng et al. [8] Let ϕ be the largest principal angle between two subspaces that are each a perturbation of the common column space within $C(\widetilde{\mathbf{G}}_1)$ and $C(\widetilde{\mathbf{G}}_2)$. Suppose θ_1, θ_2 are the respective angles for $\widetilde{\mathbf{G}}_1, \widetilde{\mathbf{G}}_2$ from Theorem .0.1. Then ϕ is bounded by

$$\sin(\phi) \leq \sin(\theta_1 + \theta_2).$$

Below we outline the algorithm employed by the AJIVE method.

1. Data blocks are centered,

$$\mathbf{X}_{k,c} = (\mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top)\mathbf{X}_k, \quad k = 1, 2$$

2. Let $r_k = r_J + r_{Ik}$ represent the total signal rank of data matrix $\mathbf{X}_{k,sc}, (k = 1, 2)$. Take the Singular Value Decomposition (SVD) of each data block and obtain $\mathbf{X}_k = \mathbf{U}_k \mathbf{D}_k \mathbf{V}_k^\top$. Concatenate the first r_k left-singular vectors of both datasets to form $\mathbf{J} = [\widetilde{\mathbf{U}}_1, \widetilde{\mathbf{U}}_2] \in \mathbb{R}^{n \times (r_1 + r_2)}$.

3. Principal Angle Analysis:

   (a) Denote the residual left singular vectors from each SVD in step 2 above as $\mathbf{U}_k^\perp \in \mathbb{R}^{n \times (\min(p_k, n) - r_k)}$. Similarly, let $\mathbf{V}_k^\perp \in \mathbb{R}^{p_k \times (\min(p_k, n) - r_k)}$. be the residual right singular vectors from the SVDs in step 1 above.

   (b) To estimate the unobserved values $||\mathbf{E}_k \widetilde{\mathbf{V}}_k||$ and $||\mathbf{E}_k^\top \widetilde{\mathbf{U}}_k||$, sample from directions orthogonal to the signal space and project the data block onto

those sampled directions. For each k, sample r_k non-zero columns without replacement from $\widetilde{\mathbf{U}}_k$, denoted $\mathbf{U}_k^*$ and compute $||\mathbf{X}_k^\top \mathbf{U}_k^*||$. Perform this sampling and computation 1000 times to approximate the distribution of $||\mathbf{E}_k^\top \widetilde{\mathbf{U}}_k||$. Similarly, approximate the distribution of $||\mathbf{E}_k \widetilde{\mathbf{V}}_k||$ with a random sampling of $||\mathbf{X}_k \mathbf{V}_k^*||$ values.

(c) (*Wedin bound*) Obtain a full-rank, $\min(n, r_1 + r_2)$, SVD of $\mathbf{J} \approx \mathbf{U}_J \mathbf{D}_J \mathbf{V}_J^\top$. Let $d_{J,i}$ represent the i^{th} singular value of $\mathbf{J}$; i.e. the i^{th} of entry of $diag(\mathbf{D}_J)$. Then the Wedin bound is estimated by the distribution of

$$2 - \sum_{k=1}^{2} \left(\frac{\max(||\mathbf{X}_k \mathbf{U}_k^*||, ||\mathbf{X}_k^\top \mathbf{V}_k^*||)}{\sigma_{\min}(\bar{\mathbf{D}}_k)} \right)^2,$$

where sampled values of $||\mathbf{X}_k^\top \mathbf{U}_k^*||$ and $||\mathbf{X}_k \mathbf{V}_k^*||$ are used to approximate the unobserved values $||\mathbf{E}_k \widetilde{\mathbf{U}}_k||$ and $||\mathbf{E}_k^\top \widetilde{\mathbf{V}}_k||$, respectively, as in the procedure explained in step (b) above. The 95^{th} percentile is used as the bound.

(d) (*Random direction bound*) The random direction bound aims to assess whether angles between directions in the proposed signal space correspond to random directions driven by noise. The distribution of principal angles generated by random subspaces is simulated as follows. Each $\widetilde{\mathbf{U}}_k$ is right-multiplied by an independent orthonormal matrix to obtain $\mathbf{U}_k^{**}$. The 95^{th} percentile of replicates of the principal angle derived from the maximum eigenvalue of $[\mathbf{U}_1^{**}, \mathbf{U}_2^{**}]$ gives the second bound. The joint rank r_J is then chosen as the number of eigenvalues $d_{J,i}$ exceeding both bounds.

4. The first r_J left singular vectors of $\mathbf{J}$ (corresponding to the r_J singular values obtained in PAA) represent a basis of the data matrices' estimated joint column space. Use them to form an orthogonal projection operator, $\mathbf{M}_J = \mathbf{U}_J \mathbf{U}_J^\top$, and project each dataset onto the estimated joint column space spanned by $\mathbf{U}_J$ to

obtain estimates of the joint signal matrices.

$$\hat{\mathbf{J}}_k = \mathbf{M}_J \mathbf{X}_{k,c}.$$

5. Lastly, the column space for each individual signal is found by computing an orthogonal projection operator onto the orthogonal complement of the joint column space with respect to the data's column space. The estimated individual signals are then given by:

$$\hat{\mathbf{A}}_k = (\mathbf{U}_k \mathbf{U}_k^\top - \mathbf{M}_J)\mathbf{X}_{k,c}.$$

CJIVE Appendix 1.1.2: Post AJIVE/R.JIVE Representations

As discussed in [8], joint scores and loadings can be computed to represent data-specific information or common information across datasets.

Let $\hat{\mathbf{J}}_{k(R)}, \hat{\mathbf{A}}_{k(R)}$ and $\hat{\mathbf{J}}_{k(A)}, \hat{\mathbf{I}}_{k(A)}$, for $k = 1, 2$ represent the signal matrices estimated using the R.JIVE and AJIVE methods, respectively. Take their SVDs:

$$\hat{\mathbf{J}}_{k(\cdot)} = \mathbf{U}_{\hat{\mathbf{J}}_{k(\cdot)}} \mathbf{\Sigma}_{\hat{\mathbf{J}}_{k(\cdot)}} \mathbf{V}_{\hat{\mathbf{J}}_{k(\cdot)}}^\top, \quad \hat{\mathbf{I}}_{k(\cdot)} = \mathbf{U}_{\hat{\mathbf{I}}_{k(\cdot)}} \mathbf{\Sigma}_{\hat{\mathbf{I}}_{k(\cdot)}} \mathbf{V}_{\hat{\mathbf{I}}_{k(\cdot)}}^\top.$$

Estimates of data-specific joint and individual subject scores are defined as $\mathbf{U}_{\hat{\mathbf{J}}_{k(\cdot)}} \mathbf{\Sigma}_{\hat{\mathbf{J}}_{k(\cdot)}}, \mathbf{U}_{\hat{\mathbf{I}}_{k(\cdot)}} \mathbf{\Sigma}_{\hat{\mathbf{I}}_{k(\cdot)}}$, respectively. Data-specific joint and individual variable loadings are $\mathbf{V}_{\hat{\mathbf{J}}_{k(\cdot)}}, \mathbf{V}_{\hat{\mathbf{I}}_{k(\cdot)}}$, respectively.

For AJIVE estimates, define $\mathbf{C}_{(A)} = [\mathbf{U}_1, \mathbf{U}_2]$ as in section A.1.1. For R.JIVE, let $\mathbf{C}_{(R)} = [\hat{\mathbf{J}}_{1(R)}, \hat{\mathbf{J}}_{2(R)}]$. Compute the rank r_J SVD of $\mathbf{C}_{(\cdot)}$:

$$\mathbf{C}_{(\cdot)} = \mathbf{U}_{\mathbf{J}_{(\cdot)}} \mathbf{\Sigma}_{\mathbf{J}_{(\cdot)}} \mathbf{V}_{\mathbf{J}_{(\cdot)}}^\top.$$

The common normalized joint scores are defined as $\mathbf{U}_{\mathbf{J}_{(\cdot)}}$. The common normalized

variable loadings for each dataset are given by $\mathbf{J}_{k(\cdot)}^{\top}\mathbf{U}_{\mathbf{J}_{(\cdot)}}$.

CJIVE Appendix 1.1.3: Proof of Equivalence between AJIVE and CCA Estimators

The main article formalizes the result showing equivalence between CCA estimators of joint subject scores and joint subject scores derived from AJIVE in the following theorem:

Theorem .0.2. *Let the columns of* $\mathbf{U}_1$ *and* $\mathbf{U}_2$ *represent orthonormal bases for the signal subspaces of* $\mathbb{R}^n$ *for* $\mathbf{X}_1$ *and* $\mathbf{X}_2$*, respectively. The* i^{th} *joint subject score from AJIVE analysis,* $\mathbf{u}_{Ji}$*, is given by the* i^{th} *column of* $\mathbf{U}_{\mathbf{J}}$*, the left singular vectors of* $\mathbf{C} = [\mathbf{U}_1, \mathbf{U}_2]$ *with singular value* σ_{Ji}*. Let* $\omega_{1,i}$ *and* $\omega_{2,i}$*, represent the canonical loadings of the signal subspaces, respectively. Then*

$$\mathbf{u}_{Ji} = \frac{1}{\sqrt{2}\sigma_{Ji}}(\mathbf{U}_1\omega_{1,i} + \mathbf{U}_2\omega_{2,i}).$$

That is, AJIVE estimates of joint subject scores are equivalent to a scaled average of the j^{th} *canonical variates of the signal subspaces.*

Proof. Let $\hat{\mathbf{G}}_k = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^{\top}$, where $\hat{\mathbf{G}}_k$ is the estimated signal matrix for the kth dataset with rank $= r_k < min(n, p_k)$. Note that $\mathbf{U}_k$ represents the scores from the PCA of the kth dataset. We first consider the first canonical variables. Define the first canonical loadings of the PC scores:

$$\{\hat{\omega}_{1,1},\ \hat{\omega}_{2,1}\} = \underset{\omega_1 \in \mathbb{R}^{r_1}, \omega_2 \in \mathbb{R}^{r_2}, ||\omega_1|| = ||\omega_2|| = 1}{\arg\max} \omega_1^{\top}\mathbf{U}_1^{\top}\mathbf{U}_2\omega_2.$$

Since $\mathbf{U}_1$ and $\mathbf{U}_2$ are both column centered and orthonormal, their variance-covariance matrices are given by $\mathbf{\Sigma}_{11} = \mathbf{U}_1^{\top}\mathbf{U}_1 = \mathbf{I}_{r_1}$ and $\mathbf{\Sigma}_{22} = \mathbf{U}_2^{\top}\mathbf{U}_2 = \mathbf{I}_{r_2}$, respectively, where $\mathbf{I}_m$ represents the $m \times m$ identity matrix. Similarly, their covariance is given by $\mathbf{\Sigma}_{12} =$

$\mathbf{U}_1^\top \mathbf{U}_2$. Mardia et al. [29] provides a closed form solution for the loadings $\hat{\omega}_{1,1},\ \hat{\omega}_{2,1}$ as the first left and first right singular vectors, respectively, of $\mathbf{\Sigma}_{11}^{-1/2}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1/2} = \mathbf{U}_1^\top \mathbf{U}_2$, up to sign. That is, $\hat{\omega}_{1,1}$ is equivalent to the first left singular vector of $\mathbf{U}_1^\top \mathbf{U}_2$, and $\hat{\omega}_{2,1}$ is equivalent to the first right singular vector, where their signs are chosen so that the singular values of $\mathbf{U}_1^\top \mathbf{U}_2$ are all positive. Then the canonical variables are $\hat{\mathbf{c}}_{1,1} = \mathbf{U}_1\hat{\omega}_{1,1}$ and $\hat{\mathbf{c}}_{2,1} = \mathbf{U}_2\hat{\omega}_{2,1}$, and note $||\hat{\mathbf{c}}_{1,1}|| = 1$. Define the first canonical correlation $\rho_1 = \hat{\mathbf{c}}_{1,1}^\top \hat{\mathbf{c}}_{2,1}$. Consider the matrix of concatenated PC scores $[\mathbf{U}_1, \mathbf{U}_2] \in \mathbb{R}^{n \times r_1 + r_2}$.

Then define

$$\hat{\mathbf{a}}_1 = [\mathbf{U}_1, \mathbf{U}_2][\hat{\omega}_{1,1}^\top, \hat{\omega}_{2,1}^\top]^\top = \hat{\mathbf{c}}_{1,1} + \hat{\mathbf{c}}_{2,1}.$$

Next we show that $\hat{\mathbf{a}}_1$ is equivalent to the JIVE solution by showing $\hat{\mathbf{a}}_1$ is the first left singular vector of $\mathbf{C} = [\mathbf{U}_1, \mathbf{U}_2]$ up to scaling. Let $\omega^* = [\hat{\omega}_{1,1}^\top, \hat{\omega}_{2,1}^\top]^\top$. Since $\hat{\omega}_{1,1}$ and $\hat{\omega}_{2,1}$ are left and right singular vectors of $\mathbf{U}_1^\top \mathbf{U}_2$, respectively, we have $\mathbf{U}_1^\top \mathbf{U}_2 \hat{\omega}_{2,1} = \rho_1 \hat{\omega}_{1,1}$ and $\mathbf{U}_2^\top \mathbf{U}_1 \hat{\omega}_{1,1} = \rho_1 \hat{\omega}_{2,1}$. Additionally, $\mathbf{C}^\top \mathbf{C} = \begin{bmatrix} \mathbf{I} & \mathbf{U}_1^\top \mathbf{U}_2 \\ \mathbf{U}_2^\top \mathbf{U}_1 & \mathbf{I} \end{bmatrix} = \mathbf{I} + \begin{bmatrix} \mathbf{0} & \mathbf{U}_1^\top \mathbf{U}_2 \\ \mathbf{U}_2^\top \mathbf{U}_1 & \mathbf{0} \end{bmatrix}$. Thus,

$$\begin{aligned} \mathbf{C}^\top \mathbf{C} \omega^* &= \omega^* + \begin{bmatrix} \mathbf{U}_1^\top \mathbf{U}_2 \hat{\omega}_{2,1} \\ \mathbf{U}_2^\top \mathbf{U}_1 \hat{\omega}_{1,1} \end{bmatrix} \\ &= (1 + \rho_1)\omega^*. \end{aligned}$$

Hence, $\omega^*/||\omega^*|| = \omega^*/\sqrt{2}$ is the first normalized eigenvector of $\mathbf{C}^T\mathbf{C}$, which is the first right singular vector of $\mathbf{C}$. Let $\mathbf{C} = \mathbf{U_J \Sigma_J V_J^\top}$ be the SVD of $\mathbf{C}$, wherein the first r_J columns of $\mathbf{U_J}$ are the joint components from the JIVE decomposition and $\mathbf{\Sigma_J}$ has diagonal elements σ_j. Let $\hat{\mathbf{u}}_{J,1}$ be the first joint component. Note the first

row of $\mathbf{V}_\mathbf{J}^\top$ is equal to ω^*. Then

$$\begin{aligned}\mathbf{C}\frac{\omega^*}{\sqrt{2}} &= \frac{1}{\sqrt{2}}[\mathbf{U}_1, \mathbf{U}_2][\hat{\omega}_{11}^\top, \hat{\omega}_{21}^\top]^\top \\ &= \frac{\hat{\mathbf{c}}_{1,1} + \hat{\mathbf{c}}_{2,1}}{\sqrt{2}} \\ &= \sigma_1 \hat{\mathbf{u}}_{J,1} \\ &= \sqrt{(1+\rho_1)}\hat{\mathbf{u}}_{J,1}\end{aligned}$$

which corresponds to the first joint component form the AJIVE decomposition and therefore leads to the following equivalence between common scores and canonical variables:

$$\hat{\mathbf{u}}_{J,1} = \frac{\hat{\mathbf{c}}_{1,1} + \hat{\mathbf{c}}_{2,1}}{\sqrt{2(1+\rho_1)}}.$$

A similar argument applies to the other joint components. □

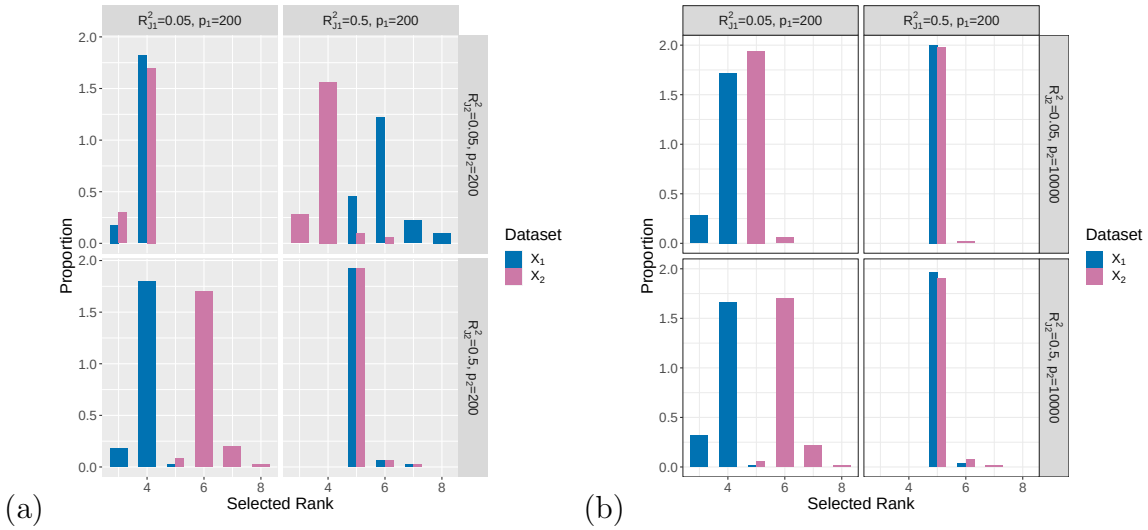
CJIVE Appendix 1.2: Simulation study

A simulation study was conducted according to a 2^3 full factorial design, in order to examine the effectiveness of AJIVE and R.JIVE for estimating the JIVE model and our proposed method for estimating the number of joint components the JIVE model.

In order to achieve the pre-described proportions of variance explained via joint and individual signals, we derive numerical solutions to:

$$R^2_{Ik} = \frac{c_k^2 \text{tr}(\mathbf{A}_k\mathbf{A}_k^\top)}{c_k^2 \text{tr}(\mathbf{A}_k\mathbf{A}_k^\top) + 2c_k\text{tr}(\mathbf{A}_k\mathbf{E}^\top) + d_k^2\text{tr}(\mathbf{J}_k\mathbf{J}_k^\top) + \text{tr}(\mathbf{E}_k\mathbf{E}_k^\top) + 2d_k\text{tr}(\mathbf{J}_k\mathbf{E}_k^\top)}$$

$$R^2_{Jk} = \frac{d_k^2 \text{tr}(\mathbf{J}_k\mathbf{J}_k^\top)}{d_k^2 \text{tr}(\mathbf{J}_k\mathbf{J}_k^\top) + 2d_k\text{tr}(\mathbf{J}_k\mathbf{E}^\top) + c_k^2\text{tr}(\mathbf{A}_k\mathbf{A}_k^\top) + \text{tr}(\mathbf{E}_k\mathbf{E}_k^\top) + 2c_k\text{tr}(\mathbf{A}_k\mathbf{E}_k^\top)}$$



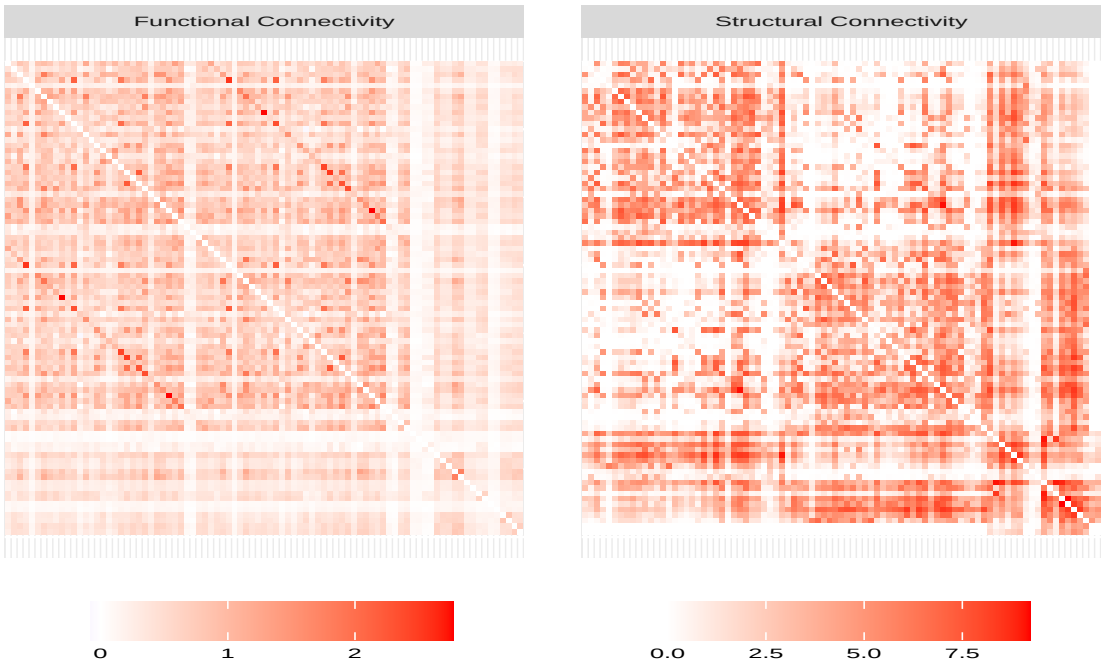
Web Figure 5: Total rank estimates from R.JIVE. The sub-figures (a) and (b) each exhibit results for $r_J = 3$, which implies that total ranks are $r_1 = r_2 = 5$.

Web Table 2: Computation Run-times (in minutes)

			R.JIVE	CJIVE-Over	CJIVE-Oracle	AJIVE-Over	AJIVE-Oracle
R^2_{J2}	R^2_{J2}	p_2			Mean (S.D.)		
0.05	0.05	200	0.4 (1.18)	0.2 (0.05)	0 (0)	1.7 (0.45)	0 (0.01)
0.5	0.05	200	6.6 (3.18)	0.1 (0)	0 (0)	1.2 (0.02)	0 (0)
0.05	0.5	200	5 (2.36)	0.1 (0)	0 (0)	1.2 (0.03)	0 (0)
0.5	0.5	200	1.5 (2.41)	0.1 (0.01)	0 (0)	0.9 (0.06)	0 (0)
0.05	0.05	10000	4.2 (0.75)	0.3 (0.09)	0.1 (0.01)	21.1 (2.95)	0.5 (0.06)
0.5	0.05	10000	11.1 (4.64)	0.2 (0.03)	0.1 (0.01)	21.4 (2.45)	0.5 (0.07)
0.05	0.5	10000	8.3 (3.88)	0.3 (0.01)	0.1 (0.01)	18 (1.69)	0.5 (0.03)
0.5	0.5	10000	7.2 (2.79)	0.2 (0.03)	0.1 (0.02)	17.5 (2.17)	0.5 (0.08)

Accuracy of joint rank selection is discussed and examined closely in the primary manuscript. However, R.JIVE also chooses individual signal ranks, and therefore total ranks. Web Figure 5 exhibits the total ranks chosen via R.JIVE, which are nearly always accurate when $R^2_{J1} = R^2_{J2} = 0.5$. When $p_2 = 10,000$ (sub-figure (a)), total rank selection is most accurate for cases with $R^2_{J2} = 0.5$. Notably, total ranks are nearly always underestimated when joint variation is small in both datasets.

Computation times (Web Table 1) show that CJIVE Computes solutions between twice and 100 times as faster than AJIVE or R.JIVE.



Web Figure 6: Mean functional connectivity (Fisher z-transformed correlations, left) and structural connectivity (log streamline counts, right) for the n = 998 HCP participants with data from both DTI and rs-fMRI available.

Web Table 3: Demographics of HCP Imaging Data

Descriptive Statistics (n = 998)					
Age		***Sex***		***Fluid Intelligence (gF)***	
21-25:	218 (21.8%)	*Female*:	532 (53.3%)	*Mean*:	17.04
26-30:	429 (43.0%)	*Male*:	466 (46.7%)	*S.D*:	4.70
31+:	351 (35.2%)			*Median*:	18.0

CJIVE Appendix 1.3: Human Connectome Project

The main article uses JIVE to simultaneously examine functional connectivity arising from resting-state fMRI scans and structural connectivity from diffusion-weighted MRI scans in the Human Connectome Project. Specifically, we used resting-state scans of the name rfMRI_REST1_LR_Atlas_hp2000_clean.dtseries.nii with subject-specific Desikan labels from <subj>.aparc.32k_fs_LR.dlabel.nii. Additional details are in the main manuscript. Here we present the images of mean SC and FC networks (Web Figure 6) and a summary of demographics (Web Table 2).

CJIVE Appendix 1.3.1: Dimension Selection

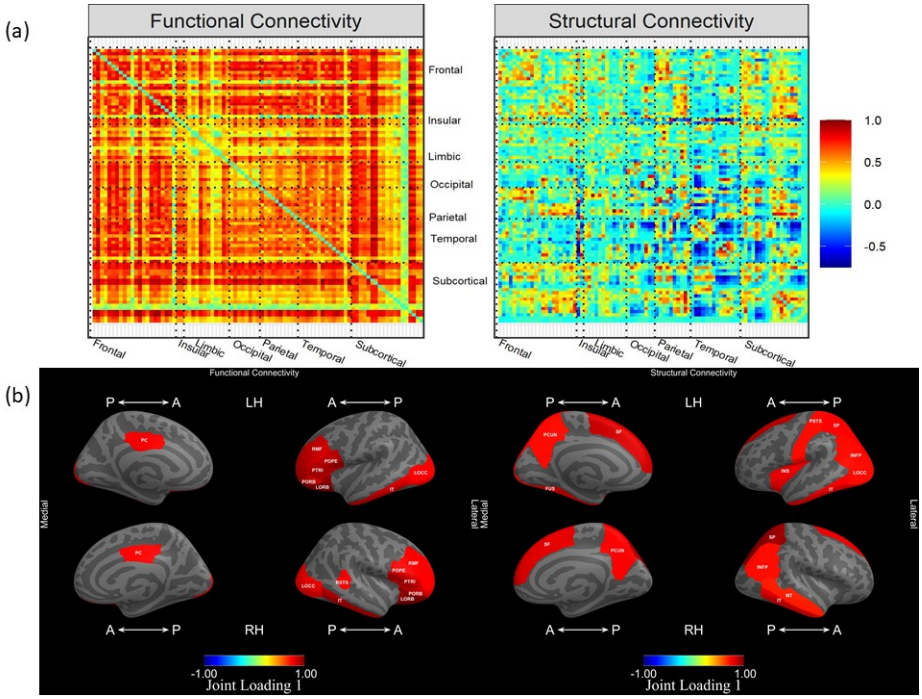
Total signal ranks (the total number of joint and individual components) were chosen using three methods: 1) "elbow" method of scree plot, 2) the number of eigenvalues which account for 95% of the sum of eigenvalues, and 3) R.JIVE permutation tests. Joint ranks were also chosen using with three methods: 1) PAA in AJIVE, 2) the permutation test presented above in R.JIVE, and 3) canonical correlation permutations, presented in the main article. Both the total signal rank chosen for each dataset and the joint signal rank are shown in the main article. It is notable that the joint rank chosen by AJIVE and CJIVE depends on the total rank chosen initially. In practice, more total signal components generally leading to fewer joint signal components.

CJIVE Appendix 1.3.2: CJIVE Variable Loadings

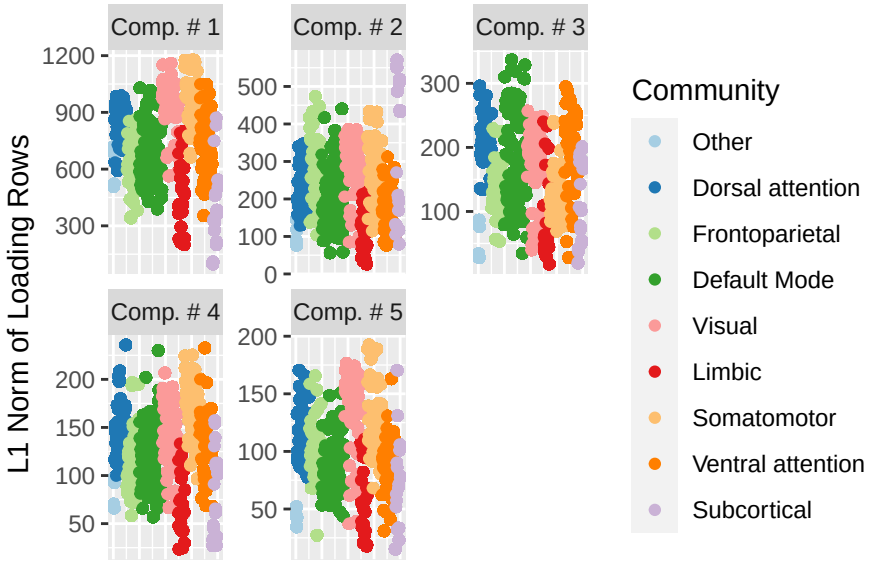
Web Figure 3 displays heatmaps and L1 norms of loadings onto the second component of the joint subspace for each data block. The second joint component was not statistically associated with fluid intelligence (gF). However, the hemispheric symmetry of the 75^{th} percentile of the loadings' L1 norms reveal biological relevance.

Web Figure 4 displays individual loadings for FC. The first component's largest values occur for edges connecting cortical regions, while the largest values for components 2 and 3 occur for edges connecting cortical and subcortical ROIs.

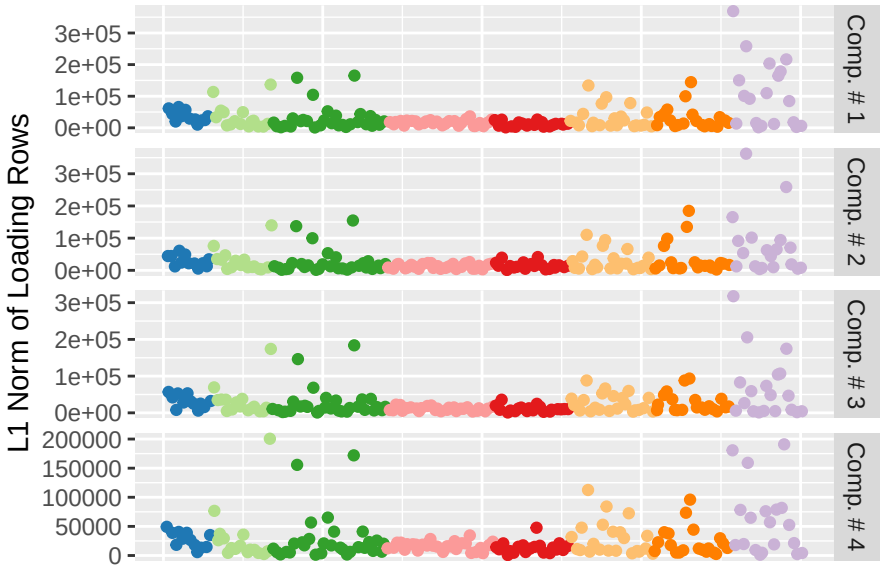
Individual loadings for SC, displayed in Web Figure 5, are much more sparse than loadings onto the first component of the joint subspace.



Web Figure 7: (a) Variable loadings for the second component of the joint signal space estimated by CJIVE and displayed on heatmaps. Sub-figure (b) displays the top 25^{th} percent of L1 norms of the variable loadings related to each cortical ROI for joint component 1. L1 norm for an ROI equals the sum of the absolute values of the rows of (a), excluding subcortical regions.



Web Figure 8: Heatmaps of variable loadings for each component of the FC individual subspace.



Web Figure 9: Heatmaps of variable loadings for each component of the SC individual subspace.

Appendix 2: Probabilistic JIVE

This appendix provides additional details and supplementary information about from the analyses conducted in Chapter 2.

ProJIVE Appendix 2.1: Summary Statistics for Brain Morphometry and Cognition/Behavior

Table 4 provides summary statistics for the cognitive measures that were included in our ProJIVE analyses of TADPOLE/ADNI data. Tables5 - 8 provide summary statistics for the measures of brain morphometry that were used. These tables are split by the type of measure being presented. Respectively, tables5 - 8 present summaries for cortical thickness, cortical volume, cortical surface area, and volume of subcortical and white-matter structures. For all five tables, we computed the mean and standard deviation for the entire sample and stratified by diagnosis. P-values are from an ANOVA that tests for differences in each measure by diagnosis, ignoring the ordinal nature of the diagnosis variable.

Table 4: Summary statistics for Cognition Measures.

	AD (N=88)	MCI (N=340)	CN (N=159)	Total (N=587)	
Cognition Score			Mean (SD)		p value
CDRSB	4.8 (1.84)	1.4 (0.97)	0.1 (0.33)	1.6 (1.79)	<0.001
ADAS11	20.4 (6.52)	9.1 (4.69)	5.5 (2.76)	9.8 (6.60)	
ADAS13	30.8 (8.06)	14.3 (7.03)	8.5 (4.35)	15.2 (9.63)	<0.001
MMSE					<0.001
MMSE	23.1 (3.05)	27.8 (2.00)	29.0 (1.14)	27.4 (2.74)	<0.001
RAVLT (immediate)	20.5 (6.18)	34.7 (10.84)	42.9 (9.35)	34.8 (12.08)	<0.001
RAVLT (learning)	1.3 (1.75)	4.4 (2.51)	5.5 (2.19)	4.3 (2.67)	<0.001
RAVLT (forgetting)	4.0 (1.61)	4.7 (2.45)	4.2 (2.80)	4.46 (2.46)	0.019
MOCA	17.8 (4.36)	23.5 (3.26)	25.9 (2.45)	23.3 (4.12)	<0.001
EcogPTMem	2.4 (0.75)	2.2 (0.74)	1.6 (0.46)	2.0 (0.74)	<0.001
EcogPTLang	1.9 (0.79)	1.8 (0.64)	1.4 (0.39)	1.7 (0.65)	<0.001
EcogPTVisspat	1.7 (0.71)	1.4 (0.54)	1.1 (0.22)	1.4 (0.53)	<0.001
EcogPTPlan	1.6 (0.72)	1.5 (0.53)	1.1 (0.26)	1.4 (0.54)	<0.001
EcogPTOrgan	1.8 (0.75)	1.6 (0.61)	1.3 (0.34)	1.5 (0.60)	<0.001
EcogPTDivatt	1.98 (0.81)	1.90 (0.76)	1.45 (0.58)	1.79 (0.75)	<0.001
EcogPTTotal	1.91 (0.66)	1.74 (0.53)	1.33 (0.30)	1.65 (0.54)	<0.001
EcogSPMem	3.32 (0.63)	2.09 (0.78)	1.29 (0.36)	2.06 (0.92)	<0.001
EcogSPLang	2.63 (0.75)	1.65 (0.68)	1.11 (0.21)	1.65 (0.77)	<0.001
EcogSPVisspat	2.47 (0.85)	1.41 (0.54)	1.07 (0.19)	1.48 (0.69)	<0.001
EcogSPPlan	2.66 (0.90)	1.55 (0.68)	1.10 (0.32)	1.60 (0.81)	<0.001
EcogSPOrgan	2.90 (0.85)	1.63 (0.78)	1.12 (0.26)	1.68 (0.88)	<0.001
EcogSPDivatt	3.09 (0.84)	1.88 (0.79)	1.24 (0.46)	1.89 (0.93)	<0.001
EcogSPTotal	2.84 (0.62)	1.71 (0.61)	1.15 (0.22)	1.73 (0.75)	<0.001

Table 5: Summary statistics for Cortical Thickness

	AD (N = 88)	MCI (N=340)	CN (N=159)	Total (N=587)	
ROI name			Mean (SD)		p value
L Bankssts	4704.5 (313)	3900.7 (1046.92)	4395.8 (388) 4155.3 (887.58)		
L Caudal Anterior Cingulate	2.2 (0.19)	2.3 (0.17)	2.3 (0.18)	2.3 (0.18)	0.01

L Caudal Middle Frontal	2.5 (0.34)	2.7 (0.33)	2.7 (0.25)	2.6 (0.32)	<0.001
L Cuneus	2.4 (0.2)	2.4 (0.15)	2.5 (0.15)	2.4 (0.16)	<0.001
L Entorhinal	2.5 (0.23)	2.5 (0.21)	2.6 (0.22)	2.5 (0.22)	0.003
L Frontal Pole	2.2 (0.17)	2.3 (0.15)	2.3 (0.15)	2.3 (0.16)	0.005
L Fusiform	1.5 (0.15)	1.5 (0.13)	1.5 (0.12)	1.5 (0.13)	0.069
L Inferior Parietal	1.8 (0.14)	1.9 (0.14)	1.9 (0.14)	1.9 (0.14)	<0.001
L Inferior Temporal	2.4 (0.2)	2.5 (0.17)	2.5 (0.19)	2.5 (0.18)	0.01
L Insula	2.3 (0.19)	2.3 (0.18)	2.4 (0.16)	2.3 (0.18)	<0.001
L Isthmus Cingulate	2.1 (0.18)	2.3 (0.15)	2.3 (0.14)	2.2 (0.16)	<0.001
L Lateral Occipital	2.8 (0.27)	2.8 (0.27)	2.8 (0.27)	2.8 (0.27)	0.828
L Lateral Orbitofrontal	2.1 (0.14)	2.2 (0.12)	2.2 (0.13)	2.2 (0.13)	0.001
L Lingual	2.4 (0.17)	2.5 (0.15)	2.5 (0.15)	2.5 (0.16)	<0.001
L Medial Orbitofrontal	2 (0.16)	2.1 (0.16)	2.1 (0.16)	2.1 (0.16)	<0.001
L Middle Temporal	2.4 (0.21)	2.6 (0.19)	2.6 (0.18)	2.6 (0.2)	<0.001
L Paracentral	2.3 (0.19)	2.4 (0.16)	2.4 (0.16)	2.4 (0.17)	<0.001
L Parahippocampal	3.2 (0.53)	3.5 (0.44)	3.7 (0.32)	3.5 (0.46)	<0.001
L Pars Opercularis	2.2 (0.25)	2.3 (0.26)	2.3 (0.23)	2.3 (0.25)	0.002
L Pars Orbitalis	2.8 (0.2)	2.9 (0.19)	3 (0.17)	2.9 (0.19)	<0.001
L Pars Triangularis	2.8 (0.22)	2.9 (0.18)	3 (0.18)	2.9 (0.19)	<0.001
L Pericalcarine	2.2 (0.21)	2.3 (0.2)	2.4 (0.17)	2.3 (0.2)	<0.001
L Postcentral	2.7 (0.35)	2.7 (0.33)	2.7 (0.32)	2.7 (0.33)	0.537
L Posterior Cingulate	2.3 (0.2)	2.4 (0.16)	2.4 (0.16)	2.4 (0.17)	<0.001

L Precentral	1.8 (0.16)	1.8 (0.14)	1.8 (0.14)	1.8 (0.15)	0.084
L Precuneus	2.6 (0.54)	3.3 (0.48)	3.5 (0.33)	3.2 (0.53)	<0.001
L Rostral Anterior Cingulate	2.5 (0.34)	2.6 (0.28)	2.7 (0.25)	2.6 (0.28)	<0.001
L Rostral Middle Frontal	2.4 (0.23)	2.6 (0.19)	2.7 (0.16)	2.6 (0.21)	<0.001
L Superior Frontal	2.1 (0.22)	2.3 (0.16)	2.3 (0.15)	2.3 (0.18)	<0.001
L Superior Parietal	2.5 (0.26)	2.7 (0.2)	2.7 (0.16)	2.7 (0.22)	<0.001
L Superior Temporal	2.3 (0.23)	2.4 (0.23)	2.5 (0.21)	2.4 (0.23)	<0.001
L Supramarginal	2 (0.2)	2.1 (0.17)	2.1 (0.16)	2.1 (0.17)	<0.001
L Temporal Pole	2.4 (0.2)	2.5 (0.16)	2.5 (0.16)	2.5 (0.17)	<0.001
L Transverse Temporal	1.8 (0.15)	1.9 (0.14)	1.9 (0.13)	1.9 (0.14)	<0.001
R Bankssts	2.3 (0.21)	2.3 (0.17)	2.4 (0.16)	2.3 (0.18)	<0.001
R Caudal Anterior Cingulate	2.5 (0.26)	2.7 (0.19)	2.8 (0.16)	2.7 (0.22)	<0.001
R Caudal Middle Frontal	2.2 (0.2)	2.3 (0.17)	2.3 (0.16)	2.3 (0.17)	0.052
R Cuneus	2.4 (0.36)	2.7 (0.39)	2.8 (0.33)	2.7 (0.38)	<0.001
R Entorhinal	2.3 (0.19)	2.4 (0.14)	2.4 (0.13)	2.4 (0.15)	<0.001
R Frontal Pole	2.5 (0.29)	2.6 (0.21)	2.6 (0.25)	2.5 (0.23)	0.006
R Fusiform	2.2 (0.19)	2.3 (0.16)	2.3 (0.17)	2.3 (0.17)	<0.001
R Inferior Parietal	1.5 (0.17)	1.5 (0.14)	1.5 (0.14)	1.5 (0.14)	0.783
R Inferior Temporal	1.9 (0.15)	1.9 (0.14)	1.9 (0.15)	1.9 (0.15)	<0.001
R Insula	2.4 (0.21)	2.5 (0.16)	2.5 (0.16)	2.5 (0.17)	0.004
R Isthmus Cingulate	2.3 (0.21)	2.4 (0.18)	2.4 (0.18)	2.4 (0.19)	<0.001
R Lateral Occipital	2.1 (0.18)	2.2 (0.16)	2.3 (0.15)	2.2 (0.16)	<0.001

R Lateral Orbitofrontal	2.8 (0.28)	2.9 (0.26)	2.9 (0.28)	2.8 (0.27)	0.027
R Lingual	2.1 (0.16)	2.2 (0.12)	2.2 (0.14)	2.2 (0.14)	<0.001
R Medial Orbitofrontal	2.4 (0.2)	2.5 (0.16)	2.6 (0.16)	2.5 (0.17)	<0.001
R Middle Temporal	2 (0.19)	2.1 (0.16)	2.1 (0.16)	2.1 (0.17)	<0.001
R Paracentral	2.4 (0.23)	2.6 (0.21)	2.6 (0.17)	2.6 (0.22)	<0.001
R Parahippocampal	2.2 (0.21)	2.4 (0.16)	2.4 (0.16)	2.4 (0.18)	<0.001
R Pars Opercularis	3.1 (0.5)	3.5 (0.39)	3.6 (0.31)	3.5 (0.42)	<0.001
R Pars Orbitalis	2.1 (0.25)	2.2 (0.25)	2.2 (0.23)	2.2 (0.25)	0.001
R Pars Triangularis	2.4 (0.22)	2.5 (0.18)	2.5 (0.18)	2.5 (0.19)	<0.001
R Pericalcarine	2.6 (0.34)	2.6 (0.3)	2.6 (0.3)	2.6 (0.31)	0.944
R Postcentral	2.3 (0.19)	2.4 (0.16)	2.4 (0.17)	2.4 (0.18)	<0.001
R Posterior Cingulate	1.8 (0.17)	1.8 (0.15)	1.8 (0.14)	1.8 (0.15)	0.397
R Precentral	2.8 (0.62)	3.4 (0.54)	3.6 (0.35)	3.4 (0.57)	<0.001
R Precuneus	2.5 (0.33)	2.5 (0.25)	2.5 (0.25)	2.5 (0.26)	0.675
R Rostral Anterior Cingulate	2.5 (0.22)	2.6 (0.2)	2.7 (0.17)	2.6 (0.21)	<0.001
R Rostral Middle Frontal	2.2 (0.2)	2.4 (0.16)	2.4 (0.15)	2.3 (0.18)	<0.001
R Superior Frontal	2.6 (0.23)	2.7 (0.2)	2.8 (0.17)	2.7 (0.21)	<0.001
R Superior Parietal	2.3 (0.21)	2.4 (0.23)	2.4 (0.23)	2.4 (0.23)	<0.001
R Superior Temporal	2.1 (0.17)	2.2 (0.17)	2.2 (0.16)	2.2 (0.17)	<0.001
R Supramarginal	2.4 (0.21)	2.5 (0.17)	2.5 (0.16)	2.5 (0.18)	<0.001
R Temporal Pole	1.9 (0.14)	1.9 (0.15)	1.9 (0.15)	1.9 (0.15)	<0.001

R Transverse Temporal	2.3 (0.24)	2.3 (0.19)	2.3 (0.16)	2.3 (0.19)	0.261

Table 6: Summary statistics for Cortical Volume

	AD (N = 88)	MCI (N=340)	CN (N=159)	Total (N=587)	
ROI name		Mean (SD)			p value
Intercranial	2.6 (0.25)	2.8 (0.18)	2.8 (0.16)	2.8 (0.2)	<0.001
L Bankssts	1453.1 (203.41)	1442.2 (198.06)	1425.6 (193.11)	1439.3 (197.42)	0.533
L Caudal Anterior Cingulate	620.1 (92.12)	639.8 (105.36)	638.8 (84.15)	636.5 (98.18)	0.232
L Caudal Middle Frontal	1351.9 (238.26)	1307.2 (215.44)	1284.4 (199.33)	1307.7 (215.48)	0.062
L Cuneus	706.7 (104.71)	725.8 (94.94)	721.4 (96.7)	721.7 (96.99)	0.261
L Entorhinal	1387.3 (239.91)	1380.2 (230.21)	1362.1 (204.15)	1376.4 (224.77)	0.622
L Frontal Pole	1448.9 (214.83)	1453.4 (219.96)	1412 (221.06)	1441.5 (219.87)	0.138
L Fusiform	3988.5 (506.39)	3949.2 (506.76)	3831.3 (433.39)	3923.2 (490.51)	0.017
L Inferior Parietal	1098 (207.84)	1106.1 (172.36)	1093.7 (161.32)	1101.6 (175.06)	0.747
L Inferior Temporal	4845.7 (555.83)	4789.6 (605.22)	4670.2 (503.67)	4765.6 (574.4)	0.035
L Insula	3624.5 (408.27)	3715.5 (494.61)	3617.2 (460.55)	3675.3 (475.08)	0.054
L Isthmus Cingulate	649.6 (147.38)	647.4 (132.29)	642.5 (136.25)	646.4 (135.5)	0.906
L Lateral Occipital	5425.2 (709.93)	5574.9 (732.02)	5475 (685)	5525.4 (717.6)	0.128

L Lateral Orbitofrontal	6497 (815.81)	6602.6 (781.71)	6443.3 (733.71)	6543.6 (776.23)	0.085
L Lingual	5145.9 (600.69)	5171.1 (572.27)	5048.3 (569.18)	5134.1 (577.22)	0.084
L Medial Orbitofrontal	3403.8 (344.07)	3412.7 (338.09)	3388 (323.51)	3404.7 (334.72)	0.745
L Middle Temporal	3444 (474.4)	3532.7 (492.17)	3447.8 (400.01)	3496.4 (467.44)	0.087
L Paracentral	413 (75.1)	421.8 (63.04)	421.5 (65.24)	420.4 (65.53)	0.52
L Parahippocampal	325.9 (55.44)	329.2 (57.18)	324 (54.88)	327.3 (56.26)	0.608
L Pars Opercularis	2166.1 (231.89)	2184.5 (262.81)	2158.3 (244.67)	2174.6 (253.42)	0.528
L Pars Orbitalis	2234.6 (275.93)	2270.1 (294.05)	2234.2 (279.46)	2255 (287.56)	0.333
L Pars Triangularis	921 (139.94)	975.1 (158.55)	968.2 (150.96)	965.1 (154.75)	0.013
L Pericalcarine	614.9 (110.93)	633.6 (163.05)	617.2 (126.74)	626.3 (147.01)	0.371
L Postcentral	2212.6 (343.27)	2182.6 (338.27)	2143.8 (346.77)	2176.6 (341.5)	0.28
L Posterior Cingulate	1409.9 (197.64)	1400.2 (201.32)	1360.9 (179.98)	1391 (195.78)	0.069
L Precentral	399.4 (84.61)	414.5 (76.43)	411.1 (78.45)	411.3 (78.28)	0.273
L Precuneus	201.6 (34.09)	198.8 (32.51)	196.5 (34.22)	198.6 (33.2)	0.501
L Rostral Anterior Cingulate	2948.7 (378.56)	3059.4 (399.52)	3048.3 (372.34)	3039.8 (390.49)	0.057
L Rostral Middle Frontal	4099.9 (622.38)	4289.2 (592.92)	4228 (471.41)	4244.3 (570.24)	0.019
L Superior Frontal	2888.2 (499.93)	3070.1 (401.3)	3110.4 (386.43)	3053.8 (419.13)	<0.001
L Superior Parietal	954.9 (167.59)	987.5 (175.89)	927.5 (147.62)	966.4 (169.21)	0.001

L Superior Temporal	4589.2 (655.81)	4609.3 (514.25)	4523.5 (534.31)	4583 (543.31)	0.257
L Supramarginal	2451.2 (320.27)	2523.3 (276.01)	2488.6 (237.08)	2503.1 (274.14)	0.066
L Temporal Pole	2889.9 (391.34)	2947.6 (402.5)	2840.3 (367.57)	2909.9 (393.81)	0.015
L Transverse Temporal	1886.4 (254.77)	1887.8 (237.44)	1843.9 (230.45)	1875.7 (238.64)	0.144
R Bankssts	2733.7 (407.71)	2894.2 (363.02)	2863.5 (343.63)	2861.8 (368.55)	0.001
R Caudal Anterior Cingulate	1304.4 (199.43)	1291.8 (185.01)	1273.6 (168.54)	1288.7 (182.95)	0.4
R Caudal Middle Frontal	609.6 (84.2)	650 (98.16)	645.1 (90.65)	642.6 (95.07)	0.002
R Cuneus	1566.2 (238.28)	1551.5 (247.03)	1555.3 (224.88)	1554.8 (239.57)	0.877
R Entorhinal	577.9 (90.85)	594.9 (78.74)	580.2 (69.39)	588.4 (78.55)	0.061
R Frontal Pole	1196.1 (173.41)	1191.4 (187.56)	1197 (168.54)	1193.6 (180.23)	0.939
R Fusiform	1301.6 (208.49)	1316.6 (214.48)	1272.9 (207.31)	1302.5 (212.15)	0.1
R Inferior Parietal	4105.3 (537.55)	4102.6 (477.46)	3996.9 (483.26)	4074.4 (489.88)	0.065
R Inferior Temporal	1063.3 (145.16)	1106.6 (182.88)	1073.1 (144.31)	1091 (168.62)	0.029
R Insula	4811.5 (592.92)	4739.6 (548.46)	4631.3 (483.45)	4721 (541.15)	0.027
R Isthmus Cingulate	3418.7 (408.85)	3528.4 (423.35)	3440.8 (395.71)	3488.2 (415.92)	0.021
R Lateral Occipital	806.3 (159.4)	811.2 (154.17)	788.8 (150.6)	804.4 (154.05)	0.314
R Lateral Orbitofrontal	5373.9 (750.12)	5385.5 (676.93)	5318.6 (666.6)	5365.6 (685.04)	0.593

R Lingual	6655.6 (870.68)	6771.4 (752.4)	6604 (709.22)	6708.7 (762.46)	0.057
R Medial Orbitofrontal	5109.9 (586.2)	5161.6 (611.85)	5034.4 (540.73)	5119.4 (591.13)	0.08
R Middle Temporal	3572 (420.14)	3608.5 (406.6)	3593.8 (374.74)	3599 (399.85)	0.734
R Paracentral	3572 (524.67)	3681.2 (512.85)	3610.6 (473.17)	3645.7 (505.24)	0.115
R Parahippocampal	436.5 (66.98)	456.7 (62.43)	458.5 (61.75)	454.1 (63.28)	0.017
R Pars Opercularis	440.3 (68.91)	442.9 (76.73)	432 (73.76)	439.5 (74.83)	0.315
R Pars Orbitalis	863 (133.15)	886.2 (128.02)	896.2 (121.85)	885.4 (127.37)	0.144
R Pars Triangularis	738.5 (155.13)	747 (184.85)	723.2 (127.03)	739.3 (166.72)	0.329
R Pericalcarine	2045.8 (327.59)	2003.9 (346.64)	1991.2 (333.25)	2006.8 (340.13)	0.47
R Postcentral	1438.5 (207.59)	1457.2 (200.48)	1418.5 (184.46)	1443.9 (197.75)	0.12
R Posterior Cingulate	355.5 (76.24)	358.4 (85.16)	349.7 (73.2)	355.6 (80.73)	0.534
R Precentral	271.6 (45.26)	272.9 (43.62)	264.7 (45.97)	270.5 (44.57)	0.159
R Precuneus	2904.4 (403.61)	3033.5 (387.84)	2984.3 (336.66)	3000.8 (379.28)	0.014
R Rostral Anterior Cingulate	5014.8 (725.8)	5128.9 (668.02)	5101.1 (652.9)	5104.3 (672.95)	0.366
R Rostral Middle Frontal	2848.6 (438.6)	3032.1 (429.49)	3012.4 (377.2)	2999.3 (421.56)	0.001
R Superior Frontal	899.9 (150.61)	898.5 (156.85)	861.4 (131.53)	888.6 (150.13)	0.027
R Superior Parietal	4423.9 (559.05)	4463.2 (548.23)	4378 (499.96)	4434.3 (537.62)	0.252

R Superior Temporal	2443.6 (362.29)	2477.9 (304.2)	2449.8 (286.08)	2465.1 (308.73)	0.497
R Supramarginal	3003.8 (408.74)	2987 (368.81)	2874.2 (355.3)	2958.9 (374.48)	0.003
R Temporal Pole	1739.6 (219)	1763.3 (217.8)	1740.8 (192.4)	1753.6 (211.36)	0.431
R Transverse Temporal	3071 (433.56)	3197.9 (413.22)	3179 (363.18)	3173.8 (405.2)	0.032

Table 7: Summary statistics for Cortical Surface Area

	AD (N = 88)	MCI (N=340)	CN (N=159)	Total (N=587)	
ROI name	Mean (SD)				p value
L Bankssts	1769.2 (324.52)	1969.8 (371.48)	2012.6 (259.29)	1951.3 (346.11)	<0.001
L Caudal Anterior Cingulate	3556 (658.45)	3540.9 (599.07)	3524.5 (567.29)	3538.7 (599.04)	0.92
L Caudal Middle Frontal	2201.8 (384.33)	2307.8 (335.09)	2332 (340.7)	2298.5 (346.33)	0.013
L Cuneus	3558.7 (679.67)	3615.3 (638.33)	3600.3 (558.25)	3602.7 (623.46)	0.749
L Entorhinal	2088.9 (430.71)	2107.9 (362.45)	2084.9 (399.49)	2098.8 (382.99)	0.794
L Frontal Pole	8066.4 (1253.95)	8335.5 (1303.77)	8026.7 (1010.71)	8211.5 (1230.27)	0.016
L Fusiform	2842.5 (555.36)	2945.9 (460.51)	2922.6 (415.97)	2924.1 (465.19)	0.178
L Inferior Parietal	1535860.5 (166861.19)	1518754.7 (156096.6)	1481550.3 (147406.12)	1511241.6 (156367.94)	0.013
L Inferior Temporal	11696.4 (1680.71)	12049 (1618.9)	11834.5 (1390.57)	11938.1 (1573.44)	0.108

L Insula	8338.9 (1133.86)	9029.6 (1298.88)	8869.5 (1022.33)	8882.7 (1227.04)	<0.001
L Isthmus Cingulate	2065 (428.92)	2057.2 (403.82)	2038.8 (414.27)	2053.4 (409.89)	0.861
L Lateral Occipital	13328.6 (1813.51)	14045.1 (1858.17)	13817.1 (1649.51)	13875.9 (1811.59)	0.004
L Lateral Orbitofrontal	18042.6 (2257.47)	19129.8 (2342.65)	18724.5 (1990.48)	18857.1 (2268.75)	<0.001
L Lingual	11297.4 (1621.04)	12071.7 (1660.01)	11808.1 (1528.64)	11884.2 (1639.5)	<0.001
L Medial Orbitofrontal	9509.2 (1243.01)	10155.9 (1276.18)	10216.1 (1139.98)	10075.2 (1256.67)	<0.001
L Middle Temporal	8598.3 (1368.88)	9267.7 (1431.91)	9143.8 (1096.22)	9133.8 (1357.15)	<0.001
L Paracentral	1847.1 (425.07)	2113.2 (389.95)	2221.9 (330.16)	2102.8 (397.38)	<0.001
L Parahippocampal	793.6 (156.84)	840.8 (170.76)	831.7 (158.57)	831.2 (166.03)	0.059
L Pars Opercularis	6248.5 (715.35)	6512.3 (802.62)	6480.8 (786.38)	6464.2 (789.8)	0.019
L Pars Orbitalis	6433.6 (806.06)	6724.2 (891.99)	6675.1 (835.1)	6667.4 (868.83)	0.02
L Pars Triangularis	2019.3 (358.4)	2288 (431.97)	2280 (403.55)	2245.5 (424.24)	<0.001
L Pericalcarine	1708.7 (405.26)	1780.1 (473.52)	1723.9 (423.55)	1754.2 (451.01)	0.255
L Postcentral	5464.9 (1022.53)	5731.9 (984.28)	5628.9 (911.82)	5664 (974.07)	0.063
L Posterior Cingulate	2704.7 (544.55)	2715.4 (455.28)	2636.1 (415.98)	2692.3 (460.2)	0.193
L Precentral	1437.2 (418.56)	1893.5 (425.04)	1994.8 (385.08)	1852.6 (450.41)	<0.001

L Precuneus	679.2 (166.91)	714.5 (145.66)	714 (137.03)	709.1 (147.07)	0.119
L Rostral Anterior Cingulate	8201.1 (1406.94)	9311.5 (1423.43)	9384.8 (1239.09)	9164.9 (1430.06)	<0.001
L Rostral Middle Frontal	9836.5 (1983.1)	11110.1 (1771.23)	11107.7 (1369.1)	10918.5 (1763.8)	<0.001
L Superior Frontal	8495.5 (1751.47)	9806.3 (1501.58)	10067.8 (1378.53)	9680.6 (1591.53)	<0.001
L Superior Parietal	2318.3 (437.28)	2556.6 (458.41)	2447 (391.95)	2491.2 (445.96)	<0.001
L Superior Temporal	10248.7 (1786.97)	10771.1 (1586.58)	10578 (1446.39)	10640.5 (1590.09)	0.019
L Supramarginal	6488.6 (798.21)	6844.6 (836.16)	6715.3 (712.72)	6756.2 (807.32)	0.001
L Temporal Pole	5763.7 (1066.78)	6083.2 (1015.28)	5851.4 (860.25)	5972.5 (991.18)	0.005
L Transverse Temporal	4714.2 (700.12)	4853.3 (660.1)	4805.9 (617.98)	4819.6 (655.89)	0.198
R Bankssts	8199.3 (1502.17)	9593 (1381.91)	9661.5 (1324.74)	9402.6 (1472.95)	<0.001
R Caudal Anterior Cingulate	3161 (568)	3203.2 (537.92)	3144.5 (455.99)	3181 (521.6)	0.467
R Caudal Middle Frontal	1787.4 (352.94)	2070.6 (385.54)	2122.6 (327.28)	2042.2 (381.23)	<0.001
R Cuneus	4162.7 (731.07)	4218.6 (721.05)	4276.8 (630.63)	4226 (699.06)	0.45
R Entorhinal	1824.2 (319.69)	1944.5 (299.41)	1888 (273.61)	1911.2 (298.57)	0.002
R Frontal Pole	2972.6 (503.22)	3069.8 (537.57)	3096 (450.73)	3062.3 (510.97)	0.176
R Fusiform	1868.8 (382.48)	1886.5 (359.7)	1830.8 (344.26)	1868.8 (359.3)	0.273

R Inferior Parietal	8516.7 (1402.34)	8875.7 (1300.54)	8601.1 (1201.54)	8747.5 (1297.32)	0.017
R Inferior Temporal	2805.4 (439.31)	2989 (492.21)	2937.5 (403.34)	2947.5 (465.54)	0.004
R Insula	11784.7 (1706.03)	12236.7 (1617.76)	12012.5 (1354.44)	12108.2 (1571.16)	0.037
R Isthmus Cingulate	7938.3 (1154.2)	8673.8 (1187.03)	8522.9 (1002.57)	8522.7 (1161.23)	<0.001
R Lateral Occipital	2500.9 (424.76)	2589.1 (443.42)	2518.8 (434.48)	2556.9 (439.2)	0.107
R Lateral Orbitofrontal	12818.6 (1760.2)	13438.5 (1762.17)	13364.8 (1696.18)	13325.6 (1754.57)	0.012
R Lingual	18321.9 (2356.1)	19735.8 (2236.85)	19424.5 (1841.65)	19439.5 (2207.2)	<0.001
R Medial Orbitofrontal	11184.4 (1647.84)	12063 (1628.59)	11716.8 (1410.92)	11837.5 (1603.66)	<0.001
R Middle Temporal	9625.7 (1487.31)	10586.7 (1415.53)	10752.3 (1250.98)	10487.5 (1430.3)	<0.001
R Paracentral	8896.6 (1563.83)	9804.6 (1418.67)	9742.2 (1148.95)	9651.6 (1408.8)	<0.001
R Parahippocampal	2003.9 (464.21)	2279.8 (383.14)	2351.7 (334.92)	2257.9 (399.39)	<0.001
R Pars Opercularis	1028.7 (188.93)	1078.9 (216.27)	1059.9 (214.53)	1066.3 (212.32)	0.128
R Pars Orbitalis	2037.9 (353.8)	2150.1 (359.43)	2211.8 (352.87)	2150 (360.3)	0.001
R Pars Triangularis	2032.5 (507.64)	2052.1 (528.66)	2004.6 (393.18)	2036.3 (491.98)	0.602
R Pericalcarine	5080.8 (901.88)	5307.6 (986.75)	5299.5 (936.16)	5271.4 (962.68)	0.131
R Postcentral	2824.6 (533.95)	2883.2 (496.11)	2797.9 (433.47)	2851.3 (486.66)	0.162

R Posterior Cingulate	1401.4 (404.45)	1774.3 (442.71)	1853.9 (349.11)	1740 (438.12)	<0.001
R Precentral	936.6 (214.25)	944.2 (179.79)	906.2 (159.12)	932.7 (180.58)	0.089
R Precuneus	8133.3 (1354.98)	9098.6 (1385.12)	9107.3 (1147.14)	8956.2 (1362.69)	<0.001
R Rostral Anterior Cingulate	12112.4 (2008.45)	13491.1 (2010.79)	13499.6 (1756.6)	13286.7 (2003.37)	<0.001
R Rostral Middle Frontal	8662.8 (1664.53)	9874.2 (1593.64)	10009.4 (1460.75)	9729.2 (1631)	<0.001
R Superior Frontal	2168.1 (409.56)	2294.2 (410.1)	2236.4 (329.64)	2259.7 (391.91)	0.018
R Superior Parietal	10209.5 (1473.15)	10797.9 (1678.19)	10592.6 (1406.29)	10654.1 (1589.77)	0.007
R Superior Temporal	6396.9 (792.09)	6673 (767.39)	6610.2 (659)	6614.6 (748.33)	0.008
R Supramarginal	6028 (1022.01)	6248 (948.84)	5987.1 (866.25)	6144.4 (945.09)	0.007
R Temporal Pole	4550.5 (633.11)	4645.4 (593.29)	4589.7 (541.09)	4616.1 (586.01)	0.321
R Transverse Temporal	9730 (1653.35)	10711.1 (1521.02)	10864.1 (1337.55)	10605.5 (1538.27)	<0.001
R Transverse Temporal	3071 (433.56)	3197.9 (413.22)	3179 (363.18)	3173.8 (405.2)	0.032

Table 8: Summary statistics for White Matter and Subcortical volumes

	AD (N = 88)	MCI (N=340)	CN (N=159)	Total (N=587)	
ROI name		Mean (SD)			p value
Brainstem	1439.9 (234.82)	1438.4 (198.64)	1425.8 (200.68)	1435.2 (204.71)	0.793

Corpus Callosum Anterior	4361 (685.61)	4633.8 (708.44)	4558.8 (574.57)	4572.6 (676.82)	0.003
Corpus Callosum Central	413.3 (91.16)	468.9 (96.14)	467.8 (86.42)	460.3 (94.79)	<0.001
Corpus Callosum Mid Anterior	5970.4 (724.81)	6178.8 (713.34)	6073.3 (664.93)	6119 (705.36)	0.03
Corpus Callosum Mid Posterior	3381.1 (418.52)	3533 (372.45)	3444.8 (358.03)	3486.3 (379.75)	0.001
Corpus Callosum Posterior	97.6 (47.4)	96.1 (50.94)	91.1 (42.52)	95 (48.25)	0.481
Csf	2191 (671.57)	1727.8 (641.64)	1617.7 (625.39)	1767.4 (666.71)	<0.001
Fourth Ventricle	7114 (6013.97)	5415.1 (5376.4)	4483.5 (5932.92)	5417.5 (5679)	0.002
L Accumbens Area	1077.8 (228.39)	1344 (258.72)	1395.7 (212.33)	1318.1 (263.28)	<0.001
L Amygdala	3327.3 (491.81)	3476.5 (588.82)	3385.6 (460.11)	3429.5 (544.92)	0.036
L Caudate	48700.7 (5363.27)	48510.3 (5250.86)	47050.4 (4435.03)	48143.4 (5096.36)	0.006
L Cerebellum Cortex	12869.6 (1956.07)	13215 (1956.25)	12843.5 (1988.49)	13062.6 (1969.8)	0.088
L Cerebellum WM	20883.3 (2533.13)	21036.5 (2375.22)	20743 (2250.96)	20934 (2366.19)	0.425
L Choroid Plexus	2091.6 (434.05)	1905.5 (412.46)	1802 (355.67)	1905.4 (410.61)	<0.001
L Hippocampus	2857.2 (560.99)	3504.6 (577.86)	3696.8 (433.85)	3459.6 (601.32)	<0.001
L Inferior Lateral Ventricle	715 (164.47)	769.4 (156.74)	767.1 (129.15)	760.6 (151.97)	0.009
L Lateral Ventricle	1726.1 (1052.84)	935.5 (737.73)	632.2 (418.44)	971.9 (802.57)	<0.001

L Pallidum	25681.4 (11197.04)	18859 (11203.15)	16624.8 (8439.07)	19276.6 (10890.02)	<0.001
L Putamen	332.1 (63.77)	359.8 (69.26)	359.8 (62.31)	355.7 (67.26)	0.002
L Thalamus	1571.2 (234.23)	1616.3 (231.44)	1605.9 (207.7)	1606.7 (225.84)	0.247
L Ventral DC	335.1 (75.68)	367.3 (80.68)	366.4 (69.23)	362.2 (77.71)	0.002
L Vessel	4503.6 (683.26)	4826.7 (721.6)	4818.4 (609.11)	4776 (695.48)	<0.001
Non WMHypo Intensities	300.2 (66.54)	333.6 (78.36)	330.2 (66.86)	327.7 (74.5)	0.001
Optic Chiasm	5948.7 (741.68)	6083.3 (653.62)	6094.2 (657.72)	6066 (669.26)	0.201
R Accumbens Area	3488.9 (430.57)	3620.5 (399.75)	3517.1 (370.01)	3572.8 (400.12)	0.003
R Amygdala	116.9 (70.84)	113.3 (65.22)	110.6 (61.78)	113.1 (65.11)	0.763
R Caudate	90.6 (61.73)	85.8 (93.24)	73.1 (64.72)	83.1 (82.24)	0.181
R Cerebellum Cortex	314.1 (74.32)	304.4 (74.95)	279.6 (65.96)	299.1 (73.46)	<0.001
R Cerebellum WM	868.2 (174.79)	900.7 (159.48)	908 (154.03)	897.8 (160.66)	0.155
R Choroid Plexus	441.1 (100.65)	498.6 (105.22)	495.7 (92.43)	489.2 (103.06)	<0.001
R Hippocampus	1177.7 (226.03)	1407.6 (253.48)	1464 (220.3)	1388.4 (257.43)	<0.001
R Inferior Lateral Ventricle	3506.7 (529.29)	3616.1 (624.67)	3531.1 (498.53)	3576.7 (580.12)	0.147
R Lateral Ventricle	49721.6 (5378.86)	49780.8 (5760.28)	48367.6 (4607.6)	49389.1 (5440.89)	0.021
R Pallidum	13032.2 (2078)	13390.4 (2089.08)	12965.4 (2139.91)	13221.6 (2107.18)	0.072
R Putamen	1779.9 (525.03)	1516.8 (412.52)	1422.8 (359.2)	1530.8 (432.16)	<0.001

R Thalamus	2535.7 (453.77)	2265.3 (540.41)	2180.4 (457.55)	2282.8 (518.26)	<0.001
R Ventral DC	2933.9 (547.89)	3593.3 (596.49)	3740 (463.6)	3534.2 (613.18)	<0.001
R Vessel	1511.4 (1047.84)	772.4 (623.67)	532.1 (405.01)	818.1 (726.53)	<0.001
Third Ventricle	22441.4 (10702.92)	17170.3 (10098.93)	15237.9 (7801.3)	17437.1 (9875.76)	<0.001
WMHypo Intensities	2234.7 (712.19)	2110.5 (689.58)	2068.3 (582.98)	2117.7 (666.91)	0.164

Appendix 3: Generalized Additive Mixed Models of Ambulatory Blood Pressure

This appendix provides additional details and supplementary information about the analyses conducted in Chapter 3. The first section reviews our model selection strategy. Section 2 reports diagnostics from our final models. Section 3 presents results of analyses with the sample size restricted to only include individuals with at least 70% of intended ABP readings. Lastly, section 4 provides graphical summaries of the number of ABP readings in each participant's profile and the time-intervals between readings.

GAMMs of ABP Appendix 3.1: Model Selection

Before examining associations between ABP profiles and the exposure of interest, we fit a generalized additive mixed model (GAMM) to describe ABP as a function of time. See section 3.3, equation (3.1). The model includes a fixed, non-parametric functional term for time (as described above) estimated estimated via thin-plate spline regression and a random intercept for each participant with an AR(1) correlation structure for the residuals.

Models were examined using ABP measurements directly and compared to the use of log-transformed ABP measurements. While log-transforming the data reduced the variability in participant's raw ABP profiles, this transformation also resulted in a time model with poorer fit. We therefore chose to leave the outcome variables untransformed for further analyses. The rank of our spline regression was also chosen based on the time model. We chose $k = 59$ as the rank that achieved sufficient model fit while maintaining sparsity. Higher ranks added additional parameters without providing meaningful improvements to the model fitting process. The final model was also compared to one with a nested random-effects structure by allowing different correlation coefficients for each day. The addition of this parameter did not provide a statistically significant improvement in model fit and was therefore removed form final consideration. AIC was used to inform model selection at each step. The estimated time curves for systolic BP (SBP) and diastolic BP (DBP), respectively, are shown in figure 3.2.

GAMMs of ABP Appendix 3.2: Model Diagnostics

Figures 10 and 10 display diagnostics plots for the models corresponding to equation (3.1). While the quantile-quantile plots (upper left corner in each) suggests a small departure from normality, each histogram of residuals (upper right corner) exhibits a symmetric distribution centered at 0. Residuals appear heteroscedastic with larger variances corresponding to larger values of the linear predictor (lower left corners). However, the responses and fitted values show a clear linear relationship. As pointed out in section 3.5, BP profiles differ greatly across individuals. Our models of mean BP profiles may not capture sufficient variability. Residuals from fitting the exposure and covariate-adjusted models were similar.

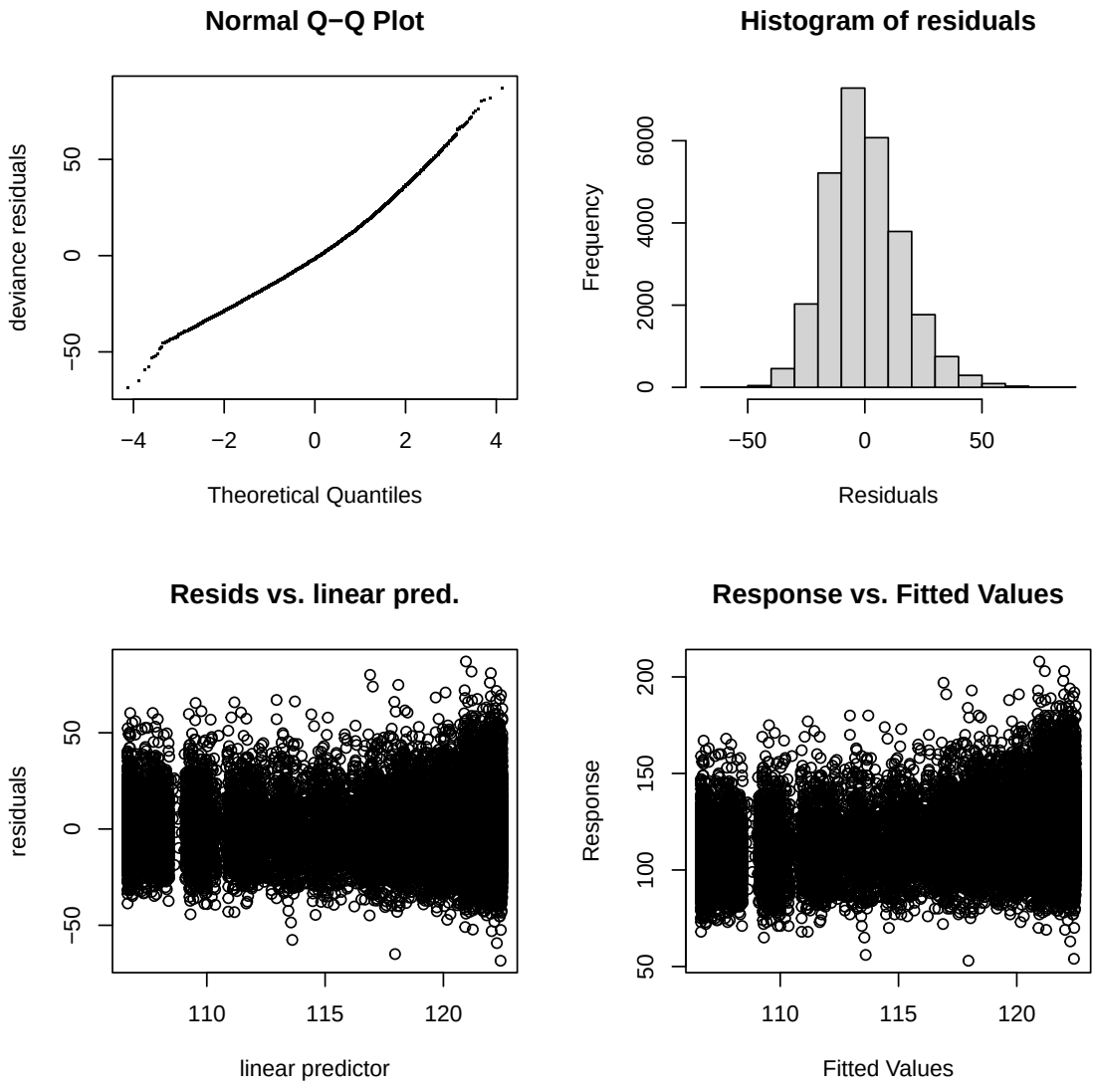


Figure 10: Model diagnostics for the time-only model shown in equation (3.1)

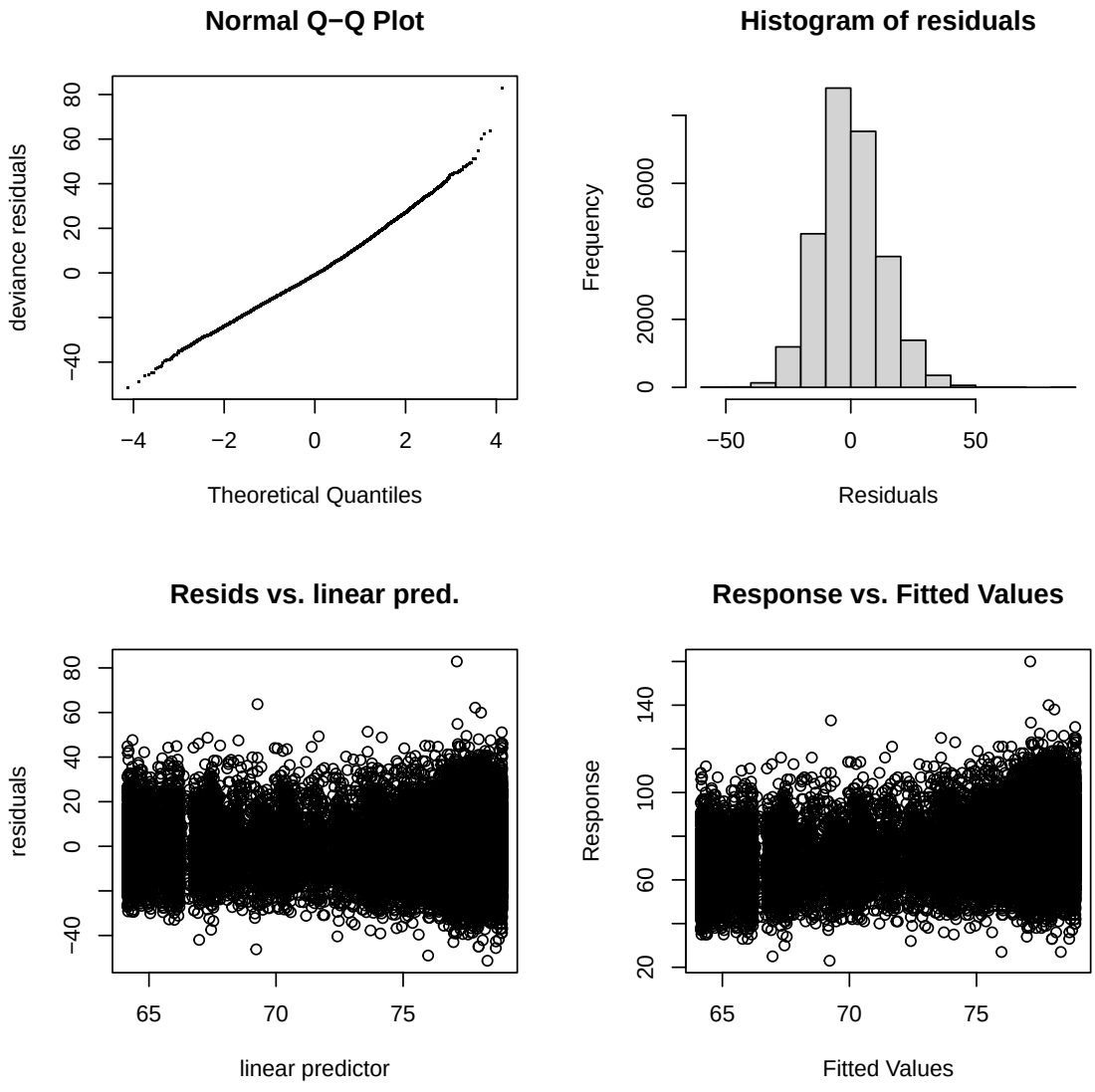


Figure 11: Model diagnostics for the time-only model shown in (3.2)

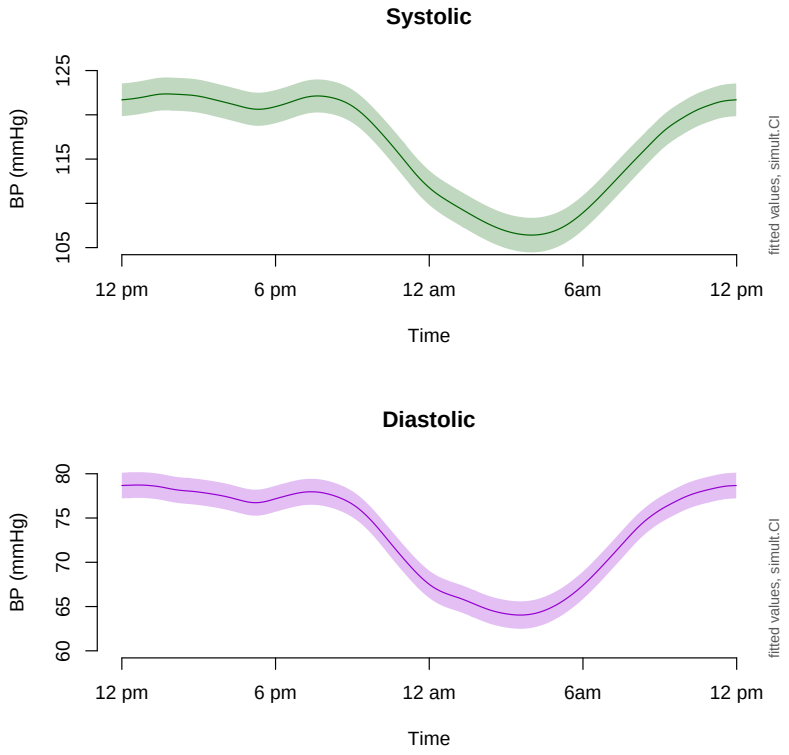


Figure 12: Estimated time model with the sample restricted to only women with 70% of intended ABP readings or more.

GAMMs of ABP Appendix 3.3: Analyses with Restricted Sample Size

Some participants did not achieve the intended number of ABP readings. To ensure that BP profiles for each participant contributed meaningfully to the estimated mean profiles, we examined results with the sample size restricted to only include those with at least 70% of intended readings. This diminished the sample size to $n = 365$ participants. Figure 12 exhibits results of fitting our time model with the restricted sample. Figure 13 shows the fitted BP profiles BWs and non-BWs, along with their difference curves. Figure 14 shows estimated profiles and differences after adjusting for meaningful covariates. Results from the restricted sample mirrored those in the full sample.

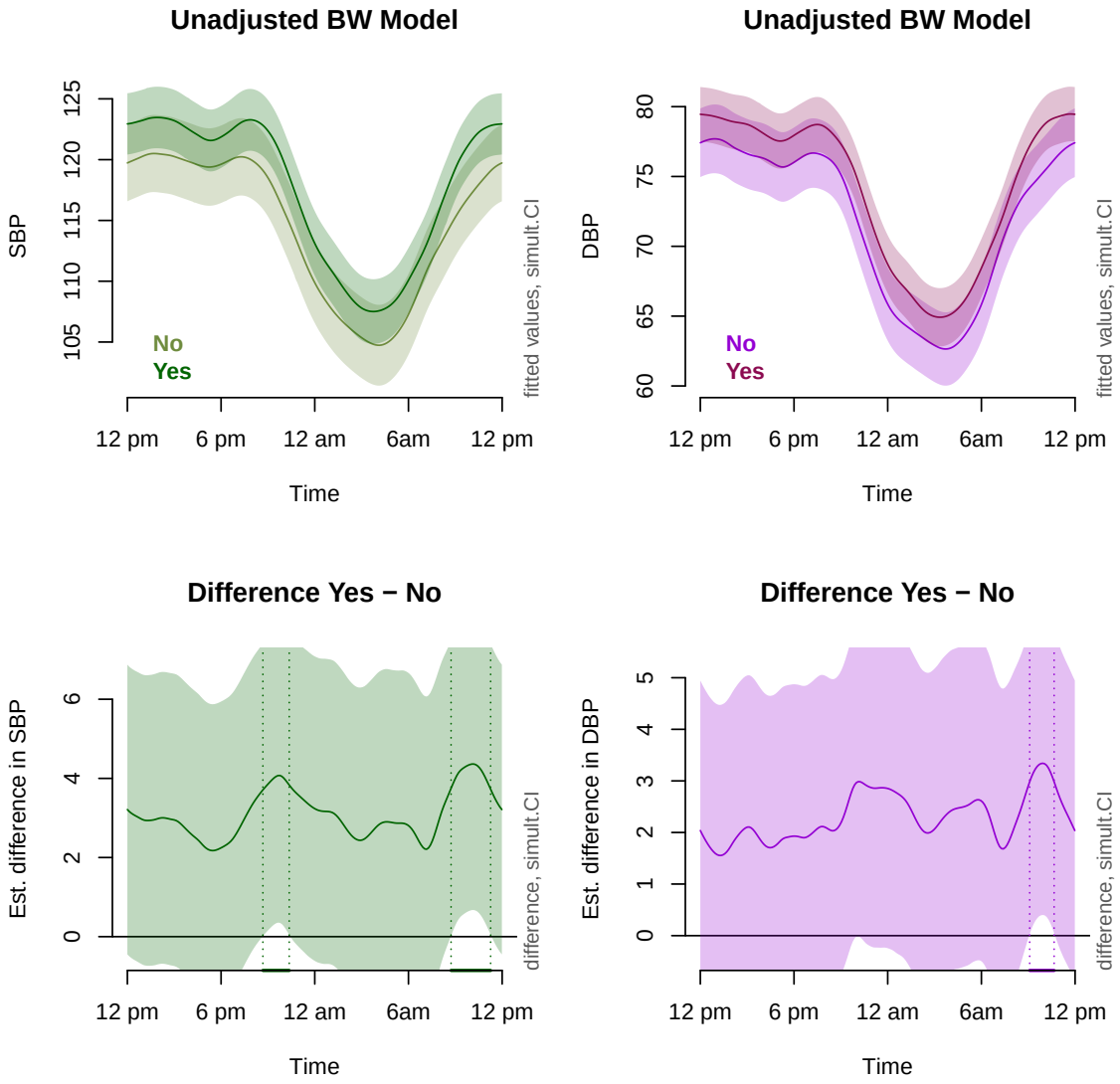


Figure 13: Estimated average ABP profiles for BWs and non-BWS (top row) and their differences across time (bottom row) exhibit consistently higher BP in BWs compared to non-BWs. Resuls presented here restrict the sample to only participants with at least 70% of intended readings. Vertical dotted lines and shaded x-axes indicate periods of time during which BWs' average BP was significantly higher than non-BWs.

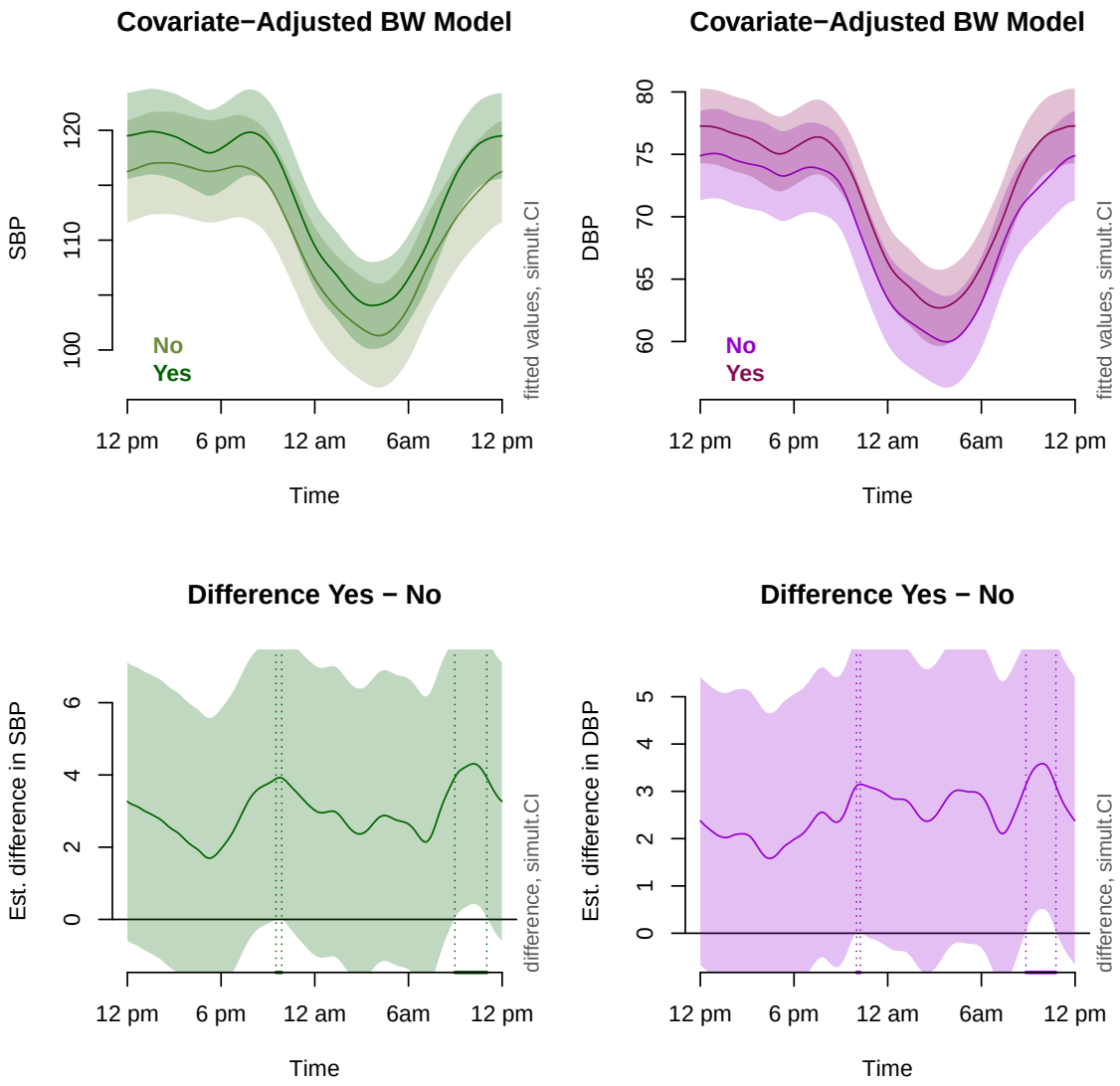


Figure 14: Estimated average ABP profiles for BWs and non-BWS (top row) and their differences across time (bottom row) exhibit consistently higher BP in BWs compared to non-BWs. Resuls presented here restrict the sample to only participants with at least 70% of intended readings. Vertical dotted lines and shaded x-axes indicate periods of time during which BWs' average BP was significantly higher than non-BWs.

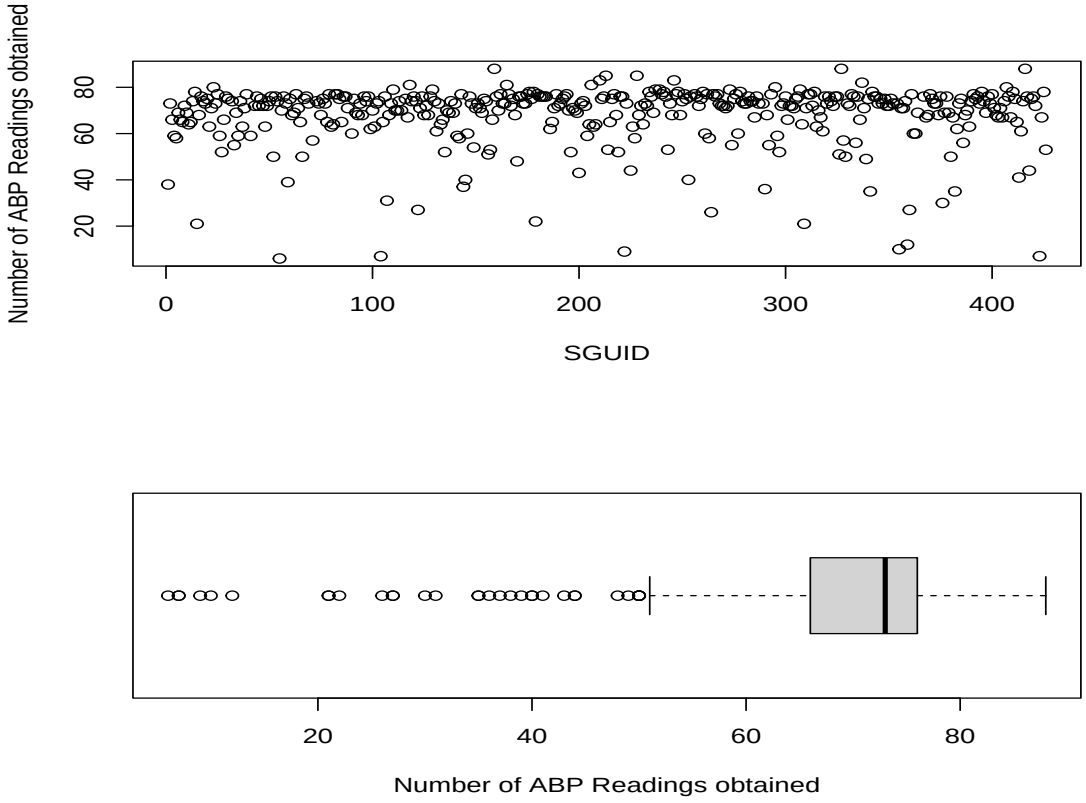


Figure 15: The scatter plot (top) shows the number of ABP readings achieved by each participant. "SGUID" refer to participants' study IDs. The box plot (bottom) shows the spread of values in the scatter plot.

GAMMs of ABP Appendix 3.4: Graphical Summaries of Selected ABP Profile characteristics

ABP profiles were not of uniform quality in the proposed. Each profile should consist of roughly 76 ABP readings. The number of readings ranged from 6 to 88. The longest interval between consecutive readings exceed 24 hours. The data analyzed in Chapter 3 uses all available profiles, regardless of their quality. The results of the full-sample analyses were strikingly similar to those from the restricted one (Appendix 3.3)

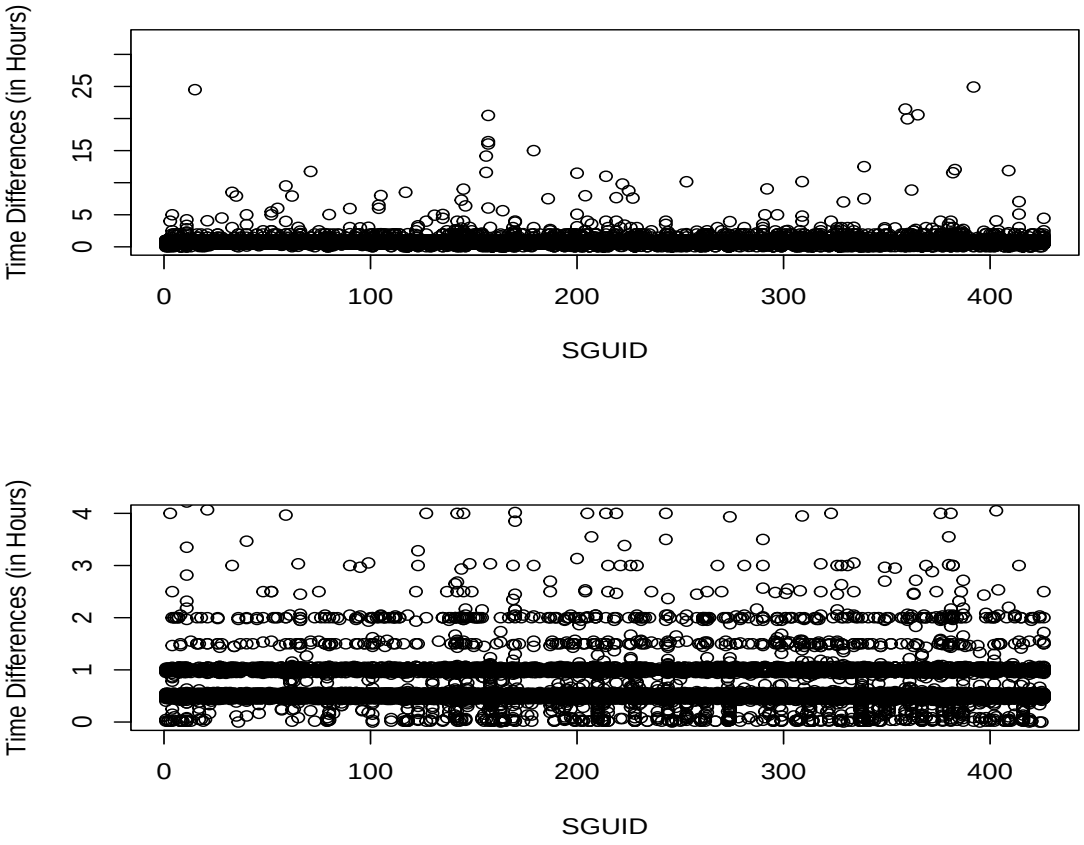


Figure 16: Study participant identifiers comprise the horizontal axis in each plot. Both show the length of time between readings for each study participant. (along vertical axes). The image on bottom limits the view of the vertical axis to values between 0 and 4.

Bibliography

[1] Lawrence J Appel, Thomas J Moore, Eva Obarzanek, William M Vollmer, Laura P Svetkey, Frank M Sacks, George A Bray, Thomas M Vogt, Jeffrey A Cutler, Marlene M Windhauser, et al. A clinical trial of the effects of dietary patterns on blood pressure. *New England journal of medicine*, 336(16):1117–1124, 1997.

[2] Aaron T Beck and Robert A Steer. Internal consistencies of the original and revised beck depression inventory. *Journal of clinical psychology*, 40(6):1365–1367, 1984.

[3] Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

[4] Dorothy Dickson and Joerg Hasford. 24-hour blood pressure measurement in antihypertensive drug trials: Data requirements and methods of analysis. *Statistics in medicine*, 11(16):2147–2158, 1992.

[5] Lloyd J Edwards and Sean L Simpson. An analysis of 24-hour ambulatory blood pressure monitoring data using orthonormal polynomials in the linear mixed model. *Blood pressure monitoring*, 19(3):153, 2014.

[6] Sarah Tomaszewski Farias, Dan Mungas, Bruce R Reed, Deborah Cahn-Weiner, William Jagust, Kathleen Baynes, and Charles DeCarli. The measurement of everyday cognition (ecog): scale development and psychometric properties. *Neuropsychology*, 22(4):531, 2008.

[7] Ashley S Felix, Amy Lehman, Timiya S Nolan, Shawnita Sealy-Jefferson, Khadijah Breathett, Darryl B Hood, Daniel Addison, Cindy M Anderson, Crystal W Cené, Barbara J Warren, et al. Stress, resilience, and cardiovascular disease risk among black women: Results from the women's health initiative. *Circulation: Cardiovascular Quality and Outcomes*, 12(4):e005284, 2019.

[8] Qing Feng, Meilei Jiang, Jan Hannig, and J. S. Marron. Angle-based joint and individual variation explained. *Journal of Multivariate Analysis*, 166:241–265, 7 2018. ISSN 10957243. doi: 10.1016/j.jmva.2018.03.008.

[9] Jeffrey H Ferguson and Carl J Shaar. The effective diagnosis and treatment of hypertension by the primary care physician: impact of ambulatory blood pressure monitoring. *The Journal of the American Board of Family Practice*, 5 (5):457–465, 1992.

[10] Emily S Finn, Xilin Shen, Dustin Scheinost, Monica D Rosenberg, Jessica Huang, Marvin M Chun, Xenophon Papademetris, and R Todd Constable. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature neuroscience*, 18(11):1664–1671, 2015.

[11] Marshal F Folstein, Susan E Folstein, and Paul R McHugh. "mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12(3):189–198, 1975.

[12] Centers for Disease COntrol and Prevention. Heart disease risk factors, 2019. URL `www.cdc.gov/heartdisease/risk_factors.htm`.

[13] Giovanni B Frisoni, Annapaola Prestia, Paul E Rasser, Matteo Bonetti, and Paul M Thompson. In vivo mapping of incremental cortical atrophy from incipient to overt alzheimer's disease. *Journal of neurology*, 256(6):916–924, 2009.

[14] Michael Gaffney, Colin Taylor, and Elizabeth Cusenza. Harmonic regression analysis of the effect of drug treatment on the diurnal rhythm of blood pressure and angina. *Statistics in medicine*, 12(2):129–142, 1993.

[15] Irina Gaynanova and Gen Li. Structural learning and integrative decomposition of multi-view data. *Biometrics*, 75(4):1121–1132, 2019. ISSN 15410420. doi: 10.1111/biom.13108.

[16] Irina Gaynanova and Gen Li. Structural learning and integrative decomposition of multi-view data. *Biometrics*, 75(4):1121–1132, 12 2019. ISSN 0006-341X. doi: 10.1111/biom.13108. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13108`.

[17] Matthew F Glasser, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R Polimeni, David C Van Essen, Mark Jenkinson, and Wuminn Hcp. NeuroImage The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, 80:105–124, 2013. doi: 10.1016/j.neuroimage.2013.04.127.

[18] Ixavier A Higgins, Suprateek Kundu, and Ying Guo. Neuroimage integrative bayesian analysis of brain functional networks incorporating anatomical knowledge. *NeuroImage*, 181(July):263–278, 2018. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2018.07.015. URL `https://doi.org/10.1016/j.neuroimage.2018.07.015`.

[19] Christopher J Honey, Olaf Sporns, Leila Cammoun, Xavier Gigandet, Jean-

Philippe Thiran, Reto Meuli, and Patric Hagmann. Predicting human resting-state functional connectivity from structural connectivity. *Proceedings of the National Academy of Sciences*, 106(6):2035–2040, 2009.

[20] Harold Hotelling. Biometrika Trust Relations Between Two Sets of Variates Author (s): Harold Hotelling Published by : Oxford University Press on behalf of Biometrika Trust Stable URL : https://www.jstor.org/stable/2333955 REFERENCES Linked references are available on JST. *Biometrika*, 28(3):321–377, 1936. URL `https://www.jstor.org/stable/2333955`.

[21] Rajan Kashyap, Ru Kong, Sagarika Bhattacharjee, Jingwei Li, Juan Zhou, and B. T. Thomas Yeo. Individual-specific fMRI-Subspaces improve functional connectivity prediction of behavior. *NeuroImage*, 189:804–812, 4 2019. ISSN 10959572. doi: 10.1016/j.neuroimage.2019.01.069.

[22] Rajan Kashyap, Ru Kong, Sagarika Bhattacharjee, Jingwei Li, Juan Zhou, and BT Thomas Yeo. Individual-specific fmri-subspaces improve functional connectivity prediction of behavior. *NeuroImage*, 189:804–812, 2019.

[23] Phebe Brenne Kemmer, Yikai Wang, F DuBois Bowman, Helen Mayberg, and Ying Guo. Evaluating the strength of structural connectivity underlying brain functional networks. *Brain Connectivity*, 8(10):579–594, 2018.

[24] Mika Kivimäki and Andrew Steptoe. Effects of stress on the development and progression of cardiovascular disease. *Nature Reviews Cardiology*, 15(4):215, 2018.

[25] Yi-ou Li, Tülay Adalı, Wei Wang, and Vince D Calhoun. Joint blind source separation by multiset canonical. *IEEE Transactions on Signal Processing*, 57 (10):3918–3929, 2009.

[26] Raphaël Liégeois, Augusto Santos, Vincenzo Matta, Dimitri Van De Ville, and Ali H Sayed. Revisiting correlation-based functional connectivity and its relationship with structural connectivity. *Network Neuroscience*, pages 1–17, 2020.

[27] Eric F. Lock, Katherine A. Hoadley, J. S. Marron, Andrew B. Nobel, Chapel Hill, Katherine A. Hoadley, Chapel Hill, Andrew B. Nobel, J. S. Marron, and Andrew B. Nobel. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Annals of Applied Statistics*, 7(1):523–542, 2013. ISSN 19326157. doi: 10.1214/12-AOAS597.

[28] JM Madden, LD Browne, Xia Li, PM Kearney, and AP Fitzgerald. Morning surge in blood pressure using a random-effects multiple-component cosinor model. *Statistics in medicine*, 37(10):1682–1695, 2018.

[29] U.K.) Mardia, K.V. (University of Leeds, U.K.) Kent, J. T. (University of Leeds, and J. M. (The Open University) Bibby. *Multivariate analysis*. Acadmic Press, Inc., 10 edition, 1979.

[30] Scott Marek, Brenden Tervo-Clemmens, Finnegan J Calabro, David F Montez, Benjamin P Kay, Alexander S Hatoum, Meghan Rose Donohue, William Foran, Ryland L Miller, Eric Feczko, et al. Towards reproducible brain-wide association studies. *BioRxiv*, 2020.

[31] John C Morris. Clinical dementia rating: a reliable and valid diagnostic and staging measure for dementia of the alzheimer type. *International psychogeriatrics*, 9(S1):173–176, 1997.

[32] Susanne G. Mueller, Michael W. Weiner, Leon J. Thal, Ronald C. Petersen, Clifford R. Jack, William Jagust, John Q. Trojanowski, Arthur W. Toga, and Laurel Beckett. Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's and Dementia*,

46(2):55–66, 2005. ISSN 1946-6242. doi: 10.1016/j.freeradbiomed.2008.10.025. The.

[33] Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699, 2005.

[34] Sean M. Nestor, Raul Rupsingh, Michael Borrie, Matthew Smith, Vittorio Accomazzi, Jennie L. Wells, Jennifer Fogarty, Robert Bartha, and the Alzheimer's Disease Neuroimaging Initiative. Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database. *Brain*, 131(9):2443–2454, 07 2008. ISSN 0006-8950. doi: 10.1093/brain/awn146. URL `https://doi.org/10.1093/brain/awn146`.

[35] E O'Brien. Twenty-four-hour ambulatory blood pressure measurement in clinical practice and research: a critical review of a technique in need of implementation. *Journal of internal medicine*, 269(5):478–495, 2011.

[36] Michael J O'Connell and Eric F Lock. R.jive for exploration of multi-source molecular data. *Bioinformatics*, 32(18):2877–2879, 2016.

[37] André Rey. L'examen psychologique dans les cas d'encéphalopathie traumatique.(les problems.). *Archives de psychologie*, 1941.

[38] Andŕe Rey. L'examen clinique en psychologie., 1958.

[39] Benjamin B Risk, Raphiel J Murden, Junjie Wu, Mary Beth Nebel, Arun Venkataraman, Zhengwu Zhang, and Deqiang Qiu. Which multiband factor

should you choose for your resting-state fmri study? *NeuroImage*, 234:117965, 2021.

[40] Wilma G Rosen, Richard C Mohs, and Kenneth L Davis. A new rating scale for alzheimer's disease. *The American journal of psychiatry*, 1984.

[41] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA, 2020. URL `http://www.rstudio.com/`.

[42] Christos Sagonas, Yannis Panagakis, Alina Leidinger, and Stefanos Zafeiriou. Robust joint and individual variance explained. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017. doi: 10.1109/CVPR.2017.608.

[43] Brian J Sandri, Adam Kaplan, Shane W Hodgson, Mark Peterson, Svetlana Avdulov, LeeAnn Higgins, Todd Markowski, Ping Yang, Andrew H Limper, Timothy J Griffin, et al. Multi-omic molecular profiling of lung cancer in copd. *European Respiratory Journal*, 52(1), 2018.

[44] Theodore D. Satterthwaite, Mark A. Elliott, Kosha Ruparel, James Loughead, Karthik Prabhakaran, Monica E. Calkins, Ryan Hopson, Chad Jackson, Jack Keefe, Marisa Riley, Frank D. Mentch, Patrick Sleiman, Ragini Verma, Christos Davatzikos, Hakon Hakonarson, Ruben C. Gur, and Raquel E. Gur. Neuroimaging of the philadelphia neurodevelopmental cohort. *NeuroImage*, 86: 544–553, 2014. ISSN 10959572. doi: 10.1016/j.neuroimage.2013.07.064. URL `http://dx.doi.org/10.1016/j.neuroimage.2013.07.064`.

[45] Joseph E Schwartz, Katherine Warren, and Thomas G Pickering. Mood, location and physical position as predictors of ambulatory blood pressure and heart rate: Application of a multi-level random effects model. *Annals of Behavioral Medicine*, 16(3):210–220, 1994.

[46] Hai Shu, Xiao Wang, and Hongtu Zhu. D-cca: A decomposition-based canonical correlation analysis for high-dimensional datasets. *Journal of the American Statistical Association*, 2019. ISSN 1537274X. doi: 10.1080/01621459.2018.1543599.

[47] Hai Shu, Xiao Wang, and Hongtu Zhu. D-cca: A decomposition-based canonical correlation analysis for high-dimensional datasets. *Journal of the American Statistical Association*, 115(529):292–306, 2020.

[48] Stephen M. Smith, Thomas E. Nichols, Diego Vidaurre, Anderson M. Winkler, Timothy E.J. Behrens, Matthew F. Glasser, Kamil Ugurbil, Deanna M. Barch, David C. Van Essen, and Karla L. Miller. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature Neuroscience*, 18(11):1565–1567, 2015. ISSN 15461726. doi: 10.1038/nn.4125. URL `http://dx.doi.org/10.1038/nn.4125`.

[49] B Streitberg and W Meyer-Sabellek. Smoothing twenty-four-hour ambulatory blood pressure profiles: a comparison of alternative methods. *Journal of hypertension. Supplement: official journal of the International Society of Hypertension*, 8(6):S21–7, 1990.

[50] B Streitberg, W Meyer-Sabellek, and P Baumgart. Statistical analysis of circadian blood pressure recordings in controlled clinical trials. *Journal of hypertension. Supplement: official journal of the International Society of Hypertension*, 7(3):S11–7, 1989.

[51] Jing Sui and Vince D Calhoun. Multimodal fusion of structural and functional brain imaging data. In *fMRI techniques and protocols*, pages 853–869. Springer, 2016.

[52] Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B: Statistical*

Methodology, 61(3):611–622, 1999. ISSN 13697412. doi: 10.1111/1467-9868.00196.

[53] Jacolien van Rij, Martijn Wieling, R. Harald Baayen, and Hedderik van Rijn. itsadug: Interpreting time series and autocorrelated data using gamms, 2020. R package version 2.4.

[54] Hilary K Wall, Matthew D Ritchey, Cathleen Gillespie, John D Omura, Ahmed Jamal, and Mary G George. Vital signs: prevalence of key cardiovascular disease risk factors for million hearts 2022—united states, 2011–2016. *Morbidity and Mortality Weekly Report*, 67(35):983, 2018.

[55] Michael A Weber, Deanna G Cheung, William F Graettinger, and Jodi L Lipson. Characterization of antihypertensive therapy by whole-day blood pressure monitoring. *Jama*, 259(22):3281–3285, 1988.

[56] Per Ake Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972. ISSN 00063835. doi: 10.1007/BF01932678.

[57] Daniela M. Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009. ISSN 14654644. doi: 10.1093/biostatistics/kxp008.

[58] Simon N Wood. Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, 62(4):1025–1036, 2006.

[59] Simon N Wood. Inference and computation with generalized additive models and their extensions. *Test*, 29(2):307–339, 2020.

[60] S.N Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2 edition, 2017.

[61] Janet S Wright, Hilary K Wall, and Matthew D Ritchey. Million hearts 2022: small steps are needed for cardiovascular disease prevention. *Jama*, 320(18): 1857–1858, 2018.

[62] Ke Ye and Lek-heng Lim. Schubert varieties and distances between subspaces of different dimensions. *SIAM Journal on Matrix Analysis and Applications*, 37 (3):1176–1197, 7 2014.

[63] Qunqun Yu, Benjamin B Risk, Kai Zhang, and J S Marron. JIVE integration of imaging and behavioral data. *NeuroImage*, 152(February):38–49, 2017. ISSN 10959572. doi: 10.1016/j.neuroimage.2017.02.072. URL `http://dx.doi.org/10.1016/j.neuroimage.2017.02.072`.

[64] Zhengwu Zhang, Maxime Descoteaux, Jingwen Zhang, Gabriel Girard, Maxime Chamberland, David Dunson, Anuj Srivastava, and Hongtu Zhu. Mapping population-based structural connectomes. *NeuroImage*, 172:130–145, 2018.

[65] Yihong Zhao, Arno Klein, F Xavier Castellanos, and Michael P Milham. Brain age prediction: Cortical and subcortical shape covariation in the developing human brain. *NeuroImage*, 202:116149, 2019.

[66] Guoxu Zhou, Andrzej Cichocki, Yu Zhang, and Danilo P Mandic. Group component analysis for multiblock data: Common and individual feature extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 27(11): 2426–2439, 2016. ISSN 21622388. doi: 10.1109/TNNLS.2015.2487364.