GENERALIZATIONS, EXTENSIONS AND APPLICATIONS FOR PRINCIPAL COMPONENT ANALYSIS

by

Chen Yue

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

Febuary, 2016

Abstract

Principal component analysis (PCA) is one of the most important dimension reduction technique. It is widely used in many applications including economics, finance and medical research. In this research, several novel generalizations of PCA are proposed to adapt the technique to more complicated scenarios. In the first project, we propose a principal surface model for manifold-like datasets in 3D space. In the second part, a new concept of graphical intra-class correlation coefficient (GICC) is defined and a Markov Chain Monte Carlo Expectation-Maximization (mcmcEM) algorithm is used for likelihood optimization. In the third part, we propose multilevel binary principal component analysis (MBPCA) models for finding the principal components of multilevel binary dataset. A variational expectation maximization algorithm is used for likelihood optimization.

Advisor:

Brian Caffo, PhD

Co-Advisor:

Vadim Zipunnikov, PhD

ABSTRACT

Committee:

Craig Hendrix, MD (chair, SOM clinical pharmacology)

Brian Caffo, PhD (advisor, SPH biostatistics)

Elizabeth Platz, PhD (SPH epidemiology)

Vadim Zipunnikov, PhD (SPH biostatistics)

Alternates:

Michelle Carlson, PhD (SPH mental health)

Martin Lindquist, PhD (SPH biostatistics)

Acknowledgments

First of all, I want to thank my advisor Dr. Brian Caffo. He not only guided me through many research projects, but also taught me how to take on more responsibility and how to lead research. To me, he is the role model for a great researcher, mentor and team leader. His impact on me went beyond just research. Dr. Caffo always taught me to be an open-minded person and always kept me optimistic about things. To me, he is one of the most wise, knowledgable, inspiring and kind people in my life. Throughout the five years, I have learned so much from him and I feel that I am so lucky to have him as my academic advisor for my PhD study.

Second, I want to thank my co-advisor Dr. Vadim Zipunnikov. I have learned so much from him through our collaborations on the PS and the MBPCA project. He gave great insights on dimension reduction related models. While always keeping me focused on the big picture of my research, he also taught me that it is important to be detail oriented. With Vadim's help, I learned so much on how to conduct high quality research.

Third, I want to thank all my collaborators. I want to thank Dr. Craig Hendrix and his lab for our collaboration on dual isotope and P24 projects. I also want to thank Dr. Martin

Pomper, Dr. Jennifer Coughlin and Dr. Yuchuan Wang for our collaboration on the NFL head concussion project. Also I want to thank Dr. Haris Sair and Dr. Raag Airan for the collaboration on the GICC project.

Fourth, I also want to thank all my committee members, Dr. Brian Caffo, Dr. Vadim Zippunikov, Dr. Craig Hendrix, Dr. Elizabeth Platz, Dr. Michelle Carlson and Dr. Martin Lindquist for all the valuable advice on my thesis. The detailed comments to the thesis significantly improve the quality of presenting my research.

Meanwhile, I want to thank the Department of Biostatistics at Hopkins for so many reasons. Dr. Scharfstein, Dr. Caffo, Dr. Crainiceanu, Dr. Ji, Dr Leek, Dr. Frangakis, and many other professors, lead all the excellent courses I have taken. These courses provide me with a solid statistics and probability background. I would also like to thank the Department Chair, Dr. Bandeen-Roche, for her understanding and support of my career choice. Also I want to thank Dr. Diener-West for her great help on my teaching assistantship. In addition, I would like to thank all the faculty/staff members who provides me with all the assistance necessary for conducting cutting edge statistical research, as well as helping create a warm and friendly environment.

I also want to thank my fellow graduate students. I have learned a lot from each of them through the many projects we have collaborated on and the many courses that we have taken together. Special thanks go to Shaojie Chen, Lei Huang, Huitong Qiu, Dengtian Deng and Yuting Xu, whom make me feel as though a part of a big extended family.

Last but not the least, I want thank my parents for their love and support. I am always

proud of them and I hope this time I make them feel proud too. Also thank my grandpa, grandma, my aunt, uncle and my cousins for their caring and support.

Thank you to my girlfriend, Xueyu Feng, for her support and love.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Statistical challenges in biomedical research

With the increasing involvement of cutting edge technology in biomedical research, innovative statistical methodologies are needed like never before. Advanced 3D scanners allow us for acquiring more precise brain images; wearable devices enable the analysis of people's daily activity. These lead to data in much higher dimension as well as much more complex structures. As a result, there is an increasing demand for new statistical methodology to analyze this new kind of data. We hope our research can help in detecting diseases in early stage, reducing mortality rates and improving the overall quality of people's life.

1.1.1 Brain imaging analysis

Several brain related diseases such as multiple sclerosis (MS), autism and attention deficit hyperactivity disorder (ADHD) have been researched for decades. The prevalence for many of them remains in a relatively high level worldwide and large efforts have been put into discovering the causes and biological consequences. In the case of MS, the total number of death caused by the disease in 2013 went up to 20,000 from 12,000 in 1990 (Naghavi et al., 2015). In the year of 2015, more than 2.5 million people are suffering from MS worldwide (Pietrangelo and Higuera, 2015). However, the underlying cause of MS remains unclear and there is no known cure so far. For such diseases, early detection can definitely help increase the survival rate, as well as the life quality of patients. New brain imaging techniques enables one to obtaining higher resolution images which may lead to

an earlier detection of such diseases. It motivates us to develop modern statistical tools for analyzing those high resolution image data and relates such data with diseases such as MS.

1.1.2 Data reproducibility

Another big challenge in biomedical research concerns data reproducibility. Low quality data usually result in misleading conclusions. For example, noise, artifacts and processing errors in brain scans or poor analysis methodologies can lead to poorly measured or incorrect brain connectivity graphs. Therefore, measuring the reproducibility of data is one of the most important tasks for biostatisticians. The intra-class correlation coefficient was propose for such purpose (Fisher et al., 1970). However, larger sized data and new formats require innovative statistical models for evaluating measurement reproducibility. This motivates us to develop a new method for measuring the reproducibility of binary graphs.

1.1.3 Daily activity

One of the most challenging tasks in biomedical research is collecting data. Researchers could only dream about collecting one's physical activity data on a daily base until the appearance of the wearable device that tracks peoples movements. Now, the explosion of wearable devices, such as wristbands, smart phones and watches makes the collection of daily activity data commonplace. As the dimension, as well as the number of subjects, dramatically increase, advanced statistical methods are needed for analysis purpose. It is very

interesting to understand the relationship between daily activity patterns and demographics features such as gender, age and body mass index (BMI). and it is also super interesting if the results could be used to improve people's health condition by changing one's activity patterns. The goal of such research is using the results to improve health and increase healthy activity.

1.2 Dimension reduction in high dimensional data

All of the above biomedical challenges require innovative statistical solutions that deal with high dimensional data. Therefore, my research focuses on the task of data dimension reduction techniques for these challenging scenarios.

A traditional common methodology for dimension reduction is principal component analysis (PCA). It serves as a powerful tool for reducing data dimension by approximating the data using several the top linear reorganizations that explain the greatest amount of variation. This method has been heavily used in biomedical research. However, as data become higher in dimension and more complex in structure, increasingly there are cases where PCA does not apply directly. For instance, original PCA does not apply to dataset with clear nonlinear structure. Another example is that the data might be in a categorical or binary format. In both cases, generalizations or extensions of PCA is necessary to achieve dimension reduction.

This research aims to make necessary generalizations and extensions for PCA as well

as applications for new types of data, as outlined in the next section.

1.3 Organizational overview

In this research, three novel statistical methodologies are developed to deal with dimension reductions for complex data.

First, we are concerned with data generated from a diffusion tensor imaging (DTI) experiment. The goal is to parameterize manifold-like white matter tracts, such as the corpus callosum, using principal surfaces. The problem is approached by finding a geometrically motivated surface-based representation of the corpus callosum and visualized fractional anisotropy (FA) values projected onto the surface; the method applies to any other diffusion summary as well as to other white matter tracts. An algorithm is proposed that 1) constructs the principal surface of a corpus callosum; 2) flattens the surface into a parametric 2D map; 3) projects associated FA values on the map. The algorithm is applied to a longitudinal study containing 466 diffusion tensor images of 176 multiple sclerosis (MS) patients observed at multiple visits. For each subject and visit the study contains a registered DTI scan of the corpus callosum at roughly 20,000 voxels. Extensive simulation studies demonstrate fast convergence and robust performance of the algorithm under a variety of challenging scenarios.

Second, the image intra-class correlation coefficient (I2C2) is generalized and the graphical intra-class correlation coefficient (GICC) is proposed for such purpose. The concept

for GICC is based on multivariate probit-linear mixed effect models. A Markov Chain Monte Carlo EM (mcmcEM) algorithm is used for estimating the GICC. Simulation results with varied settings are demonstrated and our method is applied to the KIRBY21 test-retest dataset. This proposed method, though is not a directly generalization of PCA, serves as a prerequisite of our third method, multilevel binary principal component analysis (MBPCA)

In the third method, we extend PCA to multilevel bineary data. Similar to the second method of GICC, our framework is build on a mixed effect model. We use the framework of probabilistic PCA and the models are fitted by a variational EM algorithm. The performance of the proposed method is studied in a few challenging simulation scenarios. We also apply the method to a functional magnetic resonance imaging (fMRI) data and explore a reproducibility of the results through the graphical intra-class correlation coefficient.

Chapter 2

Parametrization of white matter manifold-like structures using principal surfaces

CHAPTER 2. PARAMETRIZATION OF WHITE MATTER MANIFOLD-LIKE STRUCTURES USING PRINCIPAL SURFACES

Abstract

In this manuscript, we are concerned with data generated from a diffusion tensor imaging (DTI) experiment. The goal is to parameterize manifold-like white matter tracts, such as the corpus callosum, using principal surfaces. The problem is approached by finding a geometrically motivated surface-based representation of the corpus callosum and visualized fractional anisotropy (FA) values projected onto the surface; the method applies to any other diffusion summary as well as to other white matter tracts. An algorithm is proposed that 1) constructs the principal surface of a corpus callosum; 2) flattens the surface into a parametric 2D map; 3) projects associated FA values on the map. The algorithm is applied to a longitudinal study containing 466 diffusion tensor images of 176 multiple sclerosis (MS) patients observed at multiple visits. For each subject and visit the study contains a registered DTI scan of the corpus callosum at roughly 20,000 voxels. Extensive simulation studies demonstrate fast convergence and robust performance of the algorithm under a variety of challenging scenarios.

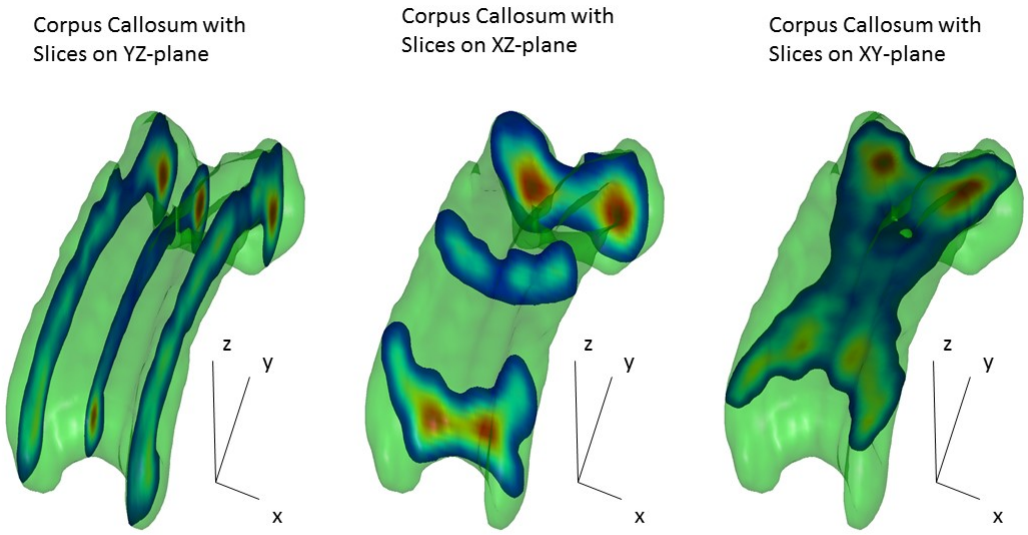
Keywords: corpus callosum, principal curves and surfaces, thin plate splines

2.1 Introduction

This research is motivated by the need to establish a parametric description of the corpus callosum structure, a major white matter fiber tract. The corpus callosum is a centrally located white matter structure connecting the left and right hemispheres of the brain. It is the largest white matter fiber bundle in the brain, facilitating inter-hemispheric communication. Corpus callosum neurons run left to right spanning the mid-sagital plane.

The three panels in Figure 2.1 display that the corpus callosum appears as a two dimensional manifold in its principal structure. It is curved towards the inferior part of the brain on both the anterior and posterior sides. Though the corpus callosum lies in a three dimensional space, its key structure is intuitively that of a "carpet" that lies in a two dimensional manifold. Therefore, dimension reduction techniques may provide strong data compression along with novel visualization and parametrization approaches that could be easy to use in practice.

Figure 2.1: 3D renderings and 2D projections of the corpus callosum of one subject. Three panels represent sagittal, frontal and horizontal 2D Slices respectively. Red and blue colors indicate higher and lower FA values on 2D slices, respectively.



In this manuscript, a method for obtaining smooth principal surfaces of the data with a consistent parametrization is proposed. The word "consistent" implies that one would achieve a similar parametrization with the same range for assembled data structures. For example, one could obtain a comparable parametrization for corpus callosum of different subjects. To achieve this, the first step is to estimate the center surface of the corpus callosum. The second is to obtain the projection of each data point; these projections can then be used to map the diffusion properties of the corpus callosum onto a 2D manifold. The procedure, similar to Tract-Based Spatial Statistics (TBSS) (Smith et al., 2006), or other medial model based methods (Yushkevich et al., 2008; Zhang et al., 2010), constructs a

consistent mapping of DTI images of the corpus callosum which has the potential to perform pixel-wise analysis, but unlike these methods, may not require non-linear registration. At the same time, our approach also has the potential to perform pixel-wise analysis of FA value, local thickness and curvature.

Before we discuss the details of our method, other potential methods are outlined. Principal component analysis (PCA) is one of the most useful tools for dimension reduction. PCA finds directions (vectors) that explain the largest variability of the data, while constraining directions to be orthogonal.

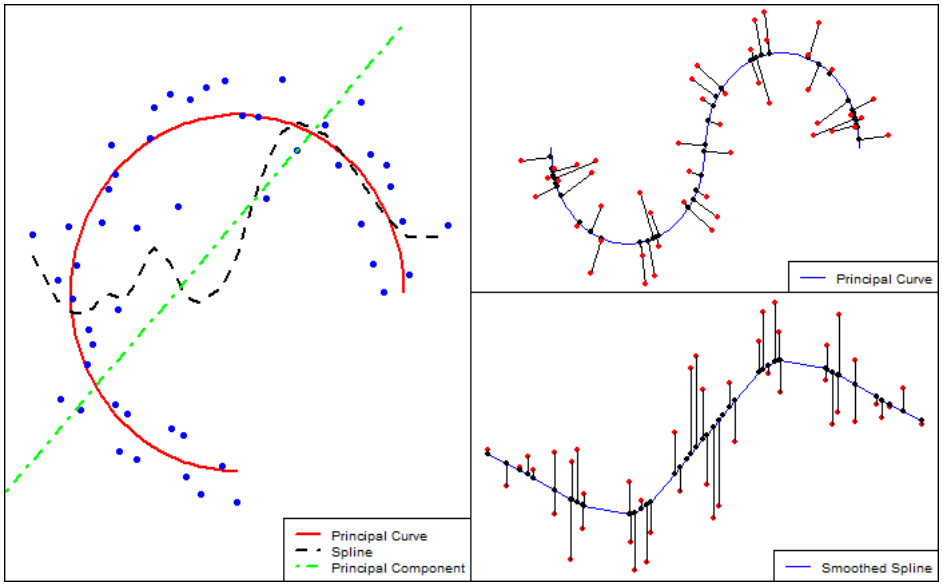
Though PCA is a widely used method for dimension reduction, it is not suitable for nonlinear situation – as shown in the left panel in Figure 2, the true curve that generates the data is part of a circle while the first principal component is a line. Direct applications of splines, wavelets and related regression methods cannot be done, as they require the mean function to be one-to-one. This is clearly violated in the corpus callosum example and many other white matter structures/tracts. Palus and Dvorak (1992) and Palus and Dvorak (1992) illustrated the pitfalls and precautions when applying linear PCA in non-linear settings. In our case, since the corpus callosum clearly is a nonlinear structure, PCA is not a viable candidate for dimension reduction.

Many non-linear methods have been proposed for fitting non-linear data structures. As an example, Gnanadesikan (1997) proposed a non-linear extension of PCA. The core idea is to include product combinations of the variables in the data matrix. Another useful tool – the self-organizing map (SOM) – was proposed in Kohonen (1990) and Kohonen (1982).

SOMs are unsupervised learning procedures which are used to discover structure in the data. Other nonlinear method methods such as non-linear principal component analysis (NLPCA) (Kramer, 1992, 1991) and principal geodesic analysis (PGA) (Fletcher et al., 2004; Fletcher, 2004) can also fit non-linear data structures.

An important concept of non-linear data compression is principal surfaces (Hastie, 1984; Hastie and Stuetzle, 1989). Principal surfaces are manifolds that pass through the middle of the data. Principal surfaces, by definition, satisfy a *self consistency* condition, in that they are the conditional expectation (local average) of the data. Hastie (1984) and Hastie (1984) shows that the principal surfaces are fit via nonparametric low-dimensional manifolds that minimize the orthogonal distance from the data to themselves. The right panel of Figure 2.2 shows the difference between the principal curves (surfaces) and regression. It highlights that the principal curve minimizes the sum of orthogonal distances, while a spline model fit tries to minimize the sum of distances parallel to the y axis. The principal surface algorithm and its extensions Dong and McAvoy (1996); Einbeck et al. (2010); Gerber et al. (2009); Goldsmith et al. (2011a); Jung et al. (2011); Leblanc and Tibshirani (1994); Ozertem and Erdogmus (2011), do not provide a consistent parametrization in the 2D space (see Section 3). ISOMAP (Tenenbaum et al., 2000) and Maximum Variance Unfolding (MVU) (Weinberger and Saul, 2006) are methods for dimension reduction. Compared to the Hastie's algorithm (denoted as "HS") , ISOMAP and MVU provide more consistent parametrization (see Section 3).

Figure 2.2: Left panel is a illustration of different dimension reduction methods. The blue points are the original data points, the dot-dashed green line is the first principal component, the dashed black curve is the spline fitting and the solid red curve is one of the principal curves. Right panel shows the difference between the principal curve and the regression method: top panel shows that the principal curve minimizes the orthogonal distance, bottom panel shows that the spline regression minimize the distance in y axis. Both panels in the middle use the same dataset.



In this manuscript, we develop a method that 1) achieves a principal surface of a target data cloud in 3D space, 2) yields a consistent parametrization in the 2D space for similar 3D data clouds. The rest of the manuscript is laid out as follows: In section 2, the principal surface concept will be introduced, and its corresponding algorithm will be shown. We will show how the principal surface algorithm works on simulated data in section 3. The algorithm then is applied to corpus callosum data and the FA maps are obtained in section 4. We will conclude the whole paper in section 5.

2.2 Methods

2.2.1 Principal Surfaces

Let $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})^T,\ i = 1, \ldots, I$ be the data points in three dimensional space, $\mathcal{R}^3$ following an underlying distribution. $\mathbf{t}_i = (t_{i1}, t_{i2})^T$ be corresponding parametrization points in two dimensional space, $\mathcal{R}^2$. In addition, It is required (without loss of generality) that the 2D coordinate space be the unit square $[0, 1] \times [0, 1] \subset \mathcal{R}^2$. Let f be the smooth principal surface function, $f : \mathbf{t}_i \mapsto f(\mathbf{t}_i)$, that maps from $\mathcal{R}^2$ to $\mathcal{R}^3$. The principal surface function satisfies the *self-consistency* condition:

$$E(\mathbf{X}|\lambda_f(\mathbf{X}) = \mathbf{t}) = f(\mathbf{t}) \text{ for all } \mathbf{t}, \tag{2.1}$$

where $\lambda_f(\mathbf{x}) = \sup_{\mathbf{t}} \left\{ \mathbf{t} : \|\mathbf{x} - f(\mathbf{t})\| = \inf_{\mu} \|\mathbf{x} - f(\mu)\| \right\}$ is the projection function with respect to f. The projection function maps a data point on to the closest principal surface point having the largest parametrization. Intuitively, the self-consistency condition implies that the principal surface is the local average of the original data cloud. Here local means that data points have the same 2D parametrization. In developing an algorithm of achieving the principal surface of a data cloud, we have found that there are two main distinctions between different fitted principal surfaces: the degree of smoothness and the method of parametrization. In most algorithms, these properties will be controlled by the specific smoother being used in the algorithm and its tuning parameters. The details of the specific

algorithm will be demonstrated in the next section.

2.2.2 Algorithm

In this paper, we put forward an algorithm that is partially based on (Hastie and Stuetzle, 1989) but heavily modified for applications where underling structures are smooth manifolds. This algorithm allows one to find the surface coordinate for each data point $(\mathbf{t}_i, i = 1, \ldots, I)$, which will be used later to create parametric summaries. However, the original principal surface algorithm can only yield surfaces which are locally flattened. Therefore, instead of local planar smoothers, thin-plate splines (TPS) are employed for fitting the surface. (Thin-plate splines were proposed by Duchon (1977) and are now widely used for bivariate smoothing.) The TPS penalize the least squares error by a high-order derivative term in order to achieve a desired degree of smoothness. Wood (2003) and Wood (2003) improved the computational efficiency when fitting TPS by using an optimal approximating basis that we employ.

Initializing. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_I]^T$ be the $I \times 3$ matrix that contains the centered coordinates. There are a few ways to initialize our algorithm:

PCA. Let $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the singular value decomposition of $\mathbf{X}$. Then $\mathbf{U}\tilde{\mathbf{\Sigma}}$ are the first two principal scores of the data matrix, where $\tilde{\mathbf{\Sigma}}$ is a submatrix of $\mathbf{\Sigma}$ containing the first two columns. Both scores are standardized to be in $[0,\ 1]$. and then used as an initial 2D parametrization.

ISOMAP. Construct a 3D to 2D ISOMAP on $\mathbf{X}$ and use its results as the initial 2D parametrization. ISOMAP usually provides a more natural initializing parametrization, but requires more computational time.

Smooth Local Averages. Bivariate thin plate splines (TPS) (Wood, 2003) are applied to the original data points using the projection coordinates obtained from the previous step. In this step, the model

$$x_{id} = f_{d1}(t_{i1}) + f_{d2}(t_{i2}) + f_{d3}(t_{i1}, t_{i2}) + \epsilon_{id}, \quad d = 1, 2, 3, \tag{2.2}$$

is fit and bivariate TPS smoothing is obtained. Here

$$f(\mathbf{t}_i) = \begin{bmatrix} \hat{f}_{11}(t_{i1}) + \hat{f}_{12}(t_{i2}) + \hat{f}_{13}(t_{i1}, t_{i2}) \\ \hat{f}_{21}(t_{i1}) + \hat{f}_{22}(t_{i2}) + \hat{f}_{23}(t_{i1}, t_{i2}) \\ \hat{f}_{31}(t_{i1}) + \hat{f}_{32}(t_{i2}) + \hat{f}_{33}(t_{i1}, t_{i2}) \end{bmatrix} \tag{2.3}$$

is the current principal surface mapping from the 2D parametrization space to 3D coordinate space. The TPS are fit using *mgcv* package in *R* (R Development Core Team, 2008).

Projection. Each data point was then projected onto the current principal surface and a new 2D parametrization was thus obtained. A grid search method was used to find the projection and constrained within the domain $[0, 1] \times [0, 1]$. Therefore, there will typically be some data points being projected onto the boundary, which raises some issues when further analyzing the 2D parametrization, see Section 5 for more discussion.

After the projection step, the procedure is iterated. The algorithm is illustrated in pseudocode below:

Input: Data in 3D coordinate, $\mathbf{X}_{I\times 3}$

Output: Principal surface function, $f_{\mathcal{R}^2\to\mathcal{R}^3}$, and the 2D parametrization of all data points, $\mathbf{T}_{I\times 2}$

Initialization: De-mean $\mathbf{X}$ for each column;

set the initial 2D parametrization as $\mathbf{T}^{(0)}$;

Set *err*$= 1$, i$= 1$;

while *($i <$max.iter* **and** *err>thres)* **do**

(1). Fit $\mathbf{X} = f(\mathbf{T}) + \epsilon$;

(2). $\mathbf{t}_i^{(new)} \longleftarrow arg\min_{\mathbf{t}} \|\mathbf{x}_i - f(\mathbf{t})\|^2$;

(3). $err \longleftarrow \|\mathbf{T^{old}} - \mathbf{T^{(new)}}\|_2^2,\ i \longleftarrow i + 1$;

end

Algorithm 1: Principal surface algorithm

To summarize, the algorithm iterates between two main stages: a smoothing step (Steps 1) and a projection step (Step 2). The TPS fitting step provides local averages with desired smoothness as well as a principal surface to be projected onto. Step 3 gives us a criterion for convergence.

The only parameter that needs to be chosen is the number of basis functions during TPS fitting step. This controls the complexity of the surface.

Generally, a principal surface is not unique. By varying the number of basis functions in the TPS fitting, one can achieve different principal surfaces with different curvature (see Section 2.3.3 for more discussion).

2.3 Simulations

2.3.1 Simulation Settings

To investigate the performance of our algorithm, four simulations studies were conducted. The number of data points was set to $I = 1500$ in all simulation settings and $500,\ 1000,\ 1500$ points in each scenarios were sub-sampled to assess the accuracy of the algorithm. We apply the proposed principal surface fitting alogirithm with both PCA and ISOMAP initializations. In addition, since ISOMAP and MVU naturally produce a 2D parametrization, we apply both of them[1] to each simulation case for comparison. HS's original principal surface algorithm was also applied. The data were centered around $(0, 0, 0)$ beforehand.

Scenario 1 The data points in the first simulation scenario are uniformly distributed around a cylinder with an open seam. Set $\theta_i \overset{i.i.d}{\sim} U(0, 2\pi - 0.5)$, $\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3} \overset{i.i.d.}{\sim} N(0, 0.15^2)$.

[1] R package "vegan" was used to run ISOMAP. Landmark MVU with one time TPS fitting was fit in "Matlab Toolbox for Dimensionality Reduction.

and $z_i \overset{i.i.d.}{\sim} U(-3, 3)$, then let

$$\mathbf{x}_i = \begin{bmatrix} \cos\theta_i + \epsilon_{i1} \\ \sin\theta_i + \epsilon_{i2} \\ z_i + \epsilon_{i3} \end{bmatrix}.$$

Scenario 2 The second scenario is a simulated corpus callosum. Set $\theta_i \overset{i.i.d.}{\sim} U(\pi/2, 3\pi/2)$, $z_{i1} = \cos(\theta_i)$, $z_{i3} = -1 + \sin(\theta_i)$ for $i = 1, \ldots, I/3$. $z_{i1} \overset{i.i.d.}{\sim} U(0, 3)$ and $z_{i3} = 0$ for $i = I/3+1, \ldots, 2I/3$. Set $\theta_i \overset{i.i.d.}{\sim} U(-\pi/2, \pi/2)$, $z_{i1} = 3+\cos(\theta_i)$, $z_{i3} = -1+\sin(\theta_i)$ for $i = 2I/3 + 1 \ldots, I$. $z_{i2} \overset{i.i.d.}{\sim} U(0, 5)$ for all i. Then let $\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3} \overset{i.i.d}{\sim} N(0, 0.16^2)$ for all i.

$$\mathbf{x}_i = \begin{bmatrix} z_{i1} + \epsilon_{i1} \\ z_{i2} + \epsilon_{i2} \\ z_{i3} + \epsilon_{i3} \end{bmatrix}.$$

Scenario 3 In the third scenario the data points form a flatten surface at the beginning and then begin to bent over towards the bottom, which looks like a half of a corpus callosum. Set $z_{i1} \overset{i.i.d.}{\sim} U(0, 2)$, $z_{i3} = 0$ when $i = 1, \ldots, I/2$; $z_{i1} = \cos(\theta_i) + 2$, $z_{i3} = -1+\sin(\theta_i)$ for $i = I/2+1, \ldots, I$, where $\theta_i \overset{i.i.d.}{\sim} U(-\pi/2, \pi/2)$. Let $z_{i2} \overset{i.i.d.}{\sim} U(0, 10)$

and $\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3} \overset{i.i.d.}{\sim} N(0, 0.4^2)$. All the random numbers are generated independently and

$$\mathbf{x}_i = \begin{bmatrix} z_{i1} + \epsilon_{i1} \\ z_{i2} + \epsilon_{i2} \\ z_{i3} + \epsilon_{i3} \end{bmatrix}.$$

Scenario 4 In the last scenario, a two dimensional stretched digit "5" was used as a basis for simulation. Let $z_{i1} \overset{i.i.d.}{\sim} U(0,1)$, $z_{i3} = 0$ when $i = 1\ldots, 3I/10$; $z_{i1} = 0$, $z_{i3} \overset{i.i.d.}{\sim} U(-1,0)$ when $i = 3I/10 + 1, \ldots, 4.5I/10$; $z_{i1} \overset{i.i.d.}{\sim} U(0, 0.5)$, $z_{i3} = -1$ when $i = 4.5I/10 + 1, \ldots, 6I/10$; $z_{i1} = \frac{1}{2} + \frac{1}{2}\cos(\theta_i)$, $z_{i3} = -\frac{3}{2} + \frac{1}{2}\sin(\theta_i)$ for $i = 6I/10 + 1, \ldots, 8.5I/10$ where $\theta_i \overset{i.i.d.}{\sim} U(-\pi/2, \pi/2)$; $z_{i1} \overset{i.i.d.}{\sim} U(0, 0.5)$, $z_{i3} = -2$ when $i = 8.5I/10 + 1, \ldots, I$. Let $z_{i2} \overset{i.i.d.}{\sim} U(0, 5)$ and $\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3} \overset{i.i.d.}{\sim} N(0, 0.1^2)$ when $i = 1, \ldots, I$. All the random numbers are generated independently and

$$\mathbf{x}_i = \begin{bmatrix} z_{i1} + \epsilon_{i1} \\ z_{i2} + \epsilon_{i2} \\ z_{i3} + \epsilon_{i3} \end{bmatrix}.$$

2.3.2 Simulation Results

For all four scenarios, the proposed algorithm converged in less than 10 steps and took under two minutes on a i7-2.4GHz PC machine with 8Gb RAM memory. The results of our principal surface fitting algorithm for four scenarios are shown in Figure 2.3. In the upper right panel in Figure 2.3, one can see that the algorithm reconstructs the cylinder very well. Since no constraints were used in fitting the surface, a closed cylinder is not obtained. Finding principal surfaces for closed cylinders or spheres remains an interesting topic for future research. The other panels of Figure 2.3 show the fitting results of the non-function shaped carpet (half corpus callosum) data cloud, the corpus callosum-shaped structure and the simulated digit "5" data cloud. The results are excellent for all scenarios. Given that the desired corpus callosum model fit is simpler and smoother than all considered examples, the simulations produce substantial evidence of the viability of the principal surface algorithm as a robust method for a wide range of problems.

Figure 2.3: Simulation results. In all panels, original data points are shown in green and the fitted surfaces are shown in blue. The panels from top-left to bottom-right show the results for our method with 1,000 sub-sampling points, initiated by ISOMAP with the three, four, four and five basis functions respectively implementing on (1) a non-function shaped "carpet" data cloud, (2) an cylinder with an open seam on one side, (3) the CC-shaped structure and (4) a simulated digit "5" data cloud.

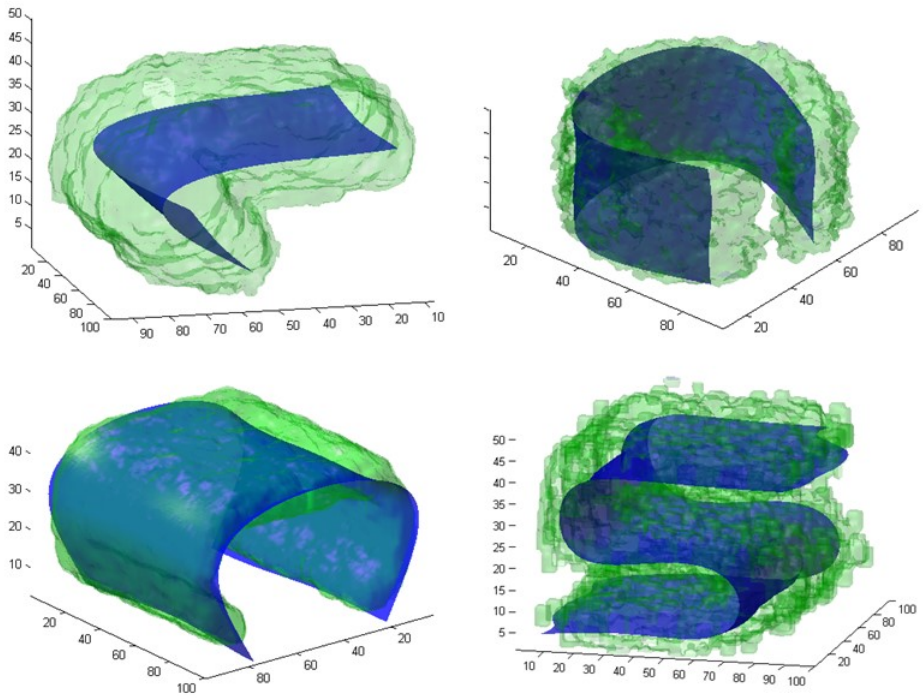


Figure 2.4 compares the fitting result using (1) proposed method with PCA initialization, (2) proposed method with ISOMAP initialization, (3) ISOMAP induced surface[2] and (4) HS's original method. The proposed principal surface algorithm generates reasonable fitting results with both initialization methods. Landmark MVU algorithm did not converge in all simulation cases, so the results are not reported.

[2]ISOMAP induced surface is created using ISOMAP 3D to 2D parametrization with TPS smoothing with each coordinates. It can be considered a surface fitting of ISOMAP result or the proposed method with only one step without any projection and further iterations.

Figure 2.4: Principal surface fitting of digit "5" results comparison between different methods. Upper left shows the fitting result of the proposed method initiated by principal scores from PCA, upper right panel shows the result of the proposed method initiated by ISOMAP, bottom left panel shows the result of ISOMAP and bottom right panel shows the result of HS's principal surfaces.

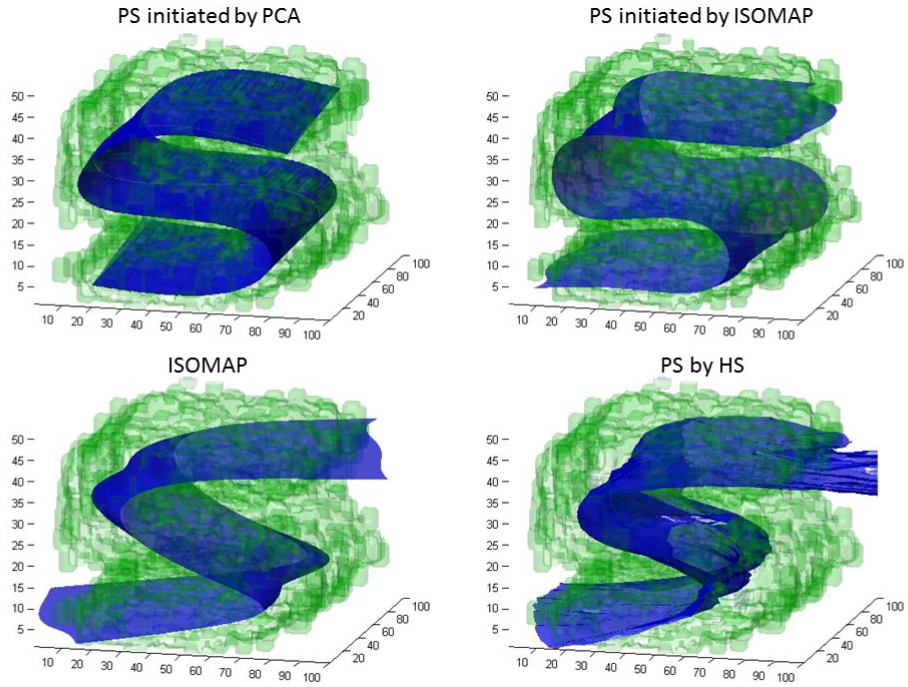


Figure 2.5 shows the results for different number basis functions in TPS fitting. Starting from upper left panel, the surface fitting with $k = 20$ yields a fitted surface with less curvature. As k increases, the fitted surface becomes increasingly wiggly. Table 2.1 shows the curvature[3] and the corresponding thickness[4] of the fitted surface. It is very clear that as the number of basis functions increases, the curvature of fitted surface also increases and the corresponding thickness decreases. In Figure 2.5, it is clear that a desired surface for this particular case is with k less than or equal to 20.

[3]Integral of the second order derivatives

[4]A moving maximum of the distance from 3D data points to the fitted surface.

Figure 2.5: Six panels demonstrate the fitting results of a Corpus Callosum-shaped structure with different number of basis in the TPS fitting. From top left to bottom right, the number of basis increases from $k = 20$ to $k = 25$.

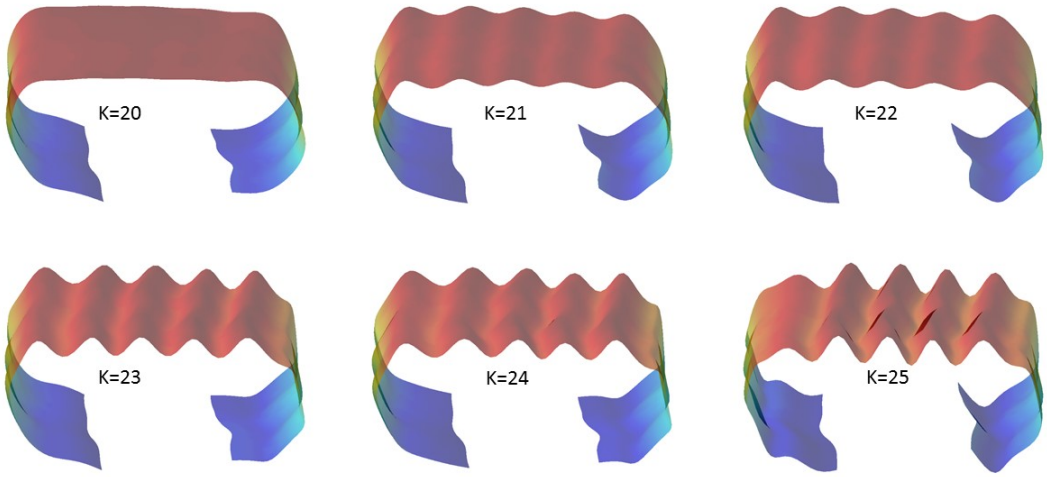


Table 2.1: Curvature and thickness comparison between different number of basis functions. Thickness is in parenthesis.

Number of Basis functions	Scenarios			
	Carpet	Cylinder	Corpus	Five
4	26.174 (0.152)	41.703 (0.035)	46.051 (0.031)	30.065 (0.017)
	0.204	0.054*	0.058	0.041*
5	32.824 (0.146)	77.866 (0.031)	48.592 (0.031)	31.228 (0.016)
	0.264	0.147	0.058*	0.046
6	69.522 (0.124)	94.234 (0.023)	56.851 (0.030)	36.454 (0.012)
	0.252	0.138	0.063	0.075
7	132.800 (0.093)	118.111 (0.020)	60.712 (0.027)	38.121 (0.012)
	0.196	0.127	0.170	0.050
8	197.029 (0.076)	136.875 (0.018)	59.546 (0.027)	43.552 (0.011)
	0.193*	0.150	0.154	0.064
9	259.639 (0.063)	164.889 (0.017)	64.241 (0.026)	45.822 (0.010)
	0.214	0.128	0.167	0.045
10	308.302 (0.056)	189.778 (0.016)	65.219 (0.026)	49.801 (0.010)
	0.237	0.127	0.143	0.065

To account for a 3D (spatial) error structure, we propose a new cross-validation method for finding the optimal number of basis functions. For each k, MSE(k) is calculated in four steps: i) a reference principal surface is fitted using the entire sample; ii) the sample is splitted into n=3 folds and n principal surfaces are fitted leaving a corresponding fold out; iii) the points from each fold are projected onto the principal surface obtained by leaving the fold out and the fold-specific MSE is calculated as the squared distance between these projections and the projections obtained at step one; iv) MSE(k) is calculated as the average of the fold-specific MSEs. The optimal k is the smallest number that minimizes MSE(k).

The key difference of the proposed cross-validation approach is step three when MSE is calculated as a difference between fold-specific principal surface projections and projections onto the all data principal surface. Note that a traditional MSE would be proportional to a thickness estimates and, in contrast to the MSE at step three, would not be informative about the stability of the fit. However, one should be careful when the data are very dense so that sub-sampling results in surfaces similar to the original surface. In such case, our suggestion is to implement the algorithm with different number of basis functions and choose the most suitable one by visually exploring the fits.

In six scenarios in Figure 2.5, the leave-one-fold surface MSEs are 0.076, 0.082, 0.083, 0.101, 0.132 and 0.126, respectively. This is consistent with our observation that no more than twenty basis functions are usually enough to provide a good fit to data. For four cases in our simulation, the optimal number of basis functions are four, four, four and five for corpus callosum shaped structure, curved carpet structure, cylinder structure and digit

"five" shaped structure respectively.

2.3.3 Mean Square Error Comparison

In evaluating the goodness of fit, we used mean squared error (MSE) that measures the distance from the fitted surface to the underlying true surface. The MSE was calculated by: 1) projecting each data point onto the fitted surface, 2) averaging the squared distance from the fitted surface to the true surfaces. Two hundred simulations were conducted for different number of sub-sampling points and the average MSE in each scenario is shown in Table 2.2. The MSE generally decreases with the increase of the sub-sample size. Of course, in each scenario, full dataset yields the lowest MSE. In the first three cases, ISOMAP initialization yields lower MSE compared to PCA initialization. However, PCA has faster computing time and lower memory demands.

Table 2.2: MSE comparisons between different number of sub-sampling points. The unit is 10^{-3}.

Sub-sample	Carpet		Cylinder		Corpus		Five	
	PCA	ISOMAP	PCA	ISOMAP	PCA	ISOMAP	PCA	ISOMAP
500	56.6	33.9	44.4	31.6	25.2	12.2	10.4	10.9
1,000	55.4	27.3	32.6	27.2	20.5	11.3	9.9	10.4
1,500	55.0	27.1	33.3	28.9	19.5	10.6	9.8	10.2

The MSE comparison between the proposed method and ISOMAP and HS is reported in Table 2.3. For each scenario, the proposed method with ISOMAP initialization has the lowest MSE. HS has a significantly higher MSE in each case compared to the proposed method. ISOMAP has the highest MSE among all cases. LMVU algorithm did not con-

verge in four cases, so no results for LMVU are reported in Table 2.3. The reason for that is that though both ISOMAP and MVU find low dimensional representations of high dimensional structures, neither of them is suitable for finding a principal surface. Intuitively, both algorithms try to preserve the local geometry such as distances or angles. However, 2D projections on the principal surface clearly do not preserve these properties. Therefore, it would be more appropriate to use ISOMAP or MVU on zero thickness surface structures, not on data cloud representing 3D structures such as corpus callosum. This is the reason why MVU, while preserving both local distance and angles between neighborhood data points, does not converges in most simulation cases.

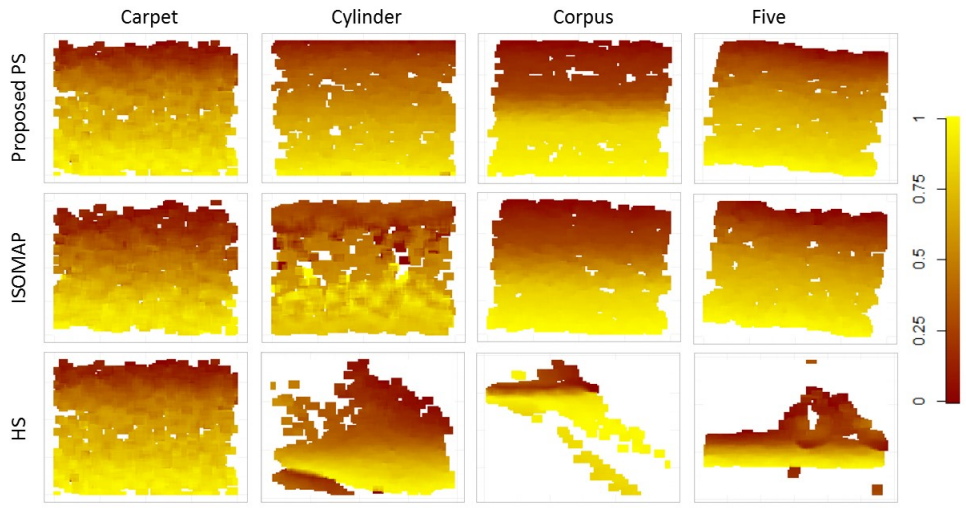
Table 2.3: MSE comparison between our method and HS, ISOMAP for all four scenarios. The unit is 10^{-3}.

	Scenarios			
Methods	Carpet	Cylinder	Corpus	Five
PS (PCA)	16.9	9.6	11.3	4.3
PS (ISOMAP)	16.3	11.5	2.5	2.1
HS	35.4	50.4	66.2	21.4
ISOMAP	94.3	476.7	296.9	164.6

Another important advantage of our method is that it obtains a more smooth parametrization than the original HS's method. In Figure 2.6, a color scale (from dark red to yellow) is used to represent the original data points. The color changes as the points move along the true surface from one side to another. After applying HS's, ISOMAP and the proposed methods, the 2D parametrizations are shown along with their original labeled colors. The proposed method as well as ISOMAP, in each scenario, shows more smooth, continuous parametrization. In contrast, the HS's method does not provide a natural parametrization

in most scenarios, and produces a clear discontinuity in the cylinder scenario. ISOMAP produces a natural parametrization, but has a disadvantage of not providing a well fitted surface. Clearly, a comprehensive cross-algorithm study is needed to systematically explore resulting parametrizations and their properties.

Figure 2.6: Parametrization results in each scenario. We colored the original points from one side to the other continuously using dark red to yellow. Each image shows the final parametrization space with each data's original labeled color. From top to bottom, *Top*: The proposed method; *Middle*: ISOMAP; *Right*: HS's method. From left to right, *Left*: Curved "Carpet" structure; *Second from left*: Cylinder structure; *Third from left*: Corpus Callosum-shaped structure; *Right*: Digit "'five'.



2.4 Application

2.4.1 Fitting the Principal Surface of a Corpus Callosum

The MS study contained 466 scans generated from a diffusion tensor imaging (DTI) experiment performed on 176 patients. For each scan, fractional anisotropy (FA) value

for each voxel of the entire corpus callosum area were calculated via tractography. The details are provided in (Ozturk et al., 2010; Reich et al., 2010). Fractional anisotropy has been associated both cross sectionally and longitudinally with multiple sclerosis diagnoses and symptoms (Goldsmith et al., 2011b, 2012; Greven et al., 2011; Reich et al., 2010; Zipunnikov et al., 2011a).

We start with the principal surface fit. For each unregistered scan, there are roughly 70,000 data points in the corpus callosum area, while for each registered scan, there are roughly 20,000 data points in the corpus callosum region of interest. For computational simplicity, 1,000 were randomly sub-sampled to build the surface. The computing time for unregistered corpus callosum is approximately 100 seconds, while the computing time for the registered ones is approximately 35 seconds. The results for two arbitrary unregistered scans are shown in Figure 2.7. The results for two arbitrary registered scans are shown in Figure 2.8. Clearly, unregistered data has a larger portion of the corpus callosum compared to the registered ones. The fitted surfaces for both registered and unregistered scans have obvious face validity and is indicative of the fits from the other scans, each inspected visually. Cross validation suggests five and seven as the optimal numbers of basis functions for registered and unregistered corpus callosum, respectively.

Figure 2.7: Principal surface fitting results of unregistered corpus callosum data. Left panels show the fitted surface with the original data cloud and right panels show the results of fitted surface.

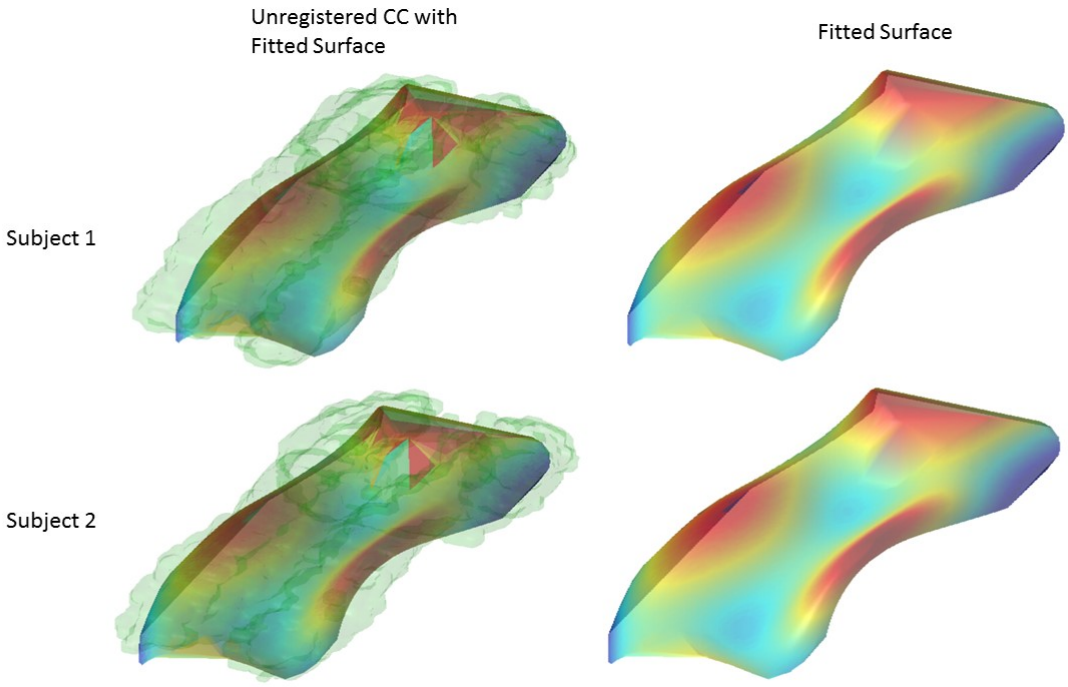
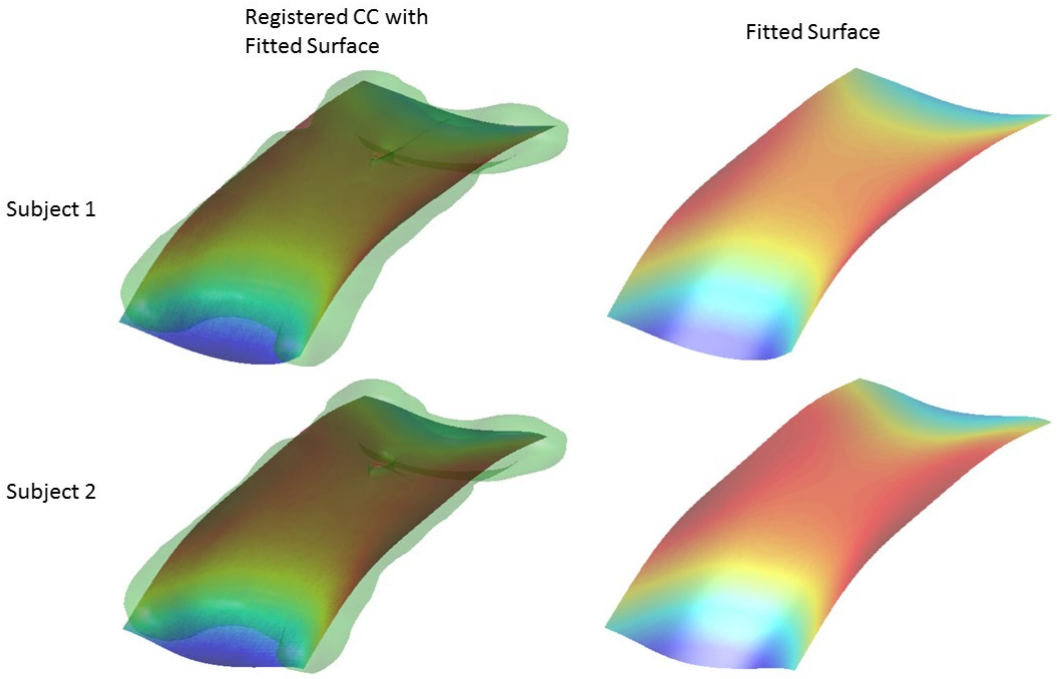


Figure 2.8: Principal surface fitting results of registered corpus callosum data. Left panels show the fitted surface with the original data cloud and right panels show the results of fitted surface.

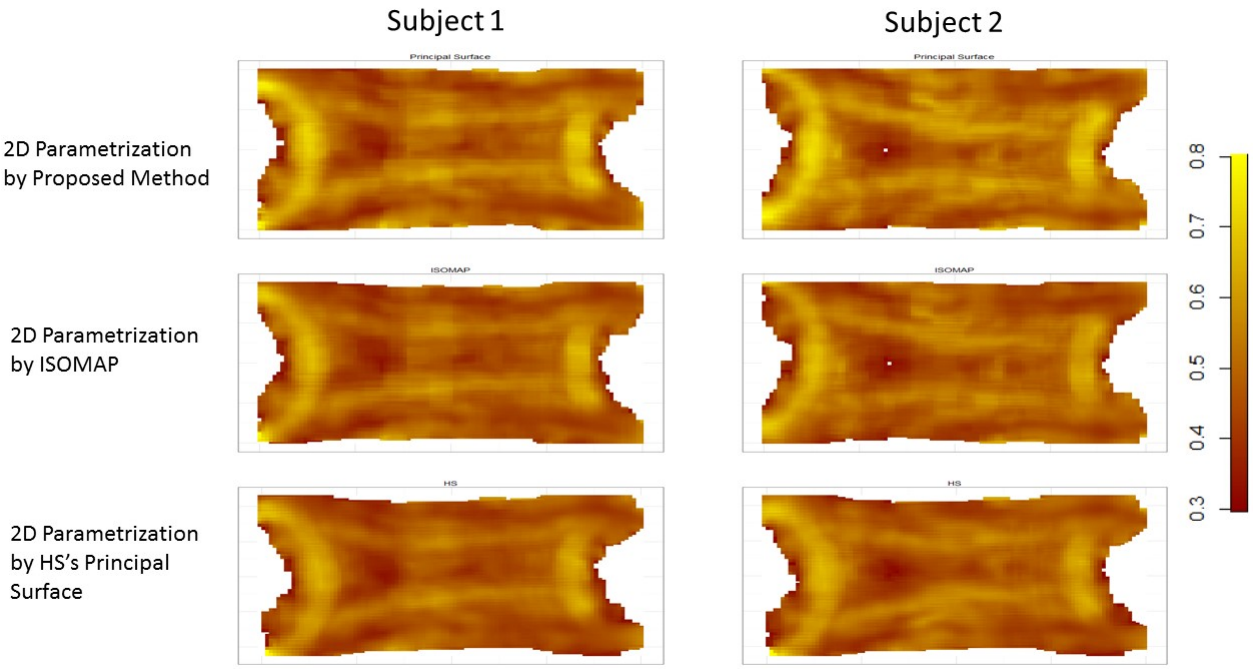


2.4.2 Flattened FA Representation

Our primary goal for the principal surface fitting is to flatten the surface to make a 2D FA image for visualization and subsequent analysis under the belief that the principal surface encodes the majority of the relevant biological information. Thus, our goal is to use DTI-based morphometric information to create 2D images of MR contrast properties, such as FA, axial diffusivity and so on. Thereby, each data point was projected onto the surface by a grid search method. Then the associated FA values were smoothed on the surface by

local averaging. Interpolating this smooth images onto a grid yields the 100×100 2D FA images, Figure 2.9 shows a few typical examples. Top panels show the diffusion properties of the flattened surfaces of the corresponding left panels. The similar 2D maps obtained from ISOMAP and HS are shown at the middle and bottom panels of Figure 2.9. For registered corpus callosum data, all three methods yield similar parametrizations. Arguably, the important FA information is retained. Thus this investigation provides the potential insight that it may be sufficient to visualize the 2D FA values instead of the original 3D FA values.

Figure 2.9: 2D parametrization of corpus callosum. Top row shows 2D parametrization from the proposed principal surface algorithm; middle row shows the result from ISOMAP and bottom row shows the result of HS's principal surface. Bright color indicates higher FA values and darker color indicates lower FA values.



2.4.3 Correspondence

For corpus callosum data, it is crucial that the fitted 2D parametrization is consistent across all subjects. In other words, it is desirable that assembled structures have aligned 2D indices after the data points got projected onto the fitted principal surfaces. We admit that stable parametrization is not guaranteed in fitting the principal surface, especially, when the registration process involves complex non-linear transformation or linear transformation in a relatively large scale. However, we provided a small simulation study to explore how sensitive our algorithm is to mild registration errors. Specifically, for an arbitrary chosen image, we created 50 "unregistered" copies by applying random rotations and random scaling along each of the three axes. We, then, ran our algorithm with ISOMAP initialization on all fifty images and compared the obtained parametrizations. Mathematically, assume $\mathbf{X} = (\mathbf{x}_1,^T, \ldots, \mathbf{x}_n^T)^T$ is the original data cloud. Let $\Phi(\mathbf{X}) := \mathbf{XUD}$, where $\mathbf{U}$ is a randomly generated 3×3 rotation matrix[5] and $\mathbf{D} = \text{diag}(d_1, d_2, d_3)$ where all $d_i's$, $i = 1, 2, 3$ are uniformly generated from $U(0.9, 1.1)$. For one scan, we randomly generated 50 transformations Φ_k, $k = 1, \ldots, 50$. Let $\mathbf{T} = (\mathbf{t}_1^T, \ldots, \mathbf{t}_n^T)^T$ be the 2D parametrization result of the original corpus callosum and $\mathbf{T}_k$ be the 2D parametrization result of $\Phi_k(\mathbf{X})$. Then $MSE_k := \frac{1}{n}\sum_i \left\{\|\mathbf{t}_{ik} - \mathbf{t}_i\|^2\right\}$ is calculated to evaluate the mismatch of the parametrization between the original data cloud and morphed data cloud. Ideally, zero MSE suggests zero

[5] 3×3 rotation matrices are generated as $\mathbf{U} = \mathbf{U}_x\mathbf{U}_y\mathbf{U}_z$, where $\mathbf{U}_x = \begin{Bmatrix} 1 & 0 & 0 \\ 0 & cos(\theta_x) & -sin(\theta_x) \\ 0 & sin(\theta_x) & cos(\theta_x) \end{Bmatrix}$, $\mathbf{U}_y = \begin{Bmatrix} cos(\theta_y) & 0 & sin(\theta_y) \\ 0 & 1 & 0 \\ -sin(\theta_y) & 0 & cos(\theta_y) \end{Bmatrix}$ and $\mathbf{U}_z = \begin{Bmatrix} cos(\theta_z) & -sin(\theta_z) & 0 \\ sin(\theta_z) & cos(\theta_z) & 0 \\ 0 & 0 & 1 \end{Bmatrix}$. θ_x, θ_y and θ_z are generated uniformly from -0.1π to 0.1π.

misalignment which means the proposed method is invariant to rotation and a restricted scaling. Notice that the parametrization in the proposed method is restricted to $[0, 1]$ and minimal mismatch in each axis is 0.01, so that a 10^{-4} MSE would suggest that on average the 2D parametrization moves one unit distance in 2D space. When the MSE is large, say over 0.09, it means that averagely the parametrization moves more than 0.3 distance (30 units) when all points are flipped over, or there are large mismatches in the 2D spaces of the morphed shape. Among 50 MSE's using ISOMAP initialization, the median is 8.97×10^{-5}, the mean is 1.81×10^{-4} and the maximum is 2.26×10^{-3}. In other words, under moderate rotation and relatively mild scaling, our algorithm can achieve a good parametrization consistency. More complex situations remain to be explored.

2.5 Discussion

In this manuscript a principal surface algorithm was introduced and used to fit the corpus callosum. The goal of this work is largely developmental, creating a handy tool for dimension reduction in morphological analysis of primary brain structures. In simulations, the proposed algorithm performed superbly. While applied to the corpus callosum, the algorithm could be applied to any other three dimensional manifold-like objects where a two dimensional surface could be embedded and is of interest, for example, all the major white matter tracts in the brain (Bazin et al., 2011). The two dimensional manifold characterizes the original data and accomplishes both dimension reduction and better visualization. The

surface that was constructed is smooth and could be easily projected onto to represent other properties of the original structure, such as the FA, mean diffusivity, parallel diffusivity, local thickness and so on. The algorithm is computationally feasible and scales well to larger images and densely measured structures.

The role of the two dimensional sub-representation needs to be further explored. Before the sub-representation routinely practised in any statistics analysis, the boundary issue mentioned in Section 3 should be carefully tackled. Due to the local averaging step, the obtained surface is a shrinkage of the original data cloud. Therefore, more points are projected onto the boundary on the domain $[0, 1] \times [0, 1]$. In applications, these boundary points could be deleted since the 2D representations on the boundary are less representative.

The corpus callosum in our application is anatomically simpler than the one considered in Yushkevich et al. (2008). Therefore, it remains to be explored how our principal surface algorithm performs in structures of higher complexity. For future work, we are developing functional data analysis tools Goldsmith et al. (2011b,c); Greven et al. (2011); Zhu et al. (2010, 2011) for relating the dimension reduced 2D manifold to outcomes of interest for the purpose of inference, biomarker creation and prediction. Of note, we are particularly interested in whether or not the 2D representation of the corpus callosum is less sensitive to issues of whole brain registration often used in the processing pipeline. In fact, it is possible that registering the 2D representation is preferable to whole brain registration a priori in certain applications. Note also that MFPCA (Di et al., 2009) and LFPCA (Zipunnikov et al., 2011a, 2014), methods have been shown to isolate registration error as a part of the model

(Lee et al., 2015), thus raising the intriguing possibility of DTI processing streams that dramatically decrease the need and importance of whole brain template-based registration.

Furthermore, current work could be easily modified to achieve dimension reduction for different input and output dimensions.[6]. For example, sagittal mid-line callosum average thickness is meaningful in the study of neurological diseases (Luders et al., 2009; Vidal et al., 2006). In addition, Zhang et al. (2010) used medial models for obtaining 2D thickness maps of corpus callosum. With the proposed method, CC mid-sagittal thickness or FA curves as well as the CC thickness or FA on principal surfaces could be achieved simultaneously.

Finally, consider that the proposed procedure can be further extended as a general problem of fitting skeleton manifolds. We are also interested in fitting surfaces with fixed boundaries. Work has been done to analyze principal curves with fixed origin and the end point (Caffo et al., 2008). The extension to surfaces seems to be quite challenging. A possible route is follows. Suppose $\underline{t} = (t_1, t_2)$ is the corresponding coordinate on the surface of the original data point, $\underline{x} = (x_1, x_2, x_3)$, which lies in 3D space. Now consider a predetermined function for $\underline{x}(\underline{t}) = (x_1(\underline{t}), x_2(\underline{t}), x_3(\underline{t}))$ when the boundary values of t are linearly constrained. Such constraints would yield cylindrical fits easily though extensions to completely closed surfaces would require more elaborate constraints.

[6] A 3D to 1D principal curve algorithm could be easily achieved by changing the input and output dimension in TPS fitting

Chapter 3

Estimating a graphical intra-class correlation coefficient (GICC) using multivariate probit-linear mixed models

CHAPTER 3. ESTIMATING A GRAPHICAL INTRA-CLASS CORRELATION COEFFICIENT (GICC) USING MULTIVARIATE PROBIT-LINEAR MIXED MODELS

Abstract

Data reproducibility is a critical issue in all scientific experiments. In this manuscript, the problem of quantifying the reproducibility of graphical measurements is considered. The image intra-class correlation coefficient (I2C2) is generalized and the graphical intra-class correlation coefficient (GICC) is proposed for such purpose. The concept for GICC is based on multivariate probit-linear mixed effect models. A Markov Chain Monte Carlo EM (mcmcEM) algorithm is used for estimating the GICC. Simulation results with varied settings are demonstrated and our method is applied to the KIRBY21 test-retest dataset.

***keywords*: graphical intra class correlation coefficient, multivariate probit-linear mixed model, MCMCEM**

3.1 Introduction

A crucial question in any statistical analysis is: how reliable is the data? Experimental replication for the purpose of measuring the reliability of measurements is the most common method for establishing reproducibility. In this paper, we consider repeated measurement of graphs and propose the concept of the graphical intra-class correlation coefficient for measuring their reliability.

The Intra-class correlation coefficient (ICC) has been proposed (Fisher et al., 1970) and used to evaluate the reliability of measurements in a variety applications (Bartko, 1966; Shrout et al., 1979). ANOVA mixed-effect models have been proposed as a framework for estimating the ICC (Stanish and Taylor, 1983). Suppose y_{ij} denotes the j^{th} measurement of subject i, x_i denotes the subject specific random effect and u_{ij} indicates the measurement error. The one-way ANOVA model is:

$$\begin{aligned} y_{ij} &= \mu + x_i + u_{ij} \\ x_i &\sim N(0, \sigma_x^2),\ u_{ij} \sim N(0, \sigma_u^2),\ i.i.d. \end{aligned} \tag{3.1}$$

The ICC is then defined as:

$$ICC = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}. \tag{3.2}$$

In (3.1) and (3.2), the total variability of the data is decomposed into subject-specific variability and measurement error; ICC represents the proportion of variability that is due to heterogeneity in subjects. In recent research, the ICC has been generalized to multivariate

cases. The work in Di et al. (2009) proposed a model analogous to (1) in functional data using multilevel functional principal component analysis (MFPCA) and an image intra-class correlation coefficient (I2C2) was subsequently proposed in Shou et al. (2013) to calculate ICC for image data.

Graphical data are becoming increasingly popular in scientific research. Notably, graphs are used in describing brain networks in neuroimaging. In such research, binary graphs are often obtained from functional magnetic resonance image (fMRI) (Di Martino et al., 2008; Guye et al., 2010; Huang et al., 2010; Salvador et al., 2005; Van Den Heuvel and Hulshoff Pol, 2010). The increasing number of graphical datasets motivates us to evaluate the reliability of binary graphs.

Figure 3.1: The left panel shows a high GICC case, where graphical measurements are similar within subjects. The right panel illustrates a low GICC case, where graphical measurements are less consistent within subjects.

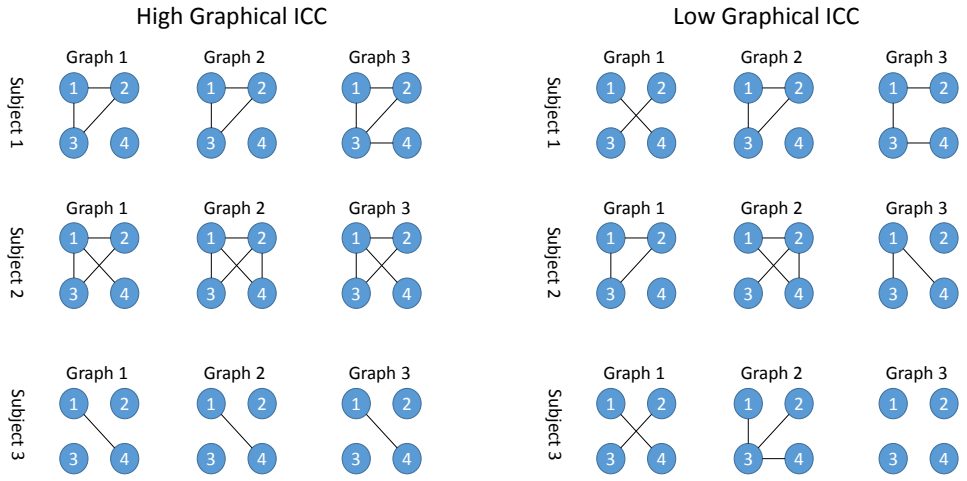


Figure.3.1 illustrates idealized graphical measurements for three different subjects. Here each subject is measured three times. The left panel shows a case where graphical measurements resemble each other within one subject. The ICC, consequentially, should be higher. The right panel, on the other hand, demonstrates the opposite situation, where the repeated measurements within one subject show poor consistency. In such case, the ICC should be relatively lower. In this manuscript, we propose the concept of the graphical ICC (GICC) to quantify the similarity between repeated measurements of binary graphs. In Figure.3.1, each binary graph is represented by a $0-1$ vector. For example, the first graph of subject 1 is represented by $(1,1,0,1,0,0)^T$[1]. Thus our goal is to define an ICC for multivariate binary data.

Many authors have discussed the ICC for single variate binary data. Ridout et al. (1999) proposed a moment based estimator. Probit linear mixed-effect models were used by Rodrıguez and Elo (2003) and Zou and Donner (2004) to estimate a confidence interval for binary data ICC.

There is also work discussing the similarity between graphs. The work in Zager and Verghese (2008) and Blondel et al. (2004) discussed the similarity between nodes and edges in graphs. One main purpose of these papers was to find assembled subgraphs between two graphs. Instead of having a fixed node-to-node or edge-to-edge match, they found the match between two graphs based on edge/node similarity score.

Our objective, on the other hand, is to estimate the ICC to evaluate the reliability of

[1]Each element of the vector is an indicator of the existence of an edge, the order of the six elements is ① − ②, ① − ③, ① − ④, ② − ③, ② − ④, ③ − ④.

replicated measurement of binary graphs. In section 3.2, a multivariate probit linear mixed model is proposed. A Monte Carlo expectation maximization (MCEM) algorithm will be discussed in section 3.3. Simulation results with various settings will be shown in section 3.4 and the results of our method being implemented on binary brain connectivity maps are in section 3.5. We will summarize the paper in section 3.6.

3.2 Model

Suppose $\{o_{ij}(d) : \ i = 1, \ldots, I; \ j = 1 \ldots, J_i; \ d = 1, \ldots, D, \}$ are binary observations representing repeated graph measurements for multiple subjects. Here, I is the total number of subjects, J_i is the number of visits for the i^{th} subject and D is the number of possible edges for all graphs. Usually, we have $D = \frac{N(N-1)}{2}$ where N is the number of nodes. In Figure.3.1, for example, we have $I = 3, \ J_i = 3, \ N = 4, \ D = 6$. The multivariate probit-linear mixed model is as follows:

$$\begin{aligned} \Phi^{-1}\big(P(o_{ij}(d)|x_i(d))\big) &= \mu(d) + x_i(d), \\ \mathbf{x}_i \sim \ \mathbf{N}(\mathbf{0}, \mathbf{\Sigma}_x), & \end{aligned} \tag{3.3}$$

where $\mathbf{x}_i = (x_i(1), \ldots, x_i(D))^T$. The GICC, is then defined as:

$$GICC = \frac{tr(\mathbf{\Sigma}_x)}{tr(\mathbf{\Sigma}_x) + D}. \tag{3.4}$$

For the purpose of estimation, the model can also be viewed as a threshold model that dichotomizes the observations from a latent Gaussian distribution. In other words:

$$\begin{aligned} o_{ij}(d) &= \mathbf{I}_{(y_{ij}(d)>0)}, \\ y_{ij}(d) &= \mu(d) + x_i(d) + u_{ij}(d), \\ \mathbf{x}_i &\sim \mathbf{N}(\mathbf{0}, \mathbf{\Sigma}_x),\ i.i.d., \\ \mathbf{u}_{ij} &\sim \mathbf{N}(\mathbf{0}, \mathbf{I}),\ i.i.d., \end{aligned} \tag{3.5}$$

where $\mathbf{x}_i = (x_i(1), \ldots, x_i(d))^T$ and $\mathbf{u}_{ij} = (u_{ij}(1), \ldots, u_{ij}(d))^T$. The equivalency of these two models can be easily shown by the following calculation:

$$\begin{aligned} P(o_{ij}(d) = 1 | x_i(d)) &= P(y_{ij}(d) > 0 | x_i(d)) \\ &= P\Big(u_{ij}(d) > -(\mu(d) + x_i(d)) \Big| x_i(d)\Big) \\ &= 1 - \Phi(-(\mu(d) + x_i(d))) \\ &= \Phi(\mu(d) + x_i(d)). \end{aligned}$$

Formula 3.4 is a direct generalization from the univariate ICC Formula 3.2. $GICC = 0$ indicates that $tr(\Sigma_x) = 0$, which means that the between group variance is zero for all dimensions, $d = 1, \ldots, D$. $GICC \approx 1$ indicates that $tr(\Sigma_x) \gg D$, implying that the variation between subjects is much larger than the variation within subject. When $GICC = 0.5$, $tr(\Sigma_x) = D$, implying that the overall between subject variation is equal to the within subject variation.

The advantage of using the $trace$ is that: (1) it is an overall statistic instead of a edge specific statistic, which provides a global measurement to quantify graphical reproducibility; (2) compared to other numerical methods (e.g. $\max(diag(\Sigma))$, $\sum(\sigma_{ij})$), the trace is invariant to orthogonal transformations, which is critical in measuring the variability of vectors.

Furthermore, using trace for measuring the global reproducibility was also proposed for the image ICC (I2C2) (see (Shou et al., 2013)) in and the functional version of ICC (see (Di et al., 2009)).

3.3 The Monte Carlo EM Algorithm

MCEM algorithms have been used in probit-linear mixed models with single variate outcomes (Chan and Kuk, 1997). Here MCEM is generalized to the multivariate case. In model 3.5, the parameters of interest are μ and $\mathbf{\Sigma}_x$. In the procedure of estimation, we treat $\mathbf{o}$ as observed data and $[\mathbf{y},\ \mathbf{x}]$ as the full data.

3.3.1 M-step

Given the full data $\mathbf{y}$ and $\mathbf{x}$, the MLE for both parameters yields an explicit form:

$$\begin{aligned} \hat{\mu} &= \frac{1}{\sum_i J_i} \sum_i \sum_j (\mathbf{y}_{ij} - \mathbf{x}_i), \\ \hat{\mathbf{\Sigma}_x} &= \frac{1}{I} \sum_i \mathbf{x}_i \mathbf{x}_i^T. \end{aligned} \tag{3.6}$$

Unlike McCulloch (1994), the estimate of μ does not involve $\mathbf{\Sigma}_x$, since $\mathbf{x}$ is also treated as part of the complete data. So $\hat{\mu}$ is obtained based on both $\mathbf{x}$ and $\mathbf{y}$, rather than only on $\mathbf{y}$.

Substituting $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{x}\mathbf{x}^T$ with $E[\mathbf{y}|\mathbf{o}]$, $E[\mathbf{x}|\mathbf{o}]$ and $E[\mathbf{x}\mathbf{x}^T|\mathbf{o}]$ respectively on the right side of 3.6, we obtain the M-step.

3.3.2 E-step

Based on 3.6, it is necessary to calculate $E(\mathbf{y}_{ij}|\mathbf{o})$, $E(\mathbf{x}_i|\mathbf{o})$ and $E(\mathbf{x}_i\mathbf{x}_i^T|\mathbf{o})$. Note that

$$\begin{aligned} E[\mathbf{x}_i|\mathbf{o}] &= E[E[\mathbf{x}_i|\mathbf{y}]|\mathbf{o}], \\ E[\mathbf{x}_i\mathbf{x}_i^T|\mathbf{o}] &= E[E[\mathbf{x}_i\mathbf{x}_i^T|\mathbf{y}]|\mathbf{o}]. \end{aligned} \tag{3.7}$$

The inner expectation can be obtained by using the joint distribution of $\{\mathbf{x}_i,\ \mathbf{y}_{i1}, \ldots, \mathbf{y}_{iJ_i}\}$.

Noticing the following fact:

$$\begin{aligned}
[\mathbf{x}_i, \mathbf{y}_{i1}, \ldots, \mathbf{y}_{iJ_i}] &= \prod_j [\mathbf{y}_{ij}|\mathbf{x}_i] \times [\mathbf{x}_i] \\
&\propto \exp\Bigg\{ -\frac{1}{2}\Big[\sum_j \big\{(\mathbf{y}_{ij} - \mu - \mathbf{x}_i\big)^T (\mathbf{y}_{ij} - \mu - \mathbf{x}_i)\big\} \\
&\quad + \mathbf{x}_i^T \boldsymbol{\Sigma}_x^{-1} \mathbf{x}_i \Big] \Bigg\} \\
&\propto \exp\Bigg\{ -\frac{1}{2}\Big[\mathbf{x}_i^T (J_i \mathbf{I} + \boldsymbol{\Sigma}_x^{-1}) \mathbf{x}_i - 2\big[\sum_j (\mathbf{y}_{ij} - \mu)\big]^T \mathbf{x}_i \Big] \Bigg\},
\end{aligned} \tag{3.8}$$

it can be derived that:

$$\mathbf{x}_i | \mathbf{y}_{i1}, \ldots \mathbf{y}_{iJ_i} \sim \mathbf{N}\Big((J_i \mathbf{I} + \boldsymbol{\Sigma}_x^{-1})^{-1} (\mathbf{y}_{i\cdot} - J_i \mu),\ \ (J_i \mathbf{I} + \boldsymbol{\Sigma}_x^{-1})^{-1} \Big), \tag{3.9}$$

where $\mathbf{y}_{i.} = \sum_j \mathbf{y}_{ij}$. Thus we have

$$
\begin{aligned}
E(\mathbf{x}_i|\mathbf{O}) =& E[E[\mathbf{x}_i|\mathbf{y}]|\mathbf{o}] \\
=& E\left[(J_i\mathbf{I} + \mathbf{\Sigma}_x^{-1})^{-1}(\mathbf{y}_{i.} - J_i\mu)|\mathbf{o}\right] \\
=& (J_i\mathbf{I} + \mathbf{\Sigma}_x^{-1})^{-1}\big(E[\mathbf{y}_{i.}|O] - J_i\mu\big), \\
E[\mathbf{x}_i\mathbf{x}_i^T|\mathbf{o}] =& E[E[\mathbf{x}_i\mathbf{x}_i^T|\mathbf{y}]|\mathbf{o}] \\
=& E\left[(J_i\mathbf{I} + \mathbf{\Sigma}_x^{-1})^{-1}(\mathbf{y}_{i.} - J_i\mu)(\mathbf{y}_{i.} - J_i\mu)^T(J_i\mathbf{I} + \mathbf{\Sigma}_x^{-1})^{-1} + (J_i\mathbf{I} + \mathbf{\Sigma}_x^{-1})^{-1}|\mathbf{o}\right] \\
=& (J_i\mathbf{I} + \mathbf{\Sigma}_x^{-1})^{-1}E\big[(\mathbf{y}_{i.} - J_i\mu)(\mathbf{y}_{i.} - J_i\mu)^T|\mathbf{o}\big](J_i\mathbf{I} + \mathbf{\Sigma}_x^{-1})^{-1} + (J_i\mathbf{I} + \mathbf{\Sigma}_x^{-1})^{-1}.
\end{aligned}
\tag{3.10}
$$

However, the term $E[\mathbf{y}_{i.}|\mathbf{o}]$ and $E[\mathbf{y}_{i.}^T\mathbf{y}_{i.}|\mathbf{o}]$ does not have an explicit form. Here we use a Gibbs sampler to approximate the conditional expectation. Notice that, given $\mathbf{o}$, the distribution of $\mathbf{y}$ is multivariate truncated normal. The Gibbs sampler for such a distribution has been discussed in Horrace (2005), Kotecha and Djuric (1999), Wilhelm and G (2013). In the Gibbs sampling cycles, we choose the burn in period to be the first $T = 200$ and treat the following $B = 500$ elements as limiting realizations from the conditional distribution of $\mathbf{y}|\mathbf{o}$. Then an empirical conditional expectation is calculated as follows:

$$
\begin{aligned}
\hat{E}[\mathbf{y}_{ij}|\mathbf{o}] &= \frac{1}{B}\sum_{b=T+1}^{T+B} \mathbf{y}_{ij}^{(b)}, \\
\hat{E}[\mathbf{y}_{i.}\mathbf{y}_{i.}^T|\mathbf{o}] &= \frac{1}{B}\sum_{b=T+1}^{T+B} \mathbf{y}_{i.}^{(b)}\mathbf{y}_{i.}^{(b)T}.
\end{aligned}
\tag{3.11}
$$

3.3.3 Observed information matrix for μ

Though we are not specifically interested in estimating μ for the graphical ICC, the estimate of μ with its standard error remains of potential interests, especially for modeling multivariate binary data using probit-linear mixed model. Louis (1982) expressed the observed information matrix in EM algorithm using the first and second derivative of the full likelihood.

Assume the observed log-likelihood is $l_o(\mathbf{o}, \theta)$ where $\theta = (\mu, \Sigma_x)$ and the full log-likelihood is $l_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \mathbf{y}, \theta)$, following Louis (1982), we have:

$$\begin{aligned} I_{\mathbf{o}}(\theta) =& E_\theta \left[-\frac{\partial^2 l_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \mathbf{y}, \theta)}{\partial \mu \partial \mu^T} \middle| \mathbf{o} \right] - E_\theta \left[\left(\frac{\partial l_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \mathbf{y}, \theta)}{\partial \mu} \right) \left(\frac{\partial l_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \mathbf{y}, \theta)}{\partial \mu} \right)^T \middle| \mathbf{o} \right] \\ &+ \left(\frac{\partial l_o(\mathbf{o}, \theta)}{\partial \mu} \right) \left(\frac{\partial l_{\mathbf{o}}(\mathbf{o}, \theta)}{\partial \mu} \right)^T . \end{aligned} \tag{3.12}$$

Let $I_{\mathbf{o}} = I_{\mathbf{o}}(\hat{\theta})$, where $\hat{\theta}$ is the maximum likelihood estimator. Then we have:

$$I_{\mathbf{o}}(\hat{\theta}) = E_{\hat{\theta}} \left[-\frac{\partial^2 l_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \mathbf{y}, \theta)}{\partial \mu \partial \mu^T} \middle| \mathbf{o} \right] - E_{\hat{\theta}} \left[\left(\frac{\partial l_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \mathbf{y}, \theta)}{\partial \mu} \right) \left(\frac{\partial l_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \mathbf{y}, \theta)}{\partial \mu} \right)^T \middle| \mathbf{o} \right] . \tag{3.13}$$

Following the same path as in the E-step, we can use Gibbs sampler and empirical averages to approximate the conditional expectation.

3.4 Simulation

3.4.1 Estimates

We set number of subjects at $I = 100,\ 200$ and each subject receives $J = 2,\ 4$ repeated measurements. The number of nodes is set to be $N = 5$ so that the number of possible undirected edges is $D = 10$. The true μ is set to be 0.5 for all elements and

$$\Sigma_x[i, j] = r\rho^{|i-j|}, \text{ where } \rho = 0.8.$$

The underlying true graphical ICC using definition 3.4 is controlled by r. We set $r = 2,\ 4$ in each setting so that the corresponding ICC's are $\frac{rD}{rD+D} = 2/3$ and $4/5$ respectively. A total of 500 simulations were run in each simulation group.

In Table. 3.1, the average estimated GICC for $r = 2$ groups are 0.702, 0.672, 0.683 for $I_{100}J_2$, $I_{100}J_4$ and $I_{200}J_4$ group respectively, comparing to an underlying truth $2/3 \approx 0.667$. As number of individuals increases, or as the number of repeated measurements increases, both the bias and the standard deviation of the estimated GICC reduces. When $r = 4$, the average estimated graphical ICCs are 0.817, 0.800 and 0.806, respectively. The MLE of GICC in each case has a positive bias, which is reduced as either I or J_i increases.

Table 3.1: Simulation results

Setting	Estimates for σ_{ii}					ICC est.
$I = 100,\ J = 2,\ r = 2$	$\sigma_{1,1}$	$\sigma_{2,2}$	$\sigma_{3,3}$	$\sigma_{4,4}$	$\sigma_{5,5}$	ICC
	2.37	2.48	2.38	2.35	2.34	0.702
	(1.18)	(1.09)	(1.11)	(1.04)	(1.01)	
$ICC_{true} = 2/3$	$\sigma_{6,6}$	$\sigma_{7,7}$	$\sigma_{8,8}$	$\sigma_{9,9}$	$\sigma_{10,10}$	
	2.45	2.36	2.43	2.43	2.38	(0.033)
	(1.09)	(1.06)	(1.13)	(1.04)	(1.18)	
$I = 100,\ J = 2,\ r = 4$	$\sigma_{1,1}$	$\sigma_{2,2}$	$\sigma_{3,3}$	$\sigma_{4,4}$	$\sigma_{5,5}$	ICC
	4.56	4.74	4.62	4.40	4.56	0.817
	(2.11)	(2.20)	(2.30)	(2.01)	(2.28)	
$ICC_{true} = 4/5$	$\sigma_{6,6}$	$\sigma_{7,7}$	$\sigma_{8,8}$	$\sigma_{9,9}$	$\sigma_{10,10}$	
	4.65	4.63	4.52	4.61	4.51	(0.025)
	(2.12)	(2.04)	(2.07)	(2.04)	(2.31)	
$I = 100,\ J = 4,\ r = 2$	$\sigma_{1,1}$	$\sigma_{2,2}$	$\sigma_{3,3}$	$\sigma_{4,4}$	$\sigma_{5,5}$	ICC
	2.02	2.10	2.07	2.04	2.10	0.672
	(0.60)	(0.64)	(0.61)	(0.65)	(0.65)	
$ICC_{true} = 2/3$	$\sigma_{6,6}$	$\sigma_{7,7}$	$\sigma_{8,8}$	$\sigma_{9,9}$	$\sigma_{10,10}$	
	2.08	2.06	2.08	2.08	2.05	(0.026)
	(0.59)	(0.60)	(0.63)	(0.65)	(0.61)	
$I = 100,\ J = 4,\ r = 4$	$\sigma_{1,1}$	$\sigma_{2,2}$	$\sigma_{3,3}$	$\sigma_{4,4}$	$\sigma_{5,5}$	ICC
	4.00	4.08	4.04	3.96	4.17	0.800
	(1.29)	(1.36)	(1.28)	(1.29)	(1.36)	
$ICC_{true} = 4/5$	$\sigma_{6,6}$	$\sigma_{7,7}$	$\sigma_{8,8}$	$\sigma_{9,9}$	$\sigma_{10,10}$	
	4.04	4.04	4.10	4.09	4.05	(0.020)
	(1.24)	(1.20)	(1.34)	(1.32)	(1.24)	
$I = 200,\ J = 2,\ r = 2$	$\sigma_{1,1}$	$\sigma_{2,2}$	$\sigma_{3,3}$	$\sigma_{4,4}$	$\sigma_{5,5}$	ICC
	2.08	2.22	2.23	2.17	2.19	0.683
	(0.68)	(0.77)	(0.74)	(0.73)	(0.70)	
$ICC_{true} = 2/3$	$\sigma_{6,6}$	$\sigma_{7,7}$	$\sigma_{8,8}$	$\sigma_{9,9}$	$\sigma_{10,10}$	
	2.17	2.17	2.16	2.20	2.13	(0.026)
	(0.76)	(0.68)	(0.71)	(0.77)	(0.74)	
$I = 200,\ J = 2,\ r = 4$	$\sigma_{1,1}$	$\sigma_{2,2}$	$\sigma_{3,3}$	$\sigma_{4,4}$	$\sigma_{5,5}$	ICC
	4.07	4.33	4.25	4.20	4.18	0.806
	(1.38)	(1.41)	(1.57)	(1.49)	(1.42)	
$ICC_{true} = 4/5$	$\sigma_{6,6}$	$\sigma_{7,7}$	$\sigma_{8,8}$	$\sigma_{9,9}$	$\sigma_{10,10}$	
	4.20	4.20	4.25	4.30	4.15	(0.020)
	(1.52)	(1.36)	(1.43)	(1.47)	(1.47)	

3.4.2 Robustness

In the proposed multivariate probit-linear mixed model, the assumption was made that the underlying within group error term is independent across edges. In other words, in

model 3.5, the off-diagonal elements of $var(\mathbf{u}_{ij})$ is set to be zero. In real applications, this may not be the case. Therefore, we conducted simulations using the above settings when $I = 100,\ J = 2$ and $r = 2$. However, instead of setting $\Sigma = I$, we set $\sigma_{ij} = \rho^{|i-j|}$, where $\rho = 0.1$ and 0.5. The resulting GICC based on 100 simulations have means 0.699(0.025) and 0.722(0.026). The results highlight that the resulting GICC is still close to the underlying truth when there are correlations between edges in the error term.

3.4.3 Comparison With Other Benchmarks

We compared our proposed method with other available methods using our first simulation case ($I = 100; J = 2; r = 2$). The work in Deuker et al. (2009), Telesford et al. (2010) and Telesford et al. (2013) used ICC derived from one-way ANOVA, Rodrıguez and Elo (2003) proposed ICC for binary data using a single variate probit model. It should be pointed out that, all the other methods are based on the single variate ICC, so the comparison is limited. We compare our method with (1) average ICC(1)'s for all edges based on one-way ANOVA, (2) ICC(1) for the mean of the binary vector and (3) average edge-wise ICC based on single variate probit model. The results are shown in Table 3.2.

First of all, the first two ICCs are all derived from the one-way ANOVA model, thus all of the binary data are considered to be continuous. The average edge-wise ICC is only 0.457, but ICC for the mean vector is 0.812. None of them are close to the truth. The edge-wise ICC treats all binary data as pure jumps instead of treating them as having underlying continuous data. The ICC for the mean of vector can only provide the ICC on the average

statistic, but can not align the data with the same edge. The average edge-wise binary ICC based on single variate probit model also yields lower ICC. It also shows the difference between an unstructured covariance matrix for x and a diagonal matrix.

Table 3.2: Compare to Benchmark

	Models			
$ICC_{true} = 0.667$	ave. ICC(1)	ICC(1) for vec.mean	1-var probit	GICC
	0.457	0.812	0.613	0.696
	(0.028)	(0.035)	(0.025)	(0.029)

3.4.4 Running Time

To evaluate the running time of the MCEM algorithm, we conducted simulations with varied settings in which the number of subjects = 50, 100, the number of replicates = 2, 3, 4 and the number of edges = 10, 20 and 30. To be consistent with all settings, we terminated the algorithm after 30 iterations. All simulations were conducted using a 2.4GHz core on PowerEdge C6145 AMD Processor-based 2U Rack Server.

Table 3.3: Running Time (in seconds)

	I=50			I=100		
	D = 10	D = 20	D = 30	D = 10	D = 20	D = 30
J = 2	17.2	28.7	62.4	33.5	57.2	124.8
J = 3	20.7	61.5	215.4	41.3	122.8	430.6
J = 4	28.0	144.5	586.5	55.9	289.0	1173.3

Table 3.3 shows that the number of subjects has a strict linear relationship with running time. Running time increases nonlinearly with either the number of replicates or the number of edges in the graph. With under 30-edge graphs and 4 replicates, the running time is less than twenty minutes. It is clear that if the number of replicates is 4, the running

time grows faster than a quadratic function of the number of edges. Therefore, the current algorithm requires a relative small number of edges for each graph. We will further discuss it in Section 3.6.

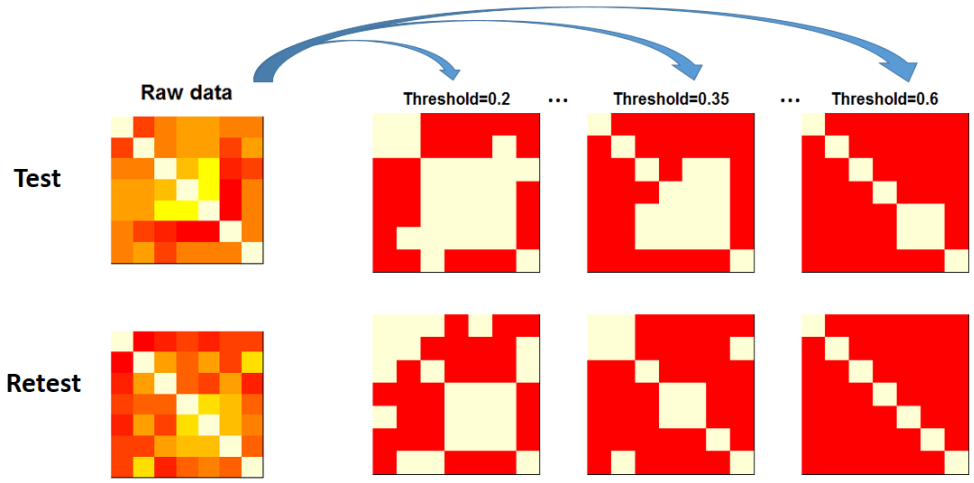
3.5 Application

Resting-state fMRI scans consisted of a test-retest dataset previously acquired at the FM Kirby Research Center at the Kennedy Krieger Institute, Johns Hopkins University Landman et al. (2011) are used to highlight the method. Twenty one healthy volunteers with no history of neurological disease each underwent two separate resting state fMRI sessions on the same scanner. A 3T MR scanner was used (Achieva, Philips Healthcare, Best, The Netherlands) utilizing a body coil with a 2D echoplanar (EPI) sequence and eight channel phased array SENSitivity Encoding (SENSE; factor of 2) with the following parameters: TR 2s, 3mm x 3mm in plane resolution, slice gap 1mm, for total imaging time of 7 minutes and 14 seconds. One subject was excluded due to technical issues at acquisition.

ICA (Independent Component Analysis) was performed using MEDOLIC (Multivariate Exploratory Linear Optimized Decomposition into Independent Components) version 3.10 in FSL (FMRIB Software Library, FMRIB, Oxford, UK). Preprocessing included removal of low-frequency drift with a highpass filter cutoff of 250s, realignment of the fMRI time series using MCFLIRT, slice timing correction, brain extraction using BET, and spa-

tial smoothing with FWHM of 6mm. Images were registered to MNI standard space with resampling resolution of 2mm. ICA was performed using multi-session temporal concatenation with automatic dimensionality estimation and time-course variance normalization implemented in MELODIC. 43 components were identified by MELODIC.

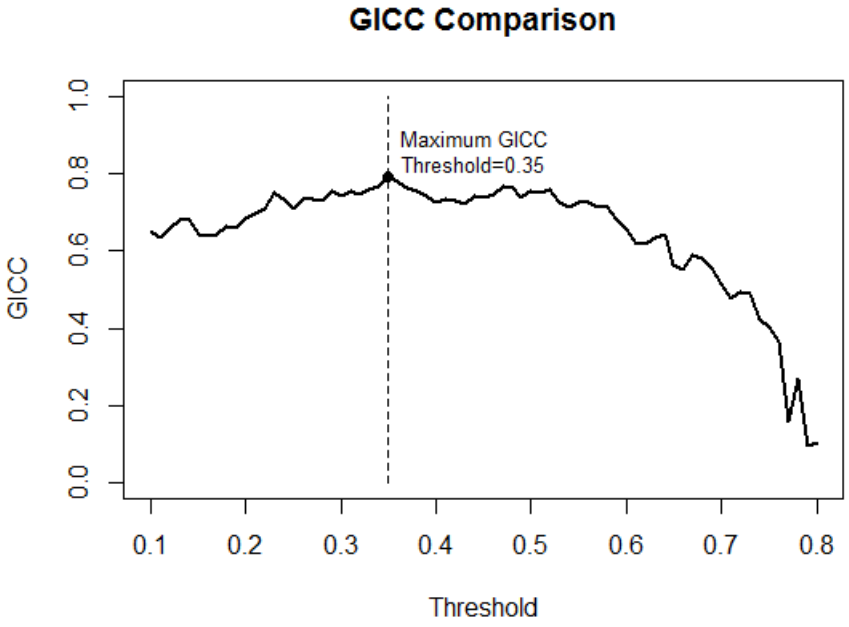
Figure 3.2: The figure illustrates two repeated measurements for one subject. On the left, raw correlations between seven nodes are illustrated. Then the raw correlations are dichotomizing using different thresholds (0.2, 0.35, 0.6 are listed here). Our algorithm is then implemented on binary graphes using each threshold. Red suggests lower value and white (yellow) suggests higher value. In the binary graph on the right, red indicates 1 and yellow indicates 0.



Relevant ICA components corresponding to known large scale brain networks were identified by a board certified neuroradiologist with experience in resting state fMRI. Seven total components were selected (default mode network, dorsal attention network, motor network, visual network, salience network, and two lateralized executive control networks),

and the 7 by 7 correlation matrix was calculated (see raw data in Figure 3.2). Different thresholds were used to dichotomize the raw graphs into binary ones, where the thresholds were chosen from 0.1 to 0.8 using grid 0.01 (see Figure 3.2). The GICC algorithm was then implemented on these binary graphs.

Figure 3.3: The calculated GICC under different thresholds. The threshold were picked equally spaced from 0.1 to 0.8 using grid 0.01. The maximized GICC is indicated in the figure, which corresponds to a 0.35 threshold.



The GICC was then calculated for each threshold (see Figure 3). The GICC remains above 0.6 when the threshold is between 0.1 and 0.6. For threshold outside of this band the GICC decreases dramatically. When the threshold is around 0.8, GICC fluctuates more significantly and the value eventually drops to 0.1. Thus the GICC shows high reproducibility of the raw data if a reasonable threshold is employed (from 0.1 to 0.6). When the threshold is too high, only few raw values will be dichotomize to 1, such that poor reproducibility

is obtained. For practical subsequent applications, one could use the value that maximizes the GICC in this data set (see Figure 3.3).

3.6 Conclusion

In this paper, we propose the concept of the graphical intra-class correlation coefficient using multivariate probit mixed-linear models. The GICC is defined as $\frac{tr(\Sigma_x)}{tr(\Sigma_x)+D}$. We used a Monte Carlo EM algorithm to obtain the MLE of Σ_x, while a Gibbs sampler was used in the E-step. We show the results of GICC in varied simulation settings and in the KIRBY21 test-retest datasets.

While providing GICC, the estimation procedure can also be generalized to multivariate probit mixed-linear model with fixed and random covariates components, which is:

$$
\begin{aligned}
o_{ij}(d) &= \mathbf{I}_{(y_{ij}(d)>0)} \\
y_{ij}(d) &= \sum_{p=1}^{P} \mu_{ip}(d)\beta_p(d) + \sum_{r=1}^{R} x_{ir}(d)\eta_{ir}(d) + u_{ij}(d) \\
\eta_{ir} &\sim \mathbf{N}(\mathbf{0}, \mathbf{\Sigma}_r),\ i.i.d. \\
\mathbf{u}_{ij} &\sim \mathbf{N}(\mathbf{0}, \mathbf{I}),\ i.i.d..
\end{aligned}
$$

In the EM algorithm, η and $\mathbf{y}$ can be treated as full data and the procedure in Section 3.3 follows. In section 3.3.3, we also calculate the observed information matrix for the fixed effects which can provide confidence intervals for β's. Moreover, the procedure can also be

used multivariate generalized mixed-linear models, such as multivariate poisson or logistic regression.

Currently, our method works for small graphs. As the number of nodes in a graph increases, the number of parameters of interests grows quadratically ($D \sim O(N^2)$ and $\#\{\sigma_{ij}\} \sim O(D^2)$). Thus a graph with 100 nodes will have tens of millions of parameters to estimate. The Gibbs sampler could not be implemented effectively in such cases. Therefore, the algorithm currently requires a relatively small number of nodes for each graph (typically less than 10). In order to achieve faster convergence rate as well as control the Monte Carlo error induced by Gibbs sampler, ascent based MCEM (Caffo et al., 2005) and acceleration EM algorithm (Varadhan and Roland, 2008) could be implemented.

Notice that from the application, GICC could also be used for choosing thresholds for dichotomizing raw graphs. The value that maximizes the GICC is a reasonable threshold, since it yields the best reproducibility of a well known benchmark data set.

In summary, GICC provides us a way to measure the reproducibility of repeated graphical measurements. The current algorithm gives us the estimates of GICC for relatively small graphs. To our knowledge, GICC for large graphs has not been addressed before and therefore deserves further investigation.

Chapter 4

Multilevel Binary Principal Component Analysis

Abstract

Principal component analysis (PCA) is a widely used dimension reduction technique. In this manuscript, we extend PCA to multilevel binary data. Our framework is built on the probabilistic PCA fitted by a variational EM algorithm. Variational Expectation Maximization (VEM) algorithm is used in optimizing the likelihood function so that the model could be implemented on large datasets with high dimensions. The performance and running time of the proposed method is studied in a few challenging simulation scenarios. We also apply the method to the National Health and Nutrition Examination Survey (NHANES) dataset as well as a functional magnetic resonance imaging (fMRI) dataset. We then explore a reproducibility of the results through the graphical intra-class correlation coefficient(GICC).

keywords: principal component analysis(PCA), binary data, variational EM, graphical intra-class correlation coefficient

4.1 Introduction

Principal component analysis (PCA) has been widely used for dimension reduction in many scientific domains including psychometrics, genomics, brain imaging among many others. Probabilistic PCA (PPCA), developed in Tipping and Bishop (1999), proposed a rigorous framework to obtain principal components as estimated parameters in a Gaussian statistical model. PPCA framework opened a way to model not only multivarite continuous measurements, but also extended PCA to categorical and binary measurements (Collins et al., 2001; Roy and Gordon, 2002). Several PCA approaches for non-Gaussian data were developed via semi-parametric in Sajama and Orlitsky (2004) and Bayesian methods in Mohamed et al. (2008).

To model binary data, Tipping (1999) proposed a probabilistic model and Jaakkola and Jordan (2000) implemented a variational approximation to the logistic link function. The binary PPCA in Tipping (1999) is modelled as follows:

$$P(x_i(d) = 1|\mathbf{v}_i) = \sigma\big(\mu(d) + \boldsymbol{\theta}(d)\mathbf{v}_i\big), \quad \text{with} \quad \mathbf{v}_i \overset{i.i.d.}{\sim} \mathbf{N}(\mathbf{0}, \mathbf{I}_{K\times K}), \tag{4.1}$$

where $\mathbf{x}_i = (x_i(1), \ldots, x_i(D))$ is a D-dimensional vector with binary entrees characterizing subject $i, i = 1, \ldots, I$, K is the total number of principal components and $\sigma(.)$ is the expit function as $\sigma(z) = \frac{1}{1+\exp(-z)}$. As an example, $\mathbf{x}_i$ could be a 256-dimensional vector representing an unfolded 16×16 black-and-white image. In this model, the columns of the $D \times K$ matrix $\boldsymbol{\Theta}$ span the principal space and the elements of the K-dimensional vector $\mathbf{v}_i$

are the principal scores of subject i. Note that binary PPCA relaxes the orthonormal constraints on the columns of $\mathbf{\Theta}$ and assumes a standard normal prior on the principal scores $\mathbf{v}_i$. Thus, the magnitude of PCs is absorbed into $\mathbf{\Theta}$. More details can be found in Tipping (1999).

In many medical and epidemiological studies participants have multiple follow-up visits during which many multivariate measurements, such as brain images, are collected. Scientifically, it is critically important to take into account the design-imposed clustering structure in the data. To accommodate both the between- and within-subject specific PCs, Di Martino et al. (2008) and Zipunnikov et al. (2011b) proposed principal component methods for functional and high-dimensional continuous observations, respectively. The framework in Di Martino et al. (2008) and Zipunnikov et al. (2011b) assumes that the between-subject PCs characterize the difference between subjects, while the within-subject PCs represent the visit-to-visit differences within subjects. Furthermore, Goldsmith et al. (2015) proposed a multilevel functional principal component analysis for exponential family. Though the model enjoys its flexibility of enriched family type, the long running time of the Gibbs sampler makes it less useful. For less than 1,000 subjects with no more than 10 basis functions and up to 20 principal components, the running time takes more than 10 days. Therefore, a huge computational improvement is strongly required.

In this manuscript, we combine the ideas of binary PPCA and the multilevel design and propose a multilevel binary principal component analysis (MBPCA) to model high-dimensional binary data collected repeatedly. Our method enjoy a significant shorter com-

puting time and can handle much larger datasets easily. Several simulation scenarios were conducted for validation. We will apply the MBPCA method on the National Health and Nutrition Examination Survey (NHANSE) dataset [1] as well as a case from resting-state fMRI experiment. Furthermore, we will apply MBPCA to quantify the reproducibility of estimated functional connectivity in an fMRI study through a graphical interclass cross-correlation coefficient (GICC) (Yue et al., 2015).

The rest of the manuscript is organized as follows. We will introduce our model and the adapted variational EM algorithm in Section 2. Simulation study will be shown in Section 3. Two application cases will be presented in Section 4. We will summarize in Section 5.

4.2 Model and Methods

Suppose $\mathbf{x}_{ij} = (x_{ij}(1)), \ldots, x_{ij}(D)$ are observed D-dimensional binary vectors, where $i = 1, \ldots, I$ indicates the subject ID and $j = 1, \ldots, J_i$ indicates the j^{th} measurement for subject i. We define Multilevel Binary Principal Component Analysis (MBPCA) model as follows,

$$P(x_{ij}(d) = 1|\mathbf{v}_i, \mathbf{u}_{ij}) = \sigma\big(\mu(d) + \boldsymbol{\theta}(d)\mathbf{v}_i + \boldsymbol{\psi}(d)\mathbf{u}_{ij}\big),$$

$$\mathbf{v}_i \overset{i.i.d.}{\sim} \mathbf{N}(\mathbf{0}, \mathbf{I}_{\mathbf{K_1}\times\mathbf{K_1}}),$$

$$\mathbf{u}_{ij} \overset{i.i.d.}{\sim} \mathbf{N}(\mathbf{0}, \mathbf{I}_{\mathbf{K_2}\times\mathbf{K_2}}),$$

[1] http://www.cdc.gov/nchs/nhanes.htm

where $\mu_{\mathbf{D}\times\mathbf{1}}$ is the population mean of the latent process, the columns of $\mathbf{\Theta}_{D\times K_1}$ span the between-subject principal space; the columns of $\mathbf{\Psi}_{D\times K_2}$ span the within-subject principal space; the rows of $\mathbf{V}_{I\times K_1}$ (denoted by $\mathbf{v}_i^T$) indicates the between-subject principal scores and the rows of $\mathbf{U}_{N\times K_2}$ (with $N=\sum J_i$) are the within-subject principal scores.

The MBPCA model assumes that each observation can be explained by the population average μ, the between subject effect $\boldsymbol{\theta}(d)\boldsymbol{v}_i$ and the within subject effect $\boldsymbol{\phi}(d)\mathbf{u}_{ij}$. The parameters of interest are $\mu,\ \boldsymbol{\theta}$ and $\boldsymbol{\psi}$. It is clear that the MBPCA model is invariant to any orthogonal transformation of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$. For instance, let $\boldsymbol{\theta}$ be one estimate of between subject principal component, then $\boldsymbol{\theta}^* := \boldsymbol{\theta P}$ yields the same likelihood with redefined $\boldsymbol{v}_i^* = \boldsymbol{P}^{-1}\boldsymbol{v}_i$ as long as $\boldsymbol{P}_{K_1\times K_1}$ is an orthogonal matrix. Therefore, for the purpose of obtaining unique solution, a singular value decomposition (SVD) is applied for both $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$. Let $\boldsymbol{\theta} = \boldsymbol{PDQ}^t$ be the SVD, we define the unique solution as $\hat{\boldsymbol{\theta}} = \boldsymbol{PD}$ and the corresponding principal scores as $\boldsymbol{v_i} = \boldsymbol{Q}^t\boldsymbol{v}_i$.

The full likelihood for observed matrix $\mathbf{X}$ and unobserved matrices $\mathbf{U}$ and $\mathbf{V}$ is given by

$$\begin{aligned} L(\mathbf{X},\mathbf{U},\mathbf{V}) \propto & \prod_{i=1}^{I}\prod_{j=1}^{J_i}\prod_{d=1}^{D}\Big\{\sigma\big((2x_{ij}(d)-1)A_{ij}(d)\big)\Big\} \\ & \times\prod_{i=1}^{I}\Big\{\exp\big(-\frac{\mathbf{v}_i^T\mathbf{v}_i}{2}\big)\Big\}\prod_{i=1}^{I}\prod_{j=1}^{J_i}\Big\{\exp\big(-\frac{\mathbf{u}_{ij}^T\mathbf{u}_{ij}}{2}\big)\Big\}, \end{aligned} \tag{4.2}$$

where $A_{ij}(d)=\mu(d)+\boldsymbol{\theta}(d)\mathbf{v}_i+\boldsymbol{\psi}(d)\mathbf{u}_{ij}$. Obtaining the marginal likelihood of $\mathbf{X}$ involves integrating out all $\mathbf{v}_i$'s and $\mathbf{u}_{ij}$'s, which cannot be computed directly. To obtain MLE, we will follow Tipping and Bishop (1999) and Schein et al. (2003) and use the variational

approximation approach which was originally proposed in Jaakkola and Jordan (2000). Next, we will discuss the approach with the necessary modification to accommodate the multilevel design.

4.2.1 Variational Approximation of the Likelihood

We use the variational approximation to $P(x_{ij}(d)|\mathbf{v}_i, \mathbf{u}_{ij})$ of the following form:

$$\begin{aligned}\tilde{P}(x_{ij}(d)|\mathbf{v}_i, \mathbf{u}_{ij}, \xi_{ij}(d)) = \sigma(\xi_{ij}(d)) \exp\Big\{((x_{ij}(d) - \frac{1}{2})A_{ij}(d) \\ - \frac{1}{2}\xi_{ij}(d)) + \lambda(\xi_{ij}(d))\big(A_{ij}(d)^2 - \xi_{ij}(d)^2\big)\Big\},\end{aligned} \tag{4.3}$$

where $\xi_{ij}(u)$ is the variational parameter and $\lambda(x) = \frac{0.5-\sigma(x)}{2x}$. Thus, the full likelihood can be approximated by

$$\begin{aligned}\tilde{L}(\mathbf{X}, \mathbf{U}, \mathbf{V}) \propto \prod_{i=1}^{I}\prod_{j=1}^{J_i}\prod_{d=1}^{D}\Big\{\tilde{P}(x_{ij}(d)|\mathbf{v}_i, \mathbf{u}_{ij}, \xi_{ij}(d))\Big\} \\ \times \prod_{i=1}^{I}\Big\{\exp\big(-\frac{\mathbf{v}_i^T\mathbf{v}_i}{2}\big)\Big\}\prod_{i=1}^{I}\prod_{j=1}^{J_i}\Big\{\exp\big(-\frac{\mathbf{u}_{ij}^T\mathbf{u}_{ij}}{2}\big)\Big\}.\end{aligned} \tag{4.4}$$

Expectation-Maximization (EM) algorithm can be used to obtain the MLE for the variationally approximated likelihood. The EM algorithm has three steps:

(1) obtaining the posterior distribution of $\mathbf{v}_i$ and $\mathbf{u}_{ij}$, which depends on $\boldsymbol{\mu}, \boldsymbol{\Theta}, \boldsymbol{\Psi}$ and $\boldsymbol{\xi}$;

(2) maximizing the variational approximation with respect to $\boldsymbol{\xi}$;

(3) maximizing the variational approximation with respect to $\boldsymbol{\mu}, \boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$.

The three steps are then iterated until the convergence is achieved. Below, we present the details of each of the steps.

(1) Joint posterior distribution of $\mathbf{v}_i$ and $\mathbf{u}_{ij}$:

The joint posterior distribution of $\mathbf{v}_i$ and $\mathbf{u}_{ij}$'s for a fixed i can be obtained from (4.4). Tipping and Bishop (1999) and Schein et al. (2003) derived the formulas for the (one-level) binary PCA. In the multilevel case, $\mathbf{v}_i$ and $\mathbf{u}_{ij}$'s are not independent in the posterior distribution. We will now derive the joint distribution. Let $\boldsymbol{m}_i$ and $\boldsymbol{C}_i$ are the posterior mean and variance covariance matrix of the vector $(\mathbf{v}_i^T, \mathbf{u}_{i1}^T, \ldots, \mathbf{u}_{iJ_i}^T)^T$. Then we have:

$$\boldsymbol{C}_i = \begin{pmatrix} H_i & B_{i1} & \ldots & B_{iJ_i} \\ B_{i1}^T & G_{i1} & \mathbf{0} & \mathbf{0} \\ \vdots & \mathbf{0} & \ddots & \mathbf{0} \\ B_{iJ_i}^T & \mathbf{0} & \mathbf{0} & G_{iJ_i} \end{pmatrix}^{-1}, \tag{4.5}$$

$$\boldsymbol{m}_i = \boldsymbol{C}_i \cdot (\boldsymbol{h}_i^T, \boldsymbol{g}_{i1}^T, \ldots, \boldsymbol{g}_{iJ_i}^T)^T, \tag{4.6}$$

where $H_i = \boldsymbol{I} - 2\sum_j \sum_d \lambda(\xi_{ij}(d))\boldsymbol{\theta}(d)^T\boldsymbol{\theta}(d)$,

$$G_{ij} = \boldsymbol{I} - 2\sum_d \lambda(\xi_{ij}(d))\boldsymbol{\psi}(d)^T\boldsymbol{\psi}(d),$$

$$B_{ij} = -\sum_d \lambda(\xi_{ij}(d))\boldsymbol{\theta}(d)^T\boldsymbol{\psi}(d),$$

$$\boldsymbol{h}_i = \sum_j \sum_d \big(x_{ij}(d) - \frac{1}{2} + 2\lambda(\xi_{ij}(d))\mu(d)\big)\boldsymbol{\theta}(d)^T,$$

and $\boldsymbol{g}_{ij} = \sum_d \big(x_{ij}(d) - \frac{1}{2} + 2\lambda(\xi_{ij}(d))\mu(d)\big)\boldsymbol{\psi}(d)^T.$

In (4.5), calculating $\boldsymbol{C}_i$ requires inverting a $K_1 + J_iK_2$ by $K_1 + J_iK_2$ matrix. This could be quite computationally demanding for large J_i. The computations can be simplified by

using the inverse block matrix formula:

$$\begin{pmatrix} A & B \\ B^T & D \end{pmatrix}^{-1} = \begin{pmatrix} (A-BD^{-1}B^T)^{-1} & -(A-BD^{-1}B^T)^{-1}BD^{-1} \\ -D^{-1}B^T(A-BD^{-1}B^T)^{-1} & D^{-1}+D^{-1}B^T(A-BD^{-1}B^T)^{-1}BD^{-1} \end{pmatrix} \tag{4.7}$$

, where $A = H_i$, $B = (B_{i1}, \ldots, B_{iJ_i})$ and $D = \text{diag}(G_{i1}, \ldots, G_{iJ_i})$. Since D is a diagonal block matrix, the highest dimension of matrix to be inverted is K by K with $K = \max(K_1, K_2)$.

(2) Maximizing the variation likelihood with respect to $\boldsymbol{\xi}$.

By construction $P(x_{ij}(d)|\mathbf{v}_i, \mathbf{u}_{ij}) > \tilde{P}(x_{ij}(d)|\mathbf{v}_i, \mathbf{u}_{ij}, \xi_{ij}(d))$, so the variational approximation can be improved by maximizing $\tilde{P}$ with respect to $\boldsymbol{\xi}$. Using the EM algorithm, we obtain that $\hat{\xi}_{ij}(d)^2 = \boldsymbol{E}_{\tilde{p}_{post}}(A_{ij}(d)^2)$. Thus we have:

$$\begin{aligned} \hat{\xi}_{ij}(d)^2 =& \boldsymbol{\theta}(d)\langle \mathbf{v}_i, \mathbf{v}_i \rangle \boldsymbol{\theta}(d)^T + \boldsymbol{\psi}(d)\langle \mathbf{u}_{ij}, \mathbf{u}_{ij} \rangle \boldsymbol{\psi}(d)^T + 2\boldsymbol{\theta}(d)\langle \mathbf{v}_i, \mathbf{u}_{ij} \rangle \boldsymbol{\psi}(d)^T \\ &+ 2\mu(d)\boldsymbol{\theta}(d)\langle \mathbf{v}_i \rangle + 2\mu(d)\boldsymbol{\psi}(d)\langle \mathbf{u}_{ij} \rangle + \mu(d)^2, \end{aligned} \tag{4.8}$$

where $\langle \mathbf{v}_i, \mathbf{v}_i \rangle = C_{iv} + \boldsymbol{m}_{iv}\boldsymbol{m}_{iv}^T$, $\langle \mathbf{u}_{ij}, \mathbf{u}_{ij} \rangle = C_{iu_{ij}} + \boldsymbol{m}_{iu_{ij}}\boldsymbol{m}_{iu_{ij}}^T$, $\langle \mathbf{v}_i, \mathbf{u}_{ij} \rangle = C_{ivu_{ij}} + \boldsymbol{m}_{iv}\boldsymbol{m}_{iu_{ij}}^T$, $\langle \mathbf{v}_i \rangle = \boldsymbol{m}_{iv}$ and $\langle \mathbf{u}_{ij} \rangle = \boldsymbol{m}_{iu_{ij}}$, and $\langle \cdot \rangle$ denotes the corresponding posterior cross-moments. The posterior distribution of $\mathbf{v}_i$ and $\mathbf{u}_{ij}$ depends on $\boldsymbol{\xi}$, so steps (1) and (2) are iterated to improve the approximation of the posterior distribution. A small number iterations is usually required to converge.

(3) Maximizing the variational likelihood with respect to $\boldsymbol{\mu}$, $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$.

Let $\mathbf{w}_{ij} = (\mathbf{u}_{ij}^T, \mathbf{v}_i^T, 1)^T$ and $\boldsymbol{\phi}(d) = (\boldsymbol{\Psi}(d)^T, \boldsymbol{\Theta}(d)^T, \mu(d))^T$. The update for $\boldsymbol{\phi}(d)$ has

a closed form and is equal to

$$\hat{\phi}(d) = -\Big[\sum_i \sum_j 2\lambda(\xi_{ij}(d))\widehat{\mathbf{w}_{ij}\mathbf{w}_{ij}^T}\Big]^{-1}\Big[\sum_i \sum_j (x_{ij}(d) - \frac{1}{2})\widehat{\mathbf{w}_{ij}}\Big],$$

$$\text{where} \quad \widehat{\mathbf{w}_{ij}\mathbf{w}_{ij}^T} = \begin{pmatrix} \langle \mathbf{u}_{ij}, \mathbf{u}_{ij} \rangle & \langle \mathbf{v}_i, \mathbf{u}_{ij} \rangle & \langle \mathbf{u}_{ij} \rangle \\ \langle \mathbf{v}_i, \mathbf{u}_{ij} \rangle & \langle \mathbf{v}_i, \mathbf{v}_i \rangle & \langle \mathbf{v}_i \rangle \\ \langle \mathbf{u}_{ij} \rangle & \langle \mathbf{v}_i \rangle & 1 \end{pmatrix},$$

$$\widehat{\mathbf{w}_{ij}} = (\langle \mathbf{u}_{ij} \rangle^T, \langle \mathbf{v}_i \rangle^T, 1)^T.$$

The EM algorithm stops when the difference between the parameters estimated at two consecutive iterations falls below a pre-specified threshold.

4.3 Simulations

4.3.1 Scenario 1

In the first scenario, the number of subjects is set to be $I = 500$ with $J_i = 6$ replications of each. Each measurement has dimension $D = 100$.

The data were generated from two between subject principal components (between-PCs) and two within subject components (within-PCs). We use a Cosine and a Sine function for between-PCs θ_1, θ_2, such that,

$$\theta_1(d) = 2 * cos(\frac{\pi}{2} * \frac{d-1}{D-1}), \qquad \theta_2(d) = sin(\frac{\pi}{2} * \frac{d-1}{D-1}), \qquad 1 \leq d \leq D$$

and the within-PCs ψ_1, ψ_2 are defined by a linear function and a quadratic function respectively:

$$\psi_1(d) = 4 * \frac{d - 0.5 * (D+1)}{D-1}, \qquad \psi_2(d) = (2 * \frac{d - 0.5 * (D+1)}{D-1})^2. \qquad 1 \leq d \leq D$$

As we discussed in Section 2, an SVD step will be performed on both true PCs (θ, ψ) and estimated PCs ($\hat{\theta}$, $\hat{\psi}$) for fair comparison.

The MBPCA models were fit with $K_1 = K_2 = 10$ and the maximum number of iterations was set to be 100.

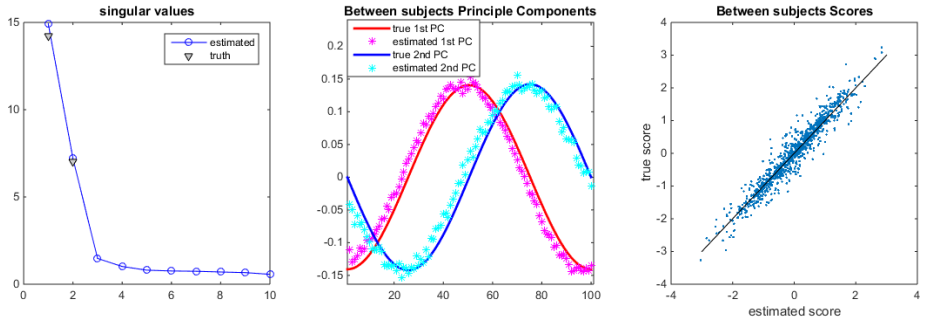


Figure 4.1: Scenario 1 simulation results for between subject PCs. Left: compares true singular values versus estimated ones. Middle: true between subject PCs in smoothed curves and estimated ones in dotted curves. Right: true PC scores versus estimated ones in dots and identity function in solid line.

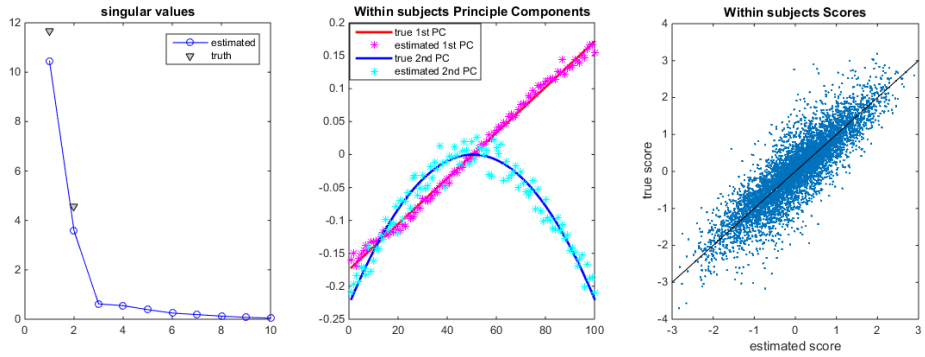


Figure 4.2: Scenario 1 simulation results for within subject PCs. Left: compares true singular values versus estimated ones. Middle: true between subject PCs in smoothed curves and estimated ones in dotted curves. Right: true PC scores versus estimated ones in dots and identity function in solid line.

The left subplot of both Figure.4.1 and Figure.4.2 compares the estimated singular value versus the corresponding true ones. It is found that the magnitudes of both between/within subjects singular values list decrease significantly after the first two true components, indicating that the first two principal components can explain the majority of variation patterns in data, which is consistent with the simulation settings.

The comparison of estimated principal components versus the true ones is illustrated in the middle subplots of Figure.4.1, 4.2 and the comparison of the principal scores is shown

in the right subplots. All figures show estimates are close to the underlying truth.

4.3.2 Scenario 2

In Scenario 2, we set $I = 500$, $J = 4$ and $D = 1,000$. There are still two true between/within subjects principal components each. The patterns of principal components are equally spaced bars, specified by:

$$\theta_1(d) = 0.447 * I_{1 \leq d \leq D/2}, \;\; \theta_2(d) = 0.224 * I_{(D/2+1) \leq d \leq D}, 1 \leq d \leq D$$

$$\psi_1(d) = 0.224 * I_{d \in [1, D/4] \cup [(D/2+1):0.75*D]}, \;\; \psi_2(d) = 0.112 * I_{d \in [(D/4+1):D/2] \cup [(0.75*D+1):D]}$$

where $I_{\cdot}$ is the indicator function. The MBPCA model was fit with the number of between and within subject principal components $K_1 = K_2 = 20$ and the maximum number of iterations 200. The results are presented in Figure.4.3 and 4.4, with the same interpretation as in simulation 1.

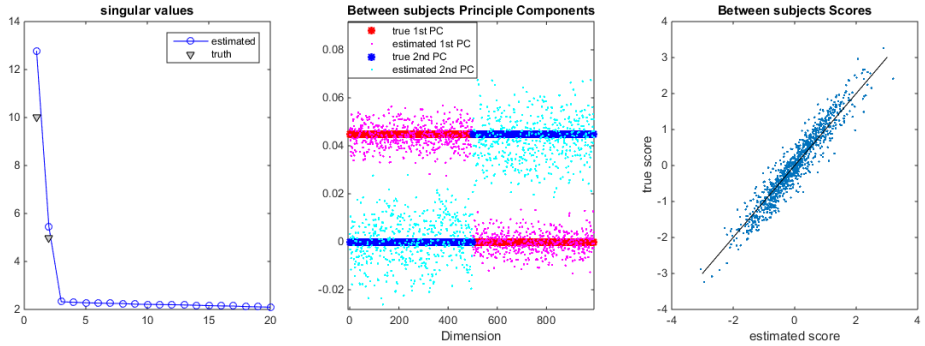


Figure 4.3: Scenario 2 simulation results for between subject PCs. Left: compares true singular values versus estimated ones. Middle: true between subject PCs in smoothed curves and estimated ones in dots. Right: true PC scores versus estimated ones in dots and identity function in solid line.

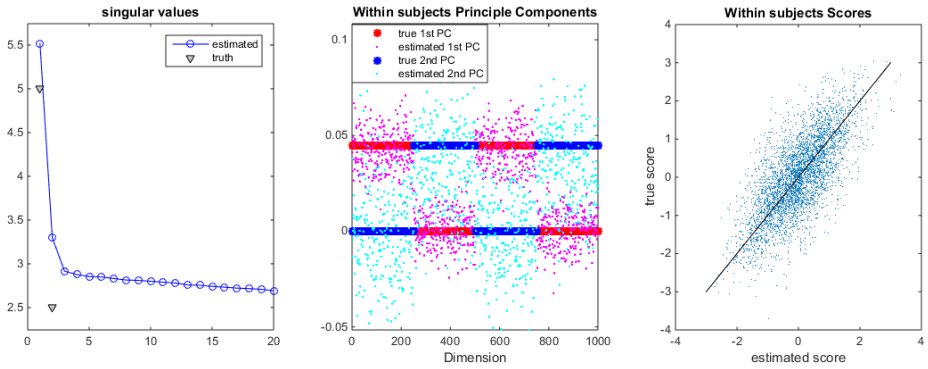


Figure 4.4: Scenario 2 simulation results for within subject PCs. Left: compares true singular values versus estimated ones. Middle: true between subject PCs in smoothed curves and estimated ones in dots. Right: true PC scores versus estimated ones in dots and identity function in solid line.

4.3.3 Runtime Analysis

To evaluate the runtime of the MBPCA algorithm, we conducted simulations according to scenario 1 with varied problem size, in which number of subject $I = 500, 1000$, number of replicates $J = 4, 8$, and dimension of each observation $D = 500, 1000$. The MBPCA model was fitted using $K_1 = K_2 = 5, 10, 15$. To be consistent with all setting,

		I = 500			I = 1000		
		K = 5	K = 10	K = 15	K = 5	K = 10	K = 15
D = 500	J = 4	167.0	297.0	440.3	315.1	551.7	830.1
	J = 8	314.1	546.9	829.3	628.0	1092.4	1666.9
D = 1000	J = 4	325.7	566.5	831.0	628.8	1113.7	1664.7
	J = 8	624.7	1085.0	1672.4	1247.4	2172.6	3262.0

Table 4.1: Runtime in seconds

we terminated the algorithm after 30 iterations, which is usually good enough for convergence in many application cases. All simulations were conducted using processor: 3.1GHz dual-core Intel Core i7 with 4MB shared L3 cache.

Table 4.1 illustrates the runtime of the VEM algorithm. First, we notice that the runtime grows equals to or slower than a linear speed with: number of subjects I, the number of replicates J, the number of dimensions of the data D and the number of PCs K. Second, if we compare the running time of our model versus the one proposed in Goldsmith et al. (2015), Table 4.1 shows huge advantages of our method. One particular setting in this table is $I = 500$, $J = 4$, $K = 10$ and $D = 500$, which is comparable to the application case in Goldsmith et al. (2015), takes less than 3 minutes in our simulation, while a similar setting took more than 10 days in Goldsmith et al. (2015).

4.3.4 The Graphical Intra-Class Correlation Coefficient (GICC)

Furthermore, we investigate the estimation quality of the GICC obtained by MBPCA, which is an important measurement of graphical data reproducibility. The GICC is pro-

posed in Yue et al. (2015) as follows:

$$GICC := \frac{\sum_{k=1}^{K_b}(\boldsymbol{d}_k^{\theta})^2}{\sum_{k=1}^{K_b}(\boldsymbol{d}_k^{\theta})^2 + \sum_{k=1}^{K_w}(\boldsymbol{d}_k^{\psi})^2}, \tag{4.9}$$

where d^{θ}s' and d^{ψ}s' are the singular values of the corresponding parameter matrices $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$.

We adopt similar simulation settings in scenario 2, in which I = 100, j = 4, D = 100, with the true GICC score varying from 0.5 to 0.9 on a 0.1 step size. The MBPCA model is fitted with $K = K_1 = K_2 = 2, 5$ and 10.

In Table 4.2, each estimated GICC is the average over 50 repeated runs. The true number of between/within subjects principal components are both 2. We found that when using $K = 2$ (equal to the truth) to fit the model, the estimated GICC is more accurate. As K increases, the GICC tends to be under-estimated. Therefore, we suggests use moderate number of PCs $K_1,\ K_2$ for better accuracy of GICC estimation. In practice, one could start with a relatively big number of K and then reduce the K_1 and K_2 to the value $\hat{K_1}$ and $\hat{K_2}$ such that the singular value starts to flatten out to zero after $\hat{K_1}$ and $\hat{K_2}$ respectively. Left subplots of Figure 4.4 indicates that $K_1 = K_2 = 2$ is a good candidate for estimating GICC.

True GICC		0.5	0.6	0.7	0.8	0.9
Estimated GICC	K = 2	0.49 (0.04)	0.60 (0.05)	0.69 (0.04)	0.80 (0.03)	0.90 (0.02)
	K = 5	0.47 (0.03)	0.56 (0.04)	0.65 (0.03)	0.75 (0.02)	0.87 (0.02)
	K = 10	0.46 (0.02)	0.53 (0.03)	0.60 (0.02)	0.70 (0.02)	0.82 (0.02)

Table 4.2: GICC Estimation. Numbers without parenthesis indicates the average GICC estimates, numbers inside parenthesis indicates the standard deviation.

4.4 Application

4.4.1 NHANES

The 2003-2004 wave of NHANES collected physical activity data for seven consecutive days on participants aged 6 and older. The Actigraph AM-7164, a uni-axial accelerometer, placed on a belt and participants were instructed to wear the device on the right hip at all times other than during any aquatic activity, including swimming and bathing, and at bedtime (Troiano et al., 2008). The participants were instructed to remove accelerometers at bedtime, therefore, we excluded the nighttime period that we defined as 11pm to 7am. In addition, non-wear time was identified using the algorithm in Van Domelen and Pittard (2014) and the days with more than 10 percent of non-wear time or non-calibrated or non-reliable data have been excluded. The remaining dataset contained 4,076 subjects having on average 3 daily physical activity profiles.

The left panel of Figure 4.8 shows daily rest/activity profiles for five randomly chosen participants. Each profile can be thought of as a 960-dimensional vector with binary components representing a daytime fragmentation of physical activity.

MBPCA model was then implemented to this dataset, the number of between and within subjects principal components were set to be 5. It took 30 steps for the algorithm to converge.

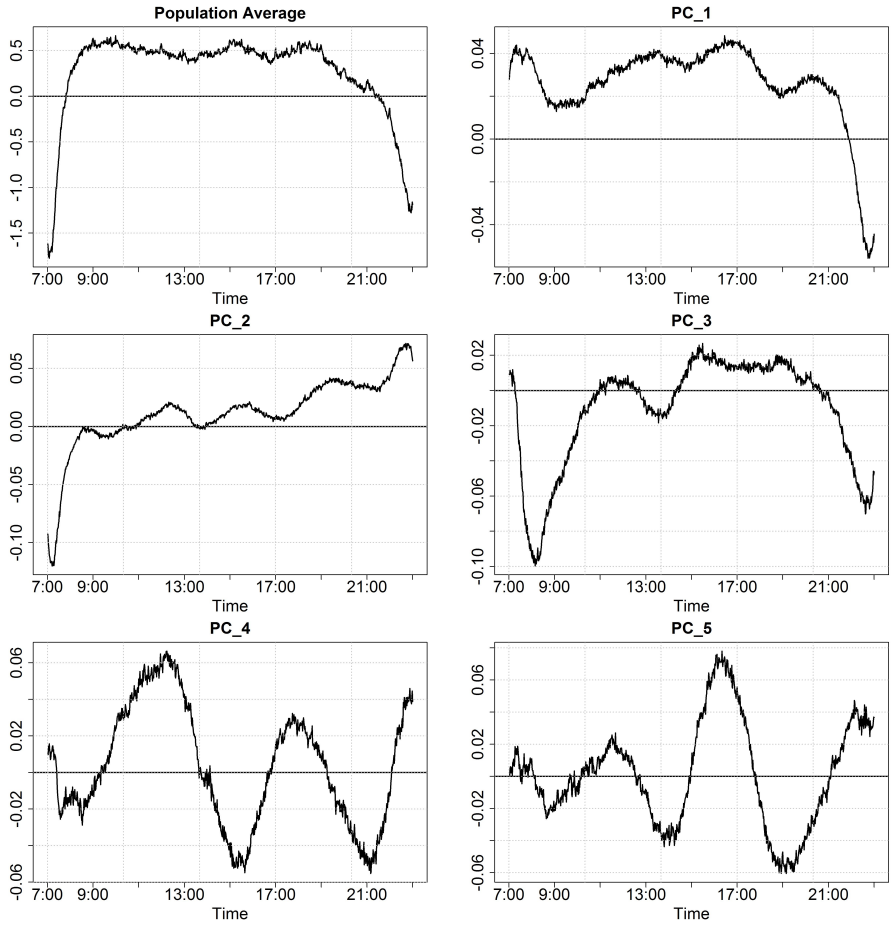


Figure 4.5: Population Averages and Between Subject PCs of NHANSE data

The population average, which is shown in the upper left subplot of Figure 4.5, shows the population average probability of being active. The sharp increasing trend in the morning and the sharp decreasing trend in the late evening illustrates the getting up and going back to sleep respectively. Three peaks could also be observed, The probability peaks in the morning around 9AM and then slowly decays until 1PM. The second peak happens in the afternoon at 3PM while the last peak suggests the evening activity at around 7PM. The first

PC illustrates the largest activity component that different people vary on. The pattern itself shows two main positive peaks and one negative peak. These large magnitude happens at 7AM, 1-5PM and 11PM. It informs us that the getting up activities, the afternoon activities and the night activities has the largest variability across different people. Meanwhile, it also suggest that the one with highest early morning and afternoon active probability also enjoys an earlier sleep time. PC2 unveils a similar pattern but only emphasizes in the early morning at 7AM and late evening at 11:00PM. The opposite sign shows that the one with low active probability in the early morning will have higher values at night. PC 3 has two negative peaks at 8AM and 11PM, while being around zero during other time period. PC4 and PC5 are much noisier comparing with the top three PCs', the magnitude of the fourth and fifth PCs are also much smaller (see Figure 4.7).

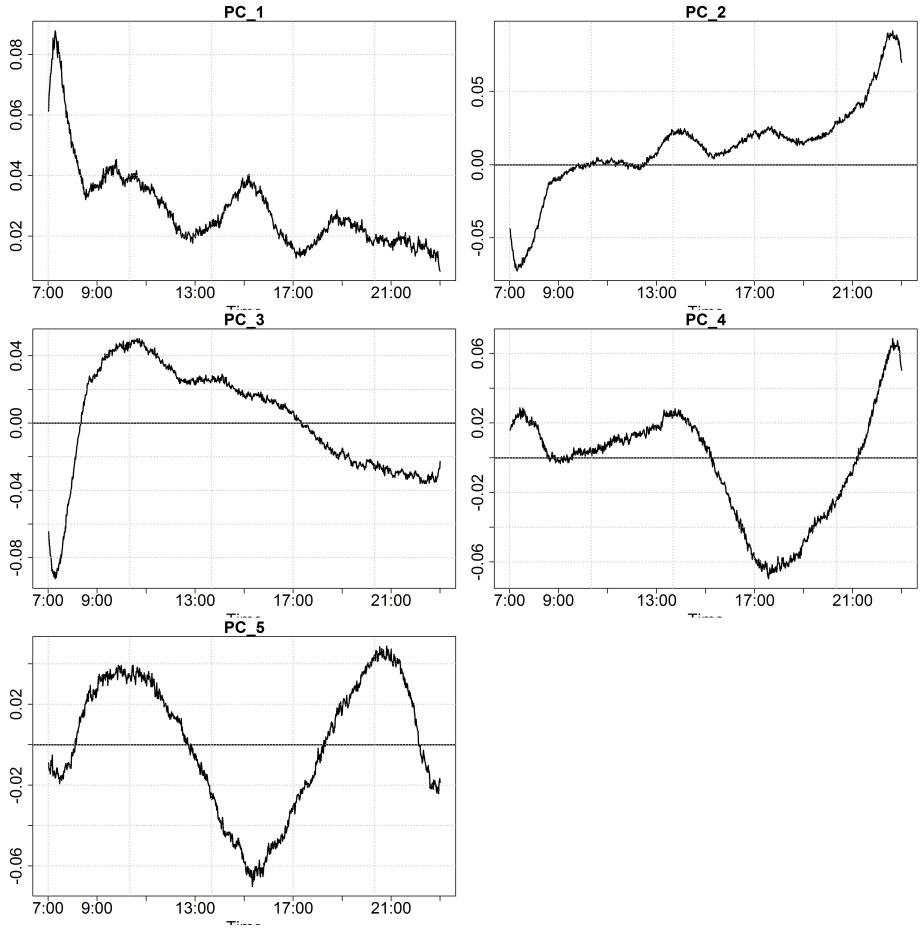


Figure 4.6: Within Subject PCs of NHANES data

The first within subject PC shows that the largest variation activity type within a subject is the early morning activity. One possible explanation is that people might get up at different times comparing weekday with weekend. PC 2 resembles the same one for between subject PC. PC 3 has one negative peak in early morning, one positive peak at 11AM which decays constantly until night. PC 4 shows the variability in the late afternoon to evening while PC 5 has positive peaks at 10AM and 9PM with a negative peak at 3PM. Both PC 4

and 5 have lower magnitude than the first three within subject PCs.

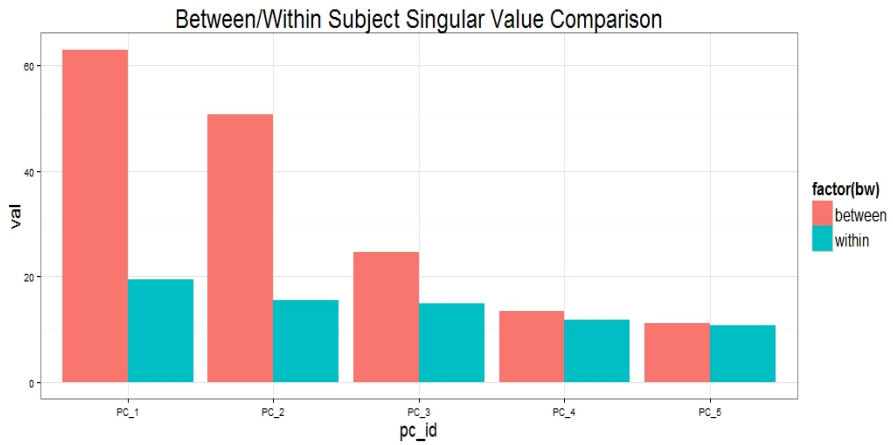


Figure 4.7: Singular Values of Between and Within PCs

Figure 4.7 illustrates the singular values of those 5 between subject PCs and within subject PCs. For the between subject PCs, the top three PCs have a much larger magnitude which explain more than 95% total between subject variability. The within subject PCs have less magnitude difference. The top three ones only take 76% total variation. The overall GICC is 87%.

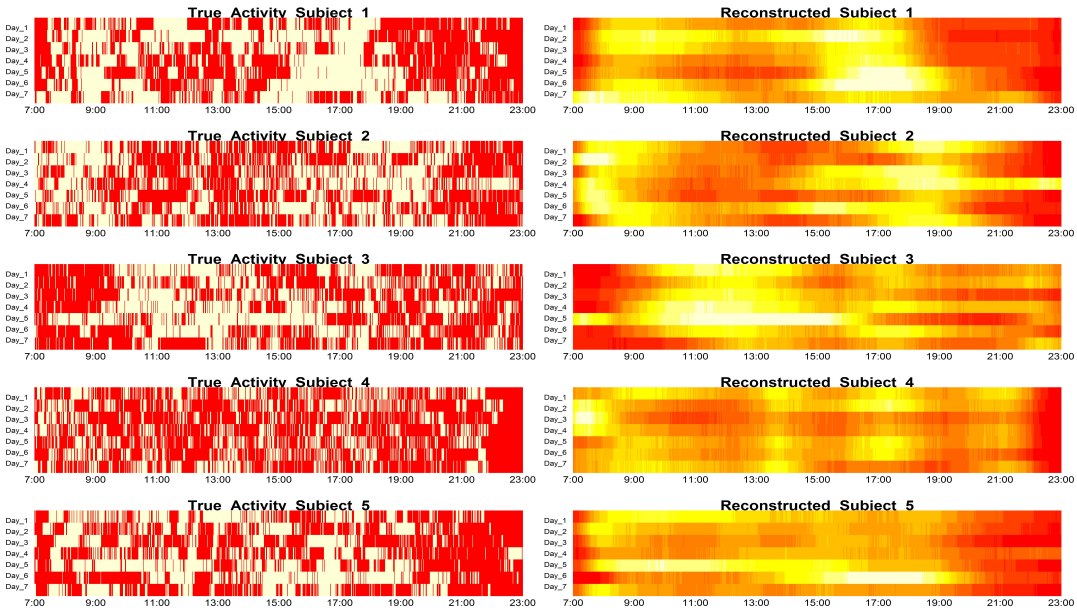


Figure 4.8: Heatmaps of the true activity map versus the reconstructed active probability map. In the left panels, yellow color suggests being active and red being nonactive. In the right panels, yellow suggests higher probabilities while red suggests lower ones.

Five people with seven repeated measurement were randomly selected and their original binary activity map were visually compared with th reconstructed probability in Figure 4.8. Overall, the reconstructed map can be visualized as a smoothed version of the binary data. At the same time, it illustrates the between subject variability and the within subject similarity. Take subject 1 and 3 for example, subject one has higher probability during morning time and close to zero probability during night time while subject 3 being the opposite. In addition, within subject variability can also be observed. Take subject 4 as an example, he/she is more active in the morning during the second to the fourth day while less active in other days. Subject 2 also has an activity during evening on the fourth day

while being less active on the rest of the days.

4.4.2 Human Connectome Project

In this section we apply MBPCA to group resting-state fMRI data with four replicate measurements per subject. We further estimate a GICC score to evaluate the reproducibility of functional connectivity (FC), which is the study of the temporal dependencies between multiple brain regions, and is usually quantified using statistical measures such as correlation (Biswal et al., 1995). In resting-state fMRI, it is common to assume that the fMRI time series follows a multivariate Gaussian distribution and measure FC by the estimated covariance, correlation or precision (inverse covariance) matrices (Varoquaux et al., 2010; Cribben et al., 2012). However, in high-dimensional settings, the estimation of covariance or correlation matrices can be difficult due to the positive definite constraint. Using the idea of variable selection, the graphical lasso (glasso) technique can be used to estimate a sparse precision matrix for high-dimensional data, by imposing an L_1 constraint to force many elements of precision matrix to zero (Friedman et al., 2008). Each of the zero elements in the precision matrix corresponds to the conditional independence between the corresponding variables or regions.

The data comes from the 2014 Human Connectome Project (HCP) data release (Van Essen et al., 2013). Resting-state fMRI data was collected for $I = 461$ subjects, with $J_i = 4$ repeated 15 min runs for each subjects. The multivariate fMRI time series has length 1200 for each run. The pre-processing procedures are described in Glasser et al. (2013),

with artifacts removed using FMRIBs ICA-based Xnoiseifier (FIX) (Griffanti et al., 2014; Salimi-Khorshidi et al., 2014). Group-PCA is applied followed by spatial-ICA with varying number of distinct ICA components (d = 25, 50, 100, 200, 300). The set of ICA spatial components were mapped onto each subject's fMRI time series by dual-regression approach (Filippini et al., 2009) to obtain a single time series per ICA component.

A sparse precision matrix Σ_{ij} was estimated for each subject i ($1 \leq i \leq 461$) and repetition $j(1 \sim 4)$ using glasso, according to Xu and Lindquist (2015):

$$\Sigma_{ij} = argmax_{\Sigma}\big\{ \log(det(\Sigma)) - tr(S_{ij}\Sigma) - \lambda\|\Sigma\|_1 \big\}$$

where S_{ij} is the empirical covariance matrix for the fMRI time series based on the multivariate Gaussian distribution assumption, and the parameter λ controls the amount of regularization in glasso.

After maximizing the penalized profile log-likelihood and obtaining a sparse estimate of Σ_{ij}, the matrix was binarized according to whether an element was zero or non-zero, and the MBPCA algorithm was applied to obtain a GICC scores for different values of λ and d. We manually choose a list of λ so that theverage proportion of non-zero elements in Σ_{ij} was approximately in a range from 0.1 to 0.9.

For illustration purposes, we only show results from the datasets with dimensions $d = 25, 50$, which corresponding to the parameter $D = d(d-1)/2$ for the MBPCA algorithm. Each D-dimension binary vector $\mathbf{x}_{ij}$ is the vectorized upper triangular part of the binarized

version of Σ_{ij}. The MBPCA model was then fit using $K_1 = K_2 = 10$ and allowing the maximum number of iterations to be 100.

Figure. 4.9 shows how the level of sparsity in the precision matrix and estimated GICC score change with the tuning parameter λ for the d=25 dataset. As λ varys from 500 to 2500, the sparsity level of the precision matrix increases substantially, but the GICC score fluctuates around 0.6 and does not appear to varying as λ changes. This indicates that GICC score is a robust measure of the reproducibility of a binarized precision matrix, without the need to fine-tune the penalty parameter λ. Figure. 4.11 shows similar results for the d =50 dataset. It reveals the same property of the GICC score, with stable values around 0.8 under a wide range of sparsity levels.

Figure. 4.10 and 4.12 shows the estimation results (mean, 1st/2nd PCs) under $\lambda = 500$ for d = 25 and d = 50 respectively. The mean figure indicates the average precision matrix pattern, and regions with larger magnitude in their principal components indicate those with larger variation between/within subjects. The clear separation of regions in the precision matrix corresponds to different functional areas, including the visual, somatomotor, and default mode networks. The mean image in Fig. 4.10 shows high positive values in regions related to the visual network and cerebellum, and in the connections between these regions. The first between-PC shows negative values throughout, particularly in cognitive-control regions. The second between-PC shows high positive values in the somatomotor regions and in its connections with all other regions, and negative values in the regions in the default mode network (DMN), as well as in their connections with all other regions, indicating high

between subject variation in these regions. The first within-PC shows negative values in regions in the somatomotor network, as well as its connections, while the second within-PC show positive values in the DMN and its connections. The interpretation for results of the same analysis on the d=50 dataset, shown in Fig. 4.12, is similar. Of particular note is the change in sign of the patterns in the second within-PC for regions in the DMN.

In summary, the results illustrate that the MCPCA algorithm can serve as an effective way of exploring the reproducibility of an fMRI study. The estimate GICC score is robust under a wide range of penalty values and corresponding sparsity levels of binary functional connectivity pattern.

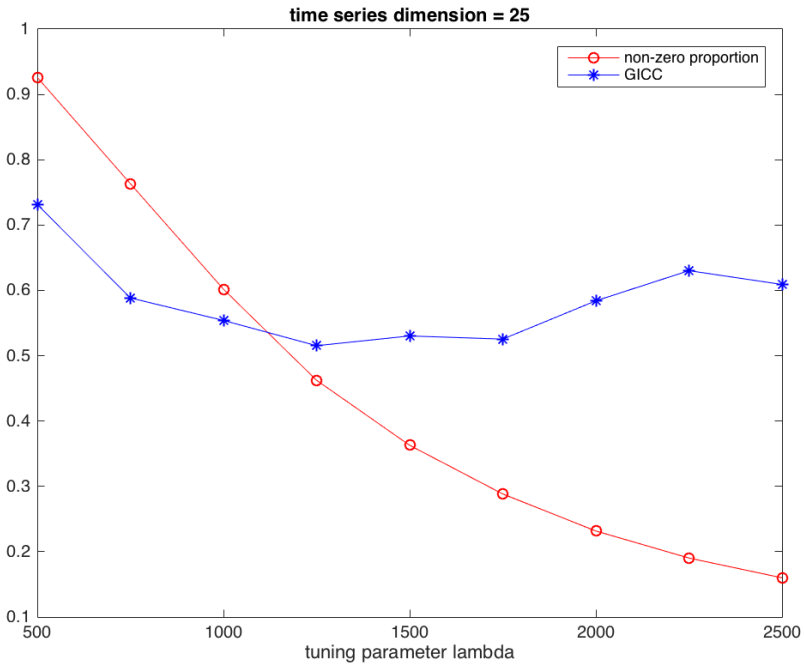


Figure 4.9: GICC score and sparsity level under different tuning parameter lambda for d = 25 dataset. The red line with circle marker denotes the average proportion of non-zero element of estimate precision matrix. The blue line with star maker denotes the GICC score computed from MBPCA results.

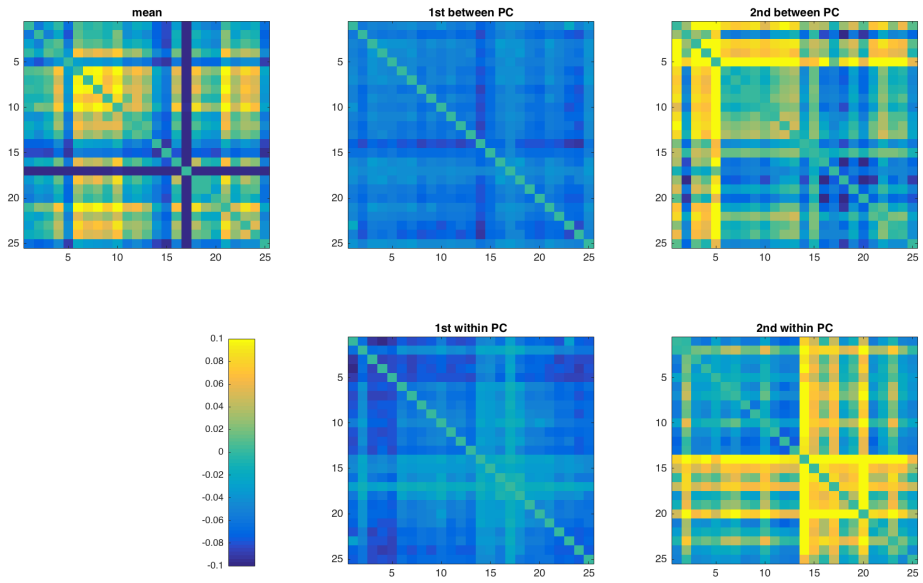


Figure 4.10: The estimated mean and 1st/2nd between/within subject principal components for the d = 25 dataset.

4.5 Conclusion

We proposed Multilevel Binary PCA model that combines multilevel designs with a binary PCA framework. A VEM algorithm was adapted for estimation and three simulation studies demonstrated a good performance of the algorithm. Moreover, we linked MBPCA and GICC, a novel reproducibility measure, and estimated GICC in both NHANSE dataset and a fMRI dataset.

It is worth noting that the proposed MBPCA model collapses to the original (single-level) binary PCA model (Tipping, 1999). Of course, one can use the original binary framework and ignore the multilevel settings. However, the resulting principal compo-

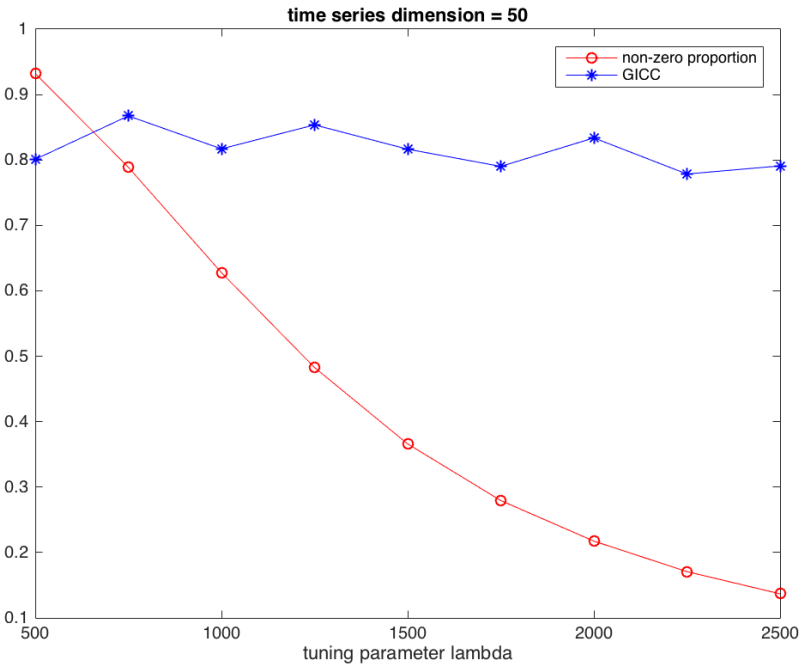


Figure 4.11: GICC score and sparsity level under different tuning parameter lambda for d = 50 dataset. The red line with circle marker denotes the average proportion of non-zero element of estimate precision matrix. The blue line with star maker denotes the GICC score computed from MBPCA results.

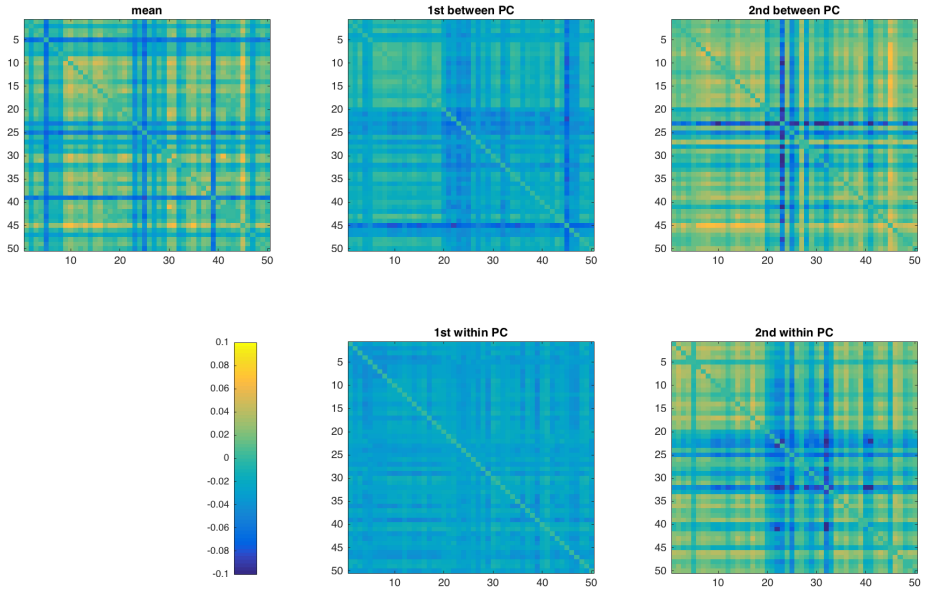


Figure 4.12: The estimated mean and 1st/2nd between/within subjects principal components for the d = 50 dataset.

nents could not be separated to between- and within-subject specific components (please see Zipunnikov et al. (2011b)). Another point to be made is that PCs estimated by MBPCA model is orthonormally invariant. Thus, we could only estimate the subspace spanned by the original PCs, not the components themselves. If orthogonality is required, a further singular value decomposition could be implemented on the estimates of $\hat{\Theta}$ and $\hat{\Psi}$. Moreover, a Gaussian prior could be added to the principal components for Bayesian analysis by fitting a Monte Carlo algorithm described in Mohamed et al. (2008).

Furthermore, the MBPCA model with VEM algorithm could also be generalized to functional data scenarios which is described in Goldsmith et al. (2015). The algorithm could be easily modified such that one can add functional terms as well as fixed or random effect regression terms at no further big computational cost.

It also remains interesting for doing the multilevel principal component analysis for other types of data. For instance, a dataset that records integer counts that follows Possion distribution or Negative Binomial distribution are also popular in application and remains to be explored.

Chapter 5

Discussion and Future Work

PCA is the best starting point for dimension reduction and has been a standard tool in the statistician's toolbox for almost a century. However, dimension reduction for new and complex dataset requires new statistical methods that generalize and extends PCA. For example, data with highly non-linear patterns and data with non-continuous types both require new techniques for meaningful dimension reduction.

In this research three novel generalizations of PCA has been proposed. All methods have both advantages and shortcomings.

First, the principal surface algorithm is a fast and simple algorithm that applies to manifold-like structures. The first advantage of the algorithm is that there is only one parameter one has to tune, compared with many other methods that have multiple tunning parameters. Second, the algorithm can be easily applied to $R^n \longleftrightarrow R^m$ scenarios. Though the algorithm was implemented in $R^3 \rightarrow R^2$ in the corpus callosum application, it could

also be applied to a principal curve algorithm which is $R^3 \to R$ or higher dimension cases. Third, the algorithm provides consistency in 2D parametrizations, which provides us the convenience of further pixel based analysis on the 2D images. Nonetheless, this algorithm also suffers from several shortcomings. First, the algorithm does not fit data structures with high complexity, due to the lack of freedom. The algorithm was designed to be simple and easily tuned so that it lacks degrees of freedom for complex structures, which may require more adaptive algorithms. Second, the current algorithm does not allow boundary restrictions. Work has been done to analyze principal curves with fixed origin in Caffo et al. (2008). The extension to surfaces seems to be quite interesting.

Second, we proposed the concept of graphical intra-class correlation coefficient (GICC). The definition is based on a multivariate probit-linear mixed model. This model provides a measurement of the reproducibility for binary graphs. Although GICC and the associate probit models provide us a novel and useful measurement, it has one major pitfall. Due to the Gibbs sampler step in the EM algorithm, this method does not have the ability for scaling up.

Third, the multilevel binary principal component analysis (MBPCA) is proposed and the variational expectation maximization (VEM) is used for optimizing the likelihood function. The MBPCA method not only provides us a fast and scalable method for modeling high dimensional multilevel binary dataset, it also serves as an alternative way of calculating the GICC. The VEM algorithm also enables one to add fixed effect regression in the model and also provides us a valid approach for modeling data follows other non-Gaussian

distribution. At the same time, functional PCA could also be implemented for smoothness.

Bibliography

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological reports*, 19(1):3–11.

Bazin, P.-L., Ye, C., Bogovic, J. A., Shiee, N., Reich, D. S., Prince, J. L., and Pham, D. L. (2011). Direct segmentation of the major white matter tracts in diffusion tensor images. *NeuroImage*, 58(2):458 – 468.

Biswal, B., Yetkin, F. Z., Haughton, V. M., and Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic resonance in medicine*, 34(4):537–541.

Blondel, V. D., Gajardo, A., Heymans, M., Senellart, P., and Van Dooren, P. (2004). A measure of similarity between graph vertices: Applications to synonym extraction and web searching. *SIAM review*, 46(4):647–666.

Caffo, B. S., Crainiceanu, C. M., Deng, L., and Hendrix, C. W. (2008). A case study in pharmacologic colon imaging using principal curves in single-photon emission computed tomography. *Journal of the American Statistical Association*, 103(484):1470–1480.

BIBLIOGRAPHY

Caffo, B. S., Jank, W., and Jones, G. L. (2005). Ascent-based monte carlo expectation–maximization. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):235–251.

Chan, J. S. and Kuk, A. Y. (1997). Maximum likelihood estimation for probit-linear mixed models with correlated random effects. *Biometrics*, pages 86–97.

Collins, M., Dasgupta, S., and Schapire, R. E. (2001). A generalization of principal component analysis to the exponential family. In *NIPS*, volume 13, page 23.

Cribben, I., Haraldsdottir, R., Atlas, L. Y., Wager, T. D., and Lindquist, M. A. (2012). Dynamic connectivity regression: determining state-related changes in brain connectivity. *Neuroimage*, 61(4):907–920.

Deuker, L., Bullmore, E. T., Smith, M., Christensen, S., Nathan, P. J., Rockstroh, B., and Bassett, D. S. (2009). Reproducibility of graph metrics of human brain functional networks. *Neuroimage*, 47(4):1460–1468.

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009). Multilevel functional principal component analysis. *Annals of Applied Statistics*, 3(1):458–488.

Di Martino, A., Scheres, A., Margulies, D., Kelly, A., Uddin, L., Shehzad, Z., Biswal, B., Walters, J., Castellanos, F., and Milham, M. (2008). Functional connectivity of human striatum: a resting state fmri study. *Cerebral cortex*, 18(12):2735–2747.

BIBLIOGRAPHY

Dong, D. and McAvoy, T. (1996). Nonlinear principal component analysis based on principal curves and neural networks. *Computers Chemical Engineering*, 20(1):65 – 78.

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in soboley spaces. *Lecture notes in mathematics*, 571/1977:85–100.

Einbeck, J., Evers, L., and Powell, B. (2010). Data compression and regression through local principal curves and surfaces. *International journal of neural systems*, 20(3):177–192.

Filippini, N., MacIntosh, B. J., Hough, M. G., Goodwin, G. M., Frisoni, G. B., Smith, S. M., Matthews, P. M., Beckmann, C. F., and Mackay, C. E. (2009). Distinct patterns of brain activity in young carriers of the apoe-ε4 allele. *Proceedings of the National Academy of Sciences*, 106(17):7209–7214.

Fisher, R. A., Genetiker, S., Fisher, R. A., Genetician, S., Britain, G., Fisher, R. A., and Généticien, S. (1970). *Statistical methods for research workers*, volume 14. Oliver and Boyd Edinburgh.

Fletcher, P., Lu, C., Pizer, S., and Joshi, S. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *Medical Imaging, IEEE Transactions on*, 23(8):995 – 1005.

Fletcher, P. T. (2004). *Statistical variability in nonlinear spaces: Application to shape analysis and DT-MRI*. PhD thesis, Citeseer.

BIBLIOGRAPHY

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Gerber, S., Tasdien, T., and Whitaker, R. (2009). Dimensionality reduction and principal surfaces via kernel map manifolds. *Computer Vision, 2009 IEEE 12th International Conference*, pages 529–536.

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

Gnanadesikan, H. (1997). *Methods for Statistical Analysis of Multivariate Observations*. Wiley, New York.

Goldsmith, J., Caffo, B., Crainiceanu, C., Reich, D., Du, Y., and Hendrix, C. (2011a). Nonlinear tube-fitting for the analysis of anatomical and functional structures. *The annals of applied statistics*, 5(1):337–363.

Goldsmith, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2012). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(3):453–469.

Goldsmith, J., Crainiceanu, C. M., Caffo, B. S., and Reich, D. S. (2011b). Penalized functional regression analysis of white-matter tract profiles in multiple sclerosis. *NeuroImage*, 57(2):431–439.

BIBLIOGRAPHY

Goldsmith, J., Wand, M., and Crainiceanu, C. (2011c). Functional regression via variational bayes. *Electronic Journal of Statistics*, 5:572–602.

Goldsmith, J., Zipunnikov, V., and Schrack, J. (2015). Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics*.

Greven, S., Crainiceanu, C., Caffo, B., and Reich, D. (2011). Longitudinal functional principal component analysis. *Recent Advances in Functional Data Analysis and Related Topics*, pages 149–154.

Griffanti, L., Salimi-Khorshidi, G., Beckmann, C. F., Auerbach, E. J., Douaud, G., Sexton, C. E., Zsoldos, E., Ebmeier, K. P., Filippini, N., Mackay, C. E., et al. (2014). Ica-based artefact removal and accelerated fmri acquisition for improved resting state network imaging. *Neuroimage*, 95:232–247.

Guye, M., Bettus, G., Bartolomei, F., and Cozzone, P. J. (2010). Graph theoretical analysis of structural and functional connectivity mri in normal and pathological brain networks. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 23(5-6):409–421.

Hastie, T. (1984). Principal curves and surfaces. *Technical report*, 11.

Hastie, T. and Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, 84(406):502–516.

Horrace, W. C. (2005). Some results on the multivariate truncated normal distribution. *Journal of Multivariate Analysis*, 94(1):209–221.

BIBLIOGRAPHY

Huang, S., Li, J., Sun, L., Ye, J., Fleisher, A., Wu, T., Chen, K., and Reiman, E. (2010). Learning brain connectivity of alzheimer's disease by sparse inverse covariance estimation. *NeuroImage*, 50(3):935–949.

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37.

Jung, S., Foskey, M., and Marron, J. (2011). Principal arc analysis on direct product manifolds. *The Annals of Applied Statistics*, 5(1):578–603.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9):1464 –1480.

Kotecha, J. H. and Djuric, P. M. (1999). Gibbs sampling approach for generation of truncated multivariate gaussian random variables. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 3, pages 1757–1760. IEEE.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243.

Kramer, M. A. (1992). Autoassociative neural networks. *Computers and Chemical Engineering*, 16(4):313 – 328. Neutral network applications in chemical engineering.

BIBLIOGRAPHY

Landman, B. A., Huang, A. J., Gifford, A., Vikram, D. S., Lim, I. A. L., Farrell, J. A., Bogovic, J. A., Hua, J., Chen, M., Jarso, S., et al. (2011). Multi-parametric neuroimaging reproducibility: A 3-t resource study. *Neuroimage*, 54(4):2854–2866.

Leblanc, M. and Tibshirani, R. (1994). Adaptive principal surfaces. *Journal of the American Statistical Association*, 89(425):53–64.

Lee, S., Zipunnikov, V., Reich, D. S., and Pham, D. L. (2015). Statistical image analysis of longitudinal ravens images. *Frontiers in neuroscience*, 9.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 226–233.

Luders, E., Narr, K. L., Hamilton, L. S., Phillips, O. R., Thompson, P. M., Valle, J. S., Del'Homme, M., Strickland, T., McCracken, J. T., Toga, A. W., and Levitt, J. G. (2009). Decreased callosal thickness in attention-deficit/hyperactivity disorder. *Biological Psychiatry*, 65(1):84 – 88.

McCulloch, C. E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, 89(425):330–335.

Mohamed, S., Ghahramani, Z., and Heller, K. A. (2008). Bayesian exponential family pca. In *Advances in Neural Information Processing Systems*, pages 1089–1096.

Naghavi, M., Wang, H., Lozano, R., Davis, A., Liang, X., Zhou, M., Vollset, S. E., Ozgoren, A. A., Abdalla, S., Abd-Allah, F., et al. (2015). Global, regional, and national age-

sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the global burden of disease study 2013. *Lancet*, 385(9963):117–171.

Ozertem, U. and Erdogmus, D. (2011). Locally defined principal curves and surfaces. *Journal of Machine learning research*, 12:1249–1286.

Ozturk, A., Smith, S., Gordon-Lipkin, E., Harrison, D., Shiee, N., Pham, D., Caffo, B., Calabresi, P., and Reich, D. (2010). Mri of the corpus callosum in multiple sclerosis: association with disability. *Multiple Sclerosis*, 16(2):166–177.

Palus, M. and Dvorak, I. (1992). Singular-value decomposition in attractor reconstruction: Pitfalls and precautions. *Physica D: Nonlinear Phenomena*, 55(12):221 – 234.

Pietrangelo, A. and Higuera, V. (2015). Multiple sclerosis by the numbers: Facts, statistics, and you. http://www.healthline.com/health/multiple-sclerosis/facts-statistics-infographic.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Reich, D., Ozturk, A., Calabresi, P., and Mori, S. (2010). Automated vs. conventional tractography in multiple sclerosis: Variability and correlation with disability. *NeuroImage*, 49:3047–3056.

Ridout, M. S., Demetrio, C. G., and Firth, D. (1999). Estimating intraclass correlation for binary data. *Biometrics*, 55(1):137–148.

Rodrıguez, G. and Elo, I. (2003). Intra-class correlation in random-effects models for binary data. *The Stata Journal*, 3(1):32–46.

Roy, N. and Gordon, G. (2002). Exponential family pca for belief compression in pomdps. In *NIPS*, volume 2, pages 1043–1049.

Sajama, S. and Orlitsky, A. (2004). Semi-parametric exponential family pca. In *Advances in Neural Information Processing Systems*, pages 1177–1184.

Salimi-Khorshidi, G., Douaud, G., Beckmann, C. F., Glasser, M. F., Griffanti, L., and Smith, S. M. (2014). Automatic denoising of functional mri data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage*, 90:449–468.

Salvador, R., Suckling, J., Coleman, M. R., Pickard, J. D., Menon, D., and Bullmore, E. (2005). Neurophysiological architecture of functional magnetic resonance images of human brain. *Cerebral Cortex*, 15(9):1332–1342.

Schein, A. I., Saul, L. K., and Ungar, L. H. (2003). A generalized linear model for principal component analysis of binary data. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, volume 38, page 46.

Shou, H., Eloyan, A., Lee, S., Zipunnikov, V., Crainiceanu, A., Nebel, M., Caffo, B., Lindquist, M., and Crainiceanu, C. (2013). Quantifying the reliability of image replica-

tion studies: The image intraclass correlation coefficient (i2c2). *Cognitive, Affective, & Behavioral Neuroscience*, 13(4):714–724.

Shrout, P. E., Fleiss, J. L., et al. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*, 86(2):420–428.

Smith, S. M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T. E., Mackay, C. E., Watkins, K. E., Ciccarelli, O., Cader, M. Z., Matthews, P. M., et al. (2006). Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *Neuroimage*, 31(4):1487–1505.

Stanish, W. M. and Taylor, N. (1983). Estimation of the intraclass correlation coefficient for the analysis of covariance model. *The American Statistician*, 37(3):221–224.

Telesford, Q. K., Burdette, J. H., and Laurienti, P. J. (2013). An exploration of graph metric reproducibility in complex brain networks. *Frontiers in neuroscience*, 7.

Telesford, Q. K., Morgan, A. R., Hayasaka, S., Simpson, S. L., Barret, W., Kraft, R. A., Mozolic, J. L., and Laurienti, P. J. (2010). Reproducibility of graph metrics in fmri networks. *Frontiers in neuroinformatics*, 4.

Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

Tipping, M. E. (1999). Probabilistic visualisation of high-dimensional binary data. *Advances in neural information processing systems*, pages 592–598.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.

Troiano, R. P., Berrigan, D., Dodd, K. W., Masse, L. C., Tilert, T., McDowell, M., et al. (2008). Physical activity in the united states measured by accelerometer. *Medicine and science in sports and exercise*, 40(1):181.

Van Den Heuvel, M. P. and Hulshoff Pol, H. E. (2010). Exploring the brain network: a review on resting-state fmri functional connectivity. *European Neuropsychopharmacology*, 20(8):519–534.

Van Domelen, D. R. and Pittard, W. S. (2014). Flexible r functions for processing accelerometer data, with emphasis on nhanes 2003–2006. *A peer-reviewed, open-access publication of the R Foundation for Statistical Computing*, page 52.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H., et al. (2013). The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79.

Varadhan, R. and Roland, C. (2008). Simple and globally convergent methods for accelerating the convergence of any em algorithm. *Scandinavian Journal of Statistics*, 35(2):335–353.

Varoquaux, G., Gramfort, A., Poline, J.-B., and Thirion, B. (2010). Brain covariance se-

lection: better individual functional connectivity models using population prior. In *Advances in Neural Information Processing Systems*, pages 2334–2342.

Vidal, C., Nicolson, R., DeVito, T., Hayashi, K., Geaga, J., Drost, D., Williamson, P., Rajakumar, N., Sui, Y., Dutton, R., et al. (2006). Mapping corpus callosum deficits in autism: an index of aberrant cortical connectivity. *Biological Psychiatry*, 60(3):218–225.

Weinberger, K. Q. and Saul, L. K. (2006). Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70(1):77–90.

Wilhelm, S. and G, M. B. (2013). *tmvtnorm: Truncated Multivariate Normal and Student t Distribution*. R package version 1.4-8.

Wood, S. (2003). Thin plate regression splines. *Journal of the royal statistical society*, 65(1):95–114.

Xu, Y. and Lindquist, M. A. (2015). Dynamic connectivity detection: an algorithm for determining functional connectivity change points in fmri data. *Frontiers in neuroscience*, 9.

Yue, C., Chen, S., Sair, H. I., Arian, R., and Caffo, B. S. (2015). Estimating a graphical intra-class correlation coefficient (gicc) using multivariate probit-linear mixed models. *Computational Statistics and Data Analysis*, 89:126–133.

Yushkevich, P. A., Zhang, H., Simon, T. J., and Gee, J. C. (2008). Structure-specific statistical mapping of white matter tracts. *NeuroImage*, 41(2):448–461.

Zager, L. A. and Verghese, G. C. (2008). Graph similarity scoring and matching. *Applied mathematics letters*, 21(1):86–94.

Zhang, H., Awate, S. P., Das, S. R., Woo, J. H., Melhem, E. R., Gee, J. C., and Yushkevich, P. A. (2010). A tract-specific framework for white matter morphometry combining macroscopic and microscopic tract features. *Medical image analysis*, 14(5):666–673.

Zhu, H., Kong, L., Li, R., Styner, M., Gerig, G., Lin, W., and Gilmore, J. (2011). Fadtts: Functional analysis of diffusion tensor tract statistics. *NeuroImage*, 56(3):1412–1425.

Zhu, H., Styner, M., Tang, N., Liu, Z., Lin, W., and Gilmore, J. (2010). Frats: Functional regression analysis of dti tract statistics. *Medical Imaging, IEEE Transactions on*, 29(4):1039–1049.

Zipunnikov, V., Caffo, B., Yousem, D. M., Davatzikos, C., Schwartz, B. S., and Crainiceanu, C. (2011a). Multilevel functional principal component analysis for high-dimensional data. *Journal of Computational and Graphical Statistics*, 20(4):852–873.

Zipunnikov, V., Caffo, B., Yousem, D. M., Davatzikos, C., Schwartz, B. S., and Crainiceanu, C. M. (2011b). Multilevel functional principal component analysis for high-dimensional data. *Journal of Computational and Graphical Statistics*, 20(4):852–873.

Zipunnikov, V., Greven, S., Shou, H., Caffo, B. S., Reich, D. S., Crainiceanu, C. M., et al. (2014). Longitudinal high-dimensional principal components analysis with appli-

cation to diffusion tensor imaging of multiple sclerosis. *The Annals of Applied Statistics*, 8(4):2175–2202.

Zou, G. and Donner, A. (2004). Confidence interval estimation of the intraclass correlation coefficient for binary outcome data. *Biometrics*, 60(3):807–811.

CURRICULUM VITAE

CHEN YUE

cyue1@jhu.edu

615 N. Wolfe St. E3030

Baltimore, MD 21205

http://www.biostat.jhsph.edu/∼cyue

Date of Birth: April 18st, 1986

Place of Birth: Beijing, China

EDUCATION

2011 - 2016 **Johns Hopkins Bloomberg School of Public Health**, Baltimore, MD
Ph.D. in Biostatistics
Thesis title: *Generalizations, Extensions and Applications for Principal Component Analysis*
Advisor: Dr. Brian Caffo
Co-advisor: Dr. Vadim Zipunnikov

2008 - 2010 **Tsinghua University**, Beijing, China
M.S. in Mathematical Science
Advisor: Dr. Ying Yang

2004 - 2008 **Tsinghua University**, Beijing, China
B.S. in Mathematical Science

PROFESSIONAL EXPERIENCE

CURRICULUM VITAE

04/2015 - 09/2015	**Quantitative Trading Intern**
	Tower Research Capital, New York, NY
06/2014 - 09/2014	**Research Scientist Intern**
	Amazon, Seattle, WA
2014 - now	**Level 2 Candidate**
	CFA Institute

HONORS AND AWARDS

Johns Hopkins University

2014	Joseph Zeger Conference Travel Award
2013	Joseph Zeger Conference Travel Award
2013	Travel awards for Workshop on Neuroimaging Data Analysis

Tsinghua University

2007	Outstanding student award second place
2006	Outstanding student award third place
2005	Outstanding student award second place

PUBLICATIONS

PUBLISHED/SUBMITTED

Li S, Chen S, **Yue C**, Caffo BS (2016) Independent Component Analysis through Fast Nonparametric Density Estimation. *Frontiers of Neuroscience*. Accepted.

CURRICULUM VITAE

Yue C, Xu Y, Chen S, Goldsmith J, Caffo BS, Zipunnikov V (2015) Multilevel Binary Principal Component Analysis with Application to Brain Activity. *In preparation*.

Hiruy H, Fuchs EJ, Marzinke MA, Bakshi RP, Breakey JC, Aung WS, Manohar M, **Yue C**, Caffo BS, et al. (2015) A Phase 1 Randomized, Blinded Comparison of the Pharmacokinetics and Colonic Distribution of Three Candidate Rectal Microbicide Formulations of Tenofovir 1% Gel with Simulated Unprotected Sex (CHARM-02). AIDS research and human retroviruses. 31(11):1098-108.

Leyva F, Fuchs EJ, Bakshi R; Carballo-Dieguez A, Ventuneac A, **Yue C**; Caffo B, et al. (2015) Simultaneous Evaluation of Safety, Acceptability, Pericoital Kinetics, and Ex Vivo Pharmacodynamics Comparing Four Rectal Microbicide Vehicle Candidates. *AIDS Research and Human Retroviruses*.31(11):1089-1097.

Coughlin J M, Wang Y, Munro CA, Ma S, **Yue C**, Chen S., et al. (2015) Neuroinflammation and brain atrophy in former NFL players: an in vivo multimodal imaging pilot study. *Neurobiology of disease*, 74, 58-65.

Yue C, Chen S, Sair H I, et al. (2015) Estimating a graphical intra-class correlation coefficient (GICC) using multivariate probit-linear mixed models. *Computational Statistics & Data Analysis*, 89: 126-133.

Coughlin J M, Wang Y, Ma S, **Yue C**, et al. (2014) Regional brain distribution of translocator protein using [11C] DPA-713 PET in individuals infected with HIV. *Journal of neurovirology*, 20(3): 219-232. *Journal of Neuro Virology*

CURRICULUM VITAE

Yue C, Zipunnikov V, Bazin P, Pham D, Reich D, Crainiceanu C, Caffo B. (2014) Parameterization of white matter manifold-like structures using principal surfaces *Under Revision by Journal of American Statistical Association*

Ye F, **Yue C**, Yang Y. (2013) Modeling time-dependent overdispersion in longitudinal count data. *Computational Statistics & Data Analysis*. 58:257-64.

PRESENTATIONS

2015 Estimating a graphical intra-class correlation coefficient (GICC) using multivariate probit-linear mixed models. ENAR Spring Meeting, Miami, FL
2014 Parameterization of white matter manifold-like structures using principal surfaces. ENAR Spring Meeting, Baltimore, MD
2013 Principal Surfaces. ENAR Spring Meeting, Orlando, FL

TEACHING

2015 Statistical Reasoning **I-II**, 140.611-612.
2014 Statistical Methods in Public Health **I-IV**, 140.621-624.
2013 Statistical Methods in Public Health **III-IV**, 140.623-624.
2013 Public Health Biostatistics 280.345.
2012 Methods in Biostatistics **I-IV**, 140.651-654.