ADVANCED STATISTICAL LEARNING METHODS FOR HETERGENEOUS MEDICAL IMAGING DATA

Chao Huang

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2019

Approved by:

Hongtu Zhu

James Stephen Marron

Marc Niethammer

Young Kinh Truong

Donglin Zeng

ABSTRACT

Chao Huang: Advanced Statistical Learning Methods for
Hetergeneous Medical Imaging Data
(Under the direction of Hongtu Zhu)

Most neuro-related diseases and disabling diseases display significant heterogeneity at the imaging and clinical scales. Characterizing such heterogeneity could transform our understanding of the etiology of these conditions and inspire new approaches to urgently needed preventions, diagnoses, and treatments. However, existing statistical methods face major challenges in delineating such heterogeneity at subject, group and study levels. In order to address these challenges, this work proposes several statistical learning methods for heterogeneous imaging data with different structures.

First, we propose a dynamic spatial random effects model for longitudinal imaging dataset, which aims at characterizing both the imaging intensity progression and the temporal-spatial heterogeneity of diseased regions across subjects and time. The key components of proposed model include a spatial random effects model and a dynamic conditional random field model. The proposed model can effectively detect the dynamic diseased regions in each patient and present a dynamic statistical disease mapping within each subpopulation of interest.

Second, to address the group level heterogeneity in non-Euclidean data, we develop a penalized model-based clustering framework to cluster high dimensional manifold data in symmetric spaces. Specifically, a mixture of geodesic factor analyzers is proposed with mixing proportions determined through a logistic model and Riemannian normal distribution in each component for data in symmetric spaces. Penalized likelihood approaches are used to realize variable selection procedures. We apply the proposed model to the ADNI hippocampal surface data, which shows excellent clustering performance and remarkably reveal meaningful clusters in the mixed population with controls and subjects with AD.

Finally, to consider the potential heterogeneity caused by unobserved environmental, demographic and technical factors, we treat the imaging data as functional responses, and set up a surrogate variable analysis framework in functional linear models. A functional latent factor regression model is proposed. The confounding factors and the bias of local linear estimators caused by the confounding factors can be estimated and removed using singular value decomposition on residuals. We further develop a test for linear hypotheses of primary coefficient functions. Both simulation studies and ADNI hippocampal surface data analysis are conducted to show the performance of proposed method.

To my mentor, parents and friends, I couldn't have done this without you. Thank you for all of your support along the way.

ACKNOWLEDGEMENTS

First, I would like to thank my academic advisor Professor Hongtu Zhu. I am grateful to him for his unwavering support for my academic activities. Apart from that, I am privileged to have had the opportunity to be acquainted with such a wonderful person. I would like to thank Professor Steve Marron for his insightful comments about my research, for his painstaking efforts in improving the quality of my writing, and for serving in my orals committee. I thank Professor Marc Niethammer for his encouragement and advice and also for serving in my orals committee. I thank Professor Donglin Zeng for serving in my orals committee. I thank Professor Young K. Truong for his advice and serving in my orals committee.

I am thankful to the entire faculty and the staff members of the department, for their efforts at resolving any problems I faced, and for providing a wonderful academic environment. I would like to thank Professor Martin Styner and Professor Rebecca Knickmeyer at Biomedical Research Imaging Center (BRIC) for their wonderful guidance and constant support. It has also been very pleasant experience as a student at University of North Carolina and for that I must thank to my colleagues Tengfei Li, Hai Shu, Ziliang Zhu, Fan Zhou, Jingwen Zhang, Yue Wang, Bingxin Zhao and Xifeng Wang at BIG-S2 lab and many friends Linglong Kong, Dehan Kong, Qiang Sun, Kaixian Yu, Ziqi Chen, Yan Jin, Zhengwu Zhang, Benjamin B. Risk, Yufeng Liu, Yang Yu and Ting Wang at the department. Finally, I lack enough words to express my gratitude towards my family- my wife Rongjie Liu and my little boy Jacob - whose encouragement, affection and sacrifices have been the biggest driving forces in this endeavor.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Most neuro-related diseases (e.g., Alzheimer's disease) and disabling diseases (e.g., osteoarthritis) display significant heterogeneity at the imaging and clinical scales. Characterizing such heterogeneity could transform our understanding of the etiology of these conditions and inspire new approaches to urgently needed preventions, diagnoses, and treatments. However, existing statistical methods face major challenges in delineating such heterogeneity at subject, group and study levels. In order to address these challenges, this work proposes several statistical learning methods for heterogeneous imaging data with different structures (e.g., longitudinal, non-Euclidean, or functional data).

First, we propose a dynamic spatial random effects model for longitudinal imaging dataset, which aims at characterizing both the imaging intensity progression and the temporal-spatial heterogeneity of diseased regions across subjects and time. The key components of proposed model include a spatial random effects model and a dynamic conditional random field model. To estimate the unknown parameters in proposed model, we employ a pseudo-likelihood function and optimize it by using an expectation-maximization algorithm. To estimated the dynamic diseased regions for each patient, the Maximum A Posteriori on Markov Random Field (MRF-MAP) method is adopted. The proposed model can effectively detect the dynamic diseased regions in each patient and present a dynamic statistical disease mapping within each subpopulation of interest.

Second, to address the group level heterogeneity in non-Euclidean data, we develop a penalized model-based clustering framework to cluster high dimensional manifold data in symmetric spaces. Specifically, a mixture of geodesic factor analyzers is proposed with mixing proportions determined through a logistic model and Riemannian normal distribution in each component for data in symmetric spaces. A geodesic factor analyzer is established to explicitly model the high dimensional features. Penalized likelihood approaches are used to realize variable selection procedures. Simulation studies are performed on data generated from Euclidean space, sphere, and shape space. We also apply the proposed model to the ADNI hippocampal surface data, which shows excellent clustering performance and remarkably reveal meaningful clusters in the mixed population with controls and subjects with AD.

Finally, to consider the potential heterogeneity caused by unobserved environmental, demographic and technical factors, we treat the imaging data as functional responses, and set up a surrogate variable analysis framework in functional linear models. In particular, a functional latent factor regression model is proposed. An estimation procedure for the proposed model is derived by using local linear regression techniques. The confounding factors and the bias of local linear estimators caused by the confounding factors can be estimated and removed using singular value decomposition on residuals. We further develop a test for linear hypotheses of primary coefficient functions. Both simulation studies and ADNI hippocampal surface data analysis are conducted to show the performance of proposed method.

CHAPTER 2: LITERATURE REVIEW

With the rapid growth of modern technology, many large-scale biomedical studies, e.g., Alzheimer's disease neuroimaging initiative (ADNI) study (Mueller et al., 2005), Osteoarthritis Initiative (OAI) study (Peterfy et al., 2008), and UK Biobank study (Sudlow et al., 2015), have been conducted to collect massive datasets with large volumes of complex information from increasingly large cohorts. Despite the numerous successes of biomedical studies, it has been difficult to unravel the disease etiology largely due to its heterogeneity at the genomic, imaging, and clinical scales. Specifically, imaging heterogeneity often represents at three different levels: subject level, group level, and study level. At the subject level, diseased regions can significantly vary across subjects and/or time in terms of their number, size, shape, and location. At the group level, due to the complexity of disease progression, distinct pathological subtypes are more likely to be found within the same patient group. At the study level, since the dataset is usually collected from multiple centers or different studies, the potential heterogeneity can result from the differences in study environment, population (e.g., race), design and protocols (e.g., imaging acquisition protocol and/or preprocessing pipeline), which are mostly unobserved (Leek and Storey, 2007). Therefore, understanding such imaging heterogeneity may be critical for the development of urgently needed approaches to the prevention, diagnosis, and treatment of these diseases, and precision medicine broadly.

Many studies have been conducting/conducted on various types of imaging data in order to investigate the underlying heterogeneity at different levels. First, the subject-level heterogeneity has been investigated through the individual disease pattern detection at different scales (e.g., diseased region or tumor cell) based on different imaging modalities, including magnetic resonance imaging (MRI), positron emission tomography (PET), computed tomography (CT), and hematoxylin and eosin (H&E) stain (Yuan et al., 2012; Huang et al., 2015; Soufi et al., 2017; Liu et al., 2018). Among the studies, the developed methods can be divided into two groups: supervised learning and unsupervised learning. For supervised learning methods, due to the rapid development of artificial intelligence in precision medicine field, deep learning algorithms, in particular convolutional networks, have already become a popular methodology of choice for pattern detection and segmentation (see Litjens et al. (2017) and references therein). However, in many practical problems, the ground truth of disease pattern is not visible in neither training nor testing dataset due to some technical difficulties. For example, in knee MRI data, due to the small volume of cartilage in relation to the rest of knee, the exact locations of lesions are unknown or difficult to delineate even by experts (Huang et al., 2015). In this case, some unsupervised learning methods, including clustering analysis and hidden Markov model (HMM) based methods, have been developed and applied in recovering the underlying pattern of interest (Geman and Geman, 1984; Li and Singh, 2009; Inano et al., 2014). In particular, besides the cross sectional imaging data analysis, Huang et al. (2015) proposed a semisupervised learning method via a HMM based regression model, i.e., spatial random effects model (SREM), to detect the diseased regions for each subject in longitudinal imaging dataset. However, for longitudinal imaging data, only spatial correlation was set up via Potts model in their proposed model, where the temporal correlation in the underlying disease pattern has not been considered yet.

Second, two main classes of approaches have been applied to assess the group-level heterogeneity. The first class consists of identifying different pathological subtypes using a supervised approach based on prior clinical, pathological, or neuroimaging criteria (Zhang et al., 2014; Byun et al., 2015; Ferreira et al., 2017). The key issue for methods within this class is that all these methods depend on an a priori disease subtype definition, which may be either difficult to obtain (e.g., from autopsy near the date of imaging), or noisy and non-specific (e.g., cognitive or clinical evaluations) (see Varol et al. (2017) and references therein). The second class includes unsupervised learning methods such as clustering (Hwang et al., 2016; Zhang et al., 2016) and semisupervised multivariate methods (Varol et al.,

2017) using voxel-based or surface-based morphometry measures. Although these studies found diverse clusters of atrophy that were partially similar to the ones previously reported in Whitwell et al. (2012), several challenges are faced for the second class of approaches. First, the imaging measurements may lie in some non-Euclidean space, e.g., directional data (Banerjee et al., 2005), shape data (Srivastava et al., 2005), and diffusion tensor data (Rohlfing et al., 2007). Thus, most clustering methods (e.g., K-means, or mean shift) in Euclidean space cannot be used anymore. Second, clustering manifold-value data are often a high-dimensional-low-sample-size problem (Dryden et al., 2005; Banerjee et al., 2005). For example, the dimension of whole brain cortical thickness data can be much larger than the sample size in most imaging studies. Third, manifold data variation is associated with some explanatory covariates (e.g., age, gender, and clinical biomarkers). Applying clustering analysis without considering these covariates will lead to potential risk of estimating clusters that reflect normal inter-individual variability from certain confounds instead of highlighting group-level heterogeneity (Varol et al., 2017). For most existing manifold clustering methods, e.g., K-subspaces (Wang et al., 2009) and nonlinear mean shift (Subbarao and Meer, 2009), they only extend standard clustering algorithms by replacing the Euclidean metric with the geodesic distance in symmetric spaces. Therefore, they are not able to address all the challenges above. To explicitly address all these challenges, Huang et al. (2015) developed a penalized model-based clustering framework to cluster landmark-based planar shape data, which may be generalized for applications on other imaging data with complex structures.

Third, in multiple imaging studies integration, there is a greater need in handling the unknown variance introduced by the study-level heterogeneity, which can hinder the detection of imaging features associated with clinical covariates of interest and cause spurious findings. However, the specific studies have only recently begun to grow substantially in the neuroimaging field (see Guillaume et al. (2018) and references therein). For example, several statistical harmonization techniques have been proposed in the context of different imaging modalities. For conventional MRI studies, intensity normalization techniques have been developed to make the image intensities comparable across studies, including histogram matching (Nyúl et al., 2000), WhiteStripe (Shinohara et al., 2014) and Removal of Artificial Voxel Effect by Linear regression (RAVEL) (Fortin et al., 2016). Another method, called source-based morphometry, adopted independent component analysis (ICA) to remove variability associated with certain scanner parameters in structural MRI (Chen et al., 2014). For diffusion tensor imaging (DTI) data, it has been proposed to use functional normalization, originally developed in Fortin et al. (2014), for harmonizing DTI scalar maps. Another DTI harmonization technique was proposed in Mirzaalian et al. (2016), which was based on rotation invariant spherical harmonics (RISH) and combined the unprocessed DTI images across scanners. However, a major drawback of this method is that it requires DTI data to have similar acquisition parameters across sites, which is often infeasible in multi-site observational analyses. The statistical harmonization was also studied on the cortical thickness measurements (Fortin et al., 2018). In these studies, several statistical approaches that were previously developed for genomics data were adopted for imaging data harmonization, including Functional normalization, Surrogate variable analysis (SVA) (Leek and Storey, 2007) and ComBat (Johnson et al., 2007). Recently, the study-level heterogeneity was considered for the mass-univariate analysis of neuroimaging data (Guillaume et al., 2018), where the unknown covariates were modeled via adopting and modifying the existing Confounder Adjusted Testing and Estimation (CATE) approach (Wang et al., 2017). However, instead of the mass-univariate analysis, the image measures across different voxels are more likely to be treated as a single functional response because the functional data analysis (FDA) is a powerful tool, which can explicitly account for the three key features of the functional data: spatial smoothness, spatial correlation, and the low-dimensional representation. Therefore, it is of great importance to investigate the study-level heterogeneity in some functional regression models, e.g., multivariate varying coefficient model (MVCM, Zhu et al. (2012)).

Next, we will review several different statistical models used to delineate the underlying heterogeneity at different levels.

2.1 Dynamic Diseased Region Detection

The Markov random field (MRF) models have been used for detecting the imaging heterogeneity at subject level. To describe the model, the following notation are introduced first. Let $\mathcal{S}$ represent the pixel (or voxel, in $3D$ problems) lattice, where one single image $\boldsymbol{y}$ is observed. The model assumes that there are K regions, $\{\mathcal{R}_1, \dots, \mathcal{R}_K\}$, such that $\mathcal{S} = \bigcup_{j=1}^{K} \mathcal{R}_j$ and $\mathcal{R}_j \cap \mathcal{R}_k = \emptyset, i \neq k$, so that the observation at pixel $\boldsymbol{s} \in \mathcal{S}$ is given by

$$\boldsymbol{y}(\boldsymbol{s}) = \sum_{j=1}^{K} \xi_j(\boldsymbol{s}) \mathbf{1}\{b(\boldsymbol{s}) = j\} + \boldsymbol{\epsilon}(\boldsymbol{s}), \tag{2.1}$$

where $\boldsymbol{\epsilon}(\boldsymbol{s})$ is a white noise field with known distribution (e.g., $\{\epsilon(\boldsymbol{s}), \boldsymbol{s} \in \mathcal{S}\}$ are zero-mean, independent, identically distributed Gaussian random variables with standard deviation σ). $\xi_j(\boldsymbol{s})$ is a parametric model that corresponds to region $\mathcal{R}_j$, and $b(\boldsymbol{s})$ indicates the corresponding label information: $b(\boldsymbol{s}) = j \iff \boldsymbol{s} \in \mathcal{R}_j$. In this model, the label field $\boldsymbol{b} = \{b(\boldsymbol{s}), \boldsymbol{s} \in \mathcal{S}\}$ is assumed to be a sample from a MRF obtained with a Gibbs model:

$$f(\boldsymbol{b}) = \frac{1}{Z_b} \exp\{- \sum_{(j,k) \in C} V(b(\boldsymbol{s}_j), b(\boldsymbol{s}_k))\}, \tag{2.2}$$

where Z_b is a normalizing constant and the sum in the exponent ranges over all pixels and the cliques of a given neighborhood system on $\mathcal{S}$, and $\{V(b(\boldsymbol{s}_j), b(\boldsymbol{s}_k)), (j,k) \in C\}$ are "potential functions", each one of which depends only on the value of $\boldsymbol{b}$ at the sites that belong to the clique C. These potential functions, together with the neighborhood system selected, control the appearance of the sample field $\boldsymbol{b}$. A potential that is often used is the generalized Ising model, which considers cliques of size 2 (e.g., pairs of sites that are one unit apart), and potentials of the form:

$$V(b(\boldsymbol{s}_j), b(\boldsymbol{s}_k)) = \begin{cases} -\eta, & \text{if } b(\boldsymbol{s}_j) = b(\boldsymbol{s}_k), \\ \eta, & \text{otherwise,} \end{cases} \tag{2.3}$$

where η is a parameter that controls the granularity of the field. Since the label field $\boldsymbol{b}$ is not directly observable, it is often called a class MRF model. The class MRF model has been explored widely in the literature. See, for example, Geman and Geman (1984), Marroquin et al. (2003) and Li and Singh (2009).

For longitudinal imaging studies, an extension of (2.1) is Gaussian hidden Markov model (GHMM, Huang et al. (2015)) which includes a spatial random effects (SRE) model and a Potts model (Besag, 1986; Qian and Titterington, 1991; Zhang et al., 2001). Assume that a longitudinal dataset is observed with imaging intensity $\{y_{ij}(s_k) : k = 1, \ldots, m\}$ measured at time t_j for $j = 1, \ldots, T_i$ and $i = 1, \ldots, n$, where n is the total number of subjects, and T_i is the total number of time points for the i-th subject. Let x_i represent disease status for each subject such that $x_i = 0$ and 1, respectively, represent normal control and diseased patient. For each subject, we assume that $\mathcal{S}$ can be decomposed into the union of normal region R_{i0}, moderately diseased region R_{i1}, and severely diseased region R_{i2}, that is $\mathcal{S} = R_{i0} \cup R_{i1} \cup R_{i2}$ and $R_{ik} \cap R_{ik'} = \emptyset$ for $k \neq k'$. It's also assumed that normal controls are expected to be perfectly healthy, i.e. to not have any diseased regions. For diseased patients, the size and location of R_{i1} and R_{i2} may vary across subjects.

In GHMM, an unobserved random effect $b_i(s_k) \in L = \{0, 1, 2\}$ is introduced to label R_{i0}, R_{i1}, and R_{i2} at each pixel s_k of $\mathcal{S}$ for the $i-$th subject. Moreover, another unobserved random effect $\boldsymbol{v}_i(s_k)$ is introduced to characterize temporal correlations among repeated measures for each subject. Given $\boldsymbol{b}_i$ and $\boldsymbol{v}_i$, a general spatial random effect model given by

$$y_{ij}(s_k) = \boldsymbol{w}_j^T \boldsymbol{\beta}(s_k) + \boldsymbol{w}_j^T \boldsymbol{v}_i(s_k) + \epsilon_{ij}(s_k) + c(x_i, \boldsymbol{w}_j, b_i(s_k), \bar{\boldsymbol{\beta}}), \tag{2.4}$$

where $\boldsymbol{w}_j$ is a $q_w \times 1$ vector of covariates (e.g., time, gender, or genetic marker) and $\boldsymbol{\beta}(s_k)$ is a $q_w \times 1$ vector of regression coefficients representing the dynamic intensity changes at pixel s_k in normal controls. Moreover, $\bar{\boldsymbol{\beta}} = (\bar{\boldsymbol{\beta}}(1), \bar{\boldsymbol{\beta}}(2))^T$ is a $2q_w \times 1$ vector of coefficients to characterize the dynamic intensity changes in the diseased regions $R_{i1} \cup R_{i2}$. The function

$c(x_i, \boldsymbol{w}_j, b_i(s_k), \bar{\boldsymbol{\beta}})$ is defined as

$$c(x_i, \boldsymbol{w}_j, b_i(s_k), \bar{\boldsymbol{\beta}}) = \begin{cases} 0, & x_i = 0, \\ \sum_{l=1}^{2} \boldsymbol{w}_j^T \bar{\boldsymbol{\beta}}(l) \delta(b_i(s_k), l) \ , & x_i \neq 0, \end{cases}$$

where $\delta(\cdot, \cdot)$ is the Kronecker function. Thus, $c(x_i, \boldsymbol{w}_j, b_i(s_k), \bar{\boldsymbol{\beta}})$ equals zero for all pixels for normal controls and the pixels with $b_i(s_k) = 0$ for diseased patients. For diseased patients, pixels in different diseased regions may have different dynamic intensity changes. Moreover, $\epsilon_{ij}(s_k)$s are independent measurement errors across subjects, time, and pixels, following $N(0, \sigma_{s_k}^2)$. For the random effects $\boldsymbol{b}_i$ and $\boldsymbol{v}_i$, it is assumed that $\boldsymbol{b}_i = (b_i(s_1), \ldots, b_i(s_m))^T$ and $\boldsymbol{v}_i = \{\boldsymbol{v}_i(s_k) : k = 1, \ldots, m\}$ are mutually independent. Moreover, $\boldsymbol{v}_i(s_k)$ are mutually independent across pixels and $\boldsymbol{v}_i(s_k)$ follows $N(\mathbf{0}, \boldsymbol{\Sigma}_{v_{s_k}})$ at pixel s_k. It is assumed that $\boldsymbol{b}_i$s are independent across subjects and each $\boldsymbol{b}_i$ follows a Potts model (Besag, 1986; Qian and Titterington, 1991; Zhang et al., 2001), whose Gibbs form is given by

$$p(\boldsymbol{b}_i|\tau) = \exp\{-U(\boldsymbol{b}_i)\tau - \log C(\tau)\}, \tag{2.5}$$

where $U(\boldsymbol{b}_i) = -\sum_{s_k \sim s_l} \delta(b_i(s_k), b_i(s_l))$ and τ is introduced to encourage spatial smoothness in homogeneous regions. Moreover, $C(\tau)$ is the partition function such that $p(\boldsymbol{b}_i|\tau)$ is a probability function. The notation "$\sum_{s_i \sim s_j}$" means that s_i is a neighbor of s_j and each neighboring pair enters the summation only once.

However, for longitudinal imaging data, the Potts model (2.5) in GHMM only considered the spatial correlation and the individual diseased regions are assumed to be unchanged across time, which is not reasonable in practice. Thus, establishing both the temporal and spatial correlations in the disease pattern is of great importance for disease early detection and precision medicine. Some dynamic models have been proposed to set up the temporal-spatial correlations, such as dynamic conditional random field models (Wang and Ji, 2005; Wang et al., 2006; Sutton et al., 2007; Yin et al., 2009).

2.2 Clustering High-Dimensional Manifold Valued Data

Model-based clustering is a popular approach for investigating the group-level heterogeneity and recovering the pathological subtypes. In general, let $\boldsymbol{X}$ be an $N \times J$ data matrix, where each row $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ is the realization of a p-dimensional vector of random variables. Model-based clustering assumes that each observation arises from a finite mixture of K probability distributions, each representing a different cluster or group (Fraley and Raftery, 2002; Bouveyron and Brunet-Saumard, 2014; McNicholas, 2016; Fop et al., 2018). The general form of a finite mixture distribution is specified as follows:

$$f(\boldsymbol{x}_i; \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k f(\boldsymbol{x}_i; \boldsymbol{\Theta}_k), \tag{2.6}$$

where the π_k are the mixing probabilities and $\boldsymbol{\Theta}_k$ is the parameter set corresponding to component k; $\boldsymbol{\Theta}$ denotes the set of all parameters of the mixture. The component densities fully characterize the group structure of the data and each observation belongs to the corresponding cluster according to a latent cluster membership indicator variable $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iK})$, such that $z_{ik} = 1$ if $\boldsymbol{x}_i$ arises from the k-th subpopulation (McLachlan and Peel, 2000).

For a fixed number of components, parameters are usually estimated using the EM algorithm (Dempster et al., 1977a). After parameters have been estimated, each observation is assigned to the corresponding cluster using the maximum a posteriori (MAP) rule (McLachlan and Peel, 2000; McNicholas, 2016). The posterior probabilities $P(z_{ik} = 1|\boldsymbol{x}_i)$ of observing cluster k given the data point $\boldsymbol{x}_i$ are estimated as follows:

$$\hat{P}(z_{ik} = 1|\boldsymbol{x}_i) = \frac{\hat{\pi}_k f(\boldsymbol{x}_i; \hat{\boldsymbol{\Theta}}_k)}{\sum_{k=1}^{K} \hat{\pi}_k f(\boldsymbol{x}_i; \hat{\boldsymbol{\Theta}}_k)}.$$

Then observation $\boldsymbol{x}_i$ is assigned to cluster k if

$$k = \text{argmax}\{\hat{P}(z_{i1} = 1|\boldsymbol{x}_i), \ldots, \hat{P}(z_{iK} = 1|\boldsymbol{x}_i)\}.$$

Nowadays, high-dimensional data are more and more common and the model based clustering approach has adapted to deal with the increasing dimensionality. In particular, a penalization term is introduced on the model parameters and variable selection is performed by inducing sparsity in the estimates. The aim is to maximize a penalized version of the log-likelihood under a mixture model and discard those variables whose parameter estimates are shrunken to zero or to a common value across the mixture components. In its general form, this penalized log-likelihood is as follows:

$$l_Q = \sum_{i=1}^{N} log \left\{ \sum_{k=1}^{K} \pi_k f(\boldsymbol{x}_i; \boldsymbol{\Theta}_k) \right\} - Q_\lambda(\boldsymbol{\Theta}), \tag{2.7}$$

where the penalization term $Q_\lambda(\boldsymbol{\Theta})$ is a function of the $\boldsymbol{\Theta}$ and λ. Generally, various methods are differentiated by the form of the function $Q_\lambda(\cdot)$, e.g., L_1 penalty function (Pan and Shen, 2007), L_2 penalty function (Xie et al., 2008), L_∞ penalty function (Wang and Zhu, 2008), and pairwise fusion penalty function (Guo et al., 2010).

Up to date, most penalized model-based clustering frameworks are based on the Gaussian mixture model and developed for investigating the group-level heterogeneity in genomics data (e.g., microarray data in Wang and Zhu (2008) and gene expression data in Guo et al. (2010)). However, compared to genomics data, the imaging data usually presents in more complex structures. For example, the contour data of some brain regions of interest after certain transformations can be treated as samples from the shape space (Srivastava et al., 2005). Another potential issue is the modeling of spatial correlation structure. For example, to reduce the dimension of parameter space, the covariance matrix in each component distribution is assumed to be common diagonal matrix (Pan and Shen, 2007). However, the spatial correlation is ignored under this assumption. To address these challenges, Huang et al. (2015) developed a penalized model-based clustering framework to cluster landmark-based planar shape data. Specifically, a mixture of offset-normal shape factor analyzers (MOSFA) is proposed with mixing proportions defined through a regression model (e.g., logistic) and an offset-normal shape distribution in each component for data in the curved shape space. A latent factor analysis model is introduced to explicitly model the complex spatial correlation. A penalized likelihood approach with both adaptive pairwise fusion Lasso penalty function and L_2 penalty function is used to automatically realize variable selection via thresholding and deliver a sparse solution.

2.3 Surrogate Variable Analysis for Multivariate Functional Responses

Surrogate variable analysis (or latent effect adjustment, confounder adjustment), proposed to tackle this study level heterogeneity, has been widely used in genomic studies. (Leek and Storey, 2007; Wang et al., 2017; Lee et al., 2017). Several existing methods, including EIGENSTRAT (Price et al., 2006), Surrogate Variable Analysis (SVA, Leek and Storey (2007, 2008); Lee et al. (2017)), Latent Effect Adjustment after Primary Projection (Sun et al., 2012), Remove Unwanted Variation (RUV, Gagnon-Bartsch et al. (2013)), and Confounder Adjusted Testing and Estimation (CATE, Wang et al. (2017)) were previously proposed to estimate unknown covariates based on the assumption that massive univariate regression models share a common set of unknown covariates.

Suppose that $\boldsymbol{Y}$ is an $n \times m$ matrix of measured features, where m is the number of features and n is the number of samples. For neuroimaging data, $\boldsymbol{Y}$ represents imaging measurements on m voxels. Further, suppose that $\boldsymbol{X}$ is an $n \times p$ matrix of observed covariates, including an intercept, and $\boldsymbol{Z}$ is an $n \times q$ matrix of unobserved hidden factors. The following model represents the true relationship between $\boldsymbol{Y}$ and $(\boldsymbol{X}, \boldsymbol{Z})$:

$$\boldsymbol{y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{Z}\boldsymbol{\delta}_j + \boldsymbol{\epsilon}_j, \tag{2.8}$$

where $\boldsymbol{y}_j$ demotes the j-th column of $\boldsymbol{Y}$, $\boldsymbol{\beta}_j$ is a $p \times 1$ vector of regression coefficients associated with $\boldsymbol{X}$, $\boldsymbol{\delta}_j$ is a $q \times 1$ vector of regression coefficients associated with $\boldsymbol{Z}$, and $\boldsymbol{\epsilon}_j$ is an $n \times 1$ random vector which follows $N(0, \sigma_j^2 \boldsymbol{I})$. In this model, $\boldsymbol{\beta}_j$ and $\boldsymbol{\delta}_j$ are assumed to be fixed and unknown.

To identify hidden factors $\boldsymbol{Z}$, principal component analysis (PCA) on the original or residualized features after removing the effects of observed dependent variables has often been used (Price et al., 2006). However, PCA based approaches are less effective for gene expression studies, where the hidden factors can affect a subset of features with relatively large effects (Leek and Storey, 2007). To overcome this limitation, surrogate variable analysis has been proposed. In particular, Leek and Storey (2007) initially developed a two-step approach which involves first identifying a subset of features that may be affected by hidden factors but not by primary variables, and then performing principal component analysis on the selected features. Later, they modified the approach to a weighted PCA, where each feature is weighted according to its probability of being affected by the hidden factors only (Leek and Storey, 2008). Surrogate variable analysis has also been extended to factor analysis (Friguet et al., 2009) and mixed-effect models (Listgarten et al., 2010). However, strong correlation between hidden factors and primary variables can prevent the two-step and weighted principal component based surrogate variable methods from identifying features that are affected by hidden factors only. To address this issue, recently a direct surrogate variable analysis (dSVA) was proposed in Lee et al. (2017). dSVA is based on the observation that naive estimators of the effects of the primary variables are biased when the effects of hidden factors are ignored in the analysis, but the bias can be estimated and removed using singular value decomposition (SVD) on residuals.

However, instead of this mass-univariate analysis, the image measures across different voxels are more likely to be treated as a single functional response, where the three key features of the functional data can be explicitly accounted for : spatial smoothness, spatial correlation, and the low-dimensional representation. Therefore, it is of great importance to investigate the study-level heterogeneity in some functional regression models.

CHAPTER 3: DYNAMIC DISEASED REGION DETECTION FOR LONGITUDINAL MEDICAL IMAGING DATA

3.1 Method

Suppose that we observe a longitudinal imaging dataset for n unrelated subjects. Let $\boldsymbol{S}_0 \subset \mathbb{R}^d$, $d = 2, 3$, be a common template from the dataset, and $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{n_v}$ be the set of n_v voxels in $\mathcal{S}$. The longitudinal imaging measurements for the i-th subject at the voxel $\boldsymbol{s}$ are denoted as $\boldsymbol{y}_i(\boldsymbol{s}) = (y_{i,1}(\boldsymbol{s}), \ldots, y_{i,m_i}(\boldsymbol{s}))^T$, $i = 1, \ldots, n$. Furthermore, let $\boldsymbol{x}_i = (\boldsymbol{x}_{i,1}, \ldots, \boldsymbol{x}_{i,m_i})$ and $\boldsymbol{x}_{i,j} = (\boldsymbol{w}_{i,j}^T, \boldsymbol{z}_{i,j}^T)^T$, where $\boldsymbol{w}_{i,j}$ is a $(p-1) \times 1$ vector including the intercept, demographic and clinical covariates (e.g., gender, age, or treatment), and $\boldsymbol{z}_{i,j}$ includes dummy variables indicating the diagnostic status at the j-th time point. In particular, we assume that there are 3 diagnostic results, i.e., normal stage ($\boldsymbol{z}_{i,j} = (0,0)^T$), early stage of disease ($\boldsymbol{z}_{i,j} = (1,0)^T$), and late stage of disease ($\boldsymbol{z}_{i,j} = (0,1)^T$). In addition, for the i-th subject at the j-th time point, we assume that $\boldsymbol{S}_0$ can be decomposed into the union of normal region $\mathcal{R}_{i,j}^0$ and diseased region $\mathcal{R}_{i,j}^1$, that is

$$\boldsymbol{S}_0 = \mathcal{R}_{i,j}^0 \cup \mathcal{R}_{i,j}^1 \quad \text{and} \quad \mathcal{R}_{i,j}^0 \cap \mathcal{R}_{i,j}^1 = \emptyset.$$

Here we also assume that: (i) subjects at the normal stage are expected to be perfectly healthy, i.e. do not have any diseased regions; (ii) for subjects with certain stage of disease, number, shape, size, and location of diseased regions $\mathcal{R}_{i,j}^1$ may vary across **subjects** and **time points**. To further illustrate the assumptions, an example including latent diseased regions for 3 subjects is presented in Figure 3.1. Subject P1 (top) is at the normal stage for the first two time points while at the early stage of disease for the follow-up two time points.

One diseased region (red) occurs at the third time point and becomes larger at the forth time point. Subject P2 (middle) is at the early stage of disease at baseline with one diseased region detected. After that, the diseased regions grow in size and number, and the diagnostic status changes to the late stage. For subject P3 (bottom), the diagnostic status is the early stage of disease at the first three time points while changes to the late stage at the forth time point. In this example, following the assumptions, subjects at the normal stage don't have any diseased regions, while the number, shape, size, and location of diseased regions are different across subjects and time points.

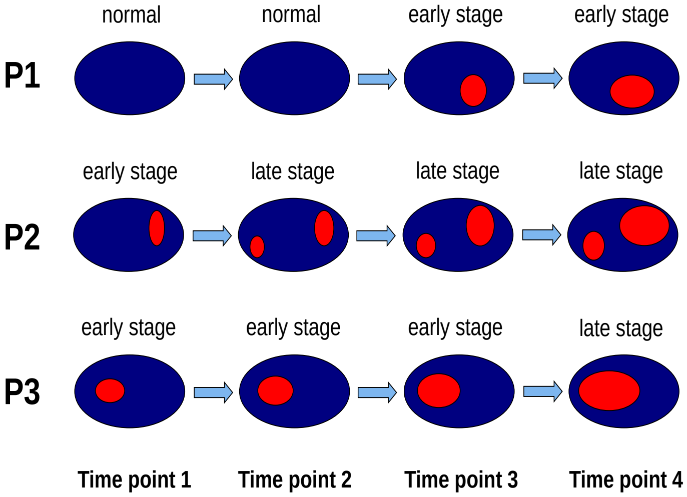


Figure 3.1: An example showing assumptions of diseased regions for 3 subjects.

3.1.1 Dynamic Spatial Random Effects Model

Our dynamic spatial random effects (DSRE) model consists of a spatial random effects (SRE) model (Besag, 1974; Geman and Geman, 1984; Diggle and Ribeiro, 2007; Li and Singh, 2009; Huang et al., 2015) and a dynamic conditional random field (DCRF) model (Wang and Ji, 2005; Wang et al., 2006; Sutton et al., 2007; Yin et al., 2009).

First, SRE model is considered to characterize the conditional distribution of the observed imaging measurements given two sets of random effects, i.e., $\{\boldsymbol{b}_i(\boldsymbol{s})\}_{i=1}^n$ and $\{\boldsymbol{\gamma}_i(\boldsymbol{s})\}_{i=1}^n$. In particular, $\boldsymbol{b}_i(\boldsymbol{s}) = (b_{i,1}(\boldsymbol{s}), \ldots, b_{i,m_i}(\boldsymbol{s}))^T$, where $b_{i,j}(\boldsymbol{s}) = 0$ if $\boldsymbol{s} \in \mathcal{R}_{i,j}^0$, otherwise $b_{i,j}(\boldsymbol{s}) = 1$. The other random effect $\boldsymbol{\gamma}_i(\boldsymbol{s})$ is a $p \times 1$ vector indicating the subject-specific random effect.

Given $b_{i,j}(\boldsymbol{s})$ and $\boldsymbol{\gamma}_i(\boldsymbol{s})$, the SRE model is given as

$$y_{i,j}(\boldsymbol{s}) = \boldsymbol{x}_{i,j}^T\boldsymbol{\beta}(\boldsymbol{s}) + \boldsymbol{x}_{i,j}^T\boldsymbol{\gamma}_i(\boldsymbol{s}) + b_{i,j}(\boldsymbol{s})\boldsymbol{x}_{i,j}^T\boldsymbol{\alpha}(\boldsymbol{s}) + \epsilon_{i,j}(\boldsymbol{s}), \tag{3.1}$$

where $\boldsymbol{\beta}(\boldsymbol{s})$ is a $p \times 1$ vector representing the fixed effect at voxel $\boldsymbol{s}$, while $\boldsymbol{\alpha}(\boldsymbol{s})$ is a $p \times 1$ vector representing the additional effect caused by the diseased regions. Denote that $\boldsymbol{\epsilon}_i(\boldsymbol{s}) = (\epsilon_{i,1}(\boldsymbol{s}), \dots, \epsilon_{i,m_i}(\boldsymbol{s}))^T, i = 1, \dots, n$, and $\{\boldsymbol{\epsilon}_i(\boldsymbol{s})\}_{i=1}^n$ are independent measurement errors across subjects and voxels, following the Gaussian distribution $N(0, \sigma^2(\boldsymbol{s})\boldsymbol{I}_{m_i})$. Based on the proposed model, the potential heterogeneity (among different **voxels**, **subjects**, and **time points**) are mainly captured by the term "$b_{i,j}(\boldsymbol{s})\boldsymbol{x}_{i,j}^T\boldsymbol{\alpha}(\boldsymbol{s})$". Additionally, for voxels in normal regions, SRE model (3.1) can be simplified into a voxel-wised linear mixed model:

$$y_{i,j}(\boldsymbol{s}) = \boldsymbol{x}_{i,j}^T\boldsymbol{\beta}(\boldsymbol{s}) + \boldsymbol{x}_{i,j}^T\boldsymbol{\gamma}_i(\boldsymbol{s}) + \epsilon_{i,j}(\boldsymbol{s}). \tag{3.2}$$

Then, we model the random effects $\boldsymbol{\gamma}_i(\boldsymbol{s})$ and $\boldsymbol{b}_i(\boldsymbol{s})$ as follows. First, it is assumed that $\boldsymbol{\gamma}_i(\boldsymbol{s})$, $\boldsymbol{b}_i(\boldsymbol{s})$ and $\boldsymbol{\epsilon}_i(\boldsymbol{s})$ are mutually independent. Second, $\{\boldsymbol{\gamma}_i(\boldsymbol{s}), \boldsymbol{s} \in \boldsymbol{S}_0\}_{i=1}^n$ are assumed to be mutually independent across subjects and voxels, following $N(\boldsymbol{0}, \boldsymbol{\Sigma}(\boldsymbol{s}))$. Moreover, to formulate both spatial and temporal dependencies of consecutive riseased regions, $\{\boldsymbol{b}_i\}_{i=1}^n$ are assumed independent across subjects and each $\boldsymbol{b}_i = \{\boldsymbol{b}_i(\boldsymbol{s}), \boldsymbol{s} \in \boldsymbol{S}_0\}$ follows a DCRF model:

$$p(\boldsymbol{b}_i|\tau, \eta) \quad \propto \quad p(\boldsymbol{b}_{i,j_{i0}}|\tau) \prod_{j=j_{i0}+1}^{m_i} p(\boldsymbol{b}_{i,j}|\boldsymbol{b}_{i,j-1}, \tau, \eta), \tag{3.3}$$

where j_{i0} is the disease baseline for the i-*th* patient. For diseased regions at disease baseline,

$$p(\boldsymbol{b}_{i,j_{i0}}|\tau) \quad = \quad \exp\Big\{ -\tau \sum_{s \in S_0} \sum_{s' \in N_s} U(b_{i,j_{i0}}(\boldsymbol{s}), b_{i,j_{i0}}(\boldsymbol{s}'))\Big\}, \tag{3.4}$$

while for diseased regions at follow-up visits,

$$\begin{aligned} p(\boldsymbol{b}_{i,j}|\boldsymbol{b}_{i,j-1},\tau,\eta) &= \exp\Big\{-\sum_{s\in S_0}\big[\tau\sum_{s'\in N_s}U(b_{i,j}(\boldsymbol{s}),b_{i,j}(\boldsymbol{s}')) \\ &\quad +\eta\sum_{s'\in M_s}U(b_{i,j}(\boldsymbol{s}),b_{i,j-1}(\boldsymbol{s}'))\big]\Big\}, \qquad (3.5) \\ U(b_{i,j}(\boldsymbol{s}),b_{i,j'}(\boldsymbol{s}')) &= \frac{1-\delta(b_{i,j}(\boldsymbol{s}),b_{i,j'}(\boldsymbol{s}'))}{||\boldsymbol{s}-\boldsymbol{s}'||^2+1}. \end{aligned}$$

Here $||\cdot||$ denotes the Euclidean distance and $\delta(\cdot)$ is the Kronecker delta function. Thus, two neighboring voxels are more likely to belong to the same region than to different ones. Both the spatial and temporal constraints become strong with decreasing distance between the neighboring voxels. τ is introduced to encourage spatial smoothness in homogeneous regions while η influences the strength of temporal dependencies. Moreover, both N_s and M_s denote the neighboring voxels of $\boldsymbol{s}$. It should be noted that M_s is not equivalent to the neighborhood N_s: (i) M_s and N_s may have different sizes, and (ii) $\boldsymbol{s} \notin N_s$ while $\boldsymbol{s} \in M_s$. To distinguish them, N_s is called the spatial neighborhood and M_s the temporal neighborhood. Throughout the paper, we consider N_s is the set of the closest $3^d - 1$ neighbors of pixel $\boldsymbol{s}$, while $M_s = N_s \bigcup\{\boldsymbol{s}\}$. Further illustrations of DSRE model and DCRF model are presented in Figure 3.2.

3.1.2 Estimation Procedure

Our next task is to estimate the random effects $\{\boldsymbol{b}_i\}_{i=1}^n$ and all unknown parameters consisting of $\tau, \eta, \boldsymbol{\beta}(\boldsymbol{s}), \boldsymbol{\alpha}(\boldsymbol{s}), \sigma^2(\boldsymbol{s})$, and $\boldsymbol{\Sigma}(\boldsymbol{s})$ for $\boldsymbol{s} \in \boldsymbol{S}_0$. We decompose these parameters into three parts: (i) $\boldsymbol{\beta}(\boldsymbol{s}), \sigma^2(\boldsymbol{s}), \boldsymbol{\Sigma}(\boldsymbol{s})$, (ii) $\boldsymbol{\alpha}(\boldsymbol{s})$, and (iii) τ, η. For parts (i) and (ii), the maximum likelihood estimate (MLE) can be calculated by using the expectation-maximization (EM) algorithm (Huang et al., 2015). In particular, the MLEs of $\boldsymbol{\beta}(\boldsymbol{s}), \sigma^2(\boldsymbol{s}), \boldsymbol{\Sigma}(\boldsymbol{s})$ can be derived based on only the normal controls for computational efficiency, while the MLE of $\boldsymbol{\alpha}(\boldsymbol{s})$ can be derived based on a subpopulation, only including the patients. For part (iii), τ and η can be predefined or dertermined by some data-driven method. In this paper, they are

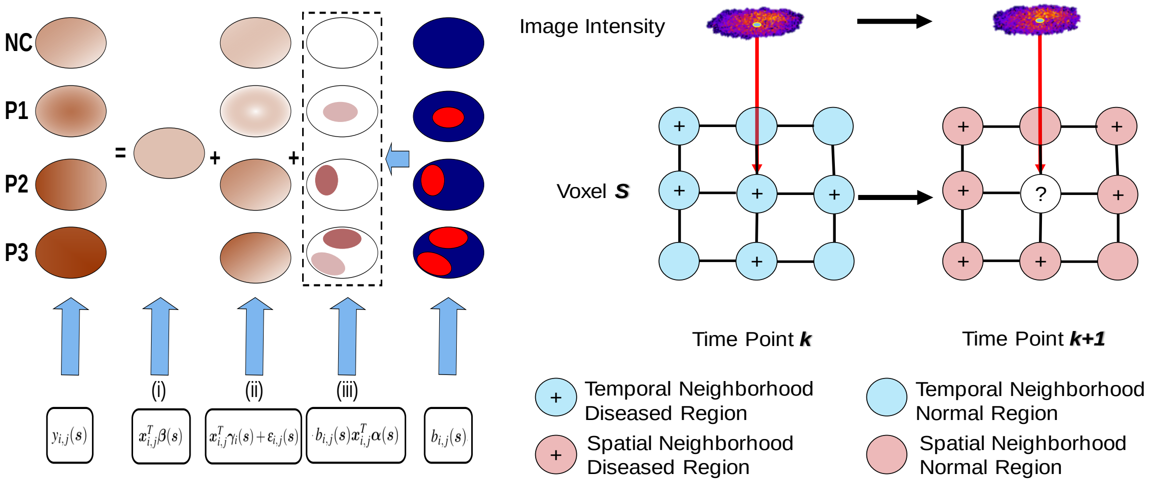

NC
P1
P2
P3
=
+
+
(i)
(ii)
(iii)
$y_{i,j}(\boldsymbol{s})$
$\boldsymbol{x}_{i,j}^T\boldsymbol{\beta}(\boldsymbol{s})$
$\boldsymbol{x}_{i,j}^T\boldsymbol{\gamma}_i(\boldsymbol{s})+\varepsilon_{i,j}(\boldsymbol{s})$
$b_{i,j}(\boldsymbol{s})\boldsymbol{x}_{i,j}^T\boldsymbol{\alpha}(\boldsymbol{s})$
$b_{i,j}(\boldsymbol{s})$
Image Intensity
Voxel ***s***
?
Time Point ***k***
Time Point ***k+1***
Temporal Neighborhood Diseased Region
Temporal Neighborhood Normal Region
Spatial Neighborhood Diseased Region
Spatial Neighborhood Normal Region

Figure 3.2: Illustrations of DSRE model and DCRF model. (left) Path diagram of DSRE model for four different subjects (one normal control and three subjects); (right) Mechanism of DCRF model.

estimated by using a pseudo-likelihood method (Geman and Graffigne, 1986) since the MLEs of τ and η are generally difficult to compute due to the normalizing part of the probability function in (3.3). In addition, the random effects $\{\boldsymbol{b}_i\}_{i=1}^n$ can be estimated via the MRF-MAP method.

EM algorithm for parameters in parts (i) and (ii) To derive the EM algorithm for parameters in part (i), we need to derive the complete-data log-likelihood function on the normal controls as follows. Recall that the distribution of $\boldsymbol{y}_i(s)$ conditional on $\boldsymbol{\gamma}_i(\boldsymbol{s})$ is given by $N(\boldsymbol{x}_i^T(\boldsymbol{\beta}(\boldsymbol{s})+\boldsymbol{\gamma}_i(\boldsymbol{s})), \sigma^2(\boldsymbol{s})\boldsymbol{I}_{m_i})$. Let $\boldsymbol{\mu}_i(\boldsymbol{s}) = \boldsymbol{y}_i(\boldsymbol{s}) - \boldsymbol{x}_i^T(\boldsymbol{\beta}(\boldsymbol{s})+\boldsymbol{\gamma}_i(\boldsymbol{s}))$, and the complete-data log-likelihood function is given by

$$\begin{aligned}\log L_0 \;\propto\;& -\frac{\sum_{i=1}^{n_0} m_i}{2}\sum_{l=1}^{n_v}\log(\sigma^2(\boldsymbol{s}_l)) - \frac{n_0}{2}\sum_{l=1}^{n_v}\log|\boldsymbol{\Sigma}(\boldsymbol{s}_l)| \\ & -\sum_{l=1}^{n_v}\frac{1}{2\sigma^2(\boldsymbol{s}_l)}\sum_{i=1}^{n_0}\boldsymbol{\mu}_i^T(\boldsymbol{s}_l)\boldsymbol{\mu}_i(\boldsymbol{s}_l) - \frac{1}{2}\sum_{l=1}^{n_v}\sum_{i=1}^{n_0}\boldsymbol{\gamma}_i^T(\boldsymbol{s}_l)\boldsymbol{\Sigma}^{-1}(\boldsymbol{s}_l)\boldsymbol{\gamma}_i(\boldsymbol{s}_l). \qquad (3.6)\end{aligned}$$

Given the current estimate of $\boldsymbol{\theta}_0 = \{\boldsymbol{\beta}(\boldsymbol{s}), \sigma^2(\boldsymbol{s}), \boldsymbol{\Sigma}(\boldsymbol{s})\}$ at iteration r, denoted as $\hat{\boldsymbol{\theta}}_0^{(r)}$, their updates are obtained via maximizing the following Q-function $Q_{\hat{\theta}^{(r)}}(\boldsymbol{\theta}_0) \doteq E_{\hat{\theta}_0^{(r)}}(\log L_0|\boldsymbol{y}, \boldsymbol{x})$

with respect to $\boldsymbol{\theta}_0$:

$$\begin{aligned} Q_{\hat{\theta}_0^{(r)}}(\boldsymbol{\theta}_0) \propto & -\frac{\sum_{i=1}^{n_0} m_i}{2} \sum_{l=1}^{n_v} \log(\sigma^2(\boldsymbol{s}_l)) - \frac{n_0}{2} \sum_{l=1}^{n_v} \log|\boldsymbol{\Sigma}(\boldsymbol{s}_l)| \\ & - \sum_{l=1}^{n_v} \frac{1}{2\sigma^2(\boldsymbol{s}_l)} \sum_{i=1}^{n_0} E\Big[\boldsymbol{\mu}_i^T(\boldsymbol{s}_l)\boldsymbol{\mu}_i(\boldsymbol{s}_l)\Big|\boldsymbol{y}_i(\boldsymbol{s}_l), \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_0^{(r)}\Big] \\ & - \frac{1}{2} \sum_{l=1}^{n_v} \sum_{i=1}^{n_0} E\Big[\boldsymbol{\gamma}_i^T(\boldsymbol{s}_l)\boldsymbol{\Sigma}^{-1}(\boldsymbol{s}_l)\boldsymbol{\gamma}_i(\boldsymbol{s}_l)\Big|\boldsymbol{y}_i(\boldsymbol{s}_l), \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_0^{(r)}\Big]. \end{aligned} \tag{3.7}$$

We consider the E-step and M-step of the EM algorithm as follows.

E-step: In the E-step, we need to calculate two conditional expectations:

$$E\big[\boldsymbol{\gamma}_i(\boldsymbol{s})\big|\boldsymbol{y}_i(\boldsymbol{s}), \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_0^{(r)}\big] \text{ and } E\big[\boldsymbol{\gamma}_i(\boldsymbol{s})\boldsymbol{\gamma}_i^T(\boldsymbol{s})\big|\boldsymbol{y}_i(\boldsymbol{s}), \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_0^{(r)}\big].$$

Recall that, given $\boldsymbol{x}_i$, $(\boldsymbol{y}_i^T(\boldsymbol{s}), \boldsymbol{\gamma}_i^T(\boldsymbol{s}))^T$ is normally distributed as

$$\begin{pmatrix} \boldsymbol{y}_i(\boldsymbol{s}) \\ \gamma_i(\boldsymbol{s}) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{x}_i^T\boldsymbol{\beta}(\boldsymbol{s}) \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{x}_i^T\boldsymbol{\Sigma}(\boldsymbol{s})\boldsymbol{x}_i + \sigma^2(\boldsymbol{s})\boldsymbol{I}_{m_i} & \boldsymbol{x}_i^T\boldsymbol{\Sigma}(\boldsymbol{s}) \\ \boldsymbol{\Sigma}(\boldsymbol{s})\boldsymbol{x}_i & \boldsymbol{\Sigma}(\boldsymbol{s}) \end{pmatrix} \right).$$

Then, given $\boldsymbol{y}_i(\boldsymbol{s})$ and $\boldsymbol{x}_i$, we have

$$E\Big[\boldsymbol{\gamma}_i(\boldsymbol{s})\Big|\boldsymbol{y}_i(\boldsymbol{s}), \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_0^{(r)}\Big] = \hat{\boldsymbol{\Sigma}}^{(r)}(\boldsymbol{s})\boldsymbol{x}_i(\boldsymbol{x}_i^T\hat{\boldsymbol{\Sigma}}^{(r)}(\boldsymbol{s})\boldsymbol{x}_i + \hat{\sigma}^{2^{(r)}}(\boldsymbol{s})\boldsymbol{I}_{m_i})^{-1}\big(\boldsymbol{y}_i(\boldsymbol{s}) - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}^{(r)}(\boldsymbol{s})\big),$$

$$Var\Big[\boldsymbol{\gamma}_i(\boldsymbol{s})\Big|\boldsymbol{y}_i(\boldsymbol{s}), \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_0^{(r)}\Big] = \hat{\boldsymbol{\Sigma}}^{(r)}(\boldsymbol{s}) - \hat{\boldsymbol{\Sigma}}^{(r)}(\boldsymbol{s})\boldsymbol{x}_i(\boldsymbol{x}_i^T\hat{\boldsymbol{\Sigma}}^{(r)}(\boldsymbol{s})\boldsymbol{x}_i + \hat{\sigma}^{2^{(r)}}(\boldsymbol{s})\boldsymbol{I}_{m_i})^{-1}\boldsymbol{x}_i^T\hat{\boldsymbol{\Sigma}}^{(r)}(\boldsymbol{s}),$$

$$E\Big[\boldsymbol{\gamma}_i(\boldsymbol{s})\boldsymbol{\gamma}_i^T(\boldsymbol{s})\Big|\boldsymbol{y}_i(\boldsymbol{s}), \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_0^{(r)}\Big] = Var\Big[\boldsymbol{\gamma}_i(\boldsymbol{s})\Big|\boldsymbol{y}_i(\boldsymbol{s}), \boldsymbol{x}_i, \boldsymbol{\theta}_0^{(r)}\Big] + E\Big[\boldsymbol{\gamma}_i(\boldsymbol{s})\Big|\boldsymbol{y}_i(\boldsymbol{s}), \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_0^{(r)}\Big]^{\otimes 2}.$$

M-step: Taking derivatives of (3.7) with respect to $\boldsymbol{\theta}_0$ and equating them to zeros, we find the updates of $\hat{\boldsymbol{\theta}}_0^{(r)}$ as follows. For $\boldsymbol{\beta}(\boldsymbol{s})$, we have

$$\hat{\boldsymbol{\beta}}^{(r+1)}(\boldsymbol{s}) = \Big[\sum_{i=1}^{n_0} \boldsymbol{x}_i(\boldsymbol{s})\boldsymbol{x}_i^T\Big]^{-1} \sum_{i=1}^{n_0} \boldsymbol{x}_i \Big(\boldsymbol{y}_i(\boldsymbol{s}) - \boldsymbol{x}_i^T E\Big[\boldsymbol{\gamma}_i(\boldsymbol{s})\Big|\boldsymbol{y}_i(\boldsymbol{s}), \boldsymbol{x}_i, \boldsymbol{\theta}_0^{(r)}\Big]\Big]\Big). \tag{3.8}$$

For the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{s})$, we have

$$\hat{\boldsymbol{\Sigma}}^{(r+1)}(\boldsymbol{s}) = \frac{1}{n_0}\sum_{i=1}^{n_0} E\Big[\boldsymbol{\gamma}_i(\boldsymbol{s})\boldsymbol{\gamma}_i^T(\boldsymbol{s})\Big|\boldsymbol{y}_i(\boldsymbol{s}), \boldsymbol{x}_i, \boldsymbol{\theta}_0^{(r)}\Big]. \tag{3.9}$$

For $\sigma^2(\boldsymbol{s})$,

$$\hat{\sigma}^{2^{(r+1)}}(\boldsymbol{s}) = \frac{1}{\sum_{i=1}^{n_0} m_i}\sum_{i=1}^{n_0} E\Big[\boldsymbol{\mu}_i^T(\boldsymbol{s})\boldsymbol{\mu}_i(\boldsymbol{s})\Big|\boldsymbol{y}_i(\boldsymbol{s}), \boldsymbol{x}_i, \boldsymbol{\theta}_0^{(r)}\Big]. \tag{3.10}$$

The E-step and M-step are alternately repeated until the difference between $\log L_0(\hat{\boldsymbol{\theta}}_0^{(r)})$ and $\log L_0(\hat{\boldsymbol{\theta}}_0^{(r+1)})$ is smaller than a desired value (e.g., 10^{-4}).

To derive the EM algorithm for parameters in part (ii), we need to derive the complete-data log-likelihood function on the patents as follows.

$$\log L_1 \propto -\sum_{l=1}^{n_v} \frac{1}{2\sigma^2(\boldsymbol{s}_l)} \sum_{i=n_0+1}^{n}\sum_{j=j_{i0}}^{m_i} \nu_{i,j}^2(\boldsymbol{s}_l), \tag{3.11}$$

where $\nu_{i,j}(\boldsymbol{s}) = y_{i,j}(\boldsymbol{s}) - \boldsymbol{x}_{i,j}^T(\boldsymbol{\beta}(\boldsymbol{s}) + \boldsymbol{\gamma}_i(\boldsymbol{s})) - b_{i,j}(\boldsymbol{s})\boldsymbol{x}_{i,j}^T\boldsymbol{\alpha}(\boldsymbol{s})$. Given the estimate of $\boldsymbol{\theta}_0$, i.e., $\hat{\boldsymbol{\theta}}_0$, the update of $\hat{\boldsymbol{\alpha}}(\boldsymbol{s})^{(r)}$ at iteration $r+1$ is

$$\begin{aligned}
\hat{\boldsymbol{\alpha}}(\boldsymbol{s})^{(r+1)} \quad = \quad & \Big[\sum_{i=n_0+1}^{n}\sum_{j=j_{i0}}^{m_i} E\Big[b_{i,j}(\boldsymbol{s})\Big|\boldsymbol{y}_i(\boldsymbol{s}), \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_0\Big]\boldsymbol{x}_{i,j}\boldsymbol{x}_{i,j}^T\Big]^{-1} \\
& \sum_{i=n_0+1}^{n}\sum_{j=j_{i0}}^{m_i} \Big\{E\Big[b_{i,j}(\boldsymbol{s})\Big|\boldsymbol{y}_i(\boldsymbol{s}), \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_0\Big]\Big[y_{i,j}(\boldsymbol{s}) - \boldsymbol{x}_{i,j}^T\hat{\boldsymbol{\beta}}(\boldsymbol{s})\Big]\boldsymbol{x}_{i,j} \\
& -\boldsymbol{x}_{i,j}^T E\Big[\boldsymbol{\gamma}_i(\boldsymbol{s})b_{i,j}(\boldsymbol{s})\Big|\boldsymbol{y}_i(\boldsymbol{s}), \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_0\Big]\boldsymbol{x}_{i,j}\Big\}.
\end{aligned} \tag{3.12}$$

In order to calculate $E\Big[b_{i,j}(\boldsymbol{s})\Big|\boldsymbol{y}_i(\boldsymbol{s}), \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_0\Big]$ and $E\Big[\boldsymbol{\gamma}_i(\boldsymbol{s})b_{i,j}(\boldsymbol{s})\Big|\boldsymbol{y}_i(\boldsymbol{s}), \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_0\Big]$ in (3.12), the class labels $\boldsymbol{b}_i$, τ and η should be estimated first. Here we consider the MRF-MAP estimation for $\boldsymbol{b}_i$, which is efficient and commonly adopted in existing literature, e.g., Zhang et al. (2001); Marroquín et al. (2002); Nie et al. (2009). For the MLEs of τ and η, the pseudo-likelihood method (Geman and Graffigne, 1986) is considered. The detialed derivation of these two estimation parts will be discussed later in next subsections.

Assumed that we have the MRF-MAP estimate of $\boldsymbol{b}_i$ and the estimates of τ, η at iteration r, i.e., $\hat{\boldsymbol{b}}_i^{(r)}, \hat{\tau}^{(r)}, \hat{\eta}^{(r)}$, the conditional expectation $E\big[b_{i,j}(\boldsymbol{s})\big|\boldsymbol{y}_i(\boldsymbol{s}), \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_0\big]$ can be calculated as

$$\frac{f(\boldsymbol{y}_i(\boldsymbol{s})|x_{i,j}, b_{i,j}(\boldsymbol{s}) = 1, \hat{\boldsymbol{b}}_i^{(r)}, \hat{\boldsymbol{\theta}}_0)P(b_{i,j}(\boldsymbol{s}) = 1|\hat{\boldsymbol{b}}_i^{(r)}, \hat{\boldsymbol{\theta}}_0, \hat{\tau}^{(r)}, \hat{\eta}^{(r)})}{\sum_{t=0}^{1} f(\boldsymbol{y}_i(\boldsymbol{s})|x_{i,j}, b_{i,j}(\boldsymbol{s}) = t, \hat{\boldsymbol{b}}_i^{(r)}, \hat{\boldsymbol{\theta}}_0)P(b_{i,j}(\boldsymbol{s}) = t|\hat{\boldsymbol{b}}_i^{(r)}, \hat{\boldsymbol{\theta}}_0, \hat{\tau}^{(r)}, \hat{\eta}^{(r)})}, \tag{3.13}$$

where $\boldsymbol{y}_i(\boldsymbol{s})|\boldsymbol{x}_i, \boldsymbol{b}_i(\boldsymbol{s}), \hat{\boldsymbol{\theta}}_0 \sim \mathcal{N}(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}(\boldsymbol{s}) + \boldsymbol{B}_i(\boldsymbol{s})\boldsymbol{x}_i^T\hat{\boldsymbol{\alpha}}(\boldsymbol{s})^{(r+1)}, \hat{\boldsymbol{\Lambda}}(\boldsymbol{s}))$, $\boldsymbol{B}_i(\boldsymbol{s}) = diag(\boldsymbol{b}_i(\boldsymbol{s}))$, and $\hat{\boldsymbol{\Lambda}}(\boldsymbol{s}) = \boldsymbol{x}_i^T\hat{\boldsymbol{\Sigma}}(\boldsymbol{s})\boldsymbol{x}_i + \hat{\sigma}^2(\boldsymbol{s})\boldsymbol{I}_{m_i}$.

If $j = j_{i0}$,

$$P(b_{i,j}(\boldsymbol{s}) = 1|\hat{\boldsymbol{b}}_i^{(r)}, \boldsymbol{\theta}^{(r)}, \hat{\tau}^{(r)}, \hat{\eta}^{(r)}) \propto \exp\Big\{-\hat{\tau}^{(r)}\sum_{s' \in N_s} U(1, \hat{b}_{i,1}^{(r)}(\boldsymbol{s}'))\Big\},$$

otherwise,

$$P(b_{i,j}(\boldsymbol{s}) = 1|\hat{\boldsymbol{b}}_i^{(r)}, \boldsymbol{\theta}^{(r)}, \hat{\tau}^{(r)}, \hat{\eta}^{(r)}) \propto \exp\Big\{-\hat{\tau}^{(r)}\sum_{s' \in N_s} U(1, \hat{b}_{i,j}^{(r)}(\boldsymbol{s}')) - \hat{\eta}^{(r)}\sum_{s' \in M_s} U(1, \hat{b}_{i,j-1}^{(r)}(\boldsymbol{s}'))\Big\}.$$

Finally, the desired expectation $E\Big[\boldsymbol{\gamma}_i(\boldsymbol{s})b_{i,j}(\boldsymbol{s})\Big|\boldsymbol{y}_i(\boldsymbol{s}), \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_0\Big]$ can be estimated as

$$E\Big[\boldsymbol{\gamma}_i(\boldsymbol{s})b_{i,j}(\boldsymbol{s})\Big|\boldsymbol{y}_i(\boldsymbol{s}), \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_0\Big] = E\Big[\boldsymbol{\gamma}_{i,j}(\boldsymbol{s})\Big|\boldsymbol{y}_i(\boldsymbol{s}), \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_0\Big]P(b_{i,j}(\boldsymbol{s}) = 1|\boldsymbol{y}_i(\boldsymbol{s}), \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_0). \tag{3.14}$$

These two expectations are updated until the difference between $\log L_1(\hat{\boldsymbol{\alpha}}(\boldsymbol{s})^{(r+1)}, \hat{\boldsymbol{\theta}}_0)$ and $\log L_1(\hat{\boldsymbol{\alpha}}(\boldsymbol{s})^{(r)}, \hat{\boldsymbol{\theta}}_0)$ is smaller than a desired value (e.g., 10^{-4}).

MRF-MAP estimation method The MRF-MAP estimation is an efficient method for many practical applications (e.g., image segmentation) and adopted in many literatures, e.g., Zhang et al. (2001); Nie et al. (2009). According to the MAP criterion, given the current estimate $\hat{\boldsymbol{\theta}}_0$, $\hat{\boldsymbol{\alpha}}(\boldsymbol{s})^{(r)}$, $\hat{\tau}^{(r)}$, and $\hat{\eta}^{(r)}$ at iteration r, the estimate of $\boldsymbol{b}_i$ is updated as

$$\begin{aligned}
\hat{\boldsymbol{b}}_i^{(r+1)} &= \arg\max_{\boldsymbol{b}_i} \Big\{ \prod_{l=1}^{n_v} f(\boldsymbol{y}_i(\boldsymbol{s}_l)|\boldsymbol{x}_i, b_i(\boldsymbol{s}_l), \hat{\boldsymbol{\theta}}_0) p(\boldsymbol{b}_i|\hat{\tau}^{(r)}, \hat{\eta}^{(r)}) \Big\} \\
&= \arg\min_{\boldsymbol{b}_i} \Big\{ \frac{1}{2} \sum_{l=1}^{n_v} \Big\{ \sum_{s \in S_0} [\boldsymbol{y}_i(\boldsymbol{s}_l) - \boldsymbol{\nu}_i^{(r)}(\boldsymbol{s}_l)]^T \hat{\boldsymbol{\Lambda}}_i(\boldsymbol{s}_l)^{-1} [\boldsymbol{y}_i(\boldsymbol{s}_l) - \boldsymbol{\nu}_i^{(r)}(\boldsymbol{s}_l)] \\
&\quad + \hat{\tau}^{(r)} \sum_{s \in S_0} \sum_{s' \in N_s} U(b_{i,j_{i0}}(\boldsymbol{s}), b_{i,j_{i0}}(\boldsymbol{s}')) + \sum_{j=j_{i0}}^{m_i} \sum_{s \in S_0} \big[\hat{\tau}^{(r)} \sum_{s' \in N_s} U(b_{i,j}(\boldsymbol{s}), b_{i,j}(\boldsymbol{s}')) \\
&\quad + \hat{\eta}^{(r)} \sum_{s' \in M_s} U(b_{i,j}(\boldsymbol{s}), b_{i,j-1}(\boldsymbol{s}')) \big] \Big\} \Big\}. \qquad (3.15)
\end{aligned}$$

To obtain the optimal solution to (3.15), in this paper, we adopt the iterated conditional modes (ICM) algorithm (Besag, 1986), which uses a greedy iterative strategy for minimization. Convergence is achieved after only a few iterations.

Pseudo-likelihood method Since τ and η in model (3.3) are not the primary parameter of interest, we use an approximate, but computationally efficient method based on a pseudo-likelihood function. A key advantage of using the pseudo-likelihood function is its computational simplicity, since it does not involve the intractable partition function. The pseudo-likelihood at iteration r is a simple product of the conditional likelihood

$$PL(\hat{\boldsymbol{b}}^{(r)}, \tau, \eta) = \prod_{\{i: z_i = 1\}} \prod_{\boldsymbol{s} \in \boldsymbol{S}_0 - \partial \boldsymbol{S}_0} PL(\hat{\boldsymbol{b}}_i^{(r)}(\boldsymbol{s})|\hat{\boldsymbol{b}}_i^{(r)}), \qquad (3.16)$$

where $\partial \boldsymbol{S}_0$ denotes the set of points at the boundaries of $\boldsymbol{S}_0$, and $PL(\hat{\boldsymbol{b}}_i^{(r)}(\boldsymbol{s})|\hat{\boldsymbol{b}}_i^{(r)})$ is given by

$$\frac{p(\hat{\boldsymbol{b}}_i^{(r)}(\boldsymbol{s})|\tau,\eta)}{\sum_{b_{i,j_{i0}}(s)=0}^{1} \cdots \sum_{b_{i,m_i}(s)=0}^{1} p(\boldsymbol{b}_i(\boldsymbol{s})|\tau,\eta)}.$$

Thus, the MPL estimates $\hat{\tau}^{(r+1)}$ and $\hat{\eta}^{(r+1)}$ can be obtained by solving

$$\frac{\partial \ln PL(\hat{\boldsymbol{b}}^{(r)},\tau,\eta)}{\partial \tau} = 0, \quad \frac{\partial \ln PL(\hat{\boldsymbol{b}}^{(r)},\tau,\eta)}{\partial \eta} = 0. \tag{3.17}$$

3.1.3 Inference Procedure

After all the parameters are estimated, we carry out formal statistical inference consisting of three different statistical tools: (1) standard errors of $\hat{\boldsymbol{\beta}}(\boldsymbol{s})$ and $\hat{\boldsymbol{\alpha}}(\boldsymbol{s})$; (2) hypothesis testing on parameters of interest; and (3) dynamic statistical disease mapping.

Standard errors of $\hat{\boldsymbol{\beta}}(\boldsymbol{s})$ and $\hat{\boldsymbol{\alpha}}(\boldsymbol{s})$ First, we calculate the standard errors of computed MLEs, $\hat{\boldsymbol{\beta}}(\boldsymbol{s})$ and $\hat{\boldsymbol{\alpha}}(\boldsymbol{s})$, at each voxel $\boldsymbol{s}$. As the $\hat{\boldsymbol{\beta}}(\boldsymbol{s})$ in (3.8) is derived only based on the normal controls with the model (3.2), the estimated covariance matrix of $\hat{\boldsymbol{\beta}}(\boldsymbol{s})$ can be approximated by the positive square root of diagonal elements in the following matrix

$$\left[\frac{1}{n_0}\sum_{i=1}^{n_0} \boldsymbol{x}_i\left(\boldsymbol{x}_i^T\hat{\boldsymbol{\Sigma}}(\boldsymbol{s})\boldsymbol{x}_i + +\hat{\sigma}^2(\boldsymbol{s})\boldsymbol{I}_{m_i}\right)^{-1}\boldsymbol{x}_i^T\right]^{-1}. \tag{3.18}$$

On the other hand, we consider the standard errors associated with $\hat{\boldsymbol{\alpha}}(\boldsymbol{s})$ conditional on the estimates of $\boldsymbol{\theta}_0$ and $\{\boldsymbol{b}_i(\boldsymbol{s})\}_{i=1}^n$. To tackle this problem, the wild bootstrap (Wu, 1986) resampling method is considered here. The idea of wild bootstrap is to leave the regressors at their sample value, but to resample the response variable based on the residuals values. The detailed procedures are listed as follows:

1. Fit the model with the original data and retain the fitted values $\widehat{\boldsymbol{y}}_i(\boldsymbol{s}) = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}(\boldsymbol{s}) +$

$\hat{\boldsymbol{B}}_i(\boldsymbol{s})\boldsymbol{x}_i^T\hat{\boldsymbol{\alpha}}(\boldsymbol{s})$ and the residuals $\hat{\boldsymbol{\epsilon}}_i(\boldsymbol{s}) = \boldsymbol{y}_i(\boldsymbol{s}) - \hat{\boldsymbol{y}}_i(\boldsymbol{s}), i = n_0 + 1, \ldots, n$;

2. Create synthetic response variables $\boldsymbol{y}_i(\boldsymbol{s})^* = \hat{\boldsymbol{y}}_i(\boldsymbol{s}) + a_i\hat{\boldsymbol{\epsilon}}_i(\boldsymbol{s}), i = n_0 + 1, \ldots, n$, where a_i is a random variable following standard normal distribution;

3. Given $\boldsymbol{\theta}_0$ and $\{\boldsymbol{b}_i(\boldsymbol{s})\}_{i=n_0+1}^n$, refit the model using the synthetic response variables $\boldsymbol{y}_i(\boldsymbol{s})^*$ and retain the estimates $\hat{\boldsymbol{\alpha}}(\boldsymbol{s})^*$ as below,

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{s})^* = \left[\sum_{i=n_0+1}^{n} \boldsymbol{x}_i\hat{\boldsymbol{B}}_i(\boldsymbol{s})\hat{\boldsymbol{\Lambda}}_i^{-1}(\boldsymbol{s})\hat{\boldsymbol{B}}_i(\boldsymbol{s})\boldsymbol{x}_i^T\right]^{-1}\sum_{i=n_0+1}^{n} \boldsymbol{x}_i\hat{\boldsymbol{B}}_i(\boldsymbol{s})\hat{\boldsymbol{\Lambda}}_i^{-1}(\boldsymbol{s})\Big[\boldsymbol{y}_i(\boldsymbol{s})^* - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}(\boldsymbol{s})\Big],$$

   in particular, let $\hat{\boldsymbol{\alpha}}(\boldsymbol{s})^* = \boldsymbol{0}$ if $\sum_{i=n_0+1}^n \boldsymbol{x}_i\hat{\boldsymbol{B}}_i(\boldsymbol{s})\hat{\boldsymbol{\Lambda}}_i^{-1}(\boldsymbol{s})\hat{\boldsymbol{B}}_i(\boldsymbol{s})\boldsymbol{x}_i^T$ is not invertible;

4. Repeat Steps 1 and 2 K times ($K = 100$ in this paper, refer to Efron and Tibshirani (1994)) to give K independent realizations of $\hat{\boldsymbol{\alpha}}(\boldsymbol{s})^*$, denoted by $\hat{\boldsymbol{\alpha}}(\boldsymbol{s})_1^*, \ldots, \hat{\boldsymbol{\alpha}}(\boldsymbol{s})_K^*$;

5. The bootstrap covariance matrix of $\hat{\boldsymbol{\alpha}}(\boldsymbol{s})$ is estimated by

$$\frac{1}{K-1}\sum_{i=1}^{K}\big(\hat{\boldsymbol{\alpha}}(\boldsymbol{s})^* - \overline{\hat{\boldsymbol{\alpha}}(\boldsymbol{s})^*}\big)\big(\hat{\boldsymbol{\alpha}}(\boldsymbol{s})_i^* - \overline{\hat{\boldsymbol{\alpha}}(\boldsymbol{s})^*}\big)^T, \qquad (3.19)$$

   where $\overline{\hat{\boldsymbol{\alpha}}(\boldsymbol{s})^*} = \frac{1}{K}\sum_{i=1}^K \hat{\boldsymbol{\alpha}}(\boldsymbol{s})_i^*$. The standard error of $\hat{\boldsymbol{\alpha}}(\boldsymbol{s})$ can be estimated by the positive square root of the diagonal element in (3.19).

Hypothesis testing In real applications, we are interested in testing (i) whether there is any intensity progression across time at each voxel in normal regions; (ii) whether there is any difference of intensity at baseline between normal regions and diseased regions; and (iii) whether there is any difference of intensity progression across time between normal regions and diseased regions. For each voxel, these hypothesis testing problems can be written in the following general forms:

$$H_0(\boldsymbol{s}):\ \boldsymbol{C}\boldsymbol{\beta}(\boldsymbol{s}) = 0 \text{ v.s. } H_1(\boldsymbol{s}):\ \boldsymbol{C}\boldsymbol{\beta}(\boldsymbol{s}) \neq 0, \tag{3.20}$$

$$H_0(\boldsymbol{s}):\ \boldsymbol{C}\boldsymbol{\alpha}(\boldsymbol{s}) = 0 \text{ v.s. } H_1(\boldsymbol{s}):\ \boldsymbol{C}\boldsymbol{\alpha}(\boldsymbol{s}) \neq 0, \tag{3.21}$$

where $\boldsymbol{C}$ is a $1 \times p$ vector. A sequence of $Wald$ tests can be used here. The test statistics for (3.20) and (3.21) can be respectively written as

$$T_\beta(\boldsymbol{s}) = \boldsymbol{C}\hat{\boldsymbol{\beta}}(\boldsymbol{s})\Big[\boldsymbol{C}Var[\hat{\boldsymbol{\beta}}(\boldsymbol{s})]\boldsymbol{C}^T\Big]^{-1}\hat{\boldsymbol{\beta}}^T(\boldsymbol{s})\boldsymbol{C}^T, \tag{3.22}$$

and

$$T_\alpha(\boldsymbol{s}) = \boldsymbol{C}\hat{\boldsymbol{\alpha}}(\boldsymbol{s})\Big[\boldsymbol{C}Var[\hat{\boldsymbol{\alpha}}(\boldsymbol{s})]\boldsymbol{C}^T\Big]^{-1}\hat{\boldsymbol{\alpha}}^T(\boldsymbol{s})\boldsymbol{C}^T, \tag{3.23}$$

where $Var[\hat{\boldsymbol{\beta}}(\boldsymbol{s})]$ and $Var[\hat{\boldsymbol{\alpha}}(\boldsymbol{s})]$ can be obtained as described in Section 2.3.1. The corresponding p-values can be derived based on the asymptotic properties of the test statistics under H_0. In particular, under the null hypothesis, when the sample size is large enough, both $T_\beta(\boldsymbol{s})$ and $T_\alpha(\boldsymbol{s})$ approximately follow χ^2 distribution with one degree of freedom. The false discovery rate (FDR) adjustment method (Yekutieli and Benjamini, 1999) is also employed here to calculate the adjusted p-values corrected for the multiple comparison problems (3.20) and (3.21).

Dynamic statistical disease mapping Third, after obtaining the diseased region labels across all voxels for each patient, we derive the dynamic statistical disease mapping at population level for patients at different disease stage. In practice, some voxels are unlikely to be affected by the disease, thus we consider a voxel-wise zero-inflated generalized linear mixed model which can predict the diseased region label for each voxel $\boldsymbol{s}$ based on the observed demographic, clinic, and disease stage information, for $1 \leq k \leq m$,

$$Pr\{b_{i,j}(\boldsymbol{s}_k) = l\} = \begin{cases} \pi_k + (1-\pi_k)\frac{1}{1+e^{\lambda_{i,j,k}}}, & l = 0, \\ (1-\pi_k)\frac{e^{\lambda_{i,j,k}}}{1+e^{\lambda_{i,j,k}}}, & l = 1. \end{cases} \tag{3.24}$$

Based on the model, $b_{i,j}(\boldsymbol{s}_k)$ is assumed to come from the point mass distribution based at zero with probability π_k and Binomial distribution $Bi(1, \frac{e^{\lambda_{i,j,k}}}{1+e^{\lambda_{i,j,k}}})$ with probability $1 - \pi_k$. Here $\lambda_{i,j,k}$ and π_k are modeled by smooth functions at $\boldsymbol{s}_k$

$$\lambda_{i,j,k} = \boldsymbol{x}_{i,j}^T \xi(\boldsymbol{s}_k). \tag{3.25}$$

Furthermore, some smoothing techniques can be adopted on $\xi(\boldsymbol{s})$ here to model both the spatial smoothness and spatial correlation within the disease map (Huang et al., 2017).

Given the estimates $\hat{\pi}_k$ and $\hat{\xi}(\boldsymbol{s})$, the conditional probability that the pixel site belongs to the diseased region given certain patient's information ($\boldsymbol{x}_0$) is calculated via the following regression equation:

$$Pr\{\boldsymbol{s} \text{ belongs to the diseased region} \mid \boldsymbol{x}_0, \boldsymbol{w}_0\} = (1-\hat{\pi}_k)\frac{\exp(\boldsymbol{x}_0^T\hat{\xi}(\boldsymbol{s}))}{1+\exp(\boldsymbol{x}_0^T\hat{\xi}(\boldsymbol{s}))}. \tag{3.26}$$

In particular, if we focuses on patients with specified age range, we can derive the dynamic changes in statistical disease mapping across age, which is of great importance in disease prevention at early stage. Also, given the age and gender information, the statistical disease mapping for patients at different stages can be compared and helpful in prediction of disease stage transition.

3.2 Simulation Studies

We examine the finite sample performance of DSRE model for dynamic diseased region detection. Here we generated the data based on two different real datasets: (i) 2D thickness maps derived from the 3D knee MRI data of normal controls in the Pfizer Longitudinal

Study (PLS-A9001140); (ii) 3D RAVENS maps derived from T1 MRI data of normal controls in ADNI study (Miranda et al., 2018). For each dataset, we first fitted the model (3.2) to the image data from normal controls. More details about the first dataset can be found in Huang et al. (2015), while the data description and processing of the second dataset will be discussed in the real data analysis section. Then, we used the obtained parameter estimators of $\boldsymbol{\beta}(s_l), \sigma^2(\boldsymbol{s}_l), \boldsymbol{\Sigma}(\boldsymbol{s}_l), k = 1, \ldots, n_v$, as the true values for simulations. The covariates $\boldsymbol{x}_i$, including intercept, age, gender, were generated according to the real dataset. Moreover, $\boldsymbol{\alpha}(\boldsymbol{s}_l)$ were set to -0.03 across all voxels within the diseased regions. We generated 30 subjects with 3 or 4 observations for each subject. In order to mimic the heterogeneity of diseased region pattern, the number, shape, size, and location of all diseased regions were predetermined and different across subject and time points. For these two datasets, the diseased regions and observed maps for two simulated subjects are presented in Figure 3.3 and Figure 3.4 respectively.

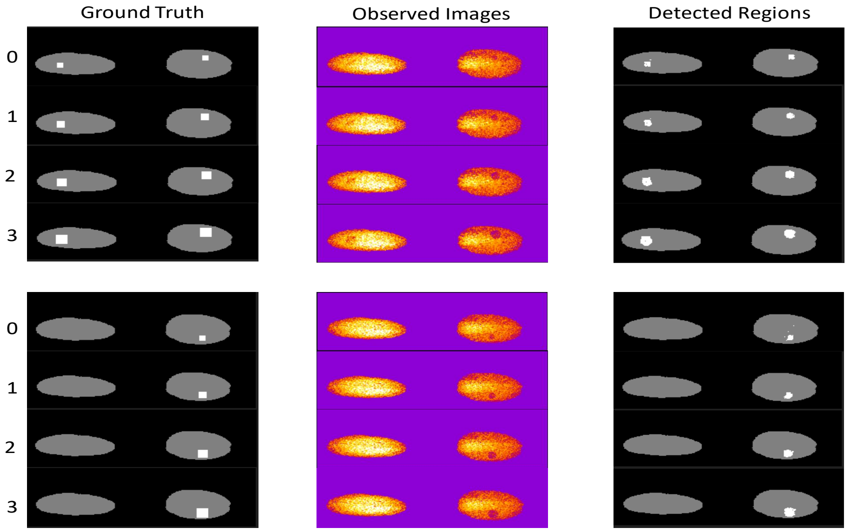


Figure 3.3: Diseased region detection on simulated 2D thickness maps: (left) ground truth for diseased regions; (middle) simulated 2D thickness maps; (right) detected diseased regions.

We applied DSRE model to detect the diseased regions for each subject at each time point. For the selected subjects, the detection results are presented in Figure 3.3 and Figure 3.4 respectively. It can be found that, the diseased regions for subjects at each time point

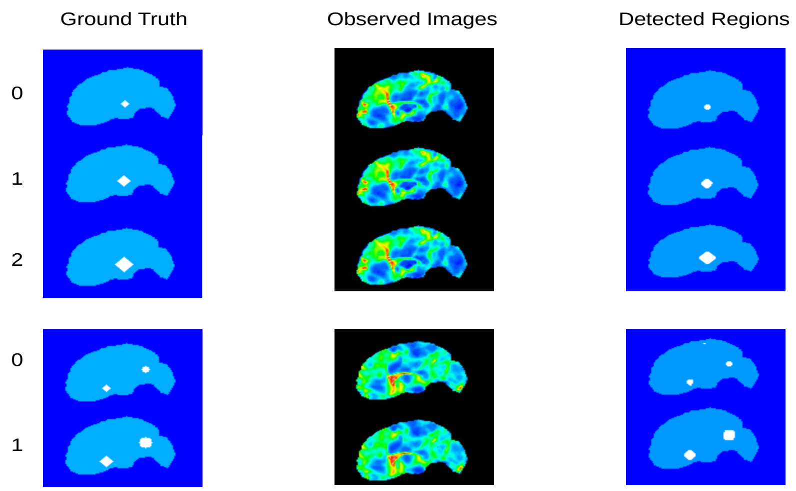


Figure 3.4: Diseased region detection on simulated 3D RAVENS maps: (left) ground truth for diseased regions; (middle) simulated 3D RAVENS maps; (right) detected diseased regions.

can be successfully detected while the detection performance at follow-up visits is better than that at baseline. The possible reason is that the diseased regions for subjects at baseline are small and the difference between signal strength in the diseased region and normal region is not significant. To compare DSRE model with other methods, we also applied the K-means clustering method and hidden Markov model (HMM) to the simulated data. The adjusted Rand index (aRI) is adopted here and reported in the boxplot (Figure 3.5). According to Figure 3.5, (i) for all the three methods, The aRIs for detection on simulated 2D thickness maps are higher than those for detection on 3D RAVENS maps; (ii) for both simulation studies, our DSRE model outperforms other two methods in terms of aRI. The possible reason is that both spectral clustering method and HMM only consider the spatial correlation Within each single image. By comparison, our DSRE model considers both spatial and temporal correlation, and the label information can be borrowed and exchanged from both spatial neighborhood and temporal neighborhood.

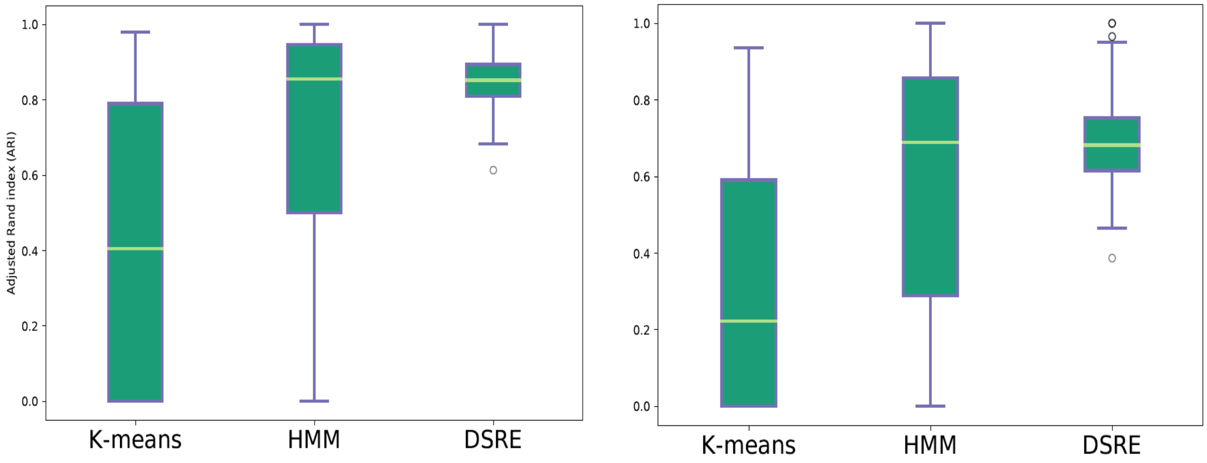


Figure 3.5: Comparison among on K-means, HMM, and DSRE model: (left) aRIs for detection on simulated 2D thickness maps; (right) aRIs for detection on simulated 3D RAVENS maps.

3.3 Real Data Analysis

3.3.1 ADNI Data Description

Data used in the preparation of this article were obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging, National Institute of Biomedical Imaging and Bioengineering, Food and Drug Administration, private pharmaceutical companies and non-profit organizations as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians in developing new treatments and monitoring their effectiveness, as well as lessening the time and cost of clinical trials. The principal investigator of this initiative is Michael W. Weiner, MD, at the VA Medical Center and University of California, San Francisco. ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The goal was to recruit 800 subjects, but the initial study (ADNI-1) has been followed by ADNI-GO and ADNI-2. To date, these three protocols have recruited over 1,500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow-up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org.

3.3.2 Data Processing

In this data analysis, we included 1179 MRI scans from healthy controls and individuals with AD (50 AD, 50 MCI, and 100 healthy controls) from ADNI-1. The scans (from 107 men and 93 women, ages 75.63 ± 6.02 years), which were performed on a variety of 1.5 Tesla MRI scanners with protocols individualized for each scanner, include standard T1-weighted images obtained using volumetric 3-dimensional sagittal MPRAGE or equivalent protocols with varying resolutions. The typical protocol includes: repetition time $=$ 2400 ms, inversion time $=$ 1000 ms, flip angle $= 8^o$, and field of view $=$ 24 cm, with a $256 \times 256 \times 170$ acquisition matrix in the $x-$, $y-$, and $z-$dimensions, which yields a voxel size of $1.25 \times 1.26 \times 1.2$ mm^3. The T1-weighted images were processed using the Hierarchical Attribute Matching Mechanism for Elastic Registration (HAMMER) pipeline. The processing steps include anterior commissure and posterior commissure correction, skull-stripping, cerebellum removal, intensity inhomogeneity correction, and segmentation. Then, registration was performed to warp the subject to the space of the Jacob template (size $256 \times 256 \times 256$ mm^3). Finally, we used the deformation field to compute the RAVENS maps. The RAVENS methodology precisely quantifies the volume of tissue in each region of the brain. The process is based on a volume-preserving spatial transformation that ensures that no volumetric information is lost during the process of spatial normalization (Davatzikos et al., 2001).

3.3.3 Data Analysis

First we applied DSRE model on the real dataset, and the estimates of coefficient functions associated to the covariates gender, age and diagnostic status (MCI or AD) are presented in Fig. 3.6. In order to test how the covarites of interest locally affect the regions, the local $Wald$ test statistics were calculated. The adjusted $-\log_{10}$ values across all vertices are shown in Fig. 3.6. It indicates that, compared to the gender effect, age and diagnostic effects are more significant in terms of local p-values.

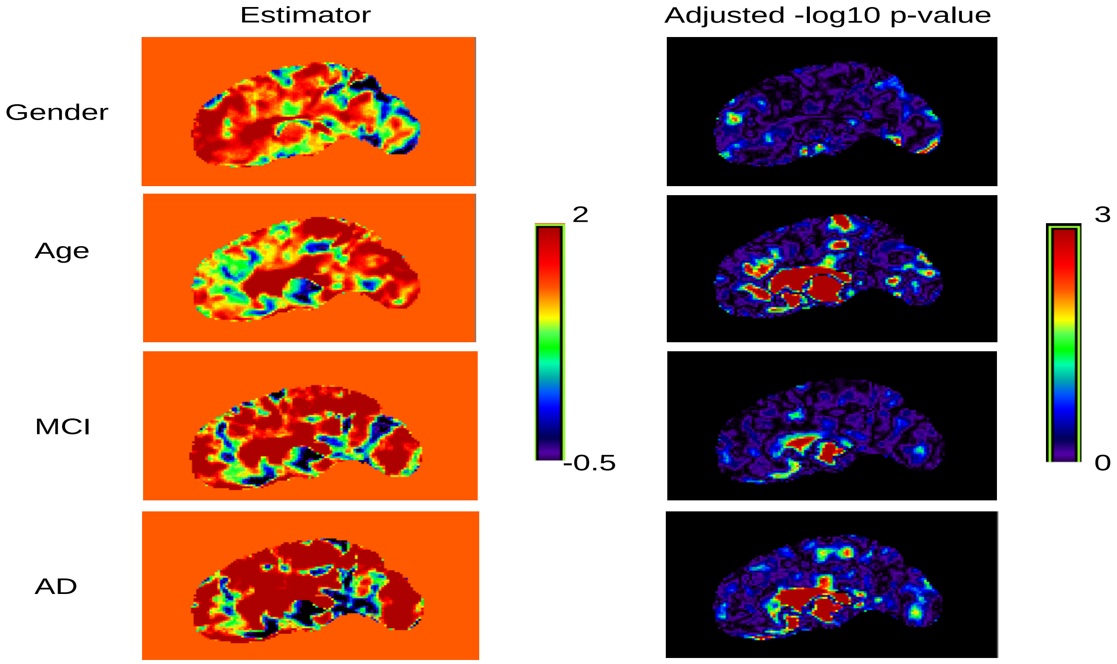


Figure 3.6: Coefficient estimators of four covariates (left); adjusted $-\log_{10}$ p-values of four covariates (right).

The inference results of $\boldsymbol{\alpha}$ are presented in Figure 3.7. For the diseased regions, the detection results of randomly selected one MCI patient and one AD patient are plotted in Figure 3.8, in which the red area indicates the detected diseased region. Both of these two patients have three observations, in which the disease status changed from normal to MCI at the second time point. The dynamic disease maps across ages were also be estimated (See Figure 3.9). As the age is getting large, the diseased regions with empirical probability larger than 0.5 include four ROIs: caudate nucleus (left and right), lingual gyrus (left and right), cingulate gyrus, and precuneus. In the existing literatures, both caudate nucleus and cingulate gyrus are found to have CMRglc reductions due to AD (Madsen et al., 2010). Also, the precuneus atrophy was found in early-onset Alzheimer's disease (Karas et al., 2007). Therefore, the detected diseased regions are meaningful and may be treated as potential imaging biomarkers for AD.

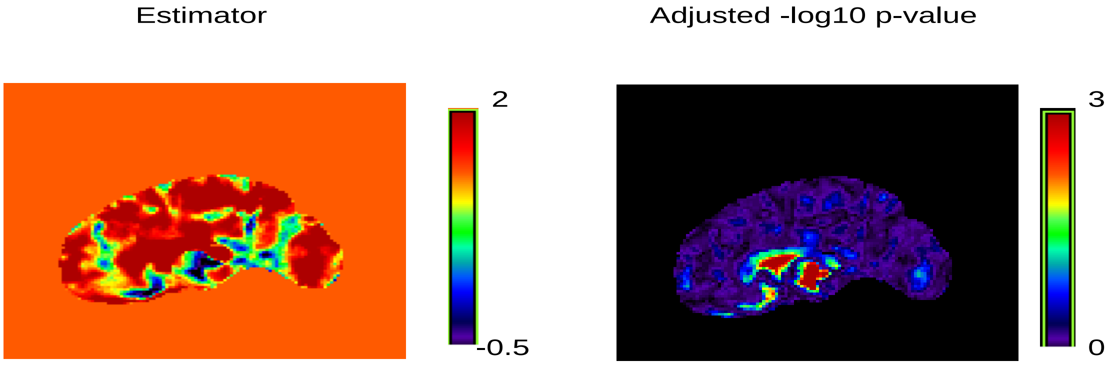


Figure 3.7: Coefficient estimator of $\boldsymbol{\alpha}(\boldsymbol{s})$ (left); adjusted $-\log_{10}$ p-values of $\boldsymbol{\alpha}(\boldsymbol{s})$ (right).

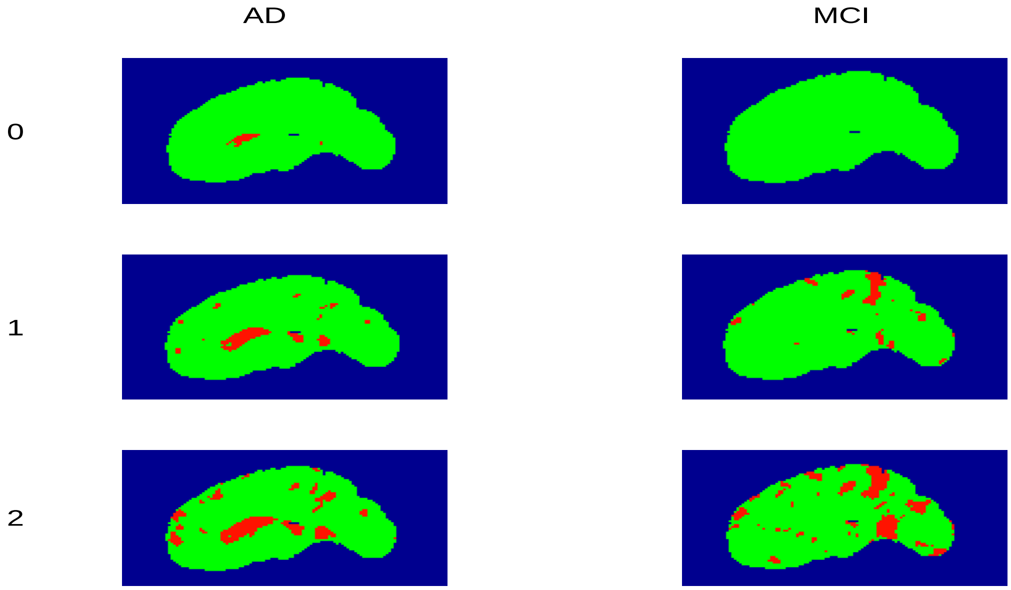


Figure 3.8: Diseased region detection for two randomly selected patients. One AD patient (left); One MCI patient (right).

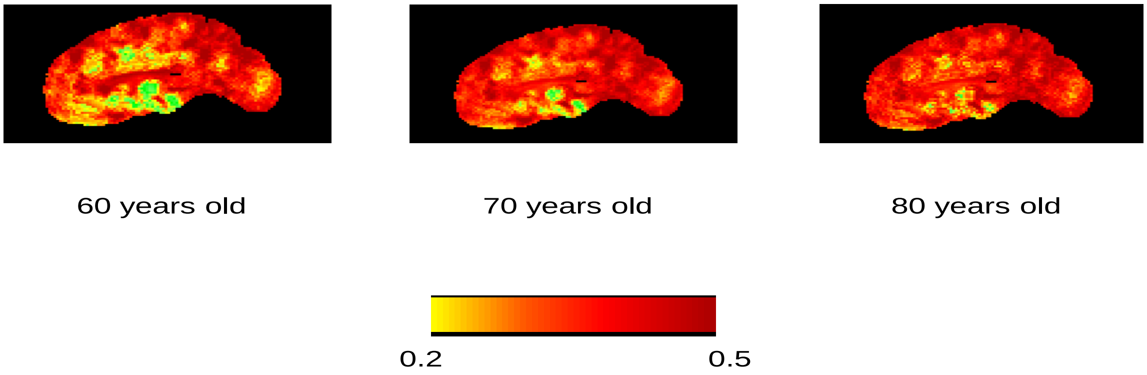


Figure 3.9: Dynamic disease maps across ages: 60 years old (left); 70 years old (middle); 80 years old (right).

CHAPTER 4: CLUSTERING HIGH-DIMENSIONAL MANIFOLD VALUED DATA IN SYMMETRIC SPACES

4.1 Method

4.1.1 Preliminaries of Riemannian Manifold and Symmetric Spaces

We review some basic results of Riemannian geometry. Let $\mathcal{M}$ be a p-dimensional complete Riemannian manifold with distance function d. We denote the tangent space at $\boldsymbol{x} \in \mathcal{M}$ by $T_x\mathcal{M}$. For any $\boldsymbol{v} \in T_x\mathcal{M}$, there is a unique geodesic curve $\gamma : [0, 1] \to \mathbb{R}$, with initial conditions $\gamma(0) = \boldsymbol{x}$ and $\gamma'(0) = \boldsymbol{v}$. It should be noted that the geodesic is only guaranteed to exist in a neighborhood of $\boldsymbol{x}$, where the largest neighborhood is denoted by $\mathcal{N}_x \in \mathcal{M}$. The exponential map at $\boldsymbol{x}$, $\mathbf{Exp}(\boldsymbol{x}, \cdot) : T_x\mathcal{M} \to \mathcal{N}_x$, is locally diffeomorphic and defined as $\mathbf{Exp}(\boldsymbol{x}, \boldsymbol{v}) = \gamma(1)$. It means that the exponential map takes the initial conditions (position $\boldsymbol{x}$ and velocity $\boldsymbol{v}$) as input and returns the point $\mathbf{Exp}(\boldsymbol{x}, \boldsymbol{v}) \in \mathcal{M}$ at time one. The log map $\mathbf{Log}(\boldsymbol{x}, \cdot) : \mathcal{N}_x \to T_x\mathcal{M}$ is defined as the inverse of exponential map. For any $\boldsymbol{x}' \in \mathcal{N}_x$, the Riemannian distance $d(\boldsymbol{x}, \boldsymbol{x}') = \|\mathbf{Log}(\boldsymbol{x}, \boldsymbol{x}')\|$, and the gradient of the squared distance function $\nabla_x d(\boldsymbol{x}, \boldsymbol{x}')^2 = -2\mathbf{Log}(\boldsymbol{x}, \boldsymbol{x}')$.

Next, we provide an overview of some necessary concepts of symmetric spaces. Recall that an isometry of a Riemannian manifold is a diffeomorphism $\kappa : \mathcal{M} \to \mathcal{M}$ that preserves the Riemannian metric, equivalently, such that $d(\boldsymbol{x}, \boldsymbol{x}') = d(\kappa(\boldsymbol{x}), \kappa(\boldsymbol{x}'))$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{M}$. Furthermore, the isometry κ is called involutive if it satisfies $\kappa = \kappa^{-1}$. Then, a Riemannian manifold $\mathcal{M}$ is called a symmetric space if, for each point $\boldsymbol{x} \in \mathcal{M}$, there exists an involutive isometry κ_x that fixes $\boldsymbol{x}$ and reserves geodesics passing through $\boldsymbol{x}$. Many useful manifolds are symmetric spaces including Euclidean spaces, spheres, the spaces of positive-definite matrices, Grassmann manifolds, Stiefel manifolds and so on (Boothby, 2003).

4.1.2 Mixture of Geodesic Factor Analyzers

It is assumed that $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ are independently and identically distributed (i.i.d.) random observations generated from a mixture model of generalized normal distributions defined on a Riemannian manifold $\mathcal{M}$, $\sum_{k=1}^{K} \pi_k g(\boldsymbol{x}|\boldsymbol{\alpha}_k, \boldsymbol{\Omega})$, where $\pi_k \geq 0$, $\sum_{k=1}^{K} \pi_k = 1$, and $g(\boldsymbol{x}|\boldsymbol{\alpha}_k, \boldsymbol{\Omega})$ is the density function of a generalized normal distribution (Pennec, 2006; Fletcher, 2013; Zhang and Fletcher, 2013) as

$$g(\boldsymbol{x}|\boldsymbol{\alpha}_k, \boldsymbol{\Omega}) = C^{-1}(\boldsymbol{\alpha}_k, \boldsymbol{\Omega}) \exp\left\{-\frac{1}{2}\mathbf{Log}^T(\boldsymbol{\alpha}_k, \boldsymbol{x})\boldsymbol{\Omega}^{-1}\mathbf{Log}(\boldsymbol{\alpha}_k, \boldsymbol{x})\right\}, \ \boldsymbol{x} \in \mathcal{M}, \tag{4.1}$$

where $C(\boldsymbol{\alpha}_k, \boldsymbol{\Omega}) = \int \exp\left\{-\frac{1}{2}\mathbf{Log}^T(\boldsymbol{\alpha}_k, \boldsymbol{x})\boldsymbol{\Omega}^{-1}\mathbf{Log}(\boldsymbol{\alpha}_k, \boldsymbol{x})\right\} d\boldsymbol{x}$, $\boldsymbol{\alpha}_k \in \mathcal{M}$ is the location parameter, and $\boldsymbol{\Omega}$ is a $p \times p$ definite positive diagonal matrix shared across clusters. When $\mathcal{M}$ is a symmetric space, this normalization term $C(\boldsymbol{\alpha}_k, \boldsymbol{\Omega})$ does not depend on $\boldsymbol{\alpha}_k$ because the distribution is invariant to isometrics. Consequently, the mixture model can be induced as

$$\sum_{k=1}^{K} \frac{\pi_k}{C(\boldsymbol{\Omega})} \exp\left\{-\frac{1}{2}\mathbf{Log}^T(\boldsymbol{\alpha}_k, \boldsymbol{x})\boldsymbol{\Omega}^{-1}\mathbf{Log}(\boldsymbol{\alpha}_k, \boldsymbol{x})\right\}. \tag{4.2}$$

In order to characterize the spatial correlation of high-dimensional manifold data, we consider the idea of factor analysis in Euclidean space to establish the geodesic factor analysis in Riemannian manifold. It is assumed that, for each observation $\boldsymbol{x}_i \in \mathcal{M}, i = 1, \ldots, n$, the location parameter $\boldsymbol{\alpha}_k$ can be formulated as follows:

$$\boldsymbol{\alpha}_k = \mathbf{Exp}(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k \boldsymbol{z}_{ki}), \ k = 1, \ldots, K, \tag{4.3}$$

where $\boldsymbol{z}_{1i}, \ldots \boldsymbol{z}_{Ki}$ are stochastic errors in $\mathbb{R}^q$, distributed independently $N(\mathbf{0}, \boldsymbol{I}_q)$, and $\boldsymbol{\Lambda}_k$ is a $p \times q$ factor loading matrix with all columns of mutually independent tangent vectors $\boldsymbol{\Lambda}_k^l, l = 1, \ldots, q$ in $T_{\mu_k}\mathcal{M}$. The idea of geodesic factor analysis is illustrated in Figure 4.1.

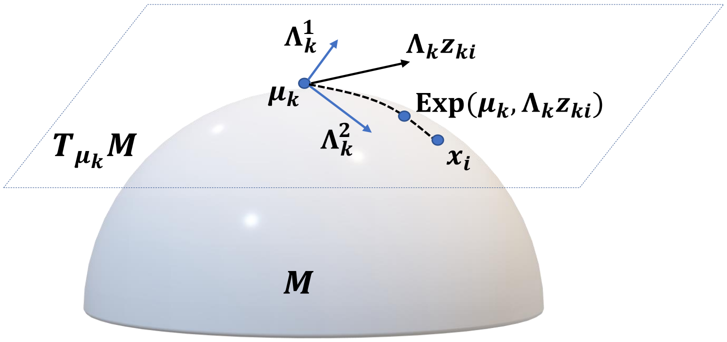


Figure 4.1: Idea of geodesic factor analysis.

Furthermore, we define υ_{ki} be a dummy variable which indicate whether $\boldsymbol{x}_i$ comes from the k-th component or not. Usually there are some covariates of interest besides the manifold data, in order to integrate these covariates, denoted by $\boldsymbol{w}_i, i = 1, \ldots, n$, into our proposed model (4.3), we consider a logistic regression model of mixing proportions $\pi_{ki} = Pr(\upsilon_{ki} = 1|\boldsymbol{w}_i)$. Specifically, given the covariates $\boldsymbol{w}_i \in \mathbb{R}^d$, the mixing proportions are defined through the logistic model given by

$$\log\left(\frac{\pi_{ki}(\boldsymbol{\beta})}{\pi_{Ki}(\boldsymbol{\beta})}\right) = \boldsymbol{w}_i^T\boldsymbol{\beta}_k \quad \text{for} \quad k = 1, \ldots, K-1 \;\; \text{and} \;\; i = 1, \ldots, n, \tag{4.4}$$

in which $\boldsymbol{w}_i = (1, w_{i,1}, \ldots, w_{i,d-1})^T$, $\boldsymbol{\beta}_k = (\beta_{k,0}, \beta_{k,1}, \ldots, \beta_{k,d-1})^T$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{K-1}^T)^T$, and $\boldsymbol{\beta}_K$ is set to $\mathbf{0}$ for identifiability. Under models (4.3) and (4.4), given the latent factor $\boldsymbol{z}_i = (\boldsymbol{z}_{1i}^T, \ldots \boldsymbol{z}_{Mi}^T)^T$, the mixture model (4.2) with respect to $\boldsymbol{x}_i$, $f(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{\theta})$, can be rewritten as

$$\begin{aligned} f(\boldsymbol{x}_i|\boldsymbol{w}_i, \boldsymbol{z}_i, \boldsymbol{\theta}) &= \sum_{k=1}^{K} \pi_{ki}(\boldsymbol{\beta}) g(\boldsymbol{x}_i|\boldsymbol{z}_{ki}, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\Omega}) \\ &= \sum_{k=1}^{K} \frac{\pi_{ki}(\boldsymbol{\beta})}{C(\boldsymbol{\Omega})} \exp\left\{-\frac{1}{2}\boldsymbol{h}^T(\boldsymbol{x}_i, \boldsymbol{z}_{ki}, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)\boldsymbol{\Omega}^{-1}\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{z}_{ki}, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)\right\}, \end{aligned} \tag{4.5}$$

where $\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{z}_{ki}, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \mathbf{Log}(\mathbf{Exp}(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k\boldsymbol{z}_{ki}), \boldsymbol{x}_i)$, $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \mathbf{diag}(\boldsymbol{\Omega})^T, \boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_K^T)^T$, and $\boldsymbol{\theta}_k$ consists of the unknown elements of $\boldsymbol{\mu}_k$, and $\boldsymbol{\Lambda}_k$, $k = 1, \ldots, K$. Here we term (4.5) a mixture of geodesic factor analyzers (MGFA).

In summary, there are several advantages of our MGFA:

(i) For each cluster, a feature space can be extracted from a high-dimensional manifold $\mathcal{M}$ to a low-dimensional latent factor space in $T_{\mu_k}\mathcal{M}$.

(ii) When $\mathcal{M} = \mathbb{R}^p$, the exponential map is an adding operation and (4.1) is a multivariate normal distribution with diagonal covariance matrix, then our MGFA reduces to a mixture of factor analyzers below

$$\boldsymbol{x}_i = \boldsymbol{\mu}_k + \boldsymbol{\Lambda}_k \boldsymbol{z}_{ki} + \boldsymbol{e}_{ki} \quad \text{with prior probabilities } \pi_{ki}(\boldsymbol{\beta}),$$

where $\boldsymbol{e}_{ki} \sim N_p(\boldsymbol{0}, \boldsymbol{\Omega})$ is independent of $\boldsymbol{z}_{ki}$ for $i = 1, \ldots, n, k = 1, \ldots, K$.

(iii) An association between mixing proportions and covariates of interest is built via a logistic regression model.

4.1.3 Estimation Procedure

EM Algorithm for the MGFA Model We first develop the EM algorithm to calculate the MLE of $\boldsymbol{\theta}$, denoted by $\tilde{\boldsymbol{\theta}}$, for low-dimensional manifold data, that is, $p \ll n$. The key idea of the EM algorithm is to introduce missing data and then maximize the conditional expectation of the complete-data log-likelihood function, called Q function. For our MGFA, we introduce υ_{ki} and $\boldsymbol{z}_{ki}$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K$ as missing data. Then, the complete-data log-likelihood function $\log L(\boldsymbol{\theta})$ is proportional to

$$\sum_{k=1}^{K}\sum_{i=1}^{n} \upsilon_{ki} \left\{ \log \pi_{ki}(\boldsymbol{\beta}) - \log C(\boldsymbol{\Omega}) - \frac{1}{2}\boldsymbol{h}^T(\boldsymbol{x}_i, \boldsymbol{z}_{ki}, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)\boldsymbol{\Omega}^{-1}\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{z}_{ki}, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) - \frac{\|\boldsymbol{z}_{ki}\|_2^2}{2} \right\}.$$

However, given $\widetilde{\boldsymbol{\theta}}$ and $\{\boldsymbol{x}_i\}_{i=1}^n$, the Q function, i.e., $Q(\boldsymbol{\theta}|\widetilde{\boldsymbol{\theta}}) = \mathbb{E}_z[\log L(\boldsymbol{\theta})|\{\boldsymbol{x}_i\}_{i=1}^n, \widetilde{\boldsymbol{\theta}}]$, does not yield a closed-form solution. So, the MCEM algorithm is considered instead to estimate $\boldsymbol{\theta}$. Similar to other MCEM procedures, there are two main steps in our algorithm, i.e., E-step and M-step.

E-step: In the E-step, given $\widetilde{\boldsymbol{\theta}}^{(r)}$ at the r-th iteration, we consider adopting the Hamiltonian Monte Carlo (HMC) sampling method (Neal, 2011) to sample $\boldsymbol{z}_{mi}$ from their posterior distribution $p(\boldsymbol{z}_{ki}|\upsilon_{ki} = 1, \{\boldsymbol{x}_i\}_{i\leq n}, \widetilde{\boldsymbol{\theta}}^{(r)})$. According to HMC method, we set up the Hamiltonian dynamic system first. The Hamiltonian function can be written as $H(\boldsymbol{z}_{ki}, \boldsymbol{r}) = U(\boldsymbol{z}_{ki}) + \frac{1}{2}\boldsymbol{r}_k^T\boldsymbol{r}_k$, where $U(\boldsymbol{z}_{ki}) = -\log p(\boldsymbol{z}_{ki}|\upsilon_{ki} = 1, \{\boldsymbol{x}_i\}_{i\leq n}, \widetilde{\boldsymbol{\theta}}^{(r)})$ is called the potential energy function. The other item $\frac{1}{2}\boldsymbol{r}_k^T\boldsymbol{r}_k$ is called the kinetic energy, where $\boldsymbol{r}_k, k = 1, \ldots, K$ are auxiliary momentum variables drawn independently from $N(\boldsymbol{0}, \boldsymbol{I}_q)$. Because of the introduction of $\boldsymbol{r}_k$, the Hamiltonian dynamics can be established as

$$\frac{d\boldsymbol{z}_{ki}}{dt} = \boldsymbol{r}_k, \quad \frac{d\boldsymbol{r}_k}{dt} = -\nabla_{z_{ki}} U(\boldsymbol{z}_{ki}). \tag{4.6}$$

Then the approximation solution to (4.6) can be obtained via the Leap Frog numerical integration method (Neal, 2011) if the item $\nabla_{z_{ki}} U(\boldsymbol{z}_{ki})$ is calculated. In fact, the gradient term $\nabla_{z_{ki}} U(\boldsymbol{z}_{ki})$ can be derived as below

$$\nabla_{z_{ki}} U(\boldsymbol{z}_{ki}) = \boldsymbol{z}_{ki}^j - \boldsymbol{\Lambda}_k^T d_v\mathbf{Exp}(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k\boldsymbol{z}_{ki}^j)^\dagger \boldsymbol{\Omega}^{-1}\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{z}_{ki}, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k), \tag{4.7}$$

where $d_v\mathbf{Exp}(\boldsymbol{u}, \boldsymbol{v})$ is the gradient of $\mathbf{Exp}(\boldsymbol{u}, \boldsymbol{v})$ with respect to $\boldsymbol{v}$, and † represents the adjoint of a linear operator. Therefore, after obtaining the samples $\boldsymbol{z}_{ki}^j, j = 1, \ldots, N_z$, the Q function at the r-th iteration, i.e., $Q(\boldsymbol{\theta}|\widetilde{\boldsymbol{\theta}}^{(r)})$, is approximated via Monte Carlo method and proportional to

$$\frac{1}{N_z}\sum_{j=1}^{N_z}\sum_{k=1}^{K}\sum_{i=1}^{n}\tilde{\tau}_{ki}^{(r)}\left\{\log \pi_{ki}(\boldsymbol{\beta}) - \log C(\boldsymbol{\Omega}) - \frac{1}{2}\boldsymbol{h}^T(\boldsymbol{x}_i, \boldsymbol{z}_{ki}^j, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)\boldsymbol{\Omega}^{-1}\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{z}_{ki}^j, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)\right\} \tag{4.8}$$

where

$$\tilde{\tau}_{ki}^{(r)} = \frac{\sum_{j=1}^{N_z}\pi_{ki}(\tilde{\boldsymbol{\beta}}^{(r)})g(\boldsymbol{x}_i|\boldsymbol{z}_{ki}^j, \tilde{\boldsymbol{\mu}}_k^{(r)}, \tilde{\boldsymbol{\Lambda}}_k^{(r)}, \tilde{\boldsymbol{\Omega}}^{(r)})}{\sum_{k=1}^{K}\sum_{j=1}^{N_z}\pi_{ki}(\tilde{\boldsymbol{\beta}}^{(r)})g(\boldsymbol{x}_i|\boldsymbol{z}_{ki}^j, \tilde{\boldsymbol{\mu}}_k^{(r)}, \tilde{\boldsymbol{\Lambda}}_k^{(r)}, \tilde{\boldsymbol{\Omega}}^{(r)})}.$$

The performance of standard HMC method is highly sensitive to two user-specified parameters: a step size ϵ and a desired number of steps L. In particular, if L is too small then the algorithm exhibits undesirable random walk behavior, while if L is too large the algorithm wastes computation. Compared with the standard HMC, the No-U-Turn Sampler (NUTS) (Hoffman and Gelman, 2014), an extension to standard HMC, can avoid setting the tuning parameter L. Specifically, NUTS uses a recursive algorithm to build a set of likely candidate points that spans a wide swath of the target distribution, stopping automatically when it starts to double back and retrace its steps. Because of this, NUTS is adopted in this paper. The details of NUTS algorithm is omitted here, and readers can refer to Section 3 and Algorithm 3 in Hoffman and Gelman (2014).

M-step: In the M-step, given the current estimate $\widetilde{\boldsymbol{\theta}}^{(r)}$, we update $\widetilde{\boldsymbol{\theta}}^{(r+1)}$ by maximizing the Q function in (4.8) with respect to $\boldsymbol{\theta}$. For $\boldsymbol{\beta}$, a update equation can be derived according to the Newton-Raphson algorithm (Huang et al., 2015). Let $\tilde{\boldsymbol{\beta}}^{(s,r+1)}$ be the value of $\tilde{\boldsymbol{\beta}}^{(r+1)}$ at the s-th iteration of the Newton-Raphson algorithm and $\tilde{\boldsymbol{\beta}}^{(0,r+1)} = \tilde{\boldsymbol{\beta}}^{(r)}$. We update $\tilde{\boldsymbol{\beta}}^{(s,r+1)}$ as follows:

$$\tilde{\boldsymbol{\beta}}^{(s+1,r+1)} = \tilde{\boldsymbol{\beta}}^{(s,r+1)} - \left[\sum_{i=1}^{n} \boldsymbol{\Upsilon}_i^T \boldsymbol{C}_i\big(\tilde{\boldsymbol{\beta}}^{(s,r+1)}\big)\boldsymbol{\Upsilon}_i\right]^{-1} \sum_{i=1}^{n} \boldsymbol{\Upsilon}_i^T \big\{\tilde{\boldsymbol{\tau}}_i^{(r)} - \boldsymbol{\pi}_i\big(\tilde{\boldsymbol{\beta}}^{(s,r+1)}\big)\big\}, \tag{4.9}$$

where $\tilde{\boldsymbol{\tau}}_i^{(r)} = (\tilde{\tau}_{1i}^{(r)}, \cdots, \tilde{\tau}_{(K-1)i}^{(r)})^T$, $\boldsymbol{\Upsilon}_i = \boldsymbol{z}_i^T \otimes \mathbf{I}_{K-1}$, $\boldsymbol{C}_i(\boldsymbol{\beta}) = \text{diag}(\boldsymbol{\pi}_i(\boldsymbol{\beta})) - \boldsymbol{\pi}_i(\boldsymbol{\beta})\boldsymbol{\pi}_i(\boldsymbol{\beta})^T$, and $\boldsymbol{\pi}_i(\boldsymbol{\beta}) = (\pi_{1i}(\boldsymbol{\beta}), \cdots, \pi_{(K-1)i}(\boldsymbol{\beta}))^T$. We update $\tilde{\boldsymbol{\beta}}^{(s+1,r+1)}$ according to (4.9) until a pre-specified tolerance is reached and then set $\tilde{\boldsymbol{\beta}}^{(s+1,r+1)}$ from the last iteration as $\tilde{\boldsymbol{\beta}}^{(r+1)}$.

For the diagonal matrix $\boldsymbol{\Omega}$ shared across clusters, the explicit update equations can also be derived at the r-th iteration. It can be updated via solving the equation below:

$$\frac{2n}{C(\boldsymbol{\Omega})}\frac{\partial C(\boldsymbol{\Omega})}{\partial \boldsymbol{\Omega}} + \frac{1}{N_z}\sum_{k=1}^{K}\sum_{j=1}^{N_z}\sum_{i=1}^{n} \tilde{\tau}_{ki}^{(r)} \boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{z}_{ki}^j, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)\boldsymbol{h}^T(\boldsymbol{x}_i, \boldsymbol{z}_{ki}^j, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = 0. \tag{4.10}$$

The above update equation requires evaluation of the normalizing constant $C(\boldsymbol{\Omega})$ and its derivative. For Gaussian distribution in Euclidean space, it's straightforward to calculate that $C(\boldsymbol{\Omega}) = \sqrt{(2\pi)^p|\boldsymbol{\Omega}|}$ and $\partial C(\boldsymbol{\Omega})/\partial\boldsymbol{\Omega} = -\frac{1}{2}C(\boldsymbol{\Omega})\boldsymbol{\Omega}$.

For the location parameter $\boldsymbol{\mu}_k$ and loading matrix $\boldsymbol{\Lambda}_k$ in each cluster, the problem on maximization of Q function at the r^{th} iteration can be written as an minimization problem on the following object function

$$\tilde{Q}_k(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \sum_{j=1}^{N_z}\sum_{i=1}^{n} \tilde{\tau}_{ki}^{(r+1)} \boldsymbol{h}^T(\boldsymbol{x}_i, \boldsymbol{z}_{ki}^j, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) \tilde{\boldsymbol{\Omega}}^{(r)^{-1}} \boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{z}_{ki}^j, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k). \tag{4.11}$$

In order to solve the optimization problem (4.11), the method of steepest descent (SD) is considered here. SD algorithm approaches the minimum in a zig-zag manner, where the new search direction is orthogonal to the previous. The choice of direction is opposite to the gradient function at $\tilde{\boldsymbol{\mu}}_k^{(r)}$ and $\tilde{\boldsymbol{\Lambda}}_k^{(r)}$. The gradient function of $\tilde{Q}_k$ with respect to $\boldsymbol{\mu}_k$ is derived as

$$\nabla_{\mu_k}\tilde{Q}_k = \sum_{j=1}^{N_z}\sum_{i=1}^{n} \tilde{\tau}_{ki}^{(r)} d_u \mathbf{Exp}(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k \boldsymbol{z}_{ki}^j)^\dagger \tilde{\boldsymbol{\Omega}}^{(r)^{-1}} \mathbf{Log}(\mathbf{Exp}(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k \boldsymbol{z}_{ki}^j), \boldsymbol{x}_i), \tag{4.12}$$

where $d_u\mathbf{Exp}(\boldsymbol{u}, \boldsymbol{v})$ is the gradient of $\mathbf{Exp}(\boldsymbol{u}, \boldsymbol{v})$ with respect to $\boldsymbol{u}$. For the loading matrix $\boldsymbol{\Lambda}_k$, the gradient term is written as

$$\nabla_{\Lambda_k}\tilde{Q}_k = \sum_{j=1}^{N_z}\sum_{i=1}^{n} \tilde{\tau}_{ki}^{(r)} d_v \mathbf{Exp}(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k \boldsymbol{z}_{ki}^j)^\dagger \tilde{\boldsymbol{\Omega}}_k^{(r)^{-1}} \mathbf{Log}(\mathbf{Exp}(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k \boldsymbol{z}_{ki}^j), \boldsymbol{x}_i) \boldsymbol{z}_{ki}^{j^T}. \tag{4.13}$$

Then, $\tilde{\boldsymbol{\mu}}_k^{(r+1)}$ and $\tilde{\boldsymbol{\Lambda}}_k^{(r+1)}$ can be updated according to (4.12) and (4.13) via SD algorithm, while the step size parameter in the algorithm is chosen based on the linear search method.

The E-step and M-step are repeated until the difference between $\log L(\tilde{\boldsymbol{\theta}}^{(r+1)})$ and $\log L(\tilde{\boldsymbol{\theta}}^{(r)})$ is smaller than a pre-specified number, say 10^{-4}. The MCEM procedure for MGFA clustering is presented in Algorithm 1.

Algorithm 1 MCEM algorithm for MGFA clustering

```
Input Data: $x_i, w_i, i = 1, \ldots, n$
Initialize tuning parameters: $K, q$
Initialize parameters of interest: $\beta, \Omega, \mu_k, \Lambda_k, k = 1 \ldots, K$
Repeat
    • Monte Carlo E-step
        • Sample $z_{ki}, 1 \leq i \leq n, 1 \leq k \leq K,$ via HMC method
    • M-step
        • Update $\beta$ and $\Omega$ based on (4.9) − (4.10)
        • Update $\mu_k$ and $\Lambda_k,\ k = 1, \ldots, K,$ via SD method
End repeat
Output: cluster membership of $x_i$ based on $\hat{v}_{ki}, 1 \leq i \leq n, 1 \leq k \leq K.$
```

EM Algorithm for the Penalized MGFA Clustering In high dimensional manifold clustering, many 'non-informative' variables exist in the manifold data, which prevent the underlying clustering structure from being uncovered. Therefore, directly using $\tilde{\boldsymbol{\theta}}$ in high dimensional manifold data clustering may not work well. Thus, it is of great importance to remove such 'non-informative' variables and use the informative ones in data clustering. In order to achieve variable selection in MGFA, we develop a penalized MGFA clustering framework below.

To realize variable selection in MGFA, we consider a penalized log-likelihood function given by

$$\log L_p(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) - \lambda_1 \sum_{l=1}^{p} \sum_{1\le k,k'\le K} a_{k,k'}^{l} |\eta_{kl} - \eta_{k'l}| - \lambda_2 \sum_{l=1}^{p}\sum_{k=1}^{K} \|\boldsymbol{\Lambda}_{kl}\|_2, \tag{4.14}$$

where η_{kl} is the l-th element in the vector $\boldsymbol{\eta}_k = \psi_k(\boldsymbol{\mu}_k)$, in which $\psi_k(\cdot) : \mathcal{M} \to \mathbb{R}^p$ is an embedding for $\boldsymbol{\mu}_k$, and $a_{k,k'}^{l}$ are the pre-specified weights. $\boldsymbol{\Lambda}_{kl}$ is the l-th row of the factor loading $\boldsymbol{\Lambda}_k$, and $\|\cdot\|_2$ denotes the L_2 norm in Euclidian space. In the second term of (4.14), the pairwise fusion Lasso penalization (Guo et al., 2010) is introduced on the position parameters $\boldsymbol{\mu}_k$ based on a chord distance on $\mathcal{M}$. Although this penalty function is not inspired from an intrinsic way, the aim of shrinking the difference between every pair of cluster centers can be achieved as well when they are close to each other. In the third term of (4.14), since the latent variable $\boldsymbol{z}_{ki}$ is defined in Euclidian space, it is reasonable to introduce a L_2 penalty on $\boldsymbol{\Lambda}_k$ to shrink small $\boldsymbol{\Lambda}_{kj}$ to be exactly zero.

In this penalty, there are $\binom{K}{2}$ terms of pairwise differences for each element of embedding in $\mathbb{R}^p$ and the total number of terms increases by an order of $O(K^2)$ given p fixed, which contains many redundant constraints and imposes great computational challenges (Ke et al., 2015; Shen and Huang, 2010; Tang and Song, 2016). To address this issue, the fusion penalty in (4.14) can be written as a simplified penalty function that uses the information on the ordering of coefficients. For the l-th element in the location parameter, let $\boldsymbol{U}_l = (U_{1l}, \ldots, U_{Kl})^T$ be the ranking with no ties of $\boldsymbol{\eta}_{.l} = (\eta_{1l}, \ldots, \eta_{Kl})^T$, from the smallest to the largest. Specifically, $U_{kl} = \sum_{k'=1}^{K} \mathbf{1}\{\eta_{k'l} \le \eta_{kl}\}$ if there are no ties in $\boldsymbol{\eta}_{.l}$; otherwise, the ties in $\boldsymbol{U}_l$ are resolved by the first-occurrence-wins rule according to k to ensure rank uniqueness. Then, the second term in (4.14) with parameter orderings $\boldsymbol{U}_l, l = 1, \ldots, p$ takes the form

$$\lambda_1 \sum_{l=1}^{p} \sum_{k=1}^{K} \sum_{k'>k}^{K} a_{k,k'}^{l} \mathbf{1}(|U_{kl} - U_{k'l}| = 1)|\eta_{kl} - \eta_{k'l}|, \tag{4.15}$$

where the constraints occur effectively only on adjacent ordered pairs. Clearly, the penalty in (4.15) only involves $K - 1$ terms, which is of an order $O(K)$ given p fixed. Furthermore, we can also encourage sparsity for the coefficient closest to zero in $\boldsymbol{\eta}_{.l}$. Specifically, let $\boldsymbol{V}_l = (V_{1l}, \ldots, V_{Kl})^T$ be the ranking with no ties, from the smallest to the largest, of the absolute values of $\boldsymbol{\eta}_{.l}$, i.e., $(|\eta_{1l}|, \ldots, |\eta_{Kl}|)^T$. Similar to $\boldsymbol{U}_l$, the ties in $\boldsymbol{V}_l$ can also be resolved by the first-occurrence-wins rule according to k. Then, for the l-th element, consider a set of transformed parameters $\boldsymbol{\zeta}_{.l} = (\zeta_{1l}, \ldots, \zeta_{Kl})^T$ defined by

$$\zeta_{1l} = \eta_{tl},\ V_{tl} = 1;\ \text{and}\ \zeta_{kl} = \eta_{(kl)} - \eta_{((k-1)l)},\ k = 2, \ldots, K, \tag{4.16}$$

where $(\eta_{(1l)}, \ldots, \eta_{(Kl)})^T$ is the ascending order of elements in $\boldsymbol{\eta}_{.l}$. Based on the definition, (4.15) can be simplified written as

$$\lambda_1 \sum_{l=1}^{p} \sum_{k=1}^{K} a_k^l |\zeta_{kl}|, \tag{4.17}$$

where a_k^l is pre-specified as

$$a_k^l = \begin{cases} \tilde{\sigma}_l^{-1} |\tilde{\mu}_{(1l)}|^{-1}, & \text{if } k = 1, \\ \tilde{\sigma}_l^{-1} |\tilde{\mu}_{(kl)} - \tilde{\mu}_{((k-1)l)}|^{-1}, & \text{if } k > 1. \end{cases} \tag{4.18}$$

Here $\tilde{\sigma}_l^2$ is the estimated l-th diagonal element of $\boldsymbol{\Omega}$ as $K = 1$ and $\boldsymbol{\Lambda} = \boldsymbol{0}$, while $\tilde{\mu}_{(kl)}$ is the estimates of $\mu_{(kl)}$ in MGFA without any penalization. Since no ties are allowed in the parameter ordering of $\boldsymbol{\eta}_{.l}$, one-to-one transformation exists between $\boldsymbol{\eta} = (\boldsymbol{\eta}_{.1}, \ldots, \boldsymbol{\eta}_{.p})^T$ and $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_{.1}, \ldots, \boldsymbol{\zeta}_{.p})^T$ by suitable sorting matrix $\boldsymbol{S}$ and reparameterization matrix $\boldsymbol{R}$; that is, $\boldsymbol{\zeta} = \boldsymbol{R}\boldsymbol{S}\boldsymbol{\eta}$ and $\boldsymbol{\eta} = (\boldsymbol{R}\boldsymbol{S})^{-1}\boldsymbol{\zeta}$ with both $\boldsymbol{S}$ and $\boldsymbol{R}$ being full-rank square matrices. For Euclidean space, the optimization problem can be solved with respect to coefficient vector $\boldsymbol{\zeta}$ and transformed observations $\boldsymbol{R}\boldsymbol{S}\boldsymbol{x}_i, i = 1, \ldots, n$.

Here the MCEM algorithm can also be adopted to calculate the maximum penalized likelihood estimate (MPLE). In fact, Since the penalty functions in (4.16) only depend on $\boldsymbol{\eta}_k$ and $\boldsymbol{\Lambda}_k$, the update equations of $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ are the same as those given in (4.9) and (4.10). To efficiently update the estimation of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Lambda}_k$, The optimization problem can be solved via Alternating direction method of multipliers (ADMM) (Boyd et al., 2011). Here the ADMM algorithm on manifold (Kovnatsky et al., 2016) is adopted, the idea of which is similar to the one in Euclidian spaces. Specifically, denote that $\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_K)$. Then, at the r-th iteration of MCEM algorithm, a minimization problem is given by

$$\begin{aligned}
&\textbf{minimize} \;\; \tilde{Q}_p(\boldsymbol{\eta}, \boldsymbol{\Lambda}) + p_{\lambda_1}(\boldsymbol{\zeta}) + p_{\lambda_2}(\boldsymbol{\nu}) \\
&\textbf{subject to} \;\; \boldsymbol{R}\boldsymbol{S}\boldsymbol{\eta}_{.l} = \boldsymbol{\zeta}_{.l}, \;\; \boldsymbol{\Lambda}_{kl} = \boldsymbol{\nu}_{kl}, \;\; l = 1, \ldots, p, \;\; k = 1, \ldots, K,
\end{aligned}$$

where $\boldsymbol{\zeta}_{.l}, \boldsymbol{\nu}_{kl}, \;\; l = 1, \ldots, p, \;\; k = 1, \ldots, K$ are a set of augmented variables. For the functions $\tilde{Q}_p(\boldsymbol{\eta}, \boldsymbol{\Lambda})$, $p_{\lambda_1}(\boldsymbol{\zeta})$, and $p_{\lambda_2}(\boldsymbol{\nu})$, we have

$$\tilde{Q}_p(\boldsymbol{\eta}, \boldsymbol{\Lambda}) = \sum_{k=1}^{K} \sum_{j=1}^{N_z} \sum_{i=1}^{n} \tilde{\tau}_{ki}^{(r+1)} \boldsymbol{h}^T(\boldsymbol{x}_i, \boldsymbol{z}_{ki}^j, \psi_k^{-1}(\boldsymbol{\eta}_k), \boldsymbol{\Lambda}_k) \tilde{\boldsymbol{\Omega}}^{(r)^{-1}} \boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{z}_{ki}^j, \psi_k^{-1}(\boldsymbol{\eta}_k), \boldsymbol{\Lambda}_k),$$

$$p_{\lambda_1}(\boldsymbol{\zeta}) = \lambda_1 \sum_{l=1}^{p} \sum_{k=1}^{K} a_k^l |\zeta_{kl}|, \text{ and } p_{\lambda_2}(\boldsymbol{\nu}) = \lambda_2 \sum_{l=1}^{p} \sum_{k=1}^{K} \|\boldsymbol{\nu}_{kl}\|_2.$$

The corresponding augmented Lagrangian function is

$$\begin{aligned}
L_\rho(\boldsymbol{\eta}, \boldsymbol{\Lambda}, \boldsymbol{\zeta}, \boldsymbol{\nu}, \boldsymbol{\delta}, \boldsymbol{\kappa}) \;\; = \;\; & \tilde{Q}_p(\boldsymbol{\eta}, \boldsymbol{\Lambda}) + p_{\lambda_1}(\boldsymbol{\zeta}) + p_{\lambda_2}(\boldsymbol{\nu}) \\
& + \frac{\rho}{2} \sum_{l=1}^{p} \left[\|\boldsymbol{R}\boldsymbol{S}\boldsymbol{\eta}_{.l} - \boldsymbol{\zeta}_{.l} + \boldsymbol{\delta}_{.l}\|_2^2 + \sum_{k=1}^{K} \|\boldsymbol{\Lambda}_{kl} - \boldsymbol{\nu}_{kl} + \boldsymbol{\kappa}_{kl}\|_2^2 \right].
\end{aligned} \quad (4.19)$$

where $\rho > 0$ and $\boldsymbol{\delta} = (\boldsymbol{\delta}_{.1}, \ldots, \boldsymbol{\delta}_{.p})^T$, $\boldsymbol{\kappa} = (\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_K)$, and $\boldsymbol{\kappa}_{kl}$ is the l-th row in $\boldsymbol{\kappa}_k$. Both $\boldsymbol{\delta}$ and $\boldsymbol{\kappa}$ have to be chosen and updated appropriately. This formulation now allows splitting the problem into two optimization sub-problems with respect to $\{\boldsymbol{\eta}, \boldsymbol{\Lambda}\}$ and $\{\boldsymbol{\zeta}, \boldsymbol{\nu}\}$, which are solved in an alternating manner, followed by an updating of $\{\boldsymbol{\delta}, \boldsymbol{\kappa}\}$. Observe that in the first sub-problem with respect to $\{\boldsymbol{\eta}, \boldsymbol{\Lambda}\}$, we minimize a smooth function with manifold constraints, and in the second sub-problem with respect to $\{\boldsymbol{\zeta}, \boldsymbol{\nu}\}$ we minimize a non-smooth function without manifold constraints. The manifold ADMM procedure is presented in

Algorithm 2.

Algorithm 2 Manifold ADMM procedure for penalized MGFA clustering

```
Initialize: $t \leftarrow 1, \boldsymbol{\zeta}^{(t)} = \boldsymbol{\eta}^{(t)}, \boldsymbol{\nu}^{(t)} = \boldsymbol{\Lambda}^{(tk)}, \boldsymbol{\delta}^{(t)} = \mathbf{0}, \boldsymbol{\kappa}^{(t)} = \mathbf{0}$.
Repeat
    • Update $\boldsymbol{\eta} : \boldsymbol{\eta}^{(t+1)} = \text{argmin}_\eta \ \tilde{Q}_p(\boldsymbol{\eta}, \boldsymbol{\Lambda}^{(t)}) + \frac{\rho}{2}\sum_{l=1}^{p} \|\boldsymbol{RS\eta}_{.l} - \boldsymbol{\zeta}_{.l}^{(t)} + \boldsymbol{\delta}_{.l}^{(t)}\|_2^2$
    • Update $\boldsymbol{\Lambda} : \boldsymbol{\Lambda}^{(t+1)} = \text{argmin}_\Lambda \ \tilde{Q}_p(\boldsymbol{\eta}^{(t+1)}, \boldsymbol{\Lambda}) + \frac{\rho}{2}\sum_{l=1}^{p}\sum_{k=1}^{K} \|\boldsymbol{\Lambda}_{kl} - \boldsymbol{\nu}_{kl}^{(t)} + \boldsymbol{\kappa}_{kl}^{(t)}\|_2^2$
    • Update $\boldsymbol{\zeta} : \boldsymbol{\zeta}_{.l}^{(t+1)} = \mathbf{ST}_{\lambda_1/\rho,a}(\boldsymbol{RS\eta}_{.l}^{(t+1)} + \boldsymbol{\delta}_{.l}^{(t)}), \ l = 1, \ldots, p$
    • Update $\boldsymbol{\nu} : \boldsymbol{\nu}_{kl}^{(t+1)} = \mathbf{VST}_{\lambda_2/\rho}(\boldsymbol{\Lambda}_{kl}^{(t+1)} + \boldsymbol{\kappa}_{kl}^{(t)}), \ l = 1, \ldots, p, \ k = 1, \ldots, K$
    • Update $\boldsymbol{\delta}_{.l} : \boldsymbol{\delta}_{.l}^{(t+1)} = \boldsymbol{\delta}_{.l}^{(t)} + \boldsymbol{RS\eta}_{.l}^{(t+1)} - \boldsymbol{\zeta}_{.l}^{(t+1)}, \ l = 1, \ldots, p$
    • Update $\boldsymbol{\kappa}_{kl} : \boldsymbol{\kappa}_{kl}^{(t+1)} = \boldsymbol{\kappa}_{kl}^{(t)} + \boldsymbol{\Lambda}_{kl}^{(t+1)} - \boldsymbol{\nu}_{kl}^{(t+1)}, \ l = 1, \ldots, p, \ k = 1, \ldots, K$
    • $t \leftarrow t + 1$
End repeat until convergence.
```

Here $\mathbf{ST}_{\lambda_1/\rho,a}(\cdot)$ is the element-wise soft thresholding operator (STO) proposed while $\mathbf{VST}_{\lambda_2/\rho}(\cdot)$ is the vector soft thresholding operator (VSTO). The definitions of STO and VSTO can be found in Huang et al. (2015).

Note that the manifold ADMM is extremely simple and easy to implement. The updates of $\{\boldsymbol{\eta}, \boldsymbol{\Lambda}\}$ can be carried out using any standard smooth manifold optimization method, e.g., SD method (Townsend et al., 2016). Similarly to common implementation of ADMM algorithms, there is no need to solve the optimization problem exactly; instead, only a few iterations of manifold optimization are done. Furthermore, for Euclidean space, these optimization problems have closed-form solutions (Huang et al., 2015). On the other hand, the updates of $\{\boldsymbol{\zeta}, \boldsymbol{\nu}\}$ also have closed-form expressions here. ρ is the only parameter of the algorithm and its choice is not critical for convergence. In our experiments, we used a rather arbitrary fixed value of ρ, though in the ADMM literature it is common to adapt ρ at each iteration, e.g. using the strategy described in Boyd et al. (2011).

4.1.4 Convergence Properties and Asymptotic Properties

Motivated by Wen et al. (2012), we can establish that, under some regularity conditions, any limit point of the iteration sequence generated by Algorithm 2, denoted as $(\boldsymbol{\eta}, \boldsymbol{\Lambda}^*, \boldsymbol{\zeta}^*, \boldsymbol{\nu}^*, \boldsymbol{\delta}^*, \boldsymbol{\kappa}^*)$, is a KKT point of $L_\rho(\boldsymbol{\eta}, \boldsymbol{\Lambda}, \boldsymbol{\zeta}, \boldsymbol{\nu}, \boldsymbol{\delta}, \boldsymbol{\kappa})$, which satisfies, for $l = 1, \ldots, p,\ k = 1, \ldots, K$,

$$\begin{cases} \nabla_{\eta_{.l}} \tilde{Q}_p(\boldsymbol{\eta}^*, \boldsymbol{\Lambda}^*) + \rho \sum_{l=1}^p \boldsymbol{S}^T \boldsymbol{R}^T (\boldsymbol{R}\boldsymbol{S}\boldsymbol{\eta}^*_{.l} - \boldsymbol{\zeta}^*_{.l} + \boldsymbol{\delta}^*_{.l}) = \mathbf{0} \\ \nabla_{\Lambda_{kl}} \tilde{Q}_p(\boldsymbol{\eta}^*, \boldsymbol{\Lambda}^*) + \rho \sum_{l=1}^p \sum_{k=1}^K (\boldsymbol{\Lambda}^*_{kl} - \boldsymbol{\nu}^*_{kl} + \boldsymbol{\kappa}^*_{kl}) = \mathbf{0} \\ \boldsymbol{\zeta}^*_{.l} = \mathbf{ST}_{\lambda_1/\rho}(\boldsymbol{R}\boldsymbol{S}\boldsymbol{\eta}^*_{.l} + \boldsymbol{\delta}^*_{.l}) \\ \boldsymbol{\nu}^*_{kl} = \mathbf{VST}_{\lambda_2/\rho}(\boldsymbol{\Lambda}^*_{kl} + \boldsymbol{\kappa}^*_{kl}) \\ \boldsymbol{R}\boldsymbol{S}\boldsymbol{\eta}^*_{.l} = \boldsymbol{\zeta}^*_{.l} \\ \boldsymbol{\Lambda}^*_{kl} = \boldsymbol{\nu}^*_{kl} \end{cases} \tag{4.20}$$

The convergence property is summarized in Theorem 4.1, and the proof is given in Appendix.

Theorem 4.1. *Assume that $\psi(\cdot) : \mathcal{M} \to \mathbb{R}^p$ is an identity embedding, and function $\tilde{Q}_p(\boldsymbol{\eta}, \boldsymbol{\Lambda})$ is nonconcave with respect to $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$. Let $\{(\boldsymbol{\eta}^t, \boldsymbol{\Lambda}^t, \boldsymbol{\zeta}^t, \boldsymbol{\nu}^t, \boldsymbol{\delta}^t, \boldsymbol{\kappa}^t)\}$ be a sequence generated by Algorithm 2. Assume that $\lim_{t\to\infty} \|\boldsymbol{\delta}^{t+1} - \boldsymbol{\delta}^t\| = 0$, $\lim_{t\to\infty} \|\boldsymbol{\kappa}^{t+1} - \boldsymbol{\kappa}^t\| = 0$, and $\{(\boldsymbol{\zeta}^t, \boldsymbol{\nu}^t)\}$ are bounded, then there exists a subsequence of $\{(\boldsymbol{\eta}^t, \boldsymbol{\Lambda}^t, \boldsymbol{\zeta}^t, \boldsymbol{\nu}^t, \boldsymbol{\delta}^t, \boldsymbol{\kappa}^t)\}$ such that it converges to a KKT point satisfying Equation (4.20).*

Second, besides the convergence properties, we can further establish the consistency of penalized estimator in MGFA for fixed number of parameters $\dim(\boldsymbol{\theta})$, fixed number of factors q and fixed number of clusters K (Khalili and Chen, 2007; Städler et al., 2010). Instead of the definition of $\boldsymbol{\theta}$ in (4.5), we re-define $\boldsymbol{\theta}$ as $(\boldsymbol{\beta}^T, \mathbf{diag}(\boldsymbol{\Omega})^T, \mathbf{vec}(\boldsymbol{\eta})^T, \mathbf{vec}(\boldsymbol{\Lambda})^T)^T$, which can be treated as an embedding in Euclidean space. Here we considered two cases: (i) the parameter orderings $\boldsymbol{U}_l$ and $\boldsymbol{V}_l$ are known; (ii) consistently estimated parameter orderings are used. Let the collection of true location parameter orderings and their absolute values ordering be $\boldsymbol{W} = \{\boldsymbol{U}_l, \boldsymbol{V}_l\}_{l=1}^p$, and the estimated orderings based on the consistent estimators in MGFA without penalization be $\hat{\boldsymbol{W}} = \{\hat{\boldsymbol{U}}_l, \hat{\boldsymbol{V}}_l\}_{l=1}^p$. Denote the penalized estimator in MGFA as $\hat{\boldsymbol{\theta}}_n^W$ when $\boldsymbol{W}$ is known, and $\hat{\boldsymbol{\theta}}_n^{\hat{W}}$ when $\hat{\boldsymbol{W}}$ is used. When the true location parameter orderings and their absolute values ordering $\boldsymbol{W}$ is known, the consistency of $\hat{\boldsymbol{\theta}}_n^W$ is presented in Theorem 4.2. Proof of Theorem 4.2 is provided in Appendix.

Theorem 4.2. *Let $(\boldsymbol{x}_i, \boldsymbol{w}_i), i = 1, 2, \ldots, n$, be a random sample drawn from $f(\boldsymbol{x}|\boldsymbol{w}, \boldsymbol{\theta}_0)p(\boldsymbol{w})$. The penalty parameters satisfy $\lambda_j = O(n^{-\frac{1}{2}})$ for $j = 1, 2$, and the initial estimates $\tilde{\sigma}_l$ and $\tilde{\mu}_{kl}$ in the weights a_k^l are $\sqrt{n}$-consistent. If the true location parameter orderings and their absolute values ordering $\boldsymbol{W}$ is known, then under some mild regularity conditions (see (C1) and (C2) in Appendix), there exists a local maximizer $\hat{\boldsymbol{\theta}}_n^W$ of the penalized log-likelihood function $\log L_p(\boldsymbol{\theta})$ such that*

$$\|\hat{\boldsymbol{\theta}}_n^W - \boldsymbol{\theta}_0\|_2 = O_p(n^{-\frac{1}{2}}), \tag{4.21}$$

where $\|\cdot\|_2$ represents the L_2 norm in Euclidean space.

Before we show the consistency based on the estimated parameter ordering $\hat{\boldsymbol{W}}$, the consistency of $\hat{\boldsymbol{W}}$ is presented in the following Lemma 4.1.

Lemma 4.1. *If $\hat{\boldsymbol{\theta}}$ is a root-n consistent estimator of $\boldsymbol{\theta}$, then we have*

$$\lim_{n\to\infty} P(\hat{\boldsymbol{U}}_l = \boldsymbol{U}_l) = 1, \text{ and } \lim_{n\to\infty} P(\hat{\boldsymbol{V}}_l = \boldsymbol{V}_l) = 1, l = 1, \ldots, p. \tag{4.22}$$

The proof of Lemma 4.1 is given in Appendix. By using Lemma 4.1, we are able to extend the properties in Theorem 4.2 to the estimator based on the estimated parameter ordering $\hat{\boldsymbol{W}}$. Proof of Theorem 4.3 is provided in Appendix.

Theorem 4.3. *Let $(\boldsymbol{x}_i, \boldsymbol{w}_i), i = 1, 2, \ldots, n$, be random samples from $f(\boldsymbol{x}|\boldsymbol{w}, \boldsymbol{\theta}_0)p(\boldsymbol{w})$. The penalty parameters satisfy $\lambda_j = O(n^{-\frac{1}{2}})$ for $j = 1, 2$, and the initial estimates $\tilde{\sigma}_l$ and $\tilde{\mu}_{kl}$ in the weights a_k^l are $\sqrt{n}$-consistent. If consistently estimated parameter orderings $\hat{\boldsymbol{W}}$ are used, then under some mild regularity conditions (see (C1) and (C2) in Appendix), there exists a local maximizer $\hat{\boldsymbol{\theta}}_n^{\hat{W}}$ of the penalized log-likelihood function $\log L_p(\boldsymbol{\theta})$ such that*

$$\|\hat{\boldsymbol{\theta}}_n^{\hat{W}} - \boldsymbol{\theta}_0\|_2 = O_p(n^{-\frac{1}{2}}), \tag{4.23}$$

Third, we present the oracle property of estimator of $\boldsymbol{\zeta} = \boldsymbol{RS\eta}$ in our penalized MGFA according to two cases mentioned above. In addition, let $\mathcal{A} = \cup\{\mathcal{A}_l\}_{l=1}^p$ be the index set of nonzero values in $\boldsymbol{\zeta}$, i.e., $\mathcal{A}_l = \{(l, k), \zeta_{kl} \neq 0\}, l = 1, \ldots, p$, and $\mathcal{A}^c$ is the complement of $\mathcal{A}$. Then $\boldsymbol{\zeta}$ can be divided into two parts, the true-zero set $\boldsymbol{\zeta}_{\mathcal{A}^c}$ and the nonzero set $\boldsymbol{\zeta}_{\mathcal{A}}$. Similarly, let $\hat{\mathcal{A}}^W$ and $\hat{\mathcal{A}}^{\hat{W}}$ be the index sets of nonzero elements in $\hat{\boldsymbol{\zeta}}^W$ and $\hat{\boldsymbol{\zeta}}^{\hat{W}}$.

Theorem 4.4. *Suppose that tuning parameters satisfy $\lambda_j = o(n^{-\frac{1}{2}}), n\lambda_j \to \infty, j = 1, 2$, and the initial estimates $\tilde{\sigma}_l$ and $\tilde{\mu}_{kl}$ in the weights a_k^l are $\sqrt{n}$-consistent. If the true location parameter orderings and their absolute values ordering $\boldsymbol{W}$ is known, then under some mild regularity conditions (see (C1) and (C2) in Appendix), the penalized estimator $\hat{\boldsymbol{\zeta}}^W$ satisfies*

- *(i) (Selection Consistency) $\lim_{n\to\infty} P(\hat{\mathcal{A}}^W = \mathcal{A}) = 1$;*
- *(ii) (Asymptotic Normality) $\sqrt{n}[\hat{\boldsymbol{\zeta}}_{\mathcal{A}}^W - \boldsymbol{\zeta}_{\mathcal{A}}] \underset{d}{\to} N(\boldsymbol{0}, \boldsymbol{I}(\boldsymbol{\zeta}_{\mathcal{A}})^{-1})$, where $\boldsymbol{I}(\boldsymbol{\zeta}_{\mathcal{A}})$ is the submatrix of Fisher information matrix $\boldsymbol{I}(\boldsymbol{\theta}_0)$ corresponding to set $\mathcal{A}$.*

Theorem 4.4 states that when the location parameter orderings $\boldsymbol{W}$ is known, under mild regularity conditions, the penalized estimator enjoys selection consistency and asymptotic normality. The proof of Theorem 4.4 follows the augments in Zou (2006), Städler et al. (2010), and Tang and Song (2016), and is given in Appendix.

Theorem 4.5. *Suppose that tuning parameters satisfy $\lambda_j = o(n^{-\frac{1}{2}}), n\lambda_j \to \infty, j = 1, 2$, and the initial estimates $\tilde{\sigma}_l$ and $\tilde{\mu}_{kl}$ in the weights a_k^l are $\sqrt{n}$-consistent. If consistently estimated* *parameter orderings $\hat{\boldsymbol{W}}$ are used, then under some mild regularity conditions (see (C1) and (C2) in Appendix), the penalized estimator $\hat{\boldsymbol{\theta}}^{\hat{W}}$ satisfies*

- *(i) (Selection Consistency)* $\lim_{n\to\infty} P(\hat{\mathcal{A}}^{\hat{W}} = \mathcal{A}) = 1$*;*
- *(ii) (Asymptotic Normality)* $\sqrt{n}[\hat{\boldsymbol{\zeta}}_{\mathcal{A}}^{\hat{W}} - \boldsymbol{\zeta}_{\mathcal{A}}] \xrightarrow[d]{} N(\mathbf{0}, \boldsymbol{I}(\boldsymbol{\zeta}_{\mathcal{A}})^{-1})$*, where* $\boldsymbol{I}(\boldsymbol{\zeta}_{\mathcal{A}})$ *is the submatrix of Fisher information matrix* $\boldsymbol{I}(\boldsymbol{\theta}_0)$ *corresponding to set* $\mathcal{A}$*.*

Theorem 4.5 states that when consistently estimated parameter orderings $\hat{\boldsymbol{W}}$ are used, under mild regularity conditions, the penalized estimator still enjoys selection consistency and asymptotic normality. The proof of Theorem 4.5 is given in Appendix. The asymptotic normality for $\boldsymbol{\mu}$ can also be derived by a simple linear transformation.

4.1.5 Model selection

We use the 2-fold cross predictive log-likelihood method as our model selection criterion to select the number of factors q, the number of components K, and the penalty parameters λ_1 and λ_2 through an exhaustive search. Specifically, in the 2-fold cross predictive log-likelihood method, the original dataset is randomly partitioned into 2 equal size sub-datasets, where one sub-dataset is retained as the testing dataset, and the other is used as the training dataset. For any given $(q, K, \lambda_1, \lambda_2)$, we estimate the penalized estimator $\hat{\boldsymbol{\theta}}$ based on the training dataset, and calculate the predictive log-likelihood function $\log L(\hat{\boldsymbol{\theta}})$ based on the testing dataset. Then we estimate $\hat{\boldsymbol{\theta}}$ based on the testing dataset and calculate the predictive log-likelihood function $\log L(\hat{\boldsymbol{\theta}})$ based on the training dataset. Consequently, these two predictive log-likelihood function values can be averaged, and the optimal $(\hat{q}, \hat{K}, \hat{\lambda}_1, \hat{\lambda}_2)$ is chosen based on the largest average predictive log-likelihood value.

We use the *random* EM algorithm to compute the penalized estimator of $\boldsymbol{\theta}$, since the EM algorithm is an iterative procedure and its performance strongly depends on its starting points. For MGFA, a good initialization is crucial for calculating $\hat{\boldsymbol{\theta}}$ due to the presence of multiple local maxima of the penalized likelihood function. Specifically, for any given value of $(q, K, \lambda_1, \lambda_2)$, multiple starting points are chosen and the relevant log-likelihood functions are calculated. The initial values that have the highest log-likelihood function are used as the starting point of the EM algorithm. In simulation studies and real data analysis, the manifold K-means method (Canas et al., 2012) is used for initializing mean parameter $\boldsymbol{\mu}_m$, while the principal geodesic analysis method (Fletcher et al., 2004) is used to initialize the factor loading matrices Λ_m and the common covariance matrix $\boldsymbol{\Omega}$.

4.2 Simulation studies

In this section, we apply our MGFA model on data simulated from three different symmetric spaces, including (i) Euclidean space $\mathbb{R}^p$; (ii) sphere $\mathbb{S}^p$; and (iii) shape space. For the data in shape space, we simulated the data from ADHD-200 Corpus Callosum Shape Data (Huang et al., 2015; Cornea et al., 2017). We set the sample size $n = 100$, and the number of cluster $K = 2$. The detailed simulation settings and performance are listed in the following subsections. For each setting, we simulated $N = 200$ data sets.

4.2.1 Euclidean space $\mathbb{R}^p$

In Euclidean space, our MGFA is equivalent to the Mixture of factor analyzers (MFA). Here we compared our method with other clustering methods established in Euclidean space, including K-means, Gaussian mixture model (GMM), and our MGFA without penalization. The following mixture model is considered.

$$\sum_{k=1}^{2} \pi_k(\boldsymbol{z}_i\boldsymbol{\beta}_k)\phi(\boldsymbol{x}_i, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), i = 1, \ldots, n, \tag{4.24}$$

where $\phi(\cdot)$ is probability density function for multivariate normal distribution ($p = 100$). In each cluster, the parameter $\boldsymbol{\mu}_k = (\boldsymbol{c}_0^T, \boldsymbol{c}_1^T, \boldsymbol{c}_2^T)^T$ for $k = 1, 2$. Here $\boldsymbol{c}_0$ is a 10×1 vector which is different across clusters, $\boldsymbol{c}_1$ is a 70×1 vector shared by different clusters, and $\boldsymbol{c}_2$ is a 20×1 vector with all elements zeros. In particular, all the elements in $\boldsymbol{c}_0$ and $\boldsymbol{c}_1$ were generated from $N(0, 0.5)$. We set $\boldsymbol{z}_i = (1, z_{i,1})$ in the logistic model of mixing proportions, in which $z_{i,1}$ were independently generated from uniform $U(-1, 1)$. We also set $\boldsymbol{\beta}_1 = (1, 2)^T$

and $\boldsymbol{\beta}_2 = (-1, 1)^T$, respectively. In order to demonstrate the robustness of our MGFA, we considered two different cases for the spatial correlation structure as follows:

- Case 1: simple diagonal matrix: $\boldsymbol{\Sigma}_k = \sigma_k^2 \boldsymbol{I}_{100}$, $k = 1, 2$;
- Case 2: AR(2) model: $\boldsymbol{\Sigma}_k^{j,j'} = \sigma_k^2 (0.25k)^{|j-j'|}$, $j, j' \leq 2, k = 1, 2$
- Case 3: latent factor analysis model: $\boldsymbol{\Sigma}_k = \boldsymbol{\Lambda}_k \boldsymbol{\Lambda}_k^T + \boldsymbol{\Omega}$, $k = 1, 2$.

The scale parameters $\sigma_k, k = 1, 2$ in Cases 1 and 2 were generated from $U(0.5, 0.6)$ and $U(0.8, 1)$, respectively. In Case 2, the number of loading factors was set as $q = 2$. The latent variable $\boldsymbol{b}_{mi}$ was generated from $N(\boldsymbol{0}, \boldsymbol{I}_2)$, while the diagonal elements in $\boldsymbol{\Omega}$, independently of $\boldsymbol{b}_{mi}$, were generated from $N(0, 0.5)$. For the loading matrices $\boldsymbol{\Lambda}_k, k = 1, 2$, the elements of the first 50 rows of each matrix were independently generated from $N(1, 2)$ and $N(2, 1)$, respectively, while the elements in the rest of rows were set as zero.

For all the three cases, we fitted GMM, MGFA, and penalized MGFA to each simulated data set with unspecified correlation structure. The Rand index (RI) (Rand, 1971b) and adjusted Rand index (aRI) (Hubert and Arabie, 1985b) were used to compare the clustering results with the ground truth and to evaluate the finite sample performance of all the four models. Table 4.1 presents the simulation results for Cases 1-3. For Case 1, all the four models show excellent clustering performance, i.e, both RI and aRI are above 0.9. From Case 2 to Case 3, as the correlation structure becomes more complex, the performance of K-means, GMM and MGFA are getting poorer and poorer. In the contrast, the penalized MGFA is stable in terms of clustering performance (all above 0.85). Therefore, our proposed model is robust to the misspecification of correlation structure and shows high clustering performance in Euclidean space.

4.2.2 Sphere $\mathbb{S}^p$

In order to demonstrate the robustness of our MGFA, we applied the R package *movMF* (Banerjee et al., 2005; Hornik and Grün, 2014) to generate the data from mixtures of von

Table 4.1: Performance of K-means, GMM, MGFA, and penalized MGFA models in Cases 1-3.

		Case 1		
Cluster $\hat{K}$	K-means	GMM	MGFA	penalized MGFA
1	0	1	0	1
2	193	195	194	198
3	7	4	6	1
RI(aRI)	0.93(0.90)	0.95(0.92)	0.95(0.92)	0.98(0.97)
		Case 2		
Cluster $\hat{K}$	K-means	GMM	MGFA	penalized MGFA
1	0	6	2	1
2	183	185	189	193
3	17	9	9	6
RI(aRI)	0.87(0.83)	0.88(0.83)	0.89(0.84)	0.97(0.93)
		Case 3		
Cluster $\hat{K}$	K-means	GMM	MGFA	penalized MGFA
1	1	8	3	2
2	173	175	177	190
3	26	17	20	8
RI(aRI)	0.70(0.64)	0.74(0.69)	0.74(0.68)	0.89(0.85)

Mises-Fisher distributions (Mardia and Jupp, 2009). The probability density function of the von Mises-Fisher distribution for $\boldsymbol{x} \in S^p$ is given by:

$$f_p(\boldsymbol{x}; \underline{\boldsymbol{\mu}}, \underline{\kappa}) = C_p(\underline{\kappa}) \exp\left(\underline{\kappa}\underline{\boldsymbol{\mu}}^T \boldsymbol{x}\right), \tag{4.25}$$

where $C_p(\underline{\kappa})$ is the normalization constant, $\underline{\boldsymbol{\mu}} \in S^p$ is called the mean direction, and $\underline{\kappa} \geq 0$ is called the concentration parameter. The smaller the concentration parameter is, the more scatteredly points drop on the sphere.

We set $\boldsymbol{w}_i = (1, w_{i,1})$ and $\boldsymbol{\beta}_k, k = 1, 2$, in the logistic model of mixing proportions, in which $w_{i,1}$ were independently generated from uniform $U(-1, 1)$ and $\boldsymbol{\beta}_k = (1, -1)^T, k = 1, 2$. For the dimension of the sphere $\mathbb{S}^p$, two cases were considered: (1) 2D sphere ($p = 2$); and (2) high dimensional hyper-sphere ($p = 49$). For Case 1, the mean directions in (4.25) were set as $\underline{\boldsymbol{\mu}}_1 = (1, 0, 0)^T$ and $\underline{\boldsymbol{\mu}}_2 = (0, 1, 0)^T$, while in Case 2 $\underline{\boldsymbol{\mu}}_1 = \boldsymbol{e}_1$ and $\underline{\boldsymbol{\mu}}_2 = \boldsymbol{e}_{25}$, in which $\boldsymbol{e}_i$ is the i-

th standard basis vector in $\mathbb{R}^{50}$. Three different settings of the concentration parameters were demonstrated for both Case 1 and Case 2: In Case 1, (i) $(\underline{\kappa}_1, \underline{\kappa}_2) = (8, 10)$, (ii) $(\underline{\kappa}_1, \underline{\kappa}_2) = (5, 6)$, and (iii) $(\underline{\kappa}_1, \underline{\kappa}_2) = (2, 3)$; in Case 1, (i) $(\underline{\kappa}_1, \underline{\kappa}_2) = (18, 22)$, (ii) $(\underline{\kappa}_1, \underline{\kappa}_2) = (16, 14)$, and (iii) $(\underline{\kappa}_1, \underline{\kappa}_2) = (8, 10)$.

We fitted mixtures of von Mises-Fisher distributions (movMF) (Banerjee et al., 2005), MGFA, and penalized MGFA to the simulated data set, where movMF and MGFA were conducted in Case 1 while movMF and penalized MGFA in Case 2. In both cases, we considered three set-ups with different values of the concentration parameters ($\underline{\kappa}_1$ and $\underline{\kappa}_2$).

In order to visualize the MGFA clustering results in Case 1, we randomly chose one data set and its clustering result. In Fig. 4.2, data points from Group 1 are labeled in '$*$', while points from Group 2 are labeled in '$\bullet$'. For clustering results, points in Cluster 1 are highlighted with symbol '$\square$', while points in Cluster 2 are in circles. It can be found out that, for all three different set-ups, no matter whether points in each group are concentrate or not, most points within the same group are clustered together, which shows the excellent performance of MGFA in low dimensional cases.

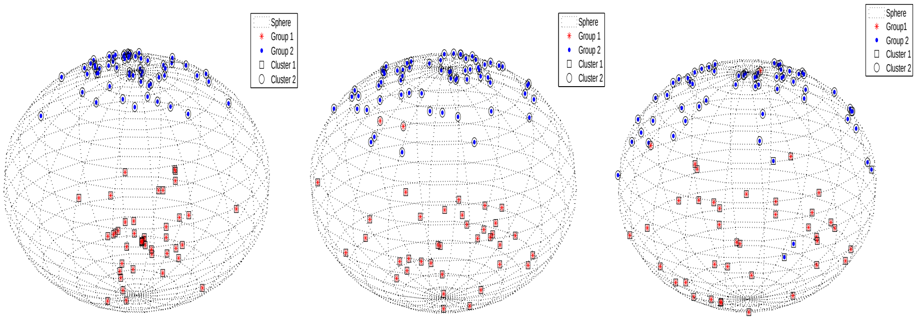


Figure 4.2: 2D Sphere data clustering for different settings of the concentration parameters.

Table 4.2 presents the simulation results corresponding to different settings of concentration parameters in both Cases 1 and 2. From Table 4.2, movMF, MGFA (Case 1), and penalized MGFA (Case 2) all perform well (Rand Index are all above 0.85) when data points from the same group are centralized, whereas the clustering performance of all three models decline when the data points belonging to the same group drop scatteredly on the sphere. Moreover, in Case 2, the performance index of movMF drops down very fast (adjusted Rand Index is below 0.1 in the third set-up) when the concentration parameters become smaller. In contrast, the performance index of penalized MGFA drops down slowly and performs much better than movMF. Therefore, even though the underlying distribution is unknown to MGFA and penalized MGFA, both MGFA and penalized MGFA perform as well as movMF when concentration parameters are of large scales, and outperform movMF when concentration parameters become smaller in Cases 1 and 2. Therefore, our proposed model is robust to the unknown underlying distribution and show excellent clustering performance in both low dimensional and high dimensional cases.

Table 4.2: Performance of movMF, MGFA, and penalized MGFA for Cases 1 and 2

		Case 1: 2D sphere ($p = 2$)		
Model	Cluster $\hat{K}$	$(\underline{\kappa}_1, \underline{\kappa}_2) = (8, 10)$	$(\underline{\kappa}_1, \underline{\kappa}_2) = (5, 6)$	$(\underline{\kappa}_1, \underline{\kappa}_2) = (2, 3)$
	1	0	1	22
movMF	2	199	187	158
	3	1	12	20
	RI(aRI)	0.982(0.943)	0.862(0.723)	0.750(0.501)
	1	0	3	13
MGFA	2	200	190	176
	3	0	7	11
	RI(aRI)	1.000(0.998)	0.888(0.776)	0.781(0.549)
		Case 2: high dimensional hyper-sphere ($p = 49$)		
Model	Cluster $\hat{K}$	$(\underline{\kappa}_1, \underline{\kappa}_2) = (18, 22)$	$(\underline{\kappa}_1, \underline{\kappa}_2) = (16, 14)$	$(\underline{\kappa}_1, \underline{\kappa}_2) = (8, 10)$
	1	8	32	73
movMF	2	177	124	65
	3	15	44	62
	RI(aRI)	0.854(0.707)	0.718(0.435)	0.526(0.052)
	1	2	21	34
penalized	2	189	160	121
MGFA	3	9	19	45
	RI(aRI)	0.922(0.883)	0.817(0.700)	0.741(0.461)

4.2.3 Shape space

For the data in shape space, we simulated the data from ADHD-200 Corpus Callosum Shape Data (Lin et al., 2017; Huang et al., 2015; Cornea et al., 2017). For comparison, along with our penalized MGFA, compared it with the mixtures of offset-normal shape (MOS) model (Kume and Welling, 2010), and penalized mixtures of offset-normal shape factor analyzers (MOSFA) (Huang et al., 2015). To show the robustness of our MGFA for shape data, we simulated CC shape data from the MOSFA, in which the offset-normal probability density function can be written as

$$f_u(\boldsymbol{u};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{|\boldsymbol{\Gamma}|^{\frac{1}{2}}\exp(-g/2)}{(2\pi)^{k-2}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\sum_{i=0}^{k-2}\binom{k-2}{i}E(l_x^{2i}|\xi_x,\sigma_x^2)E(l_y^{2k-4-2i}|\xi_y,\sigma_y^2), \tag{4.26}$$

where $\boldsymbol{\Sigma} = \Lambda\Lambda^T + \boldsymbol{\Omega}$, $\boldsymbol{\Gamma} = (\boldsymbol{W}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{W})^{-1}$, $g = vec(\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}vec(\boldsymbol{\mu}) - \boldsymbol{\nu}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\nu}$, $\boldsymbol{\nu} = \boldsymbol{\Gamma}\boldsymbol{W}^T\boldsymbol{\Sigma}^{-1}vec(\boldsymbol{\mu})$, and $(\xi_x,\xi_y)^T = \boldsymbol{\Psi}^T\boldsymbol{\nu}$, in which $\boldsymbol{\Psi}$ is the eigenvector matrix of $\boldsymbol{\Gamma}$ such that $\boldsymbol{\Gamma} = \boldsymbol{\Psi}\boldsymbol{D}\boldsymbol{\Psi}^T$ and $\boldsymbol{D} = diag(\sigma_x^2,\sigma_y^2)$. Moreover, $E(l^r|\xi,\sigma^2)$ denotes the r^{th} moment of $N(\xi,\sigma^2)$. All the simulation settings are same as those in (Huang et al., 2015). For the completeness of simulation studies, we described the settings as follows. The number of landmarks along the CC contour is 50. The contours of two randomly selected subjects (one normal control and one patient) from the ADHD-200 data were set as the mean shapes of two different clusters (see Figure 4.3). In each cluster, the landmark configuration of each subject was set as the true value of the parameter $\boldsymbol{\mu}_k$ for $k = 1, 2$. We set $\boldsymbol{z}_i = (1, z_{i,1})$ in the logistic model of mixing proportions, in which $z_{i,1}$ were independently generated from uniform $U(-1, 1)$. We also set $\boldsymbol{\beta}_1 = (1, 2)^T$ and $\boldsymbol{\beta}_2 = (-1, 1)^T$, respectively. For the factor analyzer structure, the number of loading factors was set as $q = 2$. The latent variable was generated from $N(\boldsymbol{0}, \boldsymbol{I}_2)$, while the error terms, independently of the latent variable, were generated from $N(\boldsymbol{0}, \boldsymbol{\Omega})$, where the diagonal elements in $\boldsymbol{\Omega}$ were simulated from $U(1, 2)$. For the loading matrices $\Lambda_k, k = 1, 2$, the elements of the first ℓ_0 rows of each matrix were independently generated from $N(c_1, 2)$ and $N(c_2, 1)$, respectively, while the elements in the rest of rows were set as zero.

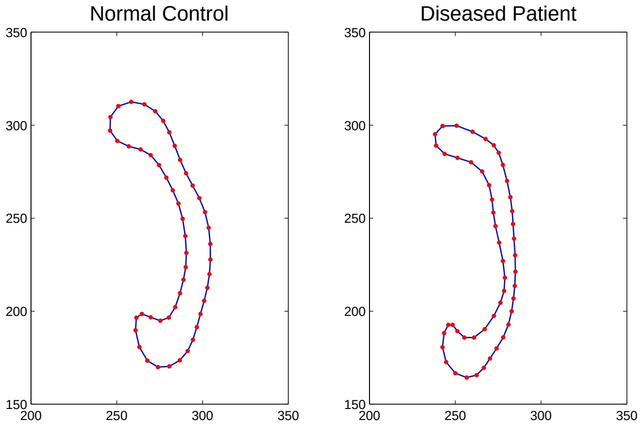


Figure 4.3: Contours and landmarks of two randomly selected subjects from the ADHD-200 data set.

Table 4.3 presents the simulation results corresponding to different values of (ℓ_0, c_1, c_2) for MOS, penalized MOSFA, and penalized MGFA. Table 4.3 shows that both MGFA and penalized MOSFA outperform MOS. Furthermore, penalized MGFA has the smallest Rand index and adjusted Rand index being larger than 0.85 for all values of ℓ_0 in the two set-ups. In contrast, penalized MOSFA performs better for $\ell_0 = 90$ in Set-up 1, whereas penalized MGFA performs better for $\ell_0 = 60$ and $\ell_0 = 90$ in Set-up 2.

4.3 Real data analysis

4.3.1 ADNI data description

Data used in the preparation of this article were obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging, National Institute of Biomedical Imaging and Bioengineering, Food and Drug Administration, private pharmaceutical companies and non-profit organizations as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression

Table 4.3: Comparison of MOS, penalized MOSFA, and penalized MGFA.

		Set-up 1: $c_1 = 2, c_2 = 1$			Set-up 2: $c_1 = 5, c_2 = 2$		
Model	Cluster $\hat{K}$	$\ell_0 = 30$	$\ell_0 = 60$	$\ell_0 = 90$	$\ell_0 = 20$	$\ell_0 = 40$	$\ell_0 = 60$
MOS	1	0	25	29	32	46	56
	2	200	172	22	139	102	31
	3	0	3	149	29	52	113
	RI(aRI)	1(1)	0.95(0.92)	0.59(0.17)	0.86(0.74)	0.76(0.54)	0.61(0.20)
penalized MOSFA	1	0	0	1	1	1	5
	2	200	199	195	198	185	169
	3	0	1	4	1	14	26
	RI(aRI)	1(1)	1(0.99)	0.98(0.93)	0.99(0.99)	0.90(0.89)	0.86(0.82)
penalized MGFA	1	0	0	2	0	2	6
	2	200	198	193	198	188	176
	3	0	2	5	2	10	18
	RI(aRI)	1(1)	1(0.99)	0.96(0.92)	0.99(0.99)	0.92(0.90)	0.88(0.85)

of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians in developing new treatments and monitoring their effectiveness, as well as lessening the time and cost of clinical trials. The principal investigator of this initiative is Michael W. Weiner, MD, at the VA Medical Center and University of California, San Francisco. ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The goal was to recruit 800 subjects, but the initial study (ADNI-1) has been followed by ADNI-GO and ADNI-2. To date, these three protocols have recruited over 1,500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow-up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org.

4.3.2 Data Processing

In this data analysis, we included 410 MRI scans from healthy controls and individuals with AD (186 AD and 224 healthy controls) from ADNI-1. The scans (from 214 men and 196 women, ages 75.88 ± 6.21 years), which were performed on a variety of 1.5 Tesla MRI scanners with protocols individualized for each scanner, include standard T1-weighted images obtained using volumetric 3-dimensional sagittal MPRAGE or equivalent protocols with varying resolutions. The typical protocol includes: repetition time = 2400 ms, inversion time = 1000 ms, flip angle = 8^o, and field of view = 24 cm, with a $256 \times 256 \times 170$ acquisition matrix in the $x-$, $y-$, and $z-$dimensions, which yields a voxel size of $1.25 \times 1.26 \times 1.2$ mm^3. We processed the MRI data by using standard steps, including anterior commissure and posterior commissure correction, skull-stripping, cerebellum removing, intensity inhomogeneity correction, segmentation, and registration. Subsequently, we carried out automatic regional labeling by labeling the template and by transferring the labels following the deformable registration of subject images. After labeling 93 ROIs, we were able to compute volumes for each of these ROIs for each subject.

We adopted a hippocampal subregional analysis package based on surface fluid registration (Shi et al., 2013; Wang et al., 2011) that uses isothermal coordinates and fluid registration to generate one-to-one hippocampal surface registration for computing the surface statistics. Then, we computed the radial distance at each vertex on the surface (p=15,000 vertices on both left and right hippocampal surfaces). In order to remove the volumetric information, we normalized the radial distance such that the radial distances of all the vertices were projected to one point lying on the high dimensional unit sphere $\mathbb{S}^{p-1}$. In the following section, we will compared our proposed penalized MGFA with some other methods in clustering the spherical data.

4.3.3 Data Analysis

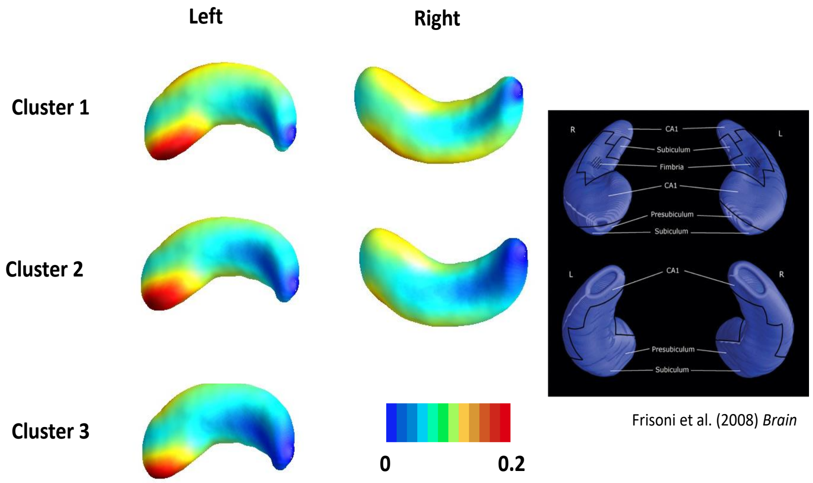
We compared our penalized MGFA model and other two methods: mixtures of von Mises-Fisher distributions (movMF) and spherical K-means (spkmeans (Buchta et al., 2012)).

In penalized MGFA, we set $\boldsymbol{w}$ =(1, Gender, Handedness, Education length, Age, RD norm), where RD norm is the norm used for radius distance normalization for each subject. The clustering results for all the three methods are summarized in Table 4.4.

Table 4.4: Comparison of movMF, spkmeans, and penalized MGFA for both left and right hippocampal surfaces.

		Left Hippocampal surface		
Model	$\hat{K}$	cluster size	NC subgroup	AD subgroup
movMF	2	(208, 202)	(153, 71) [68.3%]	(55, 131) [70.4%]
spkmeans	2	(232, 178)	(172, 52) [76.8%]	(40, 126) [67.7%]
penalized MGFA	3	(206 ,43, 161)	(198, 9, 9) [88.4%]	(8, 34, 152) [81.7%]
		Right Hippocampal surface		
	$\hat{K}$	cluster size	NC subgroup	AD subgroup
movMF	1	-	-	-
spkmeans	1	-	-	-
penalized MGFA	2	(221, 189)	(180, 45) [80.4%]	(41, 144) [77.4%]

For the left hippocampal surface, both movMF and spkmeans detect 2 clusters while our penalized MGFA detected 3 clusters. If we recalled the diagnostic information for each subjects, it can be found that our penalized MGFA outperforms other two methods in terms of consistency between clustering membership and diagnostic status. Specifically, in penalized MGFA, the first cluster contains most of normal controls (88.4%) while the third cluster contains most ADs (81.7%). For the right hippocampal surface, movMF and spkmeans fail in clustering, i.e., only one cluster was detected by either on the two method. In comparison, our penalized MGFA successfully detects 2 clusters where the first cluster includes 80.4% normal controls while the second one contains 77.4% ADs. The estimated location parameters in each cluster are presented in Figure 4.4. For both left and right hippocampal surfaces, the estimated parameters are consistent at most components across different clusters, which means the pairwise fused lasso is reasonable in this clustering task. In addition, the cluster-wise difference in estimated position parameters is consistent with the diagnosis information: most normal controls are in the first cluster, whereas most ADs are in the last cluster. To better understand the subregions where the estimated parameters are different across clusters, the cytoarchitectonic subregions mapped on blank MR-based models at 3T of the hippocampal formation (Frisoni et al., 2008) is considered here and presented in right plot of Figure 4.4. It shows that all the subregions associated with the cluster-wise difference in the estimated parameters are found in the CA1 subfield. It is interesting to note that atrophies at similar hippocampal subregions were found in AD (Frisoni et al., 2008), indicating that this finding based on our penalized MGFA is in agreement with those of previous work.



Frisoni et al. (2008) *Brain*

Figure 4.4: ADNI hippocampal surface data analysis: (left) estimated location parameters in three clusters for left hippocampal surface; (middle) estimated location parameters in two clusters for right hippocampal surface; (right) the cytoarchitectonic subregions mapped on blank MR-based models at 3T of the hippocampal formation.

4.4 Conclusions

We have developed a penalized MGFA clustering framework for clustering high-dimensional manifold data in symmetric spaces. MGFA can successfully address the major challenges including a symmetric space, a high dimensional feature space, and manifold data variation associated with some covariates. An efficient MCEM algorithm coupled with the Hamiltonian Monte Carlo algorithm has been developed to calculate the penalized MLE. Our simulations on data from diferent symmetric spaces like Euclidean space, sphere, and shape space, have confirmed that our MGFA outperforms some existing clustering methods in different scenarios. The clustering results on ADNI hippocampal surface data analysis has shown that penalized MGFA can undercover meaningful clusters which are consistent with the diagnosis information. Investigations on some other symmetric spaces, e.g., Grassmannians, and the spaces of positive-definite symmetric matrices, will be conducted in our future work.

CHAPTER 5: SURROGATE VARIABLE ANALYSIS FOR MULTIVARIATE FUNCTIONAL RESPONSES IN IMAGING DATA

5.1 Method

5.1.1 Functional latent factor regression model (FLFRM)

Suppose that we observe both the imaging data and clinical covariates from n unrelated subjects. Assumed that all the imaging data has been well registered to a common template, i.e., $\mathcal{S} \subset \mathbb{R}^d$. The template $\mathcal{S}$ includes n_v points, $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{n_v}$, which have common density $p(\boldsymbol{s})$ with support $supp(p) \subseteq \mathcal{R}^d$. For each registered image, there are J imaging measurements (or features) that have been derived, which will be treated as the functional responses later. In particular, at each point $\boldsymbol{s}_k$, the image data including J features is denoted as an $n \times J$ matrix, $\boldsymbol{y}(\boldsymbol{s}_k) = (\boldsymbol{y}_{.1}(\boldsymbol{s}_k), \ldots, \boldsymbol{y}_{.J}(\boldsymbol{s}_k))$. In addition, let $\boldsymbol{X}$ be an $n \times p$ full column rank matrix of observed covariates including the intercept. In order to build up the relationship between multivariate imaging responses and covariates of interest, a multivariate varying coefficient model (MVCM) was developed in Zhu et al. (2012):

$$\boldsymbol{y}_{.j}(\boldsymbol{s}_k) = \boldsymbol{X}\boldsymbol{\beta}_j(\boldsymbol{s}_k) + \boldsymbol{\eta}_{.j}(\boldsymbol{s}_k) + \boldsymbol{\epsilon}_{.j}(\boldsymbol{s}_k),\ j = 1, \ldots, J, \tag{5.1}$$

where $\boldsymbol{B}(\boldsymbol{s}_k) = (\boldsymbol{\beta}_1(\boldsymbol{s}_k), \ldots, \boldsymbol{\beta}_J(\boldsymbol{s}_k))$ is a $p \times J$ matrix representing the primary effect related to the observed covariates $\boldsymbol{X}$. Moreover, $\boldsymbol{\eta}(\boldsymbol{s}_k) = (\boldsymbol{\eta}_{.1}(\boldsymbol{s}_k), \ldots, \boldsymbol{\eta}_{.J}(\boldsymbol{s}_k))$ is an $n \times J$ matrix which characterizes both subject-specific and location-specific spatial variability, and $\boldsymbol{\epsilon}(\boldsymbol{s}_k) = (\boldsymbol{\epsilon}_{.1}(\boldsymbol{s}_k), \ldots, \boldsymbol{\epsilon}_{.J}(\boldsymbol{s}_k))^T$ are measurement errors. It is also assumed that the i-th row in $\boldsymbol{\eta}(\boldsymbol{s}_k)$ and that in $\boldsymbol{\epsilon}(\boldsymbol{s}_k)$ are mutually independent and identical copies of $\mathrm{SP}(\boldsymbol{0}, \boldsymbol{\Sigma}_\eta)$ and $\mathrm{SP}(\boldsymbol{0}, \boldsymbol{\Sigma}_\epsilon)$, respectively, where $\mathrm{SP}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a stochastic process vector with mean function

$\boldsymbol{\mu}(\boldsymbol{s})$ and covariance function $\boldsymbol{\Sigma}(\boldsymbol{s}, \boldsymbol{s}')$. Moreover, $\boldsymbol{\Sigma}_\epsilon(\boldsymbol{s}, \boldsymbol{s}')$ takes the form of $\boldsymbol{\Omega}_\epsilon(\boldsymbol{s})\mathbf{1}(\boldsymbol{s} = \boldsymbol{s}')$, where $\boldsymbol{\Omega}_\epsilon(\boldsymbol{s})$ is a diagonal matrix and $\mathbf{1}(\cdot)$ is the indicator function.

In MVCM, the unobserved or latent factors can be captured by the individual functions $\boldsymbol{\eta}(\boldsymbol{s})$ based on the functional PCA approach (Wang et al., 2016). Specifically, we consider a spectral decomposition of $\boldsymbol{\Sigma}_\eta(\boldsymbol{s}, \boldsymbol{s}') = (\Sigma_{\eta,jj'}(\boldsymbol{s}, \boldsymbol{s}'))$ and its approximation. According to Mercer's theorem (Mercer, 1909), if $\boldsymbol{\Sigma}_\eta(\boldsymbol{s}, \boldsymbol{s}')$ is continuous on $\mathcal{S} \times \mathcal{S}$, then $\Sigma_{\eta,jj'}(\boldsymbol{s}, \boldsymbol{s}')$ admits a spectral decomposition as

$$\Sigma_{\eta,jj'}(\boldsymbol{s}, \boldsymbol{s}') = \sum_{l=1}^{\infty} \kappa_{jl}\psi_{jl}(\boldsymbol{s})\psi_{jl}(\boldsymbol{s}')\ j = 1, \ldots, J, \tag{5.2}$$

where $\kappa_{j1} \geq \kappa_{j2} \geq \cdots \geq 0$ are ordered eigenvalues of a linear operator determined by $\Sigma_{\eta,jj}$ with $\sum_{l=1}^{\infty} \kappa_{jl} < \infty$ and the $\psi_{jl}(\boldsymbol{s})$'s are the corresponding principal components (Yao and Lee, 2006; Hall et al., 2006). Then each individual function $\boldsymbol{\eta}_i(\boldsymbol{s}) = (\eta_{ij}(\boldsymbol{s}))$ admits the Karhunen-Loeve expansion as

$$\eta_{ij}(\boldsymbol{s}) = \sum_{l=1}^{\infty} \zeta_{ijl}\psi_{jl}(\boldsymbol{s}), \tag{5.3}$$

where $\zeta_{ijl} = \int_S \eta_{ij}(\boldsymbol{s})\psi_{jl}(\boldsymbol{s})d\boldsymbol{s}$ is referred to as the jl-th functional principal component (PC) scores of the i-th subject such that $\mathbb{E}(\zeta_{ijl}) = 0$ and $\mathbb{E}(\zeta_{ijl}^2) = \kappa_{jl}$. Furthermore, all the PC scores $\{\zeta_{ijl}\}$ can be used to recover the structure of latent factors. However, according to the estimation procedure in Zhu et al. (2012), the observed covariates $\boldsymbol{X}$ are assumed to be uncorrelated with the latent factor information stored in individual functions $\boldsymbol{\eta}(\boldsymbol{s})$, which is not applicable in practice. For example, in Alzheimer's disease (AD) study, the diagnostic information is usually observed and of interest while the marital status information is usually unlikely to be included into the analysis. In fact, the association between marital status and AD has been confirmed many times in the existing literature (Helmer et al., 1999; Sundström et al., 2016; Sommerlad et al., 2018).

In order to address this issue, our FLFRM is described as below:

$$\boldsymbol{y}_{.j}(\boldsymbol{s}_k) = \boldsymbol{X}\boldsymbol{\beta}_j(\boldsymbol{s}_k) + \boldsymbol{Z}\boldsymbol{\gamma}_j(\boldsymbol{s}_k) + \boldsymbol{\eta}_{.j}(\boldsymbol{s}_k) + \boldsymbol{\epsilon}_{.j}(\boldsymbol{s}_k),\ j = 1, \ldots, J, \tag{5.4}$$

where $\boldsymbol{Z}$ is a $n \times q$ full column rank matrix of latent factors, and q is the number of latent factors, which is unknown. The $q \times J$ matrix $\boldsymbol{\Gamma}(\boldsymbol{s}_k) = (\boldsymbol{\gamma}_1(\boldsymbol{s}_k), \ldots, \boldsymbol{\gamma}_J(\boldsymbol{s}_k))$ represents the effect caused by $\boldsymbol{Z}$. Besides the distribution assumptions of $\boldsymbol{\eta}(\boldsymbol{s})$ and $\boldsymbol{\epsilon}(\boldsymbol{s})$ in FLFRM (5.4), another assumption is required here on the coefficient functions $\boldsymbol{B}(\boldsymbol{s})$ and $\boldsymbol{\Gamma}(\boldsymbol{s})$:

Assumption 5.1. *Given that $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m$ have common density $p(\boldsymbol{s})$ with support supp(p) $\subseteq \mathcal{R}^d$, the row vectors of $\boldsymbol{B}(\boldsymbol{s})$ and the row vectors of $\boldsymbol{\Gamma}(\boldsymbol{s})$ are orthogonal with respect to (w.r.t.) $p(\boldsymbol{s})$ on $\mathcal{S}$ after mean centering, i.e.,*

$$\int_{s} \boldsymbol{B}(\boldsymbol{s})(\boldsymbol{I}_J - \boldsymbol{P}_J)\boldsymbol{\Gamma}^T(\boldsymbol{s})p(\boldsymbol{s})d\boldsymbol{s} = \boldsymbol{0},$$

where $\boldsymbol{P}_J = \boldsymbol{1}_J(\boldsymbol{1}^T\boldsymbol{1}_J)^{-1}\boldsymbol{1}_J^T$, and $\boldsymbol{1}_J$ is a J-dimensional vector having each entry equal to 1.

Similar assumptions for model identification can be found in some existing methods (see Sun et al. (2012); Lee et al. (2017)). Actually, this assumption is also reasonable in practice. For example, in neuroimage data analysis, batch effects are usually caused by the study-level heterogeneity in imaging acquisition protocols. Their effect sizes would not be correlated with those of population differences (Lee et al., 2017). In the following subsections, both estimation procedure and inference procedure will be discussed, and the corresponding asymptotic properties will be investigated as well.

5.1.2 Estimation procedure

The estimation procedure can be divided into three steps here: **Step 1.** local linear kernel (LLK) smoothing on FLFRM after reparameterization; **Step 2.** singular value decomposition (SVD) on extended residual matrix; **Step 3.** bias correction of estimates in Step 1.

Step 1: LLK smoothing on FLFRM after reparameterization Our FLFRM can be reparameterized by applying the orthogonal decomposition on the matrix $\boldsymbol{Z}$ (see Figure 5.1):

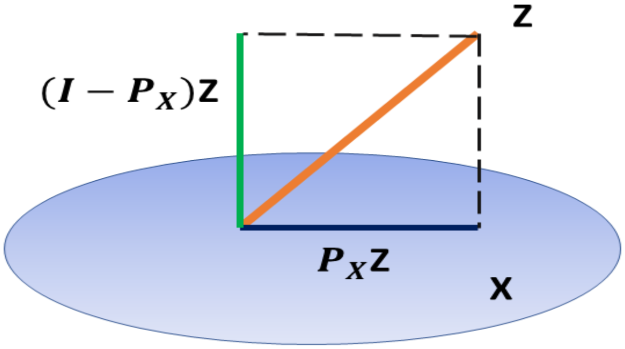


Figure 5.1: Orthogonal projection of the columns in $\boldsymbol{Z}$ onto the column space of $\boldsymbol{X}$.

$$\boldsymbol{y}_{.j}(\boldsymbol{s}_k)=\boldsymbol{X}\boldsymbol{\beta}_j^*(\boldsymbol{s}_k)+\boldsymbol{Z}^*\boldsymbol{\gamma}_j(\boldsymbol{s}_k)+\boldsymbol{\eta}_{.j}(\boldsymbol{s}_k)+\boldsymbol{\epsilon}_{.j}(\boldsymbol{s}_k),\ j=1,\ldots,J, \tag{5.5}$$

where $\boldsymbol{\beta}_j^*(\boldsymbol{s}_k)=\boldsymbol{\beta}_j(\boldsymbol{s}_k)+(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z}\boldsymbol{\gamma}_j(\boldsymbol{s}_k)$, $\boldsymbol{Z}^*=(\boldsymbol{I}_n-\boldsymbol{P}_X)\boldsymbol{Z}$, and $\boldsymbol{P}_X=\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Obviously, the columns of $\boldsymbol{X}$ are orthogonal to those of $\boldsymbol{Z}^*$. Then, given that $\boldsymbol{y}_{.j}(\boldsymbol{s}_k), j=1,\ldots,J$, and $\boldsymbol{X}$ are observed, the multivariate LLK smoothing technique (Ruppert and Wand, 1994; Fan and Gijbels, 1996; Zhang and Chen, 2007) can be applied here to derive the weighted least squares (WLS) estimator of $\boldsymbol{\beta}_j^*(\boldsymbol{s}_k)$ in (5.5). Specifically, let $K(\cdot)$ be the kernel function, and $\boldsymbol{H}_\beta$ be the bandwidth matrix, which is positive definite (e.g., a simple diagonal form). Also, denote $K_{H_\beta}(\boldsymbol{s})=|\boldsymbol{H}_\beta|^{-1}K(\boldsymbol{H}_\beta^{-1}\boldsymbol{s})$, and $\boldsymbol{z}_{H_\beta}(\boldsymbol{s}_k-\boldsymbol{s})=(1,(\boldsymbol{s}_k-\boldsymbol{s})^T\boldsymbol{H}_\beta^{-1})^T$. For each j and fixed $\boldsymbol{H}_\beta$, the WLS estimator of $\boldsymbol{\beta}_j^*(\boldsymbol{s}_k)$ is derived as

$$\hat{\boldsymbol{\beta}}_j^*(\boldsymbol{s})=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sum_{k=1}^{n_v}a_k(\boldsymbol{H}_\beta,\boldsymbol{s})\boldsymbol{y}_{.j}(\boldsymbol{s}_k), \tag{5.6}$$

where $a_k(\boldsymbol{H}_\beta,\boldsymbol{s})=(1,\boldsymbol{0}_{1\times d})[\sum_{k=1}^{n_v}K_{H_\beta}(\boldsymbol{s}_k-\boldsymbol{s})\boldsymbol{z}_{H_\beta}(\boldsymbol{s}_k-\boldsymbol{s})^{\otimes 2}]^{-1}K_{H_\beta}(\boldsymbol{s}_k-\boldsymbol{s})\boldsymbol{z}_{H_\beta}(\boldsymbol{s}_k-\boldsymbol{s})$.

Since there is no linearity assumption on the coefficient function $\boldsymbol{\beta}_j^*(\boldsymbol{s})$, the local linear smoother $\hat{\boldsymbol{\beta}}_j^*(\boldsymbol{s})$ is a biased estimator (Fan and Gijbels, 1996). To overcome this issue, a standard technique considered here is the bias correction. Following the pre-asymptotic substitution method in Fan and Gijbels (1996), the bias term can be obtained by using local cubic fit with a pilot bandwidth selected in (5.6). Furthermore, according to the definition of $\boldsymbol{\beta}_j^*(\boldsymbol{s})$, the key aim in the following two steps is to seek an estimate of $\boldsymbol{Z}\boldsymbol{\gamma}_j(\boldsymbol{s})$. Then the estimate of $\boldsymbol{\beta}_j(\boldsymbol{s})$ can be derived by subtracting the term $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\widehat{\boldsymbol{Z}\boldsymbol{\gamma}_j}(\boldsymbol{s})$ from $\hat{\boldsymbol{\beta}}_j^*(\boldsymbol{s})$.

Step 2: SVD on extended residual matrix The residual term in Step 1 is defined as

$$\boldsymbol{r}_{.j}(\boldsymbol{s}) = \boldsymbol{y}_{.j}(\boldsymbol{s}) - \boldsymbol{X}[\hat{\boldsymbol{\beta}}_j^*(\boldsymbol{s}) - \widehat{\mathbf{bias}(\hat{\boldsymbol{\beta}}_j^*(\boldsymbol{s}))}], j = 1, \ldots, J, \tag{5.7}$$

where $\widehat{\mathbf{bias}(\hat{\boldsymbol{\beta}}_j^*(\boldsymbol{s}))}$ is an estimate of the bias term in $\hat{\boldsymbol{\beta}}_j^*(\boldsymbol{s})$. Then, given $\mathcal{S}, \boldsymbol{X}$, and $\boldsymbol{Z}$, the conditional expectation of the residual term can be derived as (Ruppert and Wand, 1994):

$$\mathbb{E}[\boldsymbol{r}_{.j}(\boldsymbol{s})|\mathcal{S}, \boldsymbol{X}, \boldsymbol{Z}] = \boldsymbol{Z}^*\boldsymbol{\gamma}_j(\boldsymbol{s}) + o_p(\mathbf{Tr}(\boldsymbol{H}_\beta^2)), j = 1, \ldots, J, \tag{5.8}$$

where $\mathbf{Tr}(\cdot)$ is the trace of a given matrix. To estimate the primary term $\boldsymbol{Z}^*$ in (5.8), the SVD technique is first performed on the $n \times Jn_v$ extended residual matrix

$$\bar{\boldsymbol{R}} \doteq [\boldsymbol{r}_{.1}(\boldsymbol{s}_1), \ldots, \boldsymbol{r}_{.1}(\boldsymbol{s}_{n_v}), \ldots, \boldsymbol{r}_{.J}(\boldsymbol{s}_1), \ldots, \boldsymbol{r}_{.J}(\boldsymbol{s}_{n_v})]\,. \tag{5.9}$$

Then the corresponding SVD is denoted as $\bar{\boldsymbol{R}} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$, where the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ consist of the left and right singular vectors, respectively, and $\boldsymbol{\Lambda}$ is a diagonal matrix whose diagonal entries are the ordered singular values of $\bar{\boldsymbol{R}}$. According to the results in Lee et al. (2017), we will show that the first q columns in $\boldsymbol{U}$, $\boldsymbol{U}_{1:q}$, can be treated as an estimator of linear combinations of the columns of $\boldsymbol{Z}^*$. Then there exists a $q \times q$ orthonormal matrix $\boldsymbol{Q}$ and a function $\boldsymbol{\alpha}_j(\boldsymbol{s})$ such that

$$\boldsymbol{U}_{1:q} = (\boldsymbol{I}_n - \boldsymbol{P}_X)\boldsymbol{G} + o_p(1), \quad \boldsymbol{G} = \boldsymbol{Z}\boldsymbol{Q}, \quad \text{and} \quad \boldsymbol{\alpha}_j(\boldsymbol{s}) = \boldsymbol{Q}^T\boldsymbol{\gamma}_j(\boldsymbol{s}). \tag{5.10}$$

Step 3: bias correction of estimates in Step 1 To derive the estimate of $\boldsymbol{\alpha}_j(\boldsymbol{s})$, the residual terms in (5.7) are treated as the functional responses. Then, a new varying coefficient model is constructed via substituting the SVD results:

$$\boldsymbol{r}_{.j}(\boldsymbol{s}) = \boldsymbol{U}_{1:q}\boldsymbol{\alpha}_j(\boldsymbol{s}) + \boldsymbol{\eta}_{.j}(\boldsymbol{s}) + \boldsymbol{\epsilon}_{.j}(\boldsymbol{s}),\ j = 1, \ldots, J. \tag{5.11}$$

Similar to the derivation of $\hat{\boldsymbol{\beta}}^*_j(\boldsymbol{s})$, for each j and fixed $\boldsymbol{H}_\alpha$, the WLS estimator of $\boldsymbol{\alpha}_j(\boldsymbol{s})$ is given as

$$\hat{\boldsymbol{\alpha}}_j(\boldsymbol{s}) = \boldsymbol{U}^T_{1:q}\sum_{k=1}^{m} a_k(\boldsymbol{H}_\alpha, \boldsymbol{s})\boldsymbol{r}_{.j}(\boldsymbol{s}_k),\ j = 1, \ldots, J. \tag{5.12}$$

We further define $\widehat{\mathbf{bias}(\hat{\boldsymbol{\alpha}}_j(\boldsymbol{s}))}$ as an estimate of bias term in $\hat{\boldsymbol{\alpha}}_j(\boldsymbol{s})$, and the coefficient matrix $\boldsymbol{A}(\boldsymbol{s}) = (\boldsymbol{\alpha}_1(\boldsymbol{s}), \ldots, \boldsymbol{\alpha}_J(\boldsymbol{s}))$. Then an estimating equation can be constructed as below:

$$\boldsymbol{X}\tilde{\boldsymbol{B}}^*(\boldsymbol{s}) + \boldsymbol{U}_{1:q}\tilde{\boldsymbol{A}}(\boldsymbol{s}) = \boldsymbol{X}\boldsymbol{B}(\boldsymbol{s}) + \boldsymbol{G}\tilde{\boldsymbol{A}}(\boldsymbol{s}), \tag{5.13}$$

where $\tilde{\boldsymbol{B}}^*(\boldsymbol{s}) = \hat{\boldsymbol{B}}^*(\boldsymbol{s}) - \widehat{\mathbf{bias}(\hat{\boldsymbol{B}}^*(\boldsymbol{s}))}$ and $\tilde{\boldsymbol{A}}(\boldsymbol{s}) = \hat{\boldsymbol{A}}(\boldsymbol{s}) - \widehat{\mathbf{bias}(\hat{\boldsymbol{A}}(\boldsymbol{s}))}$. Recalling Assumption 5.1 on $\boldsymbol{B}(\boldsymbol{s})$ and $\boldsymbol{\Gamma}(\boldsymbol{s})$, we can derive the estimator of $\boldsymbol{G}$ as

$$\hat{\boldsymbol{G}} = \boldsymbol{U}_{1:q} + \boldsymbol{X}\int_s \tilde{\boldsymbol{B}}^*(\boldsymbol{s})(\boldsymbol{I}_J - \boldsymbol{P}_J)\tilde{\boldsymbol{A}}^T(\boldsymbol{s})p(\boldsymbol{s})d\boldsymbol{s}\boldsymbol{\Omega}^{-1}, \tag{5.14}$$

where $\boldsymbol{\Omega} = \int_s \tilde{\boldsymbol{A}}(\boldsymbol{s})(\boldsymbol{I}_J - \boldsymbol{P}_J)\tilde{\boldsymbol{A}}^T(\boldsymbol{s})p(\boldsymbol{s})d\boldsymbol{s}$. Since $\boldsymbol{Z}\boldsymbol{\Gamma}(\boldsymbol{s}) = \boldsymbol{G}\boldsymbol{A}(\boldsymbol{s})$, we can derive the estimator of $\boldsymbol{B}(\boldsymbol{s})$ as:

$$\hat{\boldsymbol{B}}(\boldsymbol{s}) = \hat{\boldsymbol{B}}^*(\boldsymbol{s}) - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{G}}\hat{\boldsymbol{A}}(\boldsymbol{s}). \tag{5.15}$$

5.1.3 Other issues in estimation procedure

Smoothing individual functions After the estimator of $\boldsymbol{B}(\boldsymbol{s})$ is derived and the unobserved part $\boldsymbol{Z\Gamma}(\boldsymbol{s})$ is figured out by $\hat{\boldsymbol{G}}\tilde{\boldsymbol{A}}(\boldsymbol{s})$, we consider smoothing the individual functions (rows in $\boldsymbol{\eta}(\boldsymbol{s})$, i.e., $\boldsymbol{\eta}_{i.}(\boldsymbol{s}), i = 1, \ldots, n$) based on the updated residual matrix, which can be derived as (Ruppert and Wand, 1994):

$$\hat{\boldsymbol{\eta}}(\boldsymbol{s}) = \sum_{k=1}^{m} a_k(\boldsymbol{H}_\eta, \boldsymbol{s})[\boldsymbol{y}(\boldsymbol{s}) - \boldsymbol{X}\hat{\boldsymbol{B}}(\boldsymbol{s}) - \hat{\boldsymbol{G}}\hat{\boldsymbol{A}}(\boldsymbol{s})], \tag{5.16}$$

where $\boldsymbol{H}_\eta$ is the fixed bandwidth matrix. Furthermore, their empirical covariance

$$\hat{\boldsymbol{\Sigma}}_\eta(\boldsymbol{s}, \boldsymbol{s}') = \frac{1}{n-p-q}\hat{\boldsymbol{\eta}}_{i.}(\boldsymbol{s})\hat{\boldsymbol{\eta}}_{i.}^T(\boldsymbol{s}')$$

can be straightforwardly used to estimate $\boldsymbol{\Sigma}_\eta(\boldsymbol{s}, \boldsymbol{s}')$.

Bandwidth Selection To select the optimal bandwidth in $\hat{\boldsymbol{B}}(\boldsymbol{s})$ and $\hat{\boldsymbol{A}}(\boldsymbol{s})$, we use the leave-one-curve out cross-validation (CV), while for the optimal bandwidth in $\hat{\boldsymbol{\eta}}(\boldsymbol{s})$, we use the generalized cross validation (GCV) score method. Readers interested in the details for deriving the CV and GCV scores can refer to Zhang and Chen (2007) and Zhu et al. (2012). In practice, we standardize all covariates to have mean zero and standard deviation one, and also standardize all the features in functional response data to a comparable scale. Then we can choose a common bandwidth for all covariates and features.

Determining the number of latent factors In SVD representation (5.10), the number of latent factors, q, is unknown and required to estimate. In order to obtain the estimator, four different kinds of methods are considered here: permutation version of the parallel analysis (PA) (Buja and Eyuboglu, 1992), analytical-asymptotic (AA) approach (Johnstone, 2001; Leek, 2011), eigenvalue difference (ED) method (Onatski, 2010), and bi-cross-validation (BCV) method (Owen et al., 2016). We will compare all the four different methods in the simulation studies and the one with both high detection accuracy and less computation time will be adopted in the rest data analysis.

5.1.4 Inference procedure

In this section, we study global tests for linear hypotheses of coefficient functions and simultaneous confidence bands for each varying coefficient function. They are essential for statistical inference on the coefficient functions.

Hypothesis testing We consider the linear hypothesis on $\boldsymbol{B}(\boldsymbol{s})$ as below:

$$\mathbf{H_0} : \boldsymbol{C}\mathbf{vec}(\boldsymbol{B}(\boldsymbol{s})) = \boldsymbol{b}_0(\boldsymbol{s}) \text{ for all } \boldsymbol{s} \ \text{ vs. } \ \mathbf{H_1} : \boldsymbol{C}\mathbf{vec}(\boldsymbol{B}(\boldsymbol{s})) \neq \boldsymbol{b}_0(\boldsymbol{s}), \tag{5.17}$$

where $\boldsymbol{C}$ is a $r \times Jp$ matrix with rank r and $\boldsymbol{b}_0(\boldsymbol{s})$ is a $r \times 1$ vector of functions. The global test statistic T_n is defined as:

$$T_n = \int_s T_n(\boldsymbol{s})p(\boldsymbol{s})d\boldsymbol{s}, \ \ T_n(\boldsymbol{s}) = \boldsymbol{\delta}^T(\boldsymbol{s})[\boldsymbol{C}(\hat{\boldsymbol{\Sigma}}_\eta(\boldsymbol{s},\boldsymbol{s}) \otimes [\hat{\boldsymbol{M}}\hat{\boldsymbol{M}}^T])\boldsymbol{C}^T]^{-1}\boldsymbol{\delta}(\boldsymbol{s}), \tag{5.18}$$

where $\boldsymbol{\delta}(\boldsymbol{s}) = \boldsymbol{C}\mathbf{vec}(\hat{\boldsymbol{B}}(\boldsymbol{s})) - \boldsymbol{b}_0(\boldsymbol{s})$, $\hat{\boldsymbol{M}} = (\boldsymbol{I}_p, \mathbf{0}_{q\times q})(\hat{\boldsymbol{W}}^T\hat{\boldsymbol{W}})^{-1}\hat{\boldsymbol{W}}^T$, and $\hat{\boldsymbol{W}} = [\boldsymbol{X}, \hat{\boldsymbol{G}}]$.

As the asymptotic distribution of T_n under $\mathbf{H_0}$ is quite complicated, it is difficult to derive the percentiles of T_n directly from the asymptotic result. To address this issue, the wild bootstrap method is developed here (Zhu et al., 2012), including the following four steps:

1. Fit the FLFRM under $\mathbf{H_0}$ on $\boldsymbol{X}$ and $\boldsymbol{y}(\boldsymbol{s}_k), k = 1, \ldots, n_v$, which yields $\hat{\boldsymbol{G}}$, $\hat{\boldsymbol{A}}(\boldsymbol{s})$, $\hat{\boldsymbol{B}}(\boldsymbol{s})$, $\hat{\boldsymbol{\eta}}(\boldsymbol{s})$, $\hat{\boldsymbol{\epsilon}}(\boldsymbol{s})$, and the global test statistic T_n;

2. Generate random vectors $\boldsymbol{\tau}_i^{(m)}$ and $\boldsymbol{\tau}_i^{(m)}(\boldsymbol{s}_k)$ independently from the standard normal distribution $N(\mathbf{0}, \boldsymbol{I}_\mathbf{n})$ for $k = 1, \ldots, n_v$, and then construct

$$\boldsymbol{y}^{(m)}(\boldsymbol{s}_k) = \boldsymbol{X}\hat{\boldsymbol{B}}(\boldsymbol{s}_k) + \hat{\boldsymbol{G}}\hat{\boldsymbol{A}}(\boldsymbol{s}_k) + \mathbf{diag}(\boldsymbol{\tau}_i^{(m)})\hat{\boldsymbol{\eta}}(\boldsymbol{s}_k) + \mathbf{diag}(\boldsymbol{\tau}_i^{(m)}(\boldsymbol{s}_k))\hat{\boldsymbol{\epsilon}}(\boldsymbol{s}_k),$$

where $\mathbf{diag}(\boldsymbol{a})$ denotes a diagonal matrix with the vector $\boldsymbol{a}$ lying on the diagonal;

3. Based on $\boldsymbol{X}$ and $\{\boldsymbol{y}^{(m)}(\boldsymbol{s}_k)\}_{k=1}^{n_v}$, recalculate $\hat{\boldsymbol{B}}^{(m)}(\boldsymbol{s})$ and the global test statistic $T_n^{(m)}$;

4. Repeat the previous two steps M times to obtain $\{T_n^{(1)}, \ldots, T_n^{(M)}\}$, which yields the p-value
$$p = \sum_{m=1}^{M} \mathbf{1}(T_n^{(m)} > T_n).$$

Simultaneous confidence bands Construction of simultaneous confidence bands for coefficient functions is also of great interest in statistical inference for FLFRM (5.4). For a given confidence level α, we construct the $1-\alpha$ simultaneous confidence band for $\beta_{tj}(\boldsymbol{s})$ is given by

$$\left(\hat{\beta}_{tj}(\boldsymbol{s}) - \frac{C_{tj}(\alpha)}{\sqrt{n}},\ \hat{\beta}_{tj}(\boldsymbol{s}) + \frac{C_{tj}(\alpha)}{\sqrt{n}}\right), 1 \leq t \leq p, 1 \leq j \leq J, \tag{5.19}$$

where $C_{tj}(\alpha)$ is a scalar, which is to be determined. Here an efficient resampling method is developed to approximate $C_{tj}(\alpha)$ as follows (Kosorok, 2003; Zhu et al., 2007, 2012):

1. Fit the FLFRM on $\boldsymbol{X}$ and $\boldsymbol{y}(\boldsymbol{s}_k), k = 1, \ldots, n_v$, which yields the updated residuals $\boldsymbol{\nu}_{\cdot j}(\boldsymbol{s}) = \boldsymbol{y}(\boldsymbol{s}) - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{s}) + \hat{\boldsymbol{G}}\hat{\boldsymbol{\alpha}}(\boldsymbol{s}), j = 1, \ldots, J$;

2. Generate the random vector $\boldsymbol{\tau}_i^{(m)}$ from the standard normal distribution $N(\mathbf{0}, \boldsymbol{I}_\mathbf{n})$, and then construct
$$\omega_{tj}^{(m)}(\boldsymbol{s}) = \sqrt{n}\boldsymbol{e}_t^T \hat{\boldsymbol{M}} \mathbf{diag}(\boldsymbol{\tau}_i^{(m)}) \sum_{k=1}^{n_v} a_k(\boldsymbol{H}, \boldsymbol{s}) \boldsymbol{\nu}_{\cdot j}(\boldsymbol{s}_k), j = 1, \ldots, J,$$
where $\boldsymbol{e}_t$ is a $p \times 1$ vector with the t-th element 1 and 0 otherwise;

3. Repeat the previous step M times to obtain $\{\sup_s |\omega_{tj}^{(1)}(\boldsymbol{s})|, \ldots, \sup_s |\omega_{tj}^{(M)}(\boldsymbol{s})|\}$, and use their $1-\alpha$ empirical percentile to estimate $C_{tj}(\alpha)$.

5.2 Asymptotic properties

We systematically investigate the asymptotic properties of all estimators proposed in Section 5.1.2 and several inference procedures in Section 5.1.4.

5.2.1 Assumptions

Throughout the paper, the following assumptions are used to facilitate the technical details. Some of the assumptions might be weakened but the current version simplifies the proof.

(A.1) Both n and n_v converge to ∞ with $n_v/n \to \infty$. Furthermore, let $| \cdot |_D = \mathbf{det}(\cdot)$, where $\mathbf{det}(\cdot)$ is the determinant of some given matrix. Then $|\boldsymbol{H}_\beta|_D = o(1)$, $n_v|\boldsymbol{H}_\beta|_D \to \infty$, and $|\boldsymbol{H}_\beta|_D^{-1}|\log(|\boldsymbol{H}_\beta|_D)|^{1-2/l} \leq n_v^{1-2/l}$ for $l \in (2, 4)$.

(A.2) The common density function $p(\boldsymbol{s})$ has a continuous second-order derivative and bounded support $supp(p)$. Moreover, for some $p_l > 0$ and $p_u < \infty$, $p_l < p(\boldsymbol{s}) < p_u$ for all $\boldsymbol{s} \in supp(p)$.

(A.3) Let the row vectors in $\boldsymbol{X}$ and those in $\boldsymbol{Z}$ are respectively independently and identically distributed, where both $||\boldsymbol{X}||_\infty$ and $||\boldsymbol{Z}||_\infty$ are almost surely bounded. Furthermore, let $\boldsymbol{W} = (\boldsymbol{X}, \boldsymbol{Z})$ be the matrix with $p + q$ columns formed by concatenating $\boldsymbol{X}$ and $\boldsymbol{Z}$, then $\boldsymbol{W}^T\boldsymbol{W}$ is nonsingular.

(A.4) Let ψ_l is the l-th largest singular value of $(\boldsymbol{I}_n - \boldsymbol{P}_X)\boldsymbol{Z}\bar{\boldsymbol{\Gamma}}$, where

$$\bar{\boldsymbol{\Gamma}} = [\boldsymbol{\gamma}_{.1}(\boldsymbol{s}_1), \ldots, \boldsymbol{\gamma}_{.1}(\boldsymbol{s}_{n_v}), \ldots, \boldsymbol{\gamma}_{.J}(\boldsymbol{s}_1), \ldots, \boldsymbol{\gamma}_{.J}(\boldsymbol{s}_{n_v})] \, .$$

Then $\psi_l = O(\psi_{l'})$, $\psi_{l'} = O(\psi_l)$, and $n_v^{-1/2}\psi_l \to \infty$ for $1 \leq l,\, l' \leq q$.

(A.5) Let σ_{jk}^2 be the variance of $\eta_{ij}(\boldsymbol{s}_k) + \epsilon_{ij}(\boldsymbol{s}_k)$ for $1 \leq j \leq J, 1 \leq k \leq n_v$, and $f(t) = n_v^{-1}\sum_{j,k}(\sigma_{jk}^2 - \bar{\sigma^2})^2$, where $\bar{\sigma^2} = n_v^{-1}\sum_{j,k}\sigma_{jk}^2$. Then either of the following is satisfied: (i) $f(2) = o(n^{-2}n_v)$; or (ii) $f(2) = o(n^{-3/2}n_v), f(4) = O(1)$, and $f(4) = o(n^{-4}n_v^3)$.

(A.6) The kernel function $K(t)$ is a symmetric density function with a bounded support $supp(p)$, and is Lipschitz continuous. Moreover, $\inf_{|H|<h_0, s\in S} |\boldsymbol{\Omega}_K(\boldsymbol{H}, \boldsymbol{s})|_D$ is above 0 for some

small scalar h_0, where $\boldsymbol{\Omega}_K(\boldsymbol{H}, \boldsymbol{s}) = \int_S K_H(\boldsymbol{u}-\boldsymbol{s})\boldsymbol{z}_H(\boldsymbol{u}-\boldsymbol{s})^{\otimes 2}p(\boldsymbol{u})d\boldsymbol{u}$.

(A.7) All components of $\boldsymbol{B}(\boldsymbol{s})$ and $\boldsymbol{\Gamma}(\boldsymbol{s})$ have continuous second derivatives on $\mathcal{S}$.

(A.8) For all the elements in $\boldsymbol{\epsilon}(\boldsymbol{s})$, $\mathbb{E}[\sup_s |\epsilon_{ij}(\boldsymbol{s})|^l] < \infty$ for some $l > 4$.

(A.9) Each component of $\{\boldsymbol{\eta}(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{S}\}$, $\{\boldsymbol{\eta}(\boldsymbol{s})\boldsymbol{\eta}^T(\boldsymbol{s}') : (\boldsymbol{s}, \boldsymbol{s}') \in \mathcal{S}^2\}$, and $\{\boldsymbol{W}^T\boldsymbol{\eta}(\boldsymbol{s}') : (\boldsymbol{s} \in \mathcal{S}\}$ are Donsker classes.

(A.10) $|\boldsymbol{H}_\eta|_D = o(1)$, $n_v|\boldsymbol{H}_\eta|_D \to \infty$, and $|\boldsymbol{H}_\eta|_D^{-4}(\log(n)/n)^{1-2/t} = o(1)$ for $t \in (2, \infty)$.

(A.11) The sample path of $\eta_{ij}(\boldsymbol{s})$ has continuous second-order derivative on $\mathcal{S}$ and

$$\mathbb{E}[\sup_{s\in S} \|\boldsymbol{\eta}(\boldsymbol{s})\|_2^{r_1}] < \infty,\ \mathbb{E}\{\sup_{s\in S} \|[\dot{\boldsymbol{\eta}}(\boldsymbol{s})\|_2 + \|\ddot{\boldsymbol{\eta}}(\boldsymbol{s})\|_2]^{r_2}\} < \infty$$

for some $r_1, r_2 \in (2, \infty)$, where $\| \cdot \|_2$ is the Euclidean norm.

5.2.2 Asymptotics of estimation procedure

The following theorem tackles the theoretical properties of $\hat{\boldsymbol{B}}(\boldsymbol{s})$ and $\hat{\boldsymbol{G}}$. The detailed proofs can be found in the appendix.

Theorem 5.1. *Under Assumptions 5.1 and A.1-A.9, we have the following results:*

- *(i) The columns of $\hat{\boldsymbol{G}}$ span the same column space as the columns of $\boldsymbol{Z}$ in probability.*
- *(ii)$\sqrt{n}\{[\boldsymbol{I}_J \otimes [\hat{\boldsymbol{M}}\hat{\boldsymbol{M}}^T]^{-\frac{1}{2}}]\mathbf{vec}(\hat{\boldsymbol{B}}(\boldsymbol{s}) - \mathbb{E}[\hat{\boldsymbol{B}}(\boldsymbol{s})])|\boldsymbol{s} \in \mathcal{S}\}$ weakly converges to a centered Gaussian process with covariance matrix $\boldsymbol{\Sigma}_\eta(\boldsymbol{s}, \boldsymbol{s}) \otimes \boldsymbol{I}_p$.*

5.2.3 Asymptotics of inference procedure

The following theorem derives the asymptotic distribution of global test statistic T_n under the null hypothesis and its asymptotic power under local alternative hypotheses.

Theorem 5.2. *Under Assumptions 5.1 and A.1-A.11, we have the following results:*

- *(i) $T_n \to \int_s \boldsymbol{\xi}(\boldsymbol{s})^T\boldsymbol{\xi}(\boldsymbol{s})d\boldsymbol{s}$ under the null hypothesis $\mathbf{H_0}$, where $\boldsymbol{\xi}(\boldsymbol{s})$ is a centered Gaussian process with covariance function.*

- *(ii)* $\mathbb{P}\{T_n > T_{n,\alpha}|\mathbf{H_{1n}}\} \to \mathbf{1}$ *as* $n \to \infty$ *for a sequence of local alternatives* $\mathbf{H_{1n}}$: $\boldsymbol{C}\mathbf{vec}(\boldsymbol{B}(\boldsymbol{s})) - \boldsymbol{b}_0(\boldsymbol{s}) = n^{-\tau/2}\boldsymbol{\delta}(\boldsymbol{s})$*, where* τ *is any scalar in* $[0,1)$*,* $T_{n,\alpha}$ *is the upper 100*α *percentile of* T_n *under* $\mathbf{H_0}$*, and* $0 < \| \int_s \boldsymbol{\delta}(\boldsymbol{s})d\boldsymbol{s}\| < \infty$*.*

5.3 Simulation studies

We first assessed the ability of our methodology in a toy example using synthetic curve data, which we generated from the following model:

$$y_{ij}(s_k) = \boldsymbol{x}_i^T\boldsymbol{\beta}_j(s_k) + z_i\gamma_j(s_k) + \eta_{ij}(s_k) + \epsilon_{ij}(s_k),\ j = 1, 2, \tag{5.20}$$

where $s_1 = 0 \leq s_2 \leq \cdot \leq s_m = 1$, and $s_k \sim U(0,1), k = 2, \ldots, m-1$. For the observed predictors, $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, x_{i3})$, where $x_{i1} \sim Bernoulli(0.5)$, $(x_{i2}, x_{i3})^T \sim N((0,0)^T, \boldsymbol{I}_2), i = 1, \ldots, n$. For the latent factors, z_i was constructed as follows:

$$z_i = \boldsymbol{x}_i^T\boldsymbol{\alpha} + \omega_i,\ \omega_i \sim N(0,1),\ i = 1, \ldots, n, \tag{5.21}$$

where $\boldsymbol{\alpha}$ is 4-dimensional vector to be determined later. For the random effect, $\eta_{ij}(s)$ admits the KarhunenLoeve expansion as $\eta_{ij} = \xi_{ij1}\psi_{j1}(s) + \xi_{ij2}\psi_{j2}(s)$, where $\psi_{jl}(s)$ are the eigen functions and $\xi_{ijl} \sim N(0, 0.5)$ for $j = 1, 2, l = 1, 2$. For the measurement error, $(\epsilon_{i,1}, \epsilon_{i,2})^T \sim N((0,0)^T, 0.5 * diag(\sigma_1^2, \sigma_2^2))$, where $\sigma_l^2 \sim InvGamma(10, 9)$ for $l = 1, 2$. Furthermore, we assume that $s_k, x_{i1}, x_{i2}, x_{i3}, \omega_i, \xi_{i11}, \xi_{i12}, \xi_{i21}, \xi_{i22}, \sigma_1^2$ and σ_2^2 are independent random variables. Also, we set the functional coefficients and eigenfunctions as belows:

$$\boldsymbol{\beta}_1(s) = (3s^2, 3(1-s)^2, 6s(1-s), -s^2)^T,$$
$$\boldsymbol{\beta}_2(s) = (12(s-0.5)^2, 1.5\sqrt{s}, 3s^2, -\tfrac{2}{3}s)^T,$$
$$\gamma_1(s) = -\sqrt{2}\sin(\pi s),\ \gamma_2(s) = \sqrt{2}\cos(2\pi s),$$
$$\psi_{11}(s) = 0.5,\ \psi_{12}(s) = s - 0.5,\ \psi_{21}(s) = 2s - 1,\ \psi_{22}(s) = 1.$$

Throughout this example, we set the sample size $n = 50$, and the number of location points $m = 2000$.

Here we compared our FLFRM with the two other methods, i.e., CATE (Wang et al., 2017) and MVCM (Zhu et al., 2012), where the curved data was treated as multivariate responses and the R-package **CATE** was adopted when implementing CATE method. Four different simulation scenarios are considered on the parameter $\boldsymbol{\alpha} \doteq (u_1(2b_1-1), u_2(2b_2-1), u_3(2b_3-1), u_4(2b_4-1))^T$, where $\{b_i\}$ are i.i.d. generated from $Bernoulli(0.5)$: (i) $u_l = 0$; (ii) $u_l \sim U(0, 0.2)$; (iii) $u_l \sim U(0.2, 0.5)$; and (iv) $u_l \sim U(0.5, 1)$ for $l = 1, 2, 3, 4$. These four scenarios indicate that latent factors $\boldsymbol{Z}$ are (i) independent with $\boldsymbol{X}$, (ii) weakly correlated with $\boldsymbol{X}$, (iii) moderately correlated with $\boldsymbol{X}$, and (iv) highly correlated with $\boldsymbol{X}$, respectively. For each simulation scenario, 200 datasets were generated and the performance of each method was evaluated based on the integrated square error (ISE) on the estimation of $\boldsymbol{B}(s)$: $\sum_{j=1}^{2}\int_0^1 ||\hat{\boldsymbol{\beta}}_j(s) - \boldsymbol{\beta}_j(s)||^2 ds$, where $\hat{\boldsymbol{\beta}}_j(s)$ is the estimator of $\boldsymbol{\beta}_j(s)$. In CATE and our FLFRM, the number of latent factors needs to be estimated. Specifically, for fairly comparison, the eigenvalue difference (ED) method (Onatski, 2010) was considered in both CATE and our FLFRM. In particular, the EV method estimates the number of factors as

$$\hat{q} = \max\{j < q_{max} : \lambda_j^2 - \lambda_{j+1}^2 \geq \Delta_0\},$$

where asymptotically q_{max} should be a slowly increasing function of n (which is fixed as 20 here), λ_j is the ordered singular values, and Δ_0 is calculated via a calibration method described in Onatski (2010). If $\{j < q_{max} : \lambda_j^2 - \lambda_{j+1}^2 \geq \Delta_0\}$ is empty, then $\hat{q} = 0$. The comparisons among CATE, MVCM and FLFRM for all these scenarios are presented via the boxplots in Figure 5.2.

According to the boxplots in Figure 5.2, it can be found that: ① different from MVCM, the performance for both CATE and FLFRM is stable, which is not affected too much by the correlation between $\boldsymbol{X}$ and $\boldsymbol{Z}$; ② our FLFRM outperforms CATE for all the four different

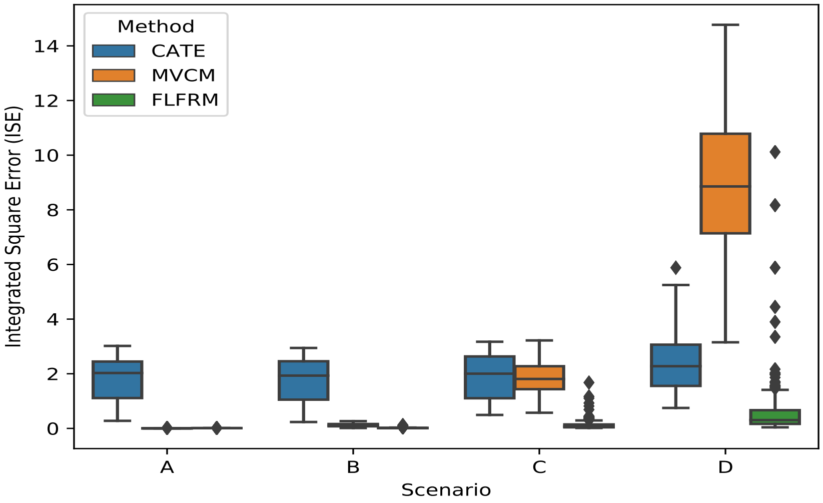


Figure 5.2: Simulation results for comparisons among CATE, MVCM and FLFRM on synthetic curve data in terms of ISE. Four scenarios were considered: the latent factors $\boldsymbol{Z}$ are **(A)** indepedent with $\boldsymbol{X}$, **(B)** weakly correlated with $\boldsymbol{X}$, **(C)** moderately correlated with $\boldsymbol{X}$, and **(D)** highly correlated with $\boldsymbol{X}$, respectively.

scenarios; ③ when the latent factors are independent with $\boldsymbol{X}$, the estimation performance of FLFRM and MVCM is almost the same; ④ when the correlation is getting higher, the performance of MVCM becomes much worse in terms of both mean ISE and standard deviation of ISE, which means our FLFRM shows advantages when the unobserved factors exist and correlated with the observed ones.

In our SVD representation (5.10), the number of latent factors, q, is required to estimate. As described previously, we derived the estimator of q based on the ED method. To check whether the ED method is a reasonable one, we would like to compare it with other three methods, i.e., permutation version of the parallel analysis (PA) (Buja and Eyuboglu, 1992), analytical-asymptotic (AA) approach (Johnstone, 2001; Leek, 2011), and bi-cross-validation (BCV) method (Owen et al., 2016). The estimation results for all the four methods are reported in Table 5.1. It can be found that, PA approach, ED method and BCV method can achieve almost 100 percent estimation accuracy. Also all the three methods outperform the analytical-asymptotic approach, which is with low estimation accuracy around 30%. In addition, in terms of average computation time, ED method ($0.8s$ on one dataset) is much more efficient than BCV method ($10s$ on one dataset) and PA approach ($70s$ on one dataset). Thus, the choice of ED method for esitmating the number of latent factors is reasonable here.

Table 5.1: Comparison of four different approaches to estimate the number of latent factors: $q = 1$. Four scenarios were considered: the latent factors $\boldsymbol{Z}$ are **(A)** indepedent with $\boldsymbol{X}$, **(B)** weakly correlated with $\boldsymbol{X}$, **(C)** moderately correlated with $\boldsymbol{X}$, and **(D)** highly correlated with $\boldsymbol{X}$, respectively.

Method	Scenario			
	A	B	C	D
PA	190/200	191/200	192/200	191/200
AA	62/200	65/200	64/200	64/200
ED	200/200	200/200	198/200	198/200
BCV	200/200	196/200	196/200	196/200

Although the great performance of our FLFRM was shown above, it should be noted that there are some outliers in terms of ISE, especially when $\boldsymbol{Z}$ and $\boldsymbol{X}$ are highly correlated. Actually, some of these outliers were caused by the failures in detection the number of latent factors, q. Therefore, it is important to investigate the sensitivity of our FLFRM with respect to the misspecification of q. We reconsidered the four scenarios and tried different choices of q in FLFRM. In particular, $q = 1$ is the true value, and $q = 2$ was also taken into account. The ISE for each choice of q on all the simulated datasets were shown in Figure 5.3.

According to the boxplots in Figure 5.3, we can conclude two findings: ① when the latent factor $\boldsymbol{Z}$ is indepedent or weakly correlated with the observed ones, if q is misspecified, the average performance of our FLFRM is still relatively stable in term of ISE. However, the performance variability is increasing in term of the standard deviation of ISE; ② when $\boldsymbol{Z}$ is moderately or even highly correlated with the observed ones: if q is misspecified as 2, the average performance of our FLFRM is somehow similar to the one when $q = 1$ in terms of ISE, while the performance variability is getting higher in terms of standard deviation of ISE.

In the estimation procedure, we are also interested in the estimated latent factors. As claimed in Section 5.2, under certain assumptions, the columns of detected latent factors span

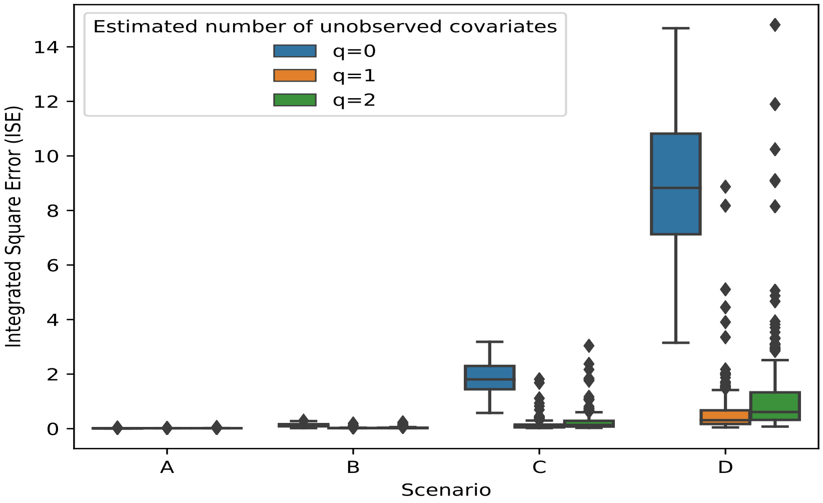


Figure 5.3: Simulation results for FLFRM with different choice of q on synthetic curve data. Four scenarios were considered: the latent factor $\boldsymbol{Z}$ is **(A)** indepedent with $\boldsymbol{X}$, **(B)** weakly correlated with $\boldsymbol{X}$, **(C)** moderately correlated with $\boldsymbol{X}$, and **(D)** highly correlated with $\boldsymbol{X}$, respectively.

the same column space as the columns of $\boldsymbol{Z}$ in probability. Here we would like to validate this asymptotic property via simulation studies. For all the four scenarios, the absolute values of Pearson correlation coefficient between estimated latent factors and $\boldsymbol{Z}$ were calculated are plotted in Figure 5.4. According to the results, the absolute values of Pearson correlation coefficient for all the four scenarios are above 0.9, which indicates the consistency between the column space of detected factors and that of the true one. In addition, when the correlation between $\boldsymbol{Z}$ and $\boldsymbol{X}$ is getting higher, the absolute values of Pearson correlation coefficient are closer to 1.

Next, we only focused on Scenario **(C)**, where $\boldsymbol{\alpha} = (u_1, -u_2, u_3, -u_4)^T$, $u_l \sim U(0.2, 0.5)$. Then, except for $\beta_{14}(s)$ and $\beta_{24}(s)$ for all s, all other parameters were fixed at the values specified above, whereas we assumed $\beta_{14}(s) = -cs^2$, $\beta_{24}(s) = -\frac{2c}{3}s$, where c is a scalar specified later. We want to test the hypotheses

$$\mathbf{H_0} : \beta_{14}(s) = \beta_{24}(s) = 0 \text{ for all } s, \; \mathbf{H_1} : \beta_{14}(s) \neq 0 \text{ or } \beta_{24}(s) \neq 0 \text{ for at lease one } s. \quad (5.22)$$

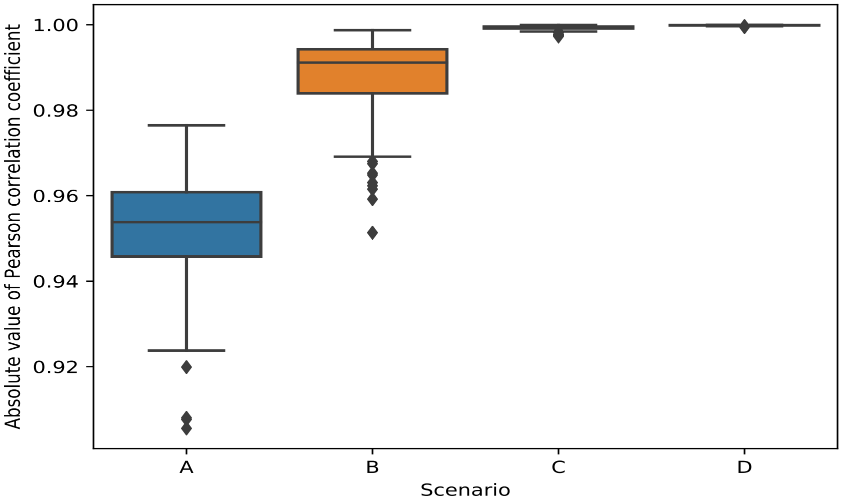


Figure 5.4: Correlation between estimated latent factors and $\boldsymbol{Z}$ on synthetic curve data. Four scenarios were considered: the latent factor $\boldsymbol{Z}$ is **(A)** indepedent with $\boldsymbol{X}$, **(B)** weakly correlated with $\boldsymbol{X}$, **(C)** moderately correlated with $\boldsymbol{X}$, and **(D)** highly correlated with $\boldsymbol{X}$, respectively.

We set $c = 0$ to assess the type I error rates for T_n, and set $c = 0.1, 0.2, 0.3, 0.4$, and 0.5 to examine the power of T_n. We set the sample size as $n = 100$ and 200. For each situation, the significance levels were set at $\alpha = 0.05$ and 0.01, and 500 bootstrap replications were generated to constructed the empirical distribution of T_n under $\mathbf{H_0}$. Figure 5.5 depicts the power curves. It can be seen that the rejection rates for T_n based on the wild bootstrap method are accurate for moderate sample sizes, such as ($n = 100$ or 200) at both significance levels ($\alpha = 0.01$ or 0.05). As expected, the power increases with the sample size.

Final, we considered the coverage probabilities of simultaneous confidence bands of the functional coefficients $\boldsymbol{B}(s)$ based on the resampling method. Here we still focused on Scenario **(C)**. In particular, the number of grid was set as $n_n = 200$ and 2000, and all other parameters were fixed at the values specified above. Based on the generated data, we calculated the simultaneous confidence bands for each component in $\boldsymbol{B}(s)$, where 200 replications were generated for the band construction. Table 5.2 summarizes the empirical coverage probabilities for $\alpha = 0.05$ and 0.01. It can be found that the coverage probabilities

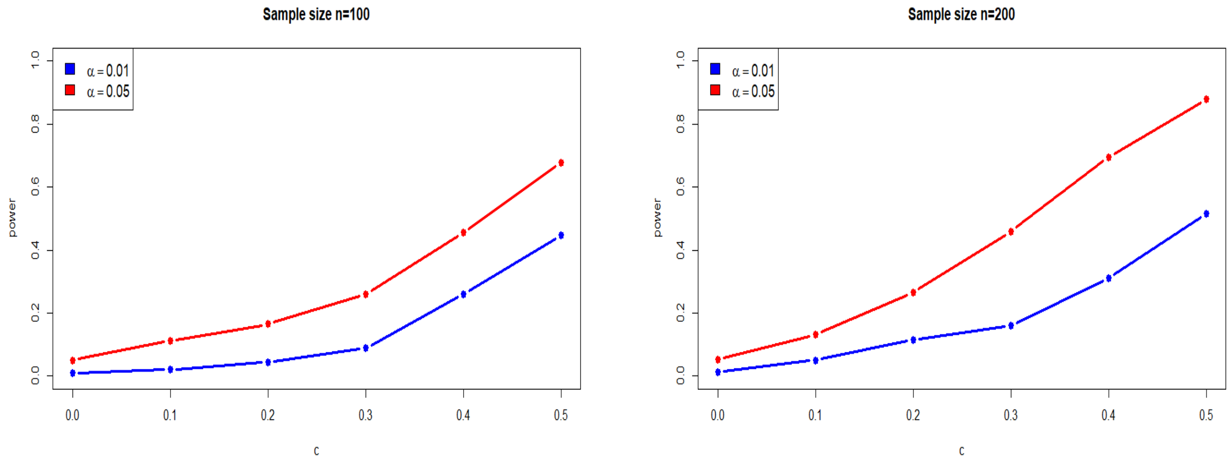


Figure 5.5: Power curves for hypothesis testing problem (5.22) based on FLFRM with different choice of c in $\beta_{14}(s)$ and $\beta_{24}(s)$.

improve with the number of grid points n_v.

Table 5.2: Empirical coverage probabilities of $1-\alpha$ simultaneous confidence bands

α	n_v	β_{11}	β_{12}	β_{13}	β_{14}
	200	0.935	0.920	0.925	0.920
0.05	2000	0.945	0.950	0.950	0.950
	200	0.985	0.990	0.995	0.980
0.01	2000	0.990	0.995	0.990	0.995
α	n_v	β_{21}	β_{22}	β_{23}	β_{24}
	200	0.915	0.915	0.930	0.940
0.05	2000	0.945	0.945	0.955	0.950
	200	0.980	0.995	0.990	0.990
0.01	2000	0.995	0.995	0.990	0.995

5.4 Real data analysis

5.4.1 ADNI data description

Data used in the preparation of this article were obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging, National Institute of Biomedical Imaging and Bioengineering, Food and Drug Administration, private pharmaceutical companies and non-profit organizations as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians in developing new treatments and monitoring their effectiveness, as well as lessening the time and cost of clinical trials. The principal investigator of this initiative is Michael W. Weiner, MD, at the VA Medical Center and University of California, San Francisco. ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The goal was to recruit 800 subjects, but the initial study (ADNI-1) has been followed by ADNI-GO and ADNI-2. To date, these three protocols have recruited over 1,500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow-up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org.

5.4.2 Data Processing

In this data analysis, we included 936 MRI scans from normal controls (NC) and individuals with MCI or AD from three different phases: ADNI-1, ADNI-GO, and ADNI-2. The demographic information of all the subjects is summarized in Table 5.3, including phase, gender, handedness, age, education length and disease status.

The scans in ADNI-1 were performed on a variety of 1.5 Tesla MRI scanners with protocols individualized for each scanner, include standard T1-weighted images obtained using volumetric 3-dimensional sagittal MPRAGE or equivalent protocols with varying resolutions. The typical protocol includes: repetition time = 2400 ms, inversion time = 1000

Table 5.3: ADNI hippocampal surface data: demographic information of all the 936 subjects, including phase, gender, handedness, age, education length and disease status.

Phase	ADNI-1	ADNI-GO	ADNI-2	Total
Size	800	24	112	936
Gender (F/M)	465/335	13/11	61/51	539/397
Handedness (R/L)	738/62	20/4	9/103	861/75
Age range (years)	[58, 95]	[55, 84]	[53, 87]	[53, 95]
Edu. length range (years)	[4, 20]	[12, 20]	[8, 20]	[4, 20]
Disease (NC/MCI/AD)	224/389/187	0/24/0	29/58/25	253/471/212

ms, flip angle = 8^o, and field of view = 24 cm, with a $256 \times 256 \times 170$ acquisition matrix in the $x-$, $y-$, and $z-$dimensions, which yields a voxel size of $1.25 \times 1.26 \times 1.2$ mm^3. The scans in ADNI-GO and ADNI-2 were performed at 3 Tesla MRI scanners with T1-weighted imaging parameters similar to those in ADNI-1. We processed the MRI data by using standard steps, including anterior commissure and posterior commissure correction, skull-stripping, cerebellum removing, intensity inhomogeneity correction, segmentation, and registration. Subsequently, we carried out automatic regional labeling by labeling the template and by transferring the labels following the deformable registration of subject images. After labeling 93 ROIs, we were able to compute volumes for each of these ROIs for each subject.

We adopted a hippocampal subregional analysis package based on surface fluid registration (Shi et al., 2013) that uses isothermal coordinates and fluid registration to generate one-to-one hippocampal surface registration for computing the surface statistics. Then, we computed the various surface statistics on the registered surface, such as multivariate tensor-based morphometry (TBM) statistics, which retain the full tensor information of the deformation Jacobian matrix, together with the radial distance, which retains information on the deformation along the surface normal direction. More details can be found in Wang et al. (2011).

5.4.3 Data Analysis

The hippocampus is believed to be involved in memory, spatial navigation and memory, and behavioral inhibition. In AD, the hippocampus is one of the first regions of the brain to be affected, leading to the confusion and loss of memory so commonly seen in the early stages of the disease (Huang et al., 2017). Recent work has revealed that the hippocampus is structurally and functionally asymmetric, and hippocampal asymmetry changes with AD progression, with the left hippocampus affected first by dementia, followed by atrophy in the right hippocampus after a time lag (Maruszak and Thuret, 2014; Shi et al., 2009; Rabl et al., 2014).

Before conducting this analysis, we would like to check if there is any batch effects caused by the phase-level heterogeneity. For both left and right hippocampal surfaces, we calculated three quantiles (i.e., Q1, Q2, and Q3) of the logged radial distances across all the vertices for each subject, which are shown in Figure 5.6. It can be found that, at each of the three levels, the pattern of calculated quantile varies across different phases (e.g., ADNI-1 v.s ADNI-GO and ADNI-2). It indicates that the phase-level heterogeneity does exist in the ADNI hippocampal surface data. Therefore, the phase information should be included as predictors in the data analysis.

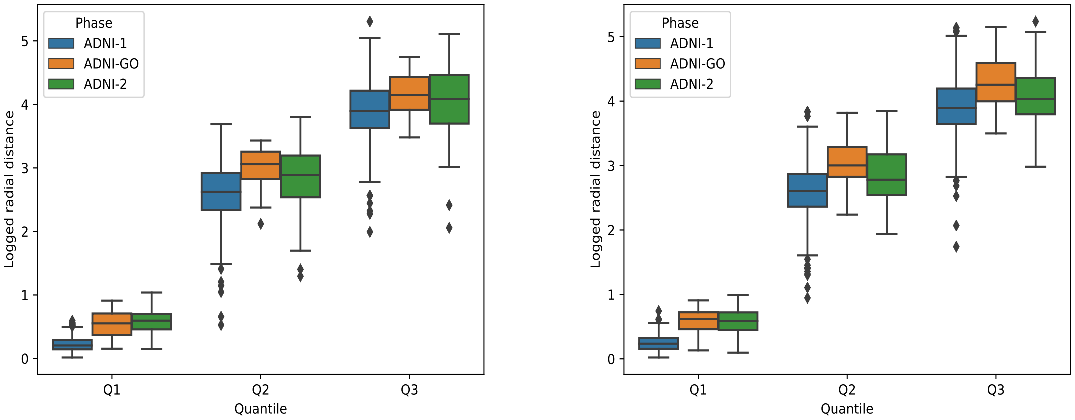


Figure 5.6: ADNI hippocampal surface data: three quantiles of the logged radial distances across all the vertices for each subject.

The object of this data analysis was to integrate the data from three different data phases (i.e., ADNI-1, ADNI-GO, and ADNI-2) and exam the effects of clinical variables and demographic variables on either the left or right hippocampus. Moreover, the latent factors were expected to be recovered and discussed. To achieve this objective, we applied FLFRM with either the left or right hippocampal surface data as the functional responses. For comparison, the MVCM was also considered here. Specifically, in model (5.4) we calculated the logged radial distance and three TBM statistics measured over 7,500 vertices on the hippocampal surface (3,750 on each side). Moreover, we included an intercept, gender, handedness, education length, age, diagnostic information (two dummy variables were introduced to represented MCI and AD), and phase information (two dummy variables were introduced to represented ADNI-GO and ADNI-2) as predictors.

After fitting the model, we statistically tested the effects of all the primary variables on the functional responses across all the vertices on hippocampal surfaces. In particular, the following hypothesis testing problems are considered: for each t, the null hypothesis is described as

$$\mathbf{H_0} : \beta_{t1}(\boldsymbol{s}) = \beta_{tj}(\boldsymbol{s}) = \beta_{tj}(\boldsymbol{s}) = \beta_{tj}(\boldsymbol{s}) = 0,\ \forall \boldsymbol{s} \in \mathcal{S}. \tag{5.23}$$

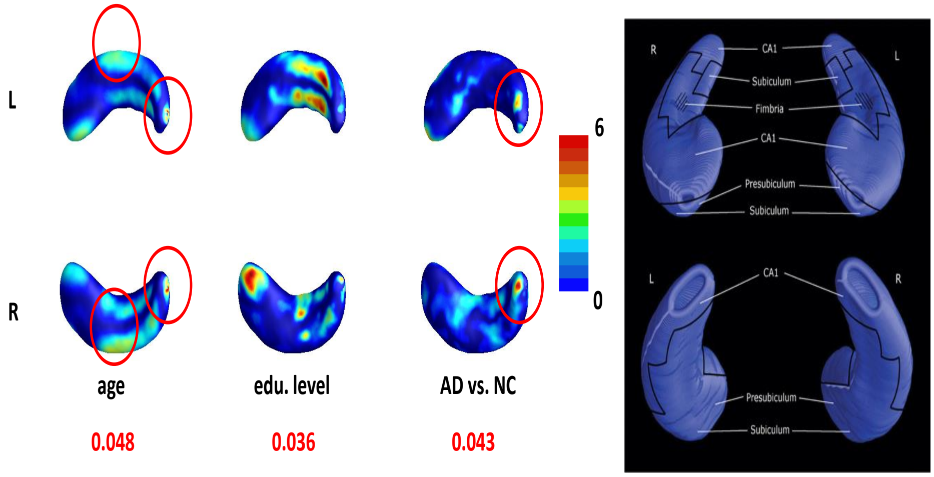
The global test statistic was calculated and 500 replications were generated in wild bootstrap approach. The corresponding p-values are summarized in Table 5.4, where p-values less than the significant level are highlighted in red. Given the significant level 0.05, both the disease effect (AD vs. NC) and age effect are found to be significant on the left hippocampal surface based on MVCM. In comparison, more variables are detected based on FLFRM. For example, significant age effect is found on the left hippocampal surface, while both education length effect and disease effect (AD vs. NC) are significant on left and right hippocampal surfaces. Among these variables, education length is the one which was detected in FLFRM but not in MVCM. In fact, education length is an important factor for the changes of hippocampus structure because of the strong correlation between education length and AD, has been found in existing literature (Arenaza-Urquijo et al., 2013; Liu et al., 2012).

Table 5.4: ADNI hippocampal surface data: comparison of p-values for primary variables between MVCM and FLFRM.

Variable	P-value			
	Left Hippocampus		Right Hippocampus	
	MVCM	FLFRM	MVCM	FLFRM
Gender	0.212	0.092	0.234	0.116
Handedness	0.652	0.102	0.704	0.082
Education length	0.132	0.036	0.244	0.048
Age	0.048	0.048	0.096	0.052
MCI vs. NC	0.156	0.066	0.082	0.064
AD vs. NC	0.046	0.034	0.054	0.040
ADNI-GO vs. ADNI-1	0.134	0.112	0.136	0.120
ADNI-2 vs. ADNI-1	0.118	0.106	0.112	0.114

For those variables detected by the global test statistic in FLFRM, we are also interested in the significant subregions detected by the local test statistic. Here the false discovery rate (FDR, Benjamini et al. (2001)) adjusted $-\log_{10}(p)$-value maps are presented in Figure 5.7. To better understand the significant subregions, the cytoarchitectonic subregions mapped on blank MR-based models at 3T of the hippocampal formation (Frisoni et al., 2008) is considered here and presented in right plot of Figure 5.7. It shows that all the significant subregions associated with age and disease are circled in red and found in the CA1 subfield, some are found on the lateral and medial aspects of the tail (CA1 subfield), and others are found on the dorsolateral aspect of the head (CA1 subfield). It is interesting to note that volumes of similar hippocampal subregions were found to be affected in AD (Frisoni et al., 2008), indicating that the findings based on our FLFRM are in agreement with those of previous work.

Besides the relationship between the functional responses and some primary variables of interest, it is also of great importance to investigate the potential hidden factors estimated by our FLFRM. By applying the ED method, three latent factors were detected, and



Frisoni et al. (2008) *Brain*

Figure 5.7: ADNI hippocampal surface data: FDR adjusted $-\log_{10}(p)$-value maps (left) and the cytoarchitectonic subregions mapped on blank MR-based models at 3T of the hippocampal formation (right).

the correlation between primary variables and detected latent factors are shown in Table 5.5, where the Pearson correlation was calculated for between two continuous variables while the polyserial correlation was calculated between a continuous variable and a discrete one. According to Table 5.5, it can be found that, on both left and right hippocampal surfaces, the detected factors are highly related to education length, age, disease status, and phase information. Recall that the latent factors and primary variables are assumed to be uncorrelated in MVCM, which is violated here. Thus, the inference results based on MVCM may not be reasonable on this dataset.

Another interesting thing is about the phase information. In the hypothesis testing problem (5.23), we don't have enough evidence to show the existence of phase-level heterogeneity in terms of the p-values associated to the phase information. While the detected latent factors are highly correlated to the phase information according to Table 5.5. Thus, it is expected to find some variables not included in the current FLFRM but strongly correlated with the latent factors. Here we considered 7 new variables in three categories here: ethnic group information (three dummy variables were introduced to represented Asian, African American, and White),

Table 5.5: ADNI hippocampal surface data: correlation between primary variables and detected latent factors.

Variable	Latent factor					
	Left Hippocampus			Right Hippocampus		
	factor 1	factor 2	factor 3	factor 1	factor 2	factor 3
Gender	-0.038	0.015	-0.048	0.006	0.023	-0.045
	(0.358)	(0.724)	(0.239)	(0.883)	(0.582)	(0.278)
Handedness	-0.013	-0.041	0.076	0.041	-0.055	0.047
	(0.835)	(0.517)	(0.209)	(0.494)	(0.382)	(0.435)
Education length	-0.021	0.024	0.090	0.058	0.014	0.074
	(0.531)	(0.466)	(0.006)	(0.078)	(0.665)	(0.025)
Age	0.120	0.089	-0.079	-0.163	0.071	-0.131
	(<0.001)	(0.007)	(0.015)	(<0.001)	(0.030)	(<0.001)
MCI vs. NC	-0.045	0.061	0.020	0.064	0.003	0.062
	(0.272)	(0.144)	(0.617)	(0.119)	(0.944)	(0.131)
AD vs. NC	0.087	-0.058	0.061	-0.094	-0.029	-0.008
	(0.041)	(0.228)	(0.507)	(0.039)	(0.530)	(0.853)
ADNI-GO vs. ADNI-1	-0.305	0.392	0.215	0.440	-0.176	0.403
	(<0.001)	(<0.001)	(0.011)	(<0.001)	(0.064)	(<0.001)
ADNI-2 vs. ADNI-1	-0.221	-0.318	0.213	0.271	-0.469	0.466
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)

marital status (three dummy variables were introduced to represented widow, divorce and no-married), and retirement status. The correlation between new variables and detected latent factors are shown in Table 5.6. It can be found that, on the left hippocampal surface, the detected latent factors are strongly correlated to ethnic group information, marital status, and retirement status, while on the right hippocampal surface, the detected latent factors are only correlated to marital status.

In addition, for each latent factor, we considered conduct a multiple regression model where the relationship between the latent factor and all the variables (both primary and new ones). The inference results are summarized in Table 5.7. It can be concluded that: education length, age, disease status, phase information, ethnic group information, marital status, and retirement status significantly affect the left hippocampal surface; age, phase information, and marital status significantly affect the right hippocampal surface. It is interesting to see that the early disease effect (MCI vs. NC) is found to be significant only on the left hippocampal

Table 5.6: ADNI hippocampal surface data: correlation between new variables and detected latent factors.

Variable	Latent factor					
	Left Hippocampus			Right Hippocampus		
	factor 1	factor 2	factor 3	factor 1	factor 2	factor 3
Asian	-0.105	-0.032	0.119	0.036	0.001	0.004
	(0.270)	(0.757)	(0.207)	(0.696)	(0.994)	(0.963)
African American	-0.187	0.015	0.017	0.045	0.113	0.042
	(0.006)	(0.830)	(0.811)	(0.526)	(0.084)	(0.544)
White	0.165	0.009	-0.047	-0.038	-0.066	-0.024
	(0.007)	(0.892)	(0.453)	(0.534)	(0.280)	(0.695)
Widow	-0.116	0.051	-0.068	0.028	0.135	-0.087
	(0.028)	(0.325)	(0.205)	(0.593)	(0.006)	(0.090)
Divorce	0.004	0.013	0.032	-0.011	-0.021	0.078
	(0.952)	(0.830)	(0.619)	(0.863)	(0.745)	(0.215)
No-married	-0.069	0.030	0.002	-0.005	0.058	0.001
	(0.352)	(0.685)	(0.980)	(0.949)	(0.420)	(0.992)
Retirement status	0.145	-0.074	0.064	-0.080	-0.027	-0.049
	(0.002)	(0.107)	(0.181)	(0.097)	(0.569)	(0.307)

surface, which is consistent with the previous finding on hippocampus asymmetry: the left hippocampus is affected first by dementia, followed by atrophy in the right hippocampus after a time lag (Maruszak and Thuret, 2014; Shi et al., 2009; Rabl et al., 2014).

5.5 Conclusions

In this paper, we proposed a functional latent factor regression model which is efficient to investigate the relationship between functional responses and primary variables of interest while adjusting the unknown factors. Both estimation procedures, hypothesis testing, and simultaneous confidence band construction have been established in the statistical inference. For the asymptotic results, the consistency of detected latent factor space and the weak convergence of estimated coefficient functions are systematically investigated. Both Monte Carlo simulations and the real data example on hippocampal surface data from ADNI study have shown that our FLFRM outperforms both traditional SVA (massive-univariate analysis) and existing functional regression models (e.g., MVCM).

Table 5.7: ADNI hippocampal surface data: regression analysis between each detect latent factor and all the variables (both primary and new ones).

Variable	*P*-value					
	Left Hippocampus			Right Hippocampus		
	factor 1	factor 2	factor 3	factor 1	factor 2	factor 3
Gender	0.842	0.414	0.354	0.626	0.908	0.258
Handedness	0.993	0.627	0.318	0.883	0.395	0.633
Education length	0.981	0.179	0.028	0.402	0.150	0.306
Age	0.025	0.377	0.432	0.016	0.251	0.716
MCI vs. NC	0.966	0.060	0.570	0.666	0.590	0.383
AD vs. NC	0.358	0.991	0.180	0.202	0.937	0.323
ADNI-GO vs. ADNI-1	0.002	<0.001	0.011	<0.001	0.014	<0.001
ADNI-2 vs. ADNI-1	<0.001	<0.001	0.001	<0.001	<0.001	<0.001
Asian	0.223	0.817	0.408	0.764	0.826	0.730
African American	0.012	0.804	0.555	0.477	0.179	0.268
White	0.750	0.565	0.736	0.895	0.102	0.727
Widow	0.015	0.811	0.984	0.121	0.016	0.600
Divorce	0.741	0.825	0.638	0.661	0.946	0.307
No-married	0.384	0.470	0.989	0.829	0.214	0.880
Retirement status	0.026	0.051	0.076	0.688	0.728	0.603

APPENDIX A: TECHNICAL DETAILS OF CHAPTER 4

In this chapter, we give the proof to the main theoretical results: Lemma 4.1, Theorem 4.1, Theorem 4.2, Theorem 4.4 and Theorem 4.5.

Proof of Theorem 4.1

Since $\lim_{t\to\infty}\|\boldsymbol{\delta}^{t+1}-\boldsymbol{\delta}^{t}\|=0$, $\lim_{t\to\infty}\|\boldsymbol{\kappa}^{t+1}-\boldsymbol{\kappa}^{t}\|=0$, we get from Algorithm 2 that

$$\lim_{t\to\infty}\|\boldsymbol{RS}\boldsymbol{\eta}^{t+1}_{.l}-\boldsymbol{\zeta}^{t+1}_{.l}\|=0,\ \lim_{t\to\infty}\|\boldsymbol{\Lambda}^{t+1}_{kl}-\boldsymbol{\nu}^{t+1}_{kl}\|=0,\ k=1,\ldots,K,\ l=1,\ldots,p. \tag{A.1}$$

Then both $\boldsymbol{\eta}^{t}_{.l}$ and $\boldsymbol{\Lambda}^{t}_{kl}$ are bounded by (A.1) and the boundedness assumption on $\{(\boldsymbol{\zeta}^{t},\boldsymbol{\nu}^{t})\}$. It follows from Algorithm 2 and the boundedness of $\{(\boldsymbol{\eta}^{t}_{.l},\boldsymbol{\Lambda}^{t}_{kl})\}$ and $\{(\boldsymbol{\zeta}^{t},\boldsymbol{\nu}^{t})\}$ that $\{(\boldsymbol{\delta}^{t},\boldsymbol{\kappa}^{t})\}$ is bounded. Since $\{(\boldsymbol{\eta}^{t},\boldsymbol{\Lambda}^{t},\boldsymbol{\zeta}^{t},\boldsymbol{\nu}^{t},\boldsymbol{\delta}^{t},\boldsymbol{\kappa}^{t})\}$ is bounded and the augmented Lagrangian function $L_\rho(\boldsymbol{\eta},\boldsymbol{\Lambda},\boldsymbol{\zeta},\boldsymbol{\nu},\boldsymbol{\delta},\boldsymbol{\kappa})$ is continuous, we can obtain that $L_\rho(\boldsymbol{\eta},\boldsymbol{\Lambda},\boldsymbol{\zeta},\boldsymbol{\nu},\boldsymbol{\delta},\boldsymbol{\kappa})$ is bounded. As the function $\tilde{Q}_p(\boldsymbol{\eta},\boldsymbol{\Lambda})$ is nonconcave, it holds that,

$$\begin{aligned}
&L_\rho(\boldsymbol{\eta}^{t},\boldsymbol{\Lambda}^{t},\boldsymbol{\zeta}^{t},\boldsymbol{\nu}^{t},\boldsymbol{\delta}^{t},\boldsymbol{\kappa}^{t})-L_\rho(\boldsymbol{\eta}^{t+1},\boldsymbol{\Lambda}^{t},\boldsymbol{\zeta}^{t},\boldsymbol{\nu}^{t},\boldsymbol{\delta}^{t},\boldsymbol{\kappa}^{t})\geq\\
&\qquad\sum_{l=1}^{p}\nabla_{\eta_{.l}}L_\rho(\boldsymbol{\eta}^{t+1},\boldsymbol{\Lambda}^{t},\boldsymbol{\zeta}^{t},\boldsymbol{\nu}^{t},\boldsymbol{\delta}^{t},\boldsymbol{\kappa}^{t})^T(\boldsymbol{\eta}^{t}_{.l}-\boldsymbol{\eta}^{t+1}_{.l})+c_\eta\sum_{l=1}^{p}\|\boldsymbol{\eta}^{t}_{.l}-\boldsymbol{\eta}^{t+1}_{.l}\|^2,\\
&L_\rho(\boldsymbol{\eta}^{t+1},\boldsymbol{\Lambda}^{t},\boldsymbol{\zeta}^{t},\boldsymbol{\nu}^{t},\boldsymbol{\delta}^{t},\boldsymbol{\kappa}^{t})-L_\rho(\boldsymbol{\eta}^{t+1},\boldsymbol{\Lambda}^{t+1},\boldsymbol{\zeta}^{t},\boldsymbol{\nu}^{t},\boldsymbol{\delta}^{t},\boldsymbol{\kappa}^{t})\geq\\
&\qquad\sum_{k=1}^{K}\sum_{l=1}^{p}\nabla_{\Lambda_{kl}}L_\rho(\boldsymbol{\eta}^{t+1},\boldsymbol{\Lambda}^{t+1},\boldsymbol{\zeta}^{t},\boldsymbol{\nu}^{t},\boldsymbol{\delta}^{t},\boldsymbol{\kappa}^{t})(\boldsymbol{\Lambda}^{t}_{kl}-\boldsymbol{\Lambda}^{t+1}_{kl})^T+c_\Lambda\sum_{k=1}^{K}\sum_{l=1}^{p}\|\boldsymbol{\Lambda}^{t+1}_{kl}-\boldsymbol{\Lambda}^{t}_{kl}\|^2,
\end{aligned}$$

where both c_η and c_Λ are constants. In addition, based on the minimization problems in Algorithm 2, we have

$$\begin{aligned}
\nabla_{\eta_{.l}}L_\rho(\boldsymbol{\eta}^{t+1},\boldsymbol{\Lambda}^{t},\boldsymbol{\zeta}^{t},\boldsymbol{\nu}^{t},\boldsymbol{\delta}^{t},\boldsymbol{\kappa}^{t})^T(\boldsymbol{\eta}^{t}_{.l}-\boldsymbol{\eta}^{t+1}_{.l})&\geq 0,\\
\nabla_{\Lambda_{kl}}L_\rho(\boldsymbol{\eta}^{t+1},\boldsymbol{\Lambda}^{t+1},\boldsymbol{\zeta}^{t},\boldsymbol{\nu}^{t},\boldsymbol{\delta}^{t},\boldsymbol{\kappa}^{t})(\boldsymbol{\Lambda}^{t}_{kl}-\boldsymbol{\Lambda}^{t+1}_{kl})^T&\geq 0.
\end{aligned}$$

Then we can obtain that

$$\begin{aligned}L_\rho(\boldsymbol{\eta}^t,\boldsymbol{\Lambda}^t,\boldsymbol{\zeta}^t,\boldsymbol{\nu}^t,\boldsymbol{\delta}^t,\boldsymbol{\kappa}^t)-L_\rho(\boldsymbol{\eta}^{t+1},\boldsymbol{\Lambda}^t,\boldsymbol{\zeta}^t,\boldsymbol{\nu}^t,\boldsymbol{\delta}^t,\boldsymbol{\kappa}^t)&\geq c_\eta\sum_{l=1}^p\|\boldsymbol{\eta}_{.l}^{t+1}-\boldsymbol{\eta}_{.l}^t\|^2,\\ L_\rho(\boldsymbol{\eta}^{t+1},\boldsymbol{\Lambda}^t,\boldsymbol{\zeta}^t,\boldsymbol{\nu}^t,\boldsymbol{\delta}^t,\boldsymbol{\kappa}^t)-L_\rho(\boldsymbol{\eta}^{t+1},\boldsymbol{\Lambda}^{t+1},\boldsymbol{\zeta}^t,\boldsymbol{\nu}^t,\boldsymbol{\delta}^t,\boldsymbol{\kappa}^t)&\geq c_\Lambda\sum_{k=1}^K\sum_{l=1}^p\|\boldsymbol{\Lambda}_{kl}^{t+1}-\boldsymbol{\Lambda}_{kl}^t\|^2.\end{aligned}\tag{A.2}$$

Moreover, based on Algorithm 2, we have

$$\begin{aligned}L_\rho(\boldsymbol{\eta}^{t+1},\boldsymbol{\Lambda}^{t+1},\boldsymbol{\zeta}^{t+1},\boldsymbol{\nu}^t,\boldsymbol{\delta}^t,\boldsymbol{\kappa}^t)&\leq L_\rho(\boldsymbol{\eta}^{t+1},\boldsymbol{\Lambda}^{t+1},\boldsymbol{\zeta}^t,\boldsymbol{\nu}^t,\boldsymbol{\delta}^t,\boldsymbol{\kappa}^t),\\ L_\rho(\boldsymbol{\eta}^{t+1},\boldsymbol{\Lambda}^{t+1},\boldsymbol{\zeta}^{t+1},\boldsymbol{\nu}^{t+1},\boldsymbol{\delta}^t,\boldsymbol{\kappa}^t)&\leq L_\rho(\boldsymbol{\eta}^{t+1},\boldsymbol{\Lambda}^{t+1},\boldsymbol{\zeta}^{t+1},\boldsymbol{\nu}^t,\boldsymbol{\delta}^t,\boldsymbol{\kappa}^t).\end{aligned}\tag{A.3}$$

Combining (A.1) to (A.3), we can get that

$$\begin{aligned}L_\rho(\boldsymbol{\eta}^t,\boldsymbol{\Lambda}^t,\boldsymbol{\zeta}^t,\boldsymbol{\nu}^t,\boldsymbol{\delta}^t,\boldsymbol{\kappa}^t)-L_\rho(\boldsymbol{\eta}^{t+1},\boldsymbol{\Lambda}^{t+1},\boldsymbol{\zeta}^{t+1},\boldsymbol{\nu}^{t+1},\boldsymbol{\delta}^{t+1},\boldsymbol{\kappa}^{t+1})+\sum_{l=1}^p\|\boldsymbol{\delta}_{.l}^t-\boldsymbol{\delta}_{.l}^{t+1}\|\\ +\sum_{k=1}^K\sum_{l=1}^p\|\boldsymbol{\kappa}_{kl}^t-\boldsymbol{\kappa}_{kl}^{t+1}\|\geq c_\eta\sum_{l=1}^p\|\boldsymbol{\eta}_{.l}^{t+1}-\boldsymbol{\eta}_{.l}^t\|^2+c_\Lambda\sum_{k=1}^K\sum_{l=1}^p\|\boldsymbol{\Lambda}_{kl}^{t+1}-\boldsymbol{\Lambda}_{kl}^t\|^2.\end{aligned}\tag{A.4}$$

Since $L_\rho(\boldsymbol{\eta}^t,\boldsymbol{\Lambda}^t,\boldsymbol{\zeta}^t,\boldsymbol{\nu}^t,\boldsymbol{\delta}^t,\boldsymbol{\kappa}^t)$ is bounded, there exists a subsequence t_j such that

$$\lim_{t_j\to\infty}L_\rho(\boldsymbol{\eta}^{t_j},\boldsymbol{\Lambda}^{t_j},\boldsymbol{\zeta}^{t_j},\boldsymbol{\nu}^{t_j},\boldsymbol{\delta}^{t_j},\boldsymbol{\kappa}^{t_j})=\underline{\lim}_{t\to\infty}L_\rho(\boldsymbol{\eta}^t,\boldsymbol{\Lambda}^t,\boldsymbol{\zeta}^t,\boldsymbol{\nu}^t,\boldsymbol{\delta}^t,\boldsymbol{\kappa}^t).\tag{A.5}$$

According to (A.4) and the assumption that $\lim_{t\to\infty}\|\boldsymbol{\delta}^{t+1}-\boldsymbol{\delta}^t\|=0$ and $\lim_{t\to\infty}\|\boldsymbol{\kappa}^{t+1}-\boldsymbol{\kappa}^t\|=0$, we can derive that

$$\text{lim}_{t_j\to\infty}\|\boldsymbol{\eta}_{.l}^{t_j+1}-\boldsymbol{\eta}_{.l}^{t_j}\|^2=0,\ \text{lim}_{t_j\to\infty}\|\boldsymbol{\Lambda}_{kl}^{t_j+1}-\boldsymbol{\Lambda}_{kl}^{t_j}\|^2=0,\tag{A.6}$$

$k=1,\dots,K,l=1,\dots,p$. Recall the result in (A.1), we have

$$\text{lim}_{t_j\to\infty}\|\boldsymbol{\zeta}_{.l}^{t_j+1}-\boldsymbol{\zeta}_{.l}^{t_j}\|^2=0,\ \text{lim}_{t_j\to\infty}\|\boldsymbol{\nu}_{kl}^{t_j+1}-\boldsymbol{\nu}_{kl}^{t_j}\|^2=0,\tag{A.7}$$

$k = 1, \dots, K, l = 1, \dots, p$. Then, by the boundedness of $\{(\boldsymbol{\eta}^t, \boldsymbol{\Lambda}^t, \boldsymbol{\zeta}^t, \boldsymbol{\nu}^t, \boldsymbol{\delta}^t, \boldsymbol{\kappa}^t)\}$, there exists a convergence subsequence, denoted by $\{t_{j'}\}$ such that $\{(\boldsymbol{\eta}^{t_{j'}}, \boldsymbol{\Lambda}^{t_{j'}}, \boldsymbol{\zeta}^{t_{j'}}, \boldsymbol{\nu}^{t_{j'}}, \boldsymbol{\delta}^{t_{j'}}, \boldsymbol{\kappa}^{t_{j'}})\}$ converges to some point $(\boldsymbol{\eta}^*, \boldsymbol{\Lambda}^*, \boldsymbol{\zeta}^*, \boldsymbol{\nu}^*, \boldsymbol{\delta}^*, \boldsymbol{\kappa}^*)$. Based on the result in (A.1), we have

$$\boldsymbol{R}\boldsymbol{S}\boldsymbol{\eta}^*_{.l} = \boldsymbol{\zeta}^*_{.l}, \ \ \boldsymbol{\Lambda}^*_{kl} = \boldsymbol{\nu}^*_{kl}, \ \ k = 1, \dots, K, l = 1, \dots, p. \tag{A.8}$$

In addition, by the assumption of function $\tilde{Q}_p(\boldsymbol{\eta}, \boldsymbol{\Lambda})$, point $(\boldsymbol{\eta}^*, \boldsymbol{\Lambda}^*, \boldsymbol{\zeta}^*, \boldsymbol{\nu}^*, \boldsymbol{\delta}^*, \boldsymbol{\kappa}^*)$ should satisfy that

$$\begin{aligned} \nabla_{\eta.l}\tilde{Q}_p(\boldsymbol{\eta}^*, \boldsymbol{\Lambda}^*) + \rho \sum_{l=1}^{p} \boldsymbol{S}^T \boldsymbol{R}^T (\boldsymbol{R}\boldsymbol{S}\boldsymbol{\eta}^*_{.l} - \boldsymbol{\zeta}^*_{.l} + \boldsymbol{\delta}^*_{.l}) = \mathbf{0} \\ \nabla_{\Lambda_{kl}}\tilde{Q}_p(\boldsymbol{\eta}^*, \boldsymbol{\Lambda}^*) + \rho \sum_{l=1}^{p}\sum_{k=1}^{K} (\boldsymbol{\Lambda}^*_{kl} - \boldsymbol{\nu}^*_{kl} + \boldsymbol{\kappa}^*_{kl}) = \mathbf{0} \end{aligned} \tag{A.9}$$

Finally, by taking the limit of both sides in updating equations of $\boldsymbol{\zeta}$ and $\boldsymbol{\nu}$ in Algorithm 2, we have

$$\begin{aligned} \boldsymbol{\zeta}^*_{.l} &= \mathbf{ST}_{\lambda_1/\rho}(\boldsymbol{R}\boldsymbol{S}\boldsymbol{\eta}^*_{.l} + \boldsymbol{\delta}^*_{.l}) \\ \boldsymbol{\nu}^*_{kl} &= \mathbf{VST}_{\lambda_2/\rho}(\boldsymbol{\Lambda}^*_{kl} + \boldsymbol{\kappa}^*_{kl}). \end{aligned} \tag{A.10}$$

Combining (A.8) with (A.9) and (A.10), we obtain that $(\boldsymbol{\eta}^*, \boldsymbol{\Lambda}^*, \boldsymbol{\zeta}^*, \boldsymbol{\nu}^*, \boldsymbol{\delta}^*, \boldsymbol{\kappa}^*)$ is a KKT point of $L_\rho(\boldsymbol{\eta}, \boldsymbol{\Lambda}, \boldsymbol{\zeta}, \boldsymbol{\nu}, \boldsymbol{\delta}, \boldsymbol{\kappa})$ satisfying (4.20). $\square$

Proof of Theorem 4.2

We introduce some notation as follows. First, we divide the true parameter $\boldsymbol{\Lambda}_{0k}$ into two parts including $\boldsymbol{\Lambda}^{(1)}_{0k}$ and $\boldsymbol{\Lambda}^{(2)}_{0k}$, where $\boldsymbol{\Lambda}^{(2)}_{0k}$ contains rows with all zeros in $\boldsymbol{\Lambda}_{0k}$. For each component $l = 1, \dots, p$, we define $\mathcal{A}_l = \{k : \zeta_{kl} = 0, \boldsymbol{\zeta}_{.l} = \boldsymbol{R}\boldsymbol{S}\boldsymbol{\mu}_{.l}\}$, which contains all pairs that do not contribute to the separation of any two clusters. We define the complement of $\mathcal{A}_l$ as $\mathcal{A}^c_l$ for $l = 1, \dots, p$. Let $\Delta_n \log L_p(\boldsymbol{u}) = \log L_p(\boldsymbol{\theta}_0 + n^{-\frac{1}{2}}\boldsymbol{u}) - \log L_p(\boldsymbol{\theta}_0)$, where $\boldsymbol{u}$ is a $\dim(\boldsymbol{\theta}) \times 1$ vector such that $||\boldsymbol{u}||_2 = O(1)$. Our aim is to show that for any given ϵ, there is a large constant $\tilde{M}_\epsilon$ such that we have

$$P\{\sup_{\|\boldsymbol{u}\|_2=\tilde{M}_\epsilon} \Delta_n \log L_p(\boldsymbol{u}) < 0\} \geq 1-\epsilon. \tag{A.11}$$

This implies that with probability tending to 1, there is a local maximum $\hat{\boldsymbol{\theta}}_n$ in the ball $\{\boldsymbol{\theta}_0 + n^{-\frac{1}{2}}\boldsymbol{u} : \|\boldsymbol{u}\|_2 \leq \tilde{M}_\epsilon\}$ such that $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2 = O_p\big(n^{-\frac{1}{2}}\big)$.

Before we prove Theorem 4.2, some mild regularity conditions are listed as follows without any detailed verification.

- (C1) The first, second, and third partial derivatives of $\log f(\boldsymbol{x}|\boldsymbol{w}, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ exist. There is a sufficiently large open set $\mathcal{O}$ including $\boldsymbol{\theta}_0$ such that $\forall \boldsymbol{\theta} \in \mathcal{O}$, all the derivatives are bounded by a non-negative function $M_2(\boldsymbol{x}, \boldsymbol{w})$ with $\mathbb{E}\{M_2(\boldsymbol{x}, \boldsymbol{w})\} < \infty$;
- (C2) Both the observed information matrix $-\frac{1}{n}\frac{\partial^2}{\partial\theta\partial\theta^T}\log L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ and the Fisher information matrix $I_n(\boldsymbol{\theta}_0)$ are finite and positive definite.

We define $\boldsymbol{u}_\Lambda^{kl}$ and $\boldsymbol{u}_\zeta^{kl}$ as the subcomponents of $\boldsymbol{u}$ corresponding to the subcomponents $\boldsymbol{\Lambda}_{0kl}^{(1)}$ and $\boldsymbol{\zeta}_{\mathcal{A}_l^c}$, respectively. Then, we have

$$\begin{aligned}
\Delta_n \log L_p(\boldsymbol{u}) \quad &\leq \quad \log L(\boldsymbol{\theta}_0 + n^{-\frac{1}{2}}\boldsymbol{u}) - \log L(\boldsymbol{\theta}_0) \\
&- \quad \lambda_1 \sum_{l=1}^{p} \sum_{k \in \mathcal{A}_l^c} a_k^l \left(|\zeta_{kl} + n^{-\frac{1}{2}}\boldsymbol{u}_\zeta^{kl}| - |\zeta_{kl}| \right) \\
&- \quad \lambda_2 \sum_{k=1}^{K} \sum_{\Lambda_{0kl}^{(1)} \subset \Lambda_{0k}^{(1)}} \left(\|\boldsymbol{\Lambda}_{0kl}^{(1)} + n^{-\frac{1}{2}}\boldsymbol{u}_\Lambda^{kl}\|_2 - \|\boldsymbol{\Lambda}_{0kl}^{(1)}\|_2 \right) \\
&\triangleq \quad E_1 + E_2 + E_3.
\end{aligned} \tag{A.12}$$

It follows from the property of convex functions, Cauchy-Schwarz inequality, triangular inequality and the consistency of MLE $\tilde{\sigma}_l$ and $\tilde{\mu}_{kl}$ in a_k^l that the last two lines on the right-hand side of (A.12) can be bounded above as follows:

$$E_2 \leq \sqrt{2K}(K-1)\bar{C}^{-1}\sqrt{p}\|\boldsymbol{u}\|_2,\ E_3 \leq \sqrt{Kp}\|\boldsymbol{u}\|_2, \tag{A.13}$$

where $\bar{C} = \min_{1\leq l\leq p}\min_{k\in\mathcal{A}_l^c}\sigma_l|\zeta_{kl}|$. Furthermore, the first line on the right-hand side of (A.12) can be written as

$$\begin{aligned} E_1 &= n^{-\frac{1}{2}}\left\{\frac{\partial}{\partial\theta}\log L(\boldsymbol{\theta})\bigg|_{\theta=\theta_0}\right\}^T\boldsymbol{u} + \frac{1}{2n}\boldsymbol{u}^T\left\{\frac{\partial^2}{\partial\theta\partial\theta^T}\log L(\boldsymbol{\theta})\bigg|_{\theta=\theta_0}\right\}\boldsymbol{u} \\ &\quad + \frac{1}{6n^{\frac{3}{2}}}\frac{\partial}{\partial\theta^T}\left(\boldsymbol{u}^T\left\{\frac{\partial^2}{\partial\theta\partial\theta^T}\log L(\boldsymbol{\theta})\right\}\boldsymbol{u}\right)\bigg|_{\theta=\check{\theta}}\boldsymbol{u} \triangleq I_1 + I_2 + I_3, \end{aligned} \tag{A.14}$$

where $\check{\boldsymbol{\theta}}$ lies between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_0 + n^{-\frac{1}{2}}\boldsymbol{u}$. By using (C1), we have

$$|I_1| = n^{-\frac{1}{2}}|\{\frac{\partial}{\partial\theta}\log L(\boldsymbol{\theta})|_{\theta=\theta_0}\}^T\boldsymbol{u}| \leq n^{-\frac{1}{2}}\|\frac{\partial}{\partial\theta}\log L(\boldsymbol{\theta})|_{\theta=\theta_0}\|_2\|\boldsymbol{u}\|_2 = O_p(1)\|\boldsymbol{u}\|_2 \tag{A.15}$$

For I_2, it follows from (C2) that

$$I_2 = -\frac{1}{2}\boldsymbol{u}^T I_n(\boldsymbol{\theta}_0)\boldsymbol{u} + o_p(1)\|\boldsymbol{u}\|_2^2. \tag{A.16}$$

For I_3, it follows from Cauchy-Schwarz inequality and condition (C1) that

$$|I_3| = \frac{1}{6n^{\frac{3}{2}}}|\sum_{i,j,k}^{p}\frac{\partial^3}{\partial\theta_l\partial\theta_j\partial\theta_k}\log L(\boldsymbol{\theta})|_{\theta=\theta_0}\delta_l\delta_j\delta_k| \leq o_p(n^{-\frac{1}{2}})\|\boldsymbol{u}\|_2^2. \tag{A.17}$$

Then, by (A.13)-(A.17), and choosing a sufficiently large $\tilde{M}_\epsilon > 0$, we know that all terms E_2, E_3, I_1, and I_3 are dominated by I_2, which is negative. Therefore, for any given $\epsilon > 0$, there exists a sufficiently large constant $\tilde{M}_\epsilon$ such that

$$\lim_{n\to\infty} P\{\sup_{\|\boldsymbol{u}\|_2=\tilde{M}_\epsilon}\log L_p(\boldsymbol{\theta}_0 + n^{-\frac{1}{2}}\boldsymbol{u}) < \log L_p(\boldsymbol{\theta}_0)\} \geq 1-\epsilon. \tag{A.18}$$

Thus, there is a local maximum in $\{\boldsymbol{\theta}_0 + n^{-\frac{1}{2}}\boldsymbol{u} : \|\boldsymbol{u}\| \leq \tilde{M}_\epsilon\}$ with high probability and the local maximizer $\widehat{\boldsymbol{\theta}}_n$ satisfies $\sqrt{n}\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2 = O_p(1)$. This completes the proof. $\square$

Proof of Lemma 4.1

The estimated ordering $\hat{\boldsymbol{U}}_l$ of $\boldsymbol{\eta}_{.l}$ is only determined by the differences between distinct parameter groups within $\boldsymbol{\eta}_{.l}, l = 1, \ldots, p$. First note that for any $0 < \epsilon < 1$, if two parameters η_{kl} and $\eta_{k'l}$ are in the same parameter group assigning arbitrary ordering between them will not affect the estimated ordering of the parameters between groups, because the ordering within the same parameter group is exchangeable. On the other hand, when two parameters η_{kl} and $\eta_{k'l}$ are from different parameter groups, without loss of generality, let $\eta_{kl} > \eta_{k'l}$, the probability of estimating a wrong ordering

$$\begin{aligned} P(\mathbf{1}\{\hat{\eta}_{kl} \geq \hat{\eta}_{k'l}\}) &= P(\hat{\eta}_{kl} \geq \hat{\eta}_{k'l}) \\ &\geq P(|\hat{\eta}_{kl} - \eta_{kl}| + |\hat{\eta}_{k'l} - \eta_{k'l}| > 0) \\ &= 1 - P(\hat{\eta}_{kl} = \eta_{kl})P(\hat{\eta}_{k'l} = \eta_{k'l}) \to 0 \end{aligned} \tag{A.19}$$

as$n \to \infty$ since $\hat{\eta}_{kl}$ and $\hat{\eta}_{k'l}$ are independent and consistent estimators. Similarly, the consistency of the estimated ordering $\hat{\boldsymbol{V}}_l$ of the absolute values in vector $\boldsymbol{\eta}_{.l}$ can be derived by taking the square of the absolute values and following the same argument as for $\hat{\boldsymbol{U}}_l$. □

Proof of Theorem 4.3

Here we assume the same regularity condition ((C1) and (C2)) as in Theorem 4.2. To complete this proof, we first define the event $\mathcal{W}$ when the orderings of all components are correctly assigned as

$$\mathcal{W} = \cap_{l=1}^{p}(\{\hat{\boldsymbol{U}}_l = \boldsymbol{U}_l\} \cap \{\hat{\boldsymbol{V}}_l = \boldsymbol{V}_l\}). \tag{A.20}$$

Let $\hat{\boldsymbol{\theta}}_n^{\hat{W}}$ be $\hat{\boldsymbol{\theta}}_{n,\mathcal{W}}$ when $\mathcal{W}$ occurs; otherwise, denote it as $\hat{\boldsymbol{\theta}}_{n,\mathcal{W}^c}$. Then, the estimator can be rewritten as

$$\hat{\boldsymbol{\theta}}_n^{\hat{W}} = \hat{\boldsymbol{\theta}}_{n,\mathcal{W}}\mathbf{1}\{\mathcal{W}\} + \hat{\boldsymbol{\theta}}_{n,\mathcal{W}^c}\mathbf{1}\{\mathcal{W}^c\}$$

and therefore

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n^{\hat{W}} - \boldsymbol{\theta}_0) = \sqrt{n}(\hat{\boldsymbol{\theta}}_{n,\mathcal{W}} - \boldsymbol{\theta}_0)\mathbf{1}\{\mathcal{W}\} + \sqrt{n}(\hat{\boldsymbol{\theta}}_{n,\mathcal{W}^c} - \boldsymbol{\theta}_0)\mathbf{1}\{\mathcal{W}^c\}. \tag{A.21}$$

By Theorem 4.2, we have $\sqrt{n}(\hat{\boldsymbol{\theta}}_{n,\mathcal{W}} - \boldsymbol{\theta}_0) = O_p(1)$ and $\sqrt{n}(\hat{\boldsymbol{\theta}}_{n,\mathcal{W}^c} - \boldsymbol{\theta}_0) = O_p(1)$ as $n \to \infty$. By Lemma 4.1, we have $P(\mathcal{W}) \to 1$ and $P(\mathcal{W}^c) \to 0$ as $n \to \infty$. Therefore, by Slutsky's Theorem, $\sqrt{n}(\hat{\boldsymbol{\theta}}_n^{\hat{W}} - \boldsymbol{\theta}_0) = O_p(1)$, which completes the proof of Theorem 4.3. $\square$

Proof of Theorem 4.4

Proof of Theorem 4.4: Selection Consistency

For all $(l,k) \in \mathcal{A}$, we easily see from consistency of $\hat{\boldsymbol{\eta}}^W$ that $P((l,k) \in \hat{\mathcal{A}}^W) \to 1$. It then remains to show that for all $(l,k) \in \mathcal{A}^c$, $P((l,k) \in [\hat{\mathcal{A}}^W]^c) \to 1$. Assume the contrary, i.e., w.l.o.g there is an $l \in \{1,\ldots,p\}$ with $\zeta_{1l} = 0$ such that $\hat{\zeta}_{1l} \neq 0$ with non-vanishing probability.

By Taylor's theorem, applied to the function $\frac{1}{n}\frac{\partial \log L_p(\boldsymbol{\theta})}{\partial \zeta_{l1}}$, there exists a (random) vector $\bar{\boldsymbol{\theta}}$ on the line segment between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_n^W$ such that

$$\begin{aligned}\frac{1}{n}\frac{\partial \log L_p(\boldsymbol{\theta})}{\partial \zeta_{1l}}|_{\theta=\hat{\theta}_n^W} &= \frac{1}{n}\frac{\partial \log L(\boldsymbol{\theta})}{\partial \zeta_{1l}}|_{\theta=\theta_0} + \frac{1}{n}\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \zeta_{1l}\partial \zeta_{1l}}(\hat{\zeta}_{1l}^W - \zeta_{1l}) \\ &+ \frac{1}{2n}\frac{\partial^3 \log L(\boldsymbol{\theta})}{\partial \zeta_{1l}\partial \zeta_{1l}\partial \zeta_{1l}}(\hat{\zeta}_{1l}^W - \zeta_{1l})^2 - \lambda_1 a_1^l \text{sgn}(\hat{\zeta}_{1l}^W),\end{aligned} \tag{A.22}$$

where sgn$(\cdot)$ is the sign function. Now, using the regularity assumptions ((C1) and (C2)), the central limit theorem and the law of large numbers, we have

$$\frac{1}{n}\frac{\partial \log L(\boldsymbol{\theta})}{\partial \zeta_{1l}}|_{\theta=\theta_0} = O_p(n^{-\frac{1}{2}}),\ \frac{1}{n}\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \zeta_{l1}\partial \zeta_{1l}} = O_p(1),\ \frac{1}{2n}\frac{\partial^3 \log L(\boldsymbol{\theta})}{\partial \zeta_{1l}\partial \zeta_{1l}\partial \zeta_{1l}} = O_p(1).$$

Since $\hat{\zeta}_{1l}^W$ is root-n consistent, we get

$$\frac{1}{n}\frac{\partial \log L_p(\boldsymbol{\theta})}{\partial \zeta_{1l}}|_{\theta=\hat{\theta}_n^W} = \frac{1}{\sqrt{n}}(\frac{n\lambda_1}{\sqrt{n}\tilde{\sigma}_l\tilde{\zeta}_{1l}^W}\text{sgn}(\hat{\zeta}_{1l}^W) + O_p(1)). \tag{A.23}$$

From the assumption on the initial estimator, we have

$$\frac{n\lambda_1}{\sqrt{n}\tilde{\sigma}_l\tilde{\zeta}_{1l}^W} = \frac{n\lambda_1}{O_p(1)} = \infty, \text{ as } n\lambda_1 \to \infty. \tag{A.24}$$

Therefore, the first term in the brackets of (A.23) dominates the second term and the probability of the event

$$\left\{\mathrm{sgn}\left(\frac{1}{n}\frac{\partial \log L_p(\boldsymbol{\theta})}{\partial \zeta_{1l}}|_{\theta=\hat{\theta}_n^W}\right) = -\mathrm{sgn}(\hat{\zeta}_{1l}^W) \neq 0\right\}$$

tends to 1. But this contradicts the assumption that $\hat{\theta}_n^W$ is a local minimizer. □

Proof of Theorem 4.4: Asymptotic Normality

Write

$$\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \mathbf{diag}(\boldsymbol{\Omega})^T, \boldsymbol{\zeta}_{\mathcal{A}}^T, \boldsymbol{\zeta}_{\mathcal{A}^c}^T, \mathbf{vec}(\boldsymbol{\Lambda})^T)^T$$

and

$$\boldsymbol{\theta}_{\hat{\zeta}_A,0} = (\boldsymbol{\beta}^T, \mathbf{diag}(\boldsymbol{\Omega})^T, (\hat{\boldsymbol{\zeta}}_{\mathcal{A}}^W)^T, \boldsymbol{\zeta}_{\mathcal{A}^c}^T, \mathbf{vec}(\boldsymbol{\Lambda})^T)^T,$$

where $\boldsymbol{\zeta}_{\mathcal{A}^c} = \mathbf{0}$. From Theorem 4.2 and the selection consistency, it follows that with probability tending to one $\boldsymbol{\theta}_{\hat{\zeta}_A,0}$ is a root-n local minimizer of $-n^{-1}\log L_p(\boldsymbol{\theta})$. By using a Taylor expansion we find,

$$\begin{aligned}
\mathbf{0} = \frac{1}{n}\frac{\partial \log L_p(\boldsymbol{\theta})}{\partial \boldsymbol{\zeta}_{\mathcal{A}}}|_{\hat{\boldsymbol{\zeta}}_{\mathcal{A}}^W} &= \frac{1}{n}\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\zeta}_{\mathcal{A}}}|_{\boldsymbol{\zeta}_{\mathcal{A}}} + \frac{1}{n}\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\zeta}_{\mathcal{A}}\partial \boldsymbol{\zeta}_{\mathcal{A}}^T}|_{\boldsymbol{\zeta}_{\mathcal{A}}}(\hat{\boldsymbol{\zeta}}_{\mathcal{A}}^W - \boldsymbol{\zeta}_{\mathcal{A}}) \\
&+ \frac{1}{2n}\sum_{(l,k)\in\mathcal{A}}(\hat{\zeta}_{kl}^W - \zeta_{kl})\frac{\partial}{\partial \zeta_{kl}}\frac{\partial^3 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\zeta}_{\mathcal{A}}\partial \boldsymbol{\zeta}_{\mathcal{A}}^T}|_{\bar{\boldsymbol{\zeta}}_{\mathcal{A}}}(\hat{\boldsymbol{\zeta}}_{\mathcal{A}}^W - \boldsymbol{\zeta}_{\mathcal{A}}) \\
&- \sqrt{n}\lambda_1\sum_{(l,k)\in\mathcal{A}} n^{-\frac{1}{2}}a_k^l\mathrm{sgn}(\hat{\zeta}_{kl}^W),
\end{aligned} \tag{A.25}$$

where $\bar{\boldsymbol{\zeta}}_{\mathcal{A}}$ is on the line segment between $\boldsymbol{\zeta}_{\mathcal{A}}$ and $\hat{\boldsymbol{\zeta}}_{\mathcal{A}}^{W}$. According to the results in (A.23) and (A.24), law of large numbers, and regularity conditions (C1) and (c2), we have

$$\frac{1}{n}\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\zeta}_{\mathcal{A}} \partial \boldsymbol{\zeta}_{\mathcal{A}}^T}|_{\boldsymbol{\zeta}_{\mathcal{A}}} = -\boldsymbol{I}(\boldsymbol{\zeta}_{\mathcal{A}}),\ \hat{\boldsymbol{\zeta}}_{\mathcal{A}}^{W} - \boldsymbol{\zeta}_{\mathcal{A}} = o_p(1),\ \frac{1}{n}\frac{\partial}{\partial \zeta_{kl}}\frac{\partial^3 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\zeta}_{\mathcal{A}} \partial \boldsymbol{\zeta}_{\mathcal{A}}^T}|_{\bar{\boldsymbol{\zeta}}_{\mathcal{A}}} = O_p(1), \tag{A.26}$$

where $\boldsymbol{I}(\boldsymbol{\zeta}_{\mathcal{A}})$ is the submatrix of Fisher information matrix $\boldsymbol{I}(\boldsymbol{\theta}_0)$ corresponding to set $\mathcal{A}$. Then we have

$$(-\boldsymbol{I}(\boldsymbol{\zeta}_{\mathcal{A}}) + O_p(1))\sqrt{n}(\hat{\boldsymbol{\zeta}}_{\mathcal{A}}^{W} - \boldsymbol{\zeta}_{\mathcal{A}}) - \sqrt{n}\lambda_1 O_p(1) = -\frac{1}{\sqrt{n}}\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\zeta}_{\mathcal{A}}}|_{\boldsymbol{\zeta}_{\mathcal{A}}}. \tag{A.27}$$

Notice that $\frac{1}{\sqrt{n}}\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\zeta}_{\mathcal{A}}}|_{\boldsymbol{\zeta}_{\mathcal{A}}} \underset{d}{\rightarrow} N(\mathbf{0}, \boldsymbol{I}(\boldsymbol{\zeta}_{\mathcal{A}})^{-1})$ by the central limit theorem. Furthermore, $\sqrt{n}\lambda_1 = o_p(1)$ as $\lambda_1 = o_p(n^{-\frac{1}{2}})$. Therefore,

$$\sqrt{n}[\hat{\boldsymbol{\zeta}}_{\mathcal{A}}^{W} - \boldsymbol{\zeta}_{\mathcal{A}}] \underset{d}{\rightarrow} N(\mathbf{0}, \boldsymbol{I}(\boldsymbol{\zeta}_{\mathcal{A}})^{-1}), \tag{A.28}$$

which completes the proof. □

Proof of Theorem 4.5

Here we assume the same regularity condition ((C1) and (C2)) as in Theorem 4.2. Similar to the proof in Theorem 4.3, we first define the event $\mathcal{W}$ when the orderings of all components are correctly assigned as

$$\mathcal{W} = \cap_{l=1}^{p}(\{\hat{\boldsymbol{U}}_l = \boldsymbol{U}_l\} \cap \{\hat{\boldsymbol{V}}_l = \boldsymbol{V}_l\}). \tag{A.29}$$

Let $\hat{\boldsymbol{\zeta}}_{\mathcal{A}}^{\hat{W}}$ be $\hat{\boldsymbol{\zeta}}_{\mathcal{A}}^{W}$ when $\mathcal{W}$ occurs; otherwise, denote it as $\hat{\boldsymbol{\zeta}}_{\mathcal{A}\cap\mathcal{W}^c}$. Then, the estimator can be rewritten as

$$\hat{\boldsymbol{\zeta}}_{\mathcal{A}}^{\hat{W}} = \hat{\boldsymbol{\zeta}}_{\mathcal{A}}^{W}\mathbf{1}\{\mathcal{W}\} + \hat{\boldsymbol{\zeta}}_{\mathcal{A}\cap\mathcal{W}^c}\mathbf{1}\{\mathcal{W}^c\}$$

and therefore

$$\sqrt{n}(\hat{\boldsymbol{\zeta}}_{\mathcal{A}}^{\hat{W}} - \boldsymbol{\zeta}_{\mathcal{A}}) = \sqrt{n}(\hat{\boldsymbol{\zeta}}_{\mathcal{A}}^{W} - \boldsymbol{\zeta}_{\mathcal{A}})\mathbf{1}\{\mathcal{W}\} + \sqrt{n}(\hat{\boldsymbol{\zeta}}_{\mathcal{A}\cap\mathcal{W}^c} - \boldsymbol{\zeta}_{\mathcal{A}})\mathbf{1}\{\mathcal{W}^c\}. \tag{A.30}$$

By Theorem 4.4, we have $\sqrt{n}(\hat{\boldsymbol{\zeta}}_{\mathcal{A}}^{W} - \boldsymbol{\zeta}_{\mathcal{A}}) = O_p(1)$ and $\sqrt{n}(\hat{\boldsymbol{\zeta}}_{\mathcal{A}\cap\mathcal{W}^c} - \boldsymbol{\zeta}_{\mathcal{A}}) = O_p(1)$ as $n \to \infty$. By Lemma 4.1, we have $P(\mathcal{W}) \to 1$ and $P(\mathcal{W}^c) \to 0$ as $n \to \infty$. Therefore, by Slutsky's Theorem, $\sqrt{n}(\hat{\boldsymbol{\zeta}}_{\mathcal{A}}^{\hat{W}} - \boldsymbol{\zeta}_{\mathcal{A}})$ converges to the same distribution as $\sqrt{n}(\hat{\boldsymbol{\zeta}}_{\mathcal{A}}^{W} - \boldsymbol{\zeta}_{\mathcal{A}})$. Similar, by results from Theorem 4.4 and Lemma 4.1, we have selection consistency

$$P(\hat{\mathcal{A}}^{\hat{W}} = \mathcal{A}) = P(\hat{\mathcal{A}}^{\hat{W}} = \mathcal{A}|\mathcal{W})P(\mathcal{W}) \to 1, n \to \infty. \tag{A.31}$$

It completes the proof of Theorem 4.5. $\square$

APPENDIX B: TECHNICAL DETAILS OF CHAPTER 5

In this chapter, we give the proof to the main theoretical results: Theorem 5.1 and Theorem 5.2. The proofs rely on the following lemmas:

Lemma B.1. *Under Assumptions 5.1 and A.1-A.9, we have the following results:*

$$\tilde{\boldsymbol{B}}^*(\boldsymbol{s}) = \boldsymbol{B}^*(\boldsymbol{s}) + o_p(1),\ \tilde{\boldsymbol{\Gamma}}(\boldsymbol{s}) = \boldsymbol{\Gamma}(\boldsymbol{s}) + o_p(1),\ \boldsymbol{U}_{1:q} = (\boldsymbol{I}_n - \boldsymbol{P}_X)\boldsymbol{G} + o_p(1). \tag{B.1}$$

Proof: By right multiplying $\boldsymbol{P}_X$ on both sides of (5.5), we have

$$\breve{\boldsymbol{y}}_{.j}(\boldsymbol{s}_k) = \boldsymbol{X}\boldsymbol{\beta}_j^*(\boldsymbol{s}_k) + \breve{\boldsymbol{\eta}}_{.j}(\boldsymbol{s}_k) + \breve{\boldsymbol{\epsilon}}_{.j}(\boldsymbol{s}_k),\ j = 1, \ldots, J,$$

where $\breve{\boldsymbol{y}}_{.j}(\boldsymbol{s}_k) = \boldsymbol{P}_X\boldsymbol{y}_{.j}(\boldsymbol{s}_k)$, $\breve{\boldsymbol{\eta}}_{.j}(\boldsymbol{s}_k) = \boldsymbol{P}_X\boldsymbol{\eta}_{.j}(\boldsymbol{s}_k)$, and $\breve{\boldsymbol{\epsilon}}_{.j}(\boldsymbol{s}_k) = \boldsymbol{P}_X\boldsymbol{\epsilon}_{.j}(\boldsymbol{s}_k)$. It is easy to check that the LLK smoother of $\boldsymbol{\beta}_j^*$ in the model above is exactly the same as $\tilde{\boldsymbol{\beta}}_j^*$. Recall the assumptions in the model above and Theorem 1 in Zhu et al. (2012), the first part in this lemma follow immediately. Next, for the residual model (5.11), if the third part in this lemma holds, similarly it can be shown that $\tilde{\boldsymbol{A}}(\boldsymbol{s}) = \boldsymbol{A}(\boldsymbol{s}) + o_p(1)$, which leads to the second part due to the fact that $\boldsymbol{\Gamma}(\boldsymbol{s}) = \boldsymbol{Q}\boldsymbol{A}(\boldsymbol{s})$. Thus, the main task here is to prove the third part. Actually, by applying SVD, we have

$$(\boldsymbol{I}_n - \boldsymbol{P}_X)\boldsymbol{Z}\bar{\boldsymbol{\Gamma}} = \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{V}}^T, \tag{B.2}$$

where $\tilde{\boldsymbol{U}}$ is a $n \times q$ orthonormal matrix, $\tilde{\boldsymbol{V}}$ is a $Jn_v \times q$ orthonormal matrix, and $\tilde{\boldsymbol{\Lambda}}$ is a $q \times q$ diagonal matrix of the ordered singular values. Then, based on the result that $\tilde{\boldsymbol{B}}^*(\boldsymbol{s}) = \boldsymbol{B}^*(\boldsymbol{s}) + o_p(1)$, the extended residual matrix $\bar{\boldsymbol{R}}$ can be written as

$$\bar{\boldsymbol{R}} = \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{V}}^T + \mathbf{M}, \tag{B.3}$$

where $\mathbf{M} = \bar{\boldsymbol{\eta}} + \bar{\boldsymbol{\epsilon}} + o_p(1)$, and $\bar{\boldsymbol{\eta}} + \bar{\boldsymbol{\epsilon}}$ are constructed in the same way as $\bar{\boldsymbol{R}}$. Then, from

Assumption A.5, Theorem 1 in Lee et al. (2014), and Lemma 1 in Lee et al. (2017), we have $\boldsymbol{U}_{1:q}^T\tilde{\boldsymbol{U}} = \boldsymbol{I}_q + o_p(1)$, which yields $\boldsymbol{U}_{1:q} = \tilde{\boldsymbol{U}} + o_p(1)$. Recall the definition of matrix $\tilde{\boldsymbol{U}}$ in (B.2), there exists a $q \times q$ orthonormal matrix $\boldsymbol{Q}$ such that $\tilde{\boldsymbol{U}} = (\boldsymbol{I}_n - \boldsymbol{P}_X)\boldsymbol{Z}\boldsymbol{Q}$, which leads to the third part of this lemma. □

Lemma B.2. *Under Assumptions 5.1 and A.1-A.9, we have that for each j, the following result holds uniformly for all $\boldsymbol{s} \in \mathcal{S}$:*

$$\sqrt{n}\boldsymbol{M}\sum_{k=1}^{n_v} a_k(\boldsymbol{H}, \boldsymbol{s})\boldsymbol{\epsilon}_{.j}(\boldsymbol{s}_k) = o_p(1), \tag{B.4}$$

where $\boldsymbol{M} = (\boldsymbol{I}_p, \boldsymbol{0}_{q\times q})(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T$.

Proof: According to the definition of $a_k(\boldsymbol{H}, \boldsymbol{s})$ and the properties of Kronecker product, the left hand side in (B.4) can be written as

$$[\boldsymbol{I}_p, \boldsymbol{0}_q]\left\{[\frac{1}{n}\boldsymbol{W}^T\boldsymbol{W}]^{-1} \otimes \left[(1, \boldsymbol{0}_{1\times d})[\frac{1}{n_v}\sum_{k=1}^{n_v} K_H(\boldsymbol{s}_k - \boldsymbol{s})\boldsymbol{z}_H(\boldsymbol{s}_k - \boldsymbol{s})^{\otimes 2}]^{-1}\right]\right\}\boldsymbol{\epsilon}_{.j}^X(\boldsymbol{s}),$$

where $\boldsymbol{\epsilon}_{.j}^X(\boldsymbol{s}) = n^{-1/2}n_v^{-1}\sum_{k=1}^{n_v}[\boldsymbol{W}^T \otimes K_H(\boldsymbol{s}_k - \boldsymbol{s})\boldsymbol{z}_H(\boldsymbol{s}_k - \boldsymbol{s})]\boldsymbol{\epsilon}_{.j}(\boldsymbol{s}_k)$. According to Lemma 1 and Lemma 2 in Zhu et al. (2012), when $d = 1$, we have that

$$\frac{1}{n_v}\sum_{k=1}^{n_v} K_H(\boldsymbol{s}_k - \boldsymbol{s})\boldsymbol{z}_H(\boldsymbol{s}_k - \boldsymbol{s})^{\otimes 2} = \boldsymbol{\Omega}_K(\boldsymbol{H}, \boldsymbol{s}) + o_p(1),\ \boldsymbol{\epsilon}_{.j}^X(\boldsymbol{s}) = o_p(1) \tag{B.5}$$

hold uniformly for all $\boldsymbol{s} \in \mathcal{S}$. In addition, all the results above can be straightforwardly extended to the situations that $d > 1$. Thus, combining the fact that $\frac{1}{n}\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{\Omega}_w + o_p(1)$, we can finish the proof of Lemma B.2. □

Lemma B.3. *Under Assumptions 5.1 and A.1-A.9, we have that for each j,*

$$\sqrt{n}[\boldsymbol{M}^T\boldsymbol{M}]^{-\frac{1}{2}}\boldsymbol{M}\sum_{k=1}^{n_v} a_k(\boldsymbol{H}, \boldsymbol{s})\boldsymbol{\eta}_{.j}(\boldsymbol{s}_k)$$

weakly converges to a centered Gaussian process with covariance function $\Sigma_{j,j}(\boldsymbol{s}, \boldsymbol{s}')\boldsymbol{I}_p$, where

$\Sigma_{j,j}(\boldsymbol{s}, \boldsymbol{s}')$ is the j-th diagonal element in $\boldsymbol{\Sigma}(\boldsymbol{s}, \boldsymbol{s}')$.

Proof: This proof consists of 2 steps. In step 1, it follows from the standard central limit theorem that for each $\boldsymbol{s} \in \mathcal{S}$,

$$\sqrt{n}[\boldsymbol{M}^T\boldsymbol{M}]^{-\frac{1}{2}}\boldsymbol{M}\sum_{k=1}^{n_v} a_k(\boldsymbol{H}, \boldsymbol{s})\boldsymbol{\eta}_{.j}(\boldsymbol{s}_k) \xrightarrow{L} N(\boldsymbol{0}, \Sigma_{j,j}(\boldsymbol{s}, \boldsymbol{s})\boldsymbol{I}_p), \tag{B.6}$$

where $\xrightarrow{L}$ denotes convergence in distribution.

In step 2, we show the asymptotic tightness of $\sqrt{n}[\boldsymbol{M}^T\boldsymbol{M}]^{-\frac{1}{2}}\boldsymbol{M}\sum_{k=1}^{n_v} a_k(\boldsymbol{H}, \boldsymbol{s})\boldsymbol{\eta}_{.j}(\boldsymbol{s}_k)$. we first define that

$$\boldsymbol{\Delta}(\boldsymbol{H}, \eta_{ij}(\boldsymbol{s})) = \frac{1}{n_v}K_H(\boldsymbol{s}_k - \boldsymbol{s})\boldsymbol{z}_H(\boldsymbol{s}_k - \boldsymbol{s})\eta_{ij}(\boldsymbol{s}_k) - \int_{\mathcal{S}} K_H(\boldsymbol{u} - \boldsymbol{s})\boldsymbol{z}_H(\boldsymbol{u} - \boldsymbol{s})\eta_{ij}(\boldsymbol{u})p(\boldsymbol{u})d\boldsymbol{u}.$$

According to the results in Lemma B.2, we can show that $\sqrt{n}\boldsymbol{M}\sum_{k=1}^{n_v} a_k(\boldsymbol{H}, \boldsymbol{s})\boldsymbol{\eta}_{.j}(\boldsymbol{s}_k)[1 + o_p(1)]$ an be approximated by three terms as follows:

$$\sqrt{n}\boldsymbol{M}\sum_{k=1}^{n_v} a_k(\boldsymbol{H}, \boldsymbol{s})\boldsymbol{\eta}_{.j}(\boldsymbol{s}_k)[1 + o_p(1)] = (\mathbf{I}) + (\mathbf{II}) + (\mathbf{III}),$$

where

$$\begin{aligned}
(\mathbf{I}) &= \boldsymbol{\Omega}_w^{-1} \otimes [(1, \boldsymbol{0}_{1\times d})\boldsymbol{\Omega}_K^{-1}(\boldsymbol{H}, \boldsymbol{s})]n^{-1/2}\sum_{i=1}^{n} \boldsymbol{w}_i \otimes \boldsymbol{\Delta}(\boldsymbol{H}, \eta_{ij}(\boldsymbol{s})), \\
(\mathbf{II}) &= n^{-1/2}\boldsymbol{\Omega}_w^{-1}\boldsymbol{W}^T\boldsymbol{\eta}_{.j}[1 + o_p(|\boldsymbol{H}|)], \\
(\mathbf{III}) &= \boldsymbol{\Omega}_w^{-1} \otimes [(1, \boldsymbol{0}_{1\times d})\boldsymbol{\Omega}_K^{-1}(\boldsymbol{H}, \boldsymbol{s})] \\
&\quad \int_{\mathcal{D}} n^{-1/2}\sum_{i=1}^{n}[\boldsymbol{w}_i(\eta_{ij}(\boldsymbol{s} + \boldsymbol{H}\boldsymbol{u}) - \eta_{ij}(\boldsymbol{s}))] \otimes [K_H(\boldsymbol{u})\boldsymbol{z}_H(\boldsymbol{u})]p(\boldsymbol{s} + \boldsymbol{H}\boldsymbol{u})d\boldsymbol{u}.
\end{aligned}$$

Here $\mathcal{D} \doteq \{\boldsymbol{u} : \boldsymbol{u} \in \mathcal{S} \text{ and } \boldsymbol{s} + \boldsymbol{H}\boldsymbol{u} \in \mathcal{S}\}$. We investigate the three terms above as follows.

For item **(I)**, it follows from Lemma 3 in Zhu et al. (2012) that

$$\sup_{\mathcal{S}} |n^{-1/2} \sum_{i=1}^{n} \boldsymbol{w}_i \otimes \boldsymbol{\Delta}(\boldsymbol{H}, \eta_{ij}(\boldsymbol{s}))| = o_p(1), \tag{B.7}$$

which yields that the term **(I)** converges to zero uniformly.

For item **(II)**, we define that

$$\boldsymbol{\varepsilon}_\eta \doteq \{f(\boldsymbol{s}, \boldsymbol{W}, \boldsymbol{\eta}_{.j}) = \boldsymbol{\Omega}_w^{-1}\boldsymbol{W}^T\boldsymbol{\eta}_{.j}(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{S}\}.$$

Due to Assumption A.9, $\boldsymbol{\varepsilon}_\eta$ is a P-Donsker class (Kosorok, 2008).

For term **(III)**, by using the same argument in the second term **(II)**, we can show that the asymptotic tightness of $n^{-1/2}\boldsymbol{W}^T\boldsymbol{\eta}_{.j}(\boldsymbol{s})$. Therefore, for any $|\boldsymbol{H}| \to 0$, we have

$$\sup_{\boldsymbol{s}\in\mathcal{S}, \boldsymbol{u}\in\mathcal{D}} |n^{-1/2} \sum_{i=1}^{n} [\boldsymbol{w}_i(\eta_{ij}(\boldsymbol{s}+\boldsymbol{H}\boldsymbol{u}) - \eta_{ij}(\boldsymbol{s}))]| = o_p(1). \tag{B.8}$$

It follows from Assumptions A.1 and A.6 and (B.8) that the term **(III)** converges to zero uniformly.

Combining (B.7), (B.8) and the Donsker property of term **(II)**, it suffices to show the asymptotic tightness. Thus, we can finish the proof of Lemma B.3. □

Proof of Theorem 5.1

Proof of Theorem 5.1 (i):

Since $\boldsymbol{G} = \boldsymbol{Z}\boldsymbol{Q}$, $\boldsymbol{G}$ can be treated as linear combinations of columns in $\boldsymbol{Z}$. Thus the column space of $\boldsymbol{G}$ is the same as that of $\boldsymbol{Z}$. To prove the first part in Theorem 5.1, we only need to show that $\hat{\boldsymbol{G}} = \boldsymbol{G} + o_p(1)$. According to Assumption A.2 and the results in Lemma B.1, $\hat{\boldsymbol{G}}$ in (5.14) can be derived as

$$\hat{\boldsymbol{G}} = (\boldsymbol{I} - \boldsymbol{P}_x)\boldsymbol{G} + \boldsymbol{X}\int_s \boldsymbol{B}^*(\boldsymbol{s})(\boldsymbol{I}_J - \boldsymbol{P}_J)\boldsymbol{A}^T(\boldsymbol{s})p(\boldsymbol{s})d\boldsymbol{s}\bar{\boldsymbol{\Omega}}^{-1} + o_p(1), \tag{B.9}$$

where $\bar{\boldsymbol{\Omega}} = \int_s \boldsymbol{A}(\boldsymbol{s})(\boldsymbol{I}_J - \boldsymbol{P}_J)\boldsymbol{A}^T(\boldsymbol{s})p(\boldsymbol{s})d\boldsymbol{s}$. Substituting the definition of $\boldsymbol{B}^*$ into (B.9), we have

$$\hat{\boldsymbol{G}} = \boldsymbol{G} + \int_s \boldsymbol{B}(\boldsymbol{s})(\boldsymbol{I}_J - \boldsymbol{P}_J)\boldsymbol{\Gamma}^T\boldsymbol{Q}d\boldsymbol{s}\bar{\boldsymbol{\Omega}}^{-1} + o_p(1), \tag{B.10}$$

which yields that $\hat{\boldsymbol{G}} = \boldsymbol{G} + o_p(1)$ according to Assumption 5.1. This completes the proof. □

Proof of Theorem 5.1 (ii):

According to the expressions of $\hat{\boldsymbol{\beta}}_j^*(\boldsymbol{s})$ in (5.6) and $\hat{\boldsymbol{\alpha}}_j(\boldsymbol{s})$ in (5.12), we have

$$\tilde{\boldsymbol{\beta}}_j(\boldsymbol{s}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{I}_n - \hat{\boldsymbol{G}}\boldsymbol{U}_{1:q}^T)\sum_{k=1}^{n_v} a_k(\boldsymbol{H}, \boldsymbol{s})\boldsymbol{y}_{.j}(\boldsymbol{s}_k), j = 1, \ldots, J, \tag{B.11}$$

which holds when $\boldsymbol{H}_\beta = \boldsymbol{H}_\alpha = \boldsymbol{H}$. Recall the results that $\boldsymbol{U}_{1:q} = (\boldsymbol{I}_n - \boldsymbol{P}_X)\boldsymbol{G} + o_p(1)$ in Lemma B.1 and $\hat{\boldsymbol{G}} = \boldsymbol{G} + o_p(1)$ in proving Theorem 5.1 (i), $\tilde{\boldsymbol{\beta}}_j(\boldsymbol{s}), j = 1, \ldots, J$, can be written as

$$\tilde{\boldsymbol{\beta}}_j(\boldsymbol{s}) = \boldsymbol{M}\sum_{k=1}^{n_v} a_k(\boldsymbol{H}, \boldsymbol{s})\boldsymbol{y}_{.j}(\boldsymbol{s}_k) + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sum_{k=1}^{n_v} a_k(\boldsymbol{H}, \boldsymbol{s})\boldsymbol{y}_{.j}(\boldsymbol{s}_k)o_p(1), \tag{B.12}$$

where $\boldsymbol{M} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T[\boldsymbol{I}_n - \boldsymbol{G}\boldsymbol{G}^T(\boldsymbol{I}_n - \boldsymbol{P}_X)]$. According to the partitioned matrix inversion theory (Bhatia, 2013), it's easy to show that $\boldsymbol{M} = (\boldsymbol{I}_p, \boldsymbol{0}_{q\times q})(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T$. Here we define

$$\sqrt{n}(\tilde{\boldsymbol{\beta}}_j(\boldsymbol{s}) - \mathbb{E}[\tilde{\boldsymbol{\beta}}_j(\boldsymbol{s})]) \doteq \boldsymbol{T}_{1,j}(\boldsymbol{s}) + \boldsymbol{T}_{2,j}(\boldsymbol{s}) + \boldsymbol{T}_{3,j}(\boldsymbol{s}), j = 1, \ldots, J,$$

where

$$\boldsymbol{T}_{1,j}(\boldsymbol{s}) = \sqrt{n}\boldsymbol{M}\sum_{k=1}^{n_v} a_k(\boldsymbol{H},\boldsymbol{s})\boldsymbol{\epsilon}_{.j}(\boldsymbol{s}_k),\ \boldsymbol{T}_{2,j}(\boldsymbol{s}) = \sqrt{n}\boldsymbol{M}\sum_{k=1}^{n_v} a_k(\boldsymbol{H},\boldsymbol{s})\boldsymbol{\eta}_{.j}(\boldsymbol{s}_k),$$

$$\boldsymbol{T}_{3,j}(\boldsymbol{s}) = \sqrt{n}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sum_{k=1}^{n_v} a_k(\boldsymbol{H},\boldsymbol{s})[\boldsymbol{\eta}_{.j}(\boldsymbol{s}_k)+\boldsymbol{\epsilon}_{.j}(\boldsymbol{s}_k)]o_p(1). \qquad \text{(B.13)}$$

Next, we can investigate the three terms above. First, according to Lemma B.2, $\boldsymbol{T}_{1,j}(\boldsymbol{s}) = o_p(1)$ holds uniformly for all $\boldsymbol{s} \in \mathcal{S}$. Second, according to Lemma B.3, $[\boldsymbol{M}\boldsymbol{M}^T]^{-\frac{1}{2}}\boldsymbol{T}_{2,j}(\boldsymbol{s})$ weakly converges to a centered Gaussian process with covariance matrix $\Sigma_{j,j}(\boldsymbol{s},\boldsymbol{s}')\boldsymbol{I}_p$. Third, by using the same argument in proving Lemma B.1 and Lemma Lemma B.3, $\boldsymbol{T}_{3,j}(\boldsymbol{s}) = o_p(1)$ holds uniformly for all $\boldsymbol{s} \in \mathcal{S}$. Combining the properties of the three terms, it is easy to show the weak convergence in the second part of Theorem 5.1. This completes the proof. □

Proof of Theorem 5.2

Theorem 5.2 (i):

Theorem 5.2 (i) is similar to Theorem 4 in Zhu et al. (2012) and Theorem 7 in Zhang and Chen (2007). All of the three theorems characterize the asymptotic distribution of the global test statistic under null hypothesis. In particular, the asymptotic distribution is delineated as a χ^2-type mixture in Zhang and Chen (2007). All discussions and proof associated with Theorem 7 in Zhang and Chen (2007) are valid here, and therefore, we do not repeat them for the sake of space.

Proof of Theorem 5.2 (ii):

First we define that $\tilde{\boldsymbol{\delta}}(\boldsymbol{s}) \doteq [\boldsymbol{C}(\hat{\boldsymbol{\Sigma}}_\eta(\boldsymbol{s},\boldsymbol{s}) \otimes [\hat{\boldsymbol{M}}\hat{\boldsymbol{M}}^T])\boldsymbol{C}^T]^{-1/2}\boldsymbol{\delta}(\boldsymbol{s})$. Under $\mathbf{H_{1n}}$, we have

$$\tilde{\boldsymbol{\delta}}(\boldsymbol{s}) \overset{\text{asymp}}{\sim} GP(\boldsymbol{\mu}_{1n}(\boldsymbol{s}), \boldsymbol{\Lambda}_{1n}(\boldsymbol{s},\boldsymbol{s}')), \qquad \text{(B.14)}$$

where $\boldsymbol{\mu}_{1n}(\boldsymbol{s}) = [\boldsymbol{C}(\hat{\boldsymbol{\Sigma}}_\eta(\boldsymbol{s},\boldsymbol{s}) \otimes [\hat{\boldsymbol{M}}\hat{\boldsymbol{M}}^T])\boldsymbol{C}^T]^{-1/2}n^{-\tau/2}[\boldsymbol{C}\mathbf{vec}(\boldsymbol{B}(\boldsymbol{s})) - \boldsymbol{b}_0(\boldsymbol{s})]$, and $\boldsymbol{\Lambda}_{1n}(\boldsymbol{s},\boldsymbol{s}') = \mathbf{cov}(\tilde{\boldsymbol{\delta}}(\boldsymbol{s}), \tilde{\boldsymbol{\delta}}(\boldsymbol{s}'))$. We consider a Hilbert space of r-dimensional vectors of functions in $L_2(\boldsymbol{s})$

denoted by $\mathbb{H}$. Define the corresponding inner product as

$$< \boldsymbol{f}(\boldsymbol{s}), \boldsymbol{g}(\boldsymbol{s}) >_{\mathbb{H}} = \sum_{t=1}^{r} < f_t(\boldsymbol{s}), g_t(\boldsymbol{s}) >, \quad < f_t(\boldsymbol{s}), g_t(\boldsymbol{s}) > = \int_S f_t(\boldsymbol{s}) g_t(\boldsymbol{s}) p(\boldsymbol{s}) d\boldsymbol{s}.$$

By the multivariate version of Mercer's theorem, there exists a set of orthonormal basis functions $\boldsymbol{\phi}_l(\boldsymbol{s}) = (\phi_{l1}(\boldsymbol{s}), \ldots, \phi_{lr}(\boldsymbol{s}))$ in $\mathbb{H}$ such that

$$\boldsymbol{\Lambda}_{1n}(\boldsymbol{s}, \boldsymbol{s}') = \sum_{l=1}^{\infty} \lambda_l \boldsymbol{\phi}_l(\boldsymbol{s}) \boldsymbol{\phi}_l^T(\boldsymbol{s}').$$

Let $\xi_{lt} =< \tilde{\delta}_t(\boldsymbol{s}), \boldsymbol{\phi}_{lt}(\boldsymbol{s}) >$, in which $\tilde{\delta}_t(\boldsymbol{s})$ is the t-th element in $\tilde{\boldsymbol{\delta}}(\boldsymbol{s})$. Then we have $\xi_{lt} \sim N(\nu_{lt}, \lambda_l)$, where $\nu_{lt} =< \mu_{1n,t}(\boldsymbol{s}), \phi_{lt}(\boldsymbol{s}) >$, and $\mu_{1n,t}(\boldsymbol{s})$ is the t-th element in $\boldsymbol{\mu}_{1n}(\boldsymbol{s})$. It is assumed that the eigenvalues are ordered in decreasing values. Without loss of generality, the first m eigenvalues are assumed to be positive. If all eigenvalues are positive, we set $m = \infty$. It is easy to see that

$$T_n = \int_S \tilde{\boldsymbol{\delta}}(\boldsymbol{s})^T \tilde{\boldsymbol{\delta}}(\boldsymbol{s}) p(\boldsymbol{s}) d\boldsymbol{s} = \sum_{l=1}^{\infty} \sum_{t=1}^{r} \xi_{lt}^2 = \sum_{l=1}^{m} \lambda_l A_l + \sum_{l=m+1}^{\infty} \sum_{t=1}^{r} \nu_{lt}^2,$$

where $A_l \sim \chi_r^2(\nu_l^2/\lambda_l)$, in which $\nu_l^2 = \sum_{t=1}^{r} \nu_{lt}^2$. Similar results have been obtained and discussed in Zhang and Chen (2007) and Zhang (2011).

Under $\mathbf{H_{1n}}$, we have $\nu_l^2 = n^{1-\tau} \zeta_{1n,l}^2$, where $\zeta_{1n,l}$ is given by

$$\zeta_{1n,l} = \sum_{t=1}^{r} \int_S \{[\boldsymbol{C}(\hat{\boldsymbol{\Sigma}}_\eta(\boldsymbol{s}, \boldsymbol{s}) \otimes [\hat{\boldsymbol{M}} \hat{\boldsymbol{M}}^T]) \boldsymbol{C}^T]^{-1/2} \boldsymbol{\delta}(\boldsymbol{s})\}_t \phi_{lt}(\boldsymbol{s}) p(\boldsymbol{s}) d\boldsymbol{s}.$$

Note that $A_l \overset{\text{d}}{=} \sum_{t=1}^{r-1} a_{lt}^2 + [a_{lr} + n^{(1-\tau)/2} \lambda_l^{-1/2} \zeta_{1n,l}]^2$, where $a_{lt} \sim N(0, 1)$. Thus, we have

$$T_n \overset{\text{d}}{=} \sum_{l=1}^{m} \lambda_l A^* + 2n^{(1-\tau)/2} \sum_{l=1}^{m} \lambda_l^{1/2} \zeta_{1n,l} a_{lr} + n^{1-\tau} \sum_{l=1}^{m} \zeta_{1n,l}^2$$

by dropping higher order terms, where $A^* \sim \chi_r^2$. As $n \to \infty$, the last two terms on the right-hand side above dominate the first term. Therefore, T_n is asymptotically normally distributed under $\mathbf{H_{1n}}$ with mean $n^{1-\tau}\sum_{l=1}^{m}\zeta_{1n,l}^2$ and variance $4n^{1-\tau}\sum_{l=1}^{m}\lambda_l a_{lr}^2$. Therefore, we have

$$\mathbb{P}\{T_n > T_{n,\alpha}|\mathbf{H_{1n}}\} = \Phi\left(n^{(1-\tau)/2}\frac{\sum_{l=1}^{m}\zeta_{1n,l}^2}{2\sqrt{\sum_{l=1}^{m}\lambda_l a_{lr}^2}}\right) + o_p(1)$$

which tends to 1 as $n \to \infty$. This completes the proof. □

REFERENCES

Amaral, G. A., Dryden, I., and Wood, A. T. A. (2007). Pivotal bootstrap methods for k-sample problems in directional statistics and shape analysis. *Journal of the American Statistical Association* **102,** 695–707.

Arenaza-Urquijo, E. M., Landeau, B., La Joie, R., Mevel, K., Mézenge, F., Perrotin, A., Desgranges, B., Bartrés-Faz, D., Eustache, F., and Chételat, G. (2013). Relationships between years of education and gray matter volume, metabolism and functional connectivity in healthy elders. *Neuroimage* **83,** 450–457.

Bae, K., Shim, H., Tao, C., Chang, S., Wang, J., Boudreau, R., and Kwoh, C. (2009). Intra- and inter-observer reproducibility of volume measurement of knee cartilage segmented from the oai mr image set using a novel semi-automated segmentation method. *Osteoarthritis and Cartilage* **17,** 1589–1597.

Banerjee, A., Dhillon, I. S., Ghosh, J., and Sra, S. (2005). Clustering on the unit hypersphere using von mises-fisher distributions. In *Journal of Machine Learning Research*, pages 1345–1382.

Benjamini, Y., Yekutieli, D., et al. (2001). The control of the false discovery rate in multiple testing under dependency. *The annals of statistics* **29,** 1165–1188.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)* **36,** 192–236.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B. Methodological* **48,** 259–302.

Bhatia, R. (2013). *Matrix analysis*, volume 169. Springer Science & Business Media.

Biswal, S., Hastie, T., Andriacchi, T. P., Bergman, G. A., Dillingham, M. F., and Lang, P. (2002a). Risk factors for progressive cartilage loss in the knee. *Arthritis & Rheumatism* **46,** 2884–2892.

Biswal, S., Hastie, T., Andriacchi, T. P., Bergman, G. A., Dillingham, M. F., and Lang, P. (2002b). Risk factors for progressive cartilage loss in the knee. *Arthritis & Rheumatism* **46,** 2884–2892.

Boothby, W. M. (2003). *An Introduction to Differentiable Manifolds and Riemannian Geometry*, volume 120. Gulf Professional Publishing.

Bouveyron, C. and Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis* **71,** 52–78.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning* **3,** 1–122.

Brem, M., Lang, P., Neumann, G., Schlechtweg, P., Schneider, E., Jackson, R., Yu, J., Eaton, C., Hennig, F., Yoshioka, H., et al. (2009). Magnetic resonance image segmentation using semi-automated software for quantification of knee articular cartilageÂ¡Âªinitial evaluation of a technique for paired scans. *Skeletal radiology* **38,** 505–511.

Buchta, C., Kober, M., Feinerer, I., and Hornik, K. (2012). Spherical k-means clustering. *Journal of Statistical Software* **50,** 1–22.

Buck, R. J., Dreher, D., and Eckstein, F. (2012). Femorotibial cartilage thickness change distributions for subjects without signs, symptoms, or risk factors of knee osteoarthritis. *Cartilage* **3,** 305–313.

Buck, R. J., Wyman, B. T., Hudelmaier, M., Wirth, W., Eckstein, F., et al. (2009). Does the use of ordered values of subregional change in cartilage thickness improve the detection of disease progression in longitudinal studies of osteoarthritis? *Arthritis Care & Research* **61,** 917–924.

Buja, A. and Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate behavioral research* **27,** 509–540.

Byun, M. S., Kim, S. E., Park, J., Yi, D., Choe, Y. M., Sohn, B. K., Choi, H. J., Baek, H., Han, J. Y., Woo, J. I., et al. (2015). Heterogeneity of regional brain atrophy patterns associated with distinct progression rates in alzheimer's disease. *PLoS One* **10,** e0142756.

Canas, G., Poggio, T., and Rosasco, L. (2012). Learning manifolds with k-means and k-flats. In *Advances in Neural Information Processing Systems*, pages 2465–2473.

Cetingul, H. E. and Vidal, R. (2009). Intrinsic mean shift for clustering on stiefel and grassmann manifolds. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1896–1902. IEEE.

Chang, A., Moisio, K., Chmiel, J. S., Eckstein, F., Guermazi, A., Almagor, O., Cahue, S., Wirth, W., Prasad, P., and Sharma, L. (2011). Subregional effects of meniscal tears on cartilage loss over 2 years in knee osteoarthritis. *Annals of the rheumatic diseases* **70,** 74–79.

Chen, J., Liu, J., Calhoun, V. D., Arias-Vasquez, A., Zwiers, M. P., Gupta, C. N., Franke, B., and Turner, J. A. (2014). Exploration of scanning effects in multi-site structural mri studies. *Journal of neuroscience methods* **230,** 37–50.

Cicuttini, F., Hankin, J., Jones, G., and Wluka, A. (2005). Comparison of conventional standing knee radiographs and magnetic resonance imaging in assessing progression of tibiofemoral joint osteoarthritis. *Osteoarthritis and cartilage* **13,** 722–727.

Cornea, E., Zhu, H., Kim, P., Ibrahim, J. G., and Initiative, A. D. N. (2017). Regression models on riemannian symmetric spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79,** 463–482.

Davatzikos, C. (2018). Machine learning in neuroimaging: Progress and challenges. *NeuroImage* .

Davatzikos, C., Genc, A., Xu, D., and Resnick, S. M. (2001). Voxel-based morphometry using the ravens maps: methods and validation using simulated longitudinal atrophy. *NeuroImage* **14,** 1361–1369.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977a). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39,** 1–38.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977b). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* pages 1–38.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology* **26,** 297–302.

Diggle, P. and Ribeiro, P. J. (2007). *Model-based geostatistics.* Springer.

Dryden, I. and Mardia, K. (1998). *Statistical Analysis of Shape.* Wiley.

Dryden, I. L. et al. (2005). Statistical analysis on high-dimensional spheres and shape spaces. *The Annals of Statistics* **33,** 1643–1665.

Eckstein, F., Buck, R. J., Wyman, B. T., Kotyk, J. J., Graverand, L., Hellio, M.-P., Remmers, A. E., Evelhoch, J. L., Hudelmaier, M., and Charles, H. C. (2007). Quantitative imaging of cartilage morphology at 3.0 tesla in the presence of gadopentate dimeglumine (gd-dtpa). *Magnetic Resonance in Medicine* **58,** 402–406.

Eckstein, F., Cicuttini, F., Raynauld, J.-P., Waterton, J., and Peterfy, C. (2006). Magnetic resonance imaging (mri) of articular cartilage in knee osteoarthritis (oa): morphological assessment. *Osteoarthritis and cartilage* **14,** 46–75.

Eckstein, F., Gavazzeni, A., Sittek, H., Haubner, M., Lösch, A., Milz, S., Englmeier, K.-H., Schulte, E., Putz, R., and Reiser, M. (1996). Determination of knee joint cartilage thickness using three-dimensional magnetic resonance chondro-crassometry (3d mr-ccm). *Magnetic Resonance in Medicine* **36,** 256–265.

Eckstein, F., Wirth, W., and Nevitt, M. C. (2012). Recent advances in osteoarthritis imagingÂ¡Âªthe osteoarthritis initiative. *Nature Reviews Rheumatology* **8,** 622–630.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*, volume 57. CRC press.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman and Hall, London.

Fan, J., Peng, H., et al. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* **32,** 928–961.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics* **27,** 1491–1518.

Felson, D. T., Lawrence, R. C., Dieppe, P. A., Hirsch, R., Helmick, C. G., Jordan, J. M., Kington, R. S., Lane, N. E., Nevitt, M. C., Zhang, Y., et al. (2000). Osteoarthritis: new insights. part 1: the disease and its risk factors. *Annals of internal medicine* **133,** 635–646.

Felson, D. T., Niu, J., Yang, T., Torner, J., Lewis, C. E., Aliabadi, P., Sack, B., Sharma, L., Guermazi, A., Goggins, J., et al. (2013). Physical activity, alignment and knee osteoarthritis: data from most and the oai. *Osteoarthritis and Cartilage* **21,** 789–795.

Ferreira, D., Verhagen, C., Hernández-Cabrera, J. A., Cavallin, L., Guo, C.-J., Ekman, U., Muehlboeck, J.-S., Simmons, A., Barroso, J., Wahlund, L.-O., et al. (2017). Distinct subtypes of alzheimer's disease based on patterns of brain atrophy: longitudinal trajectories and clinical applications. *Scientific reports* **7,** 46263.

Fletcher, P. T. (2013). Geodesic regression and the theory of least squares on riemannian manifolds. *International journal of computer vision* **105,** 171–185.

Fletcher, P. T., Lu, C., Pizer, S. M., and Joshi, S. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE transactions on medical imaging* **23,** 995–1005.

Fokoué, E. (2005). Mixtures of factor analyzers: an extension with covariates. *Journal of Multivariate Analysis* **95,** 370–384.

Fop, M., Murphy, T. B., et al. (2018). Variable selection methods for model-based clustering. *Statistics Surveys* **12,** 18–65.

Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., Adams, P., Cooper, C., Fava, M., McGrath, P. J., et al. (2018). Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage* **167,** 104–120.

Fortin, J.-P., Labbe, A., Lemire, M., Zanke, B. W., Hudson, T. J., Fertig, E. J., Greenwood, C. M., and Hansen, K. D. (2014). Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome biology* **15,** 503.

Fortin, J.-P., Parker, D., Tunc, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E., et al. (2017). Harmonization of multi-site diffusion tensor imaging data. *Neuroimage* **161,** 149–170.

Fortin, J.-P., Sweeney, E. M., Muschelli, J., Crainiceanu, C. M., Shinohara, R. T., Initiative, A. D. N., et al. (2016). Removing inter-subject technical variability in magnetic resonance imaging studies. *NeuroImage* **132,** 198–212.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association* **97,** 611–631.

Friguet, C., Kloareg, M., and Causeur, D. (2009). A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association* **104,** 1406–1415.

Frisoni, G., Sabattoli, F., Lee, A., Dutton, R., Toga, A., and Thompson, P. (2006). In vivo neuropathology of the hippocampal formation in AD: a radial mapping MR-based study. *Neuroimage* **32,** 104–110.

Frisoni, G. B., Ganzola, R., Canu, E., Rüb, U., Pizzini, F. B., Alessandrini, F., Zoccatelli, G., Beltramello, A., Caltagirone, C., and Thompson, P. M. (2008). Mapping local hippocampal changes in alzheimer's disease and normal ageing with mri at 3 tesla. *Brain* **131,** 3266–3276.

Gagnon-Bartsch, J. A., Jacob, L., and Speed, T. P. (2013). Removing unwanted variation from high dimensional data with negative controls. *Berkeley: Tech Reports from Dep Stat Univ California* pages 1–112.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **6,** 721–741.

Geman, S. and Graffigne, C. (1986). Markov random field image models and their applications to computer vision. In *Proceedings of the International Congress of Mathematicians*, volume 1, page 2. AMS, Providence, RI.

Goh, A. and Vidal, R. (2008). Clustering and dimensionality reduction on riemannian manifolds. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on.* IEEE.

Gougoutas, A. J., Wheaton, A. J., Borthakur, A., Shapiro, E. M., Kneeland, J. B., Udupa, J. K., and Reddy, R. (2004). Cartilage volume quantification via live wire segmentation< sup> 1</sup>. *Academic radiology* **11,** 1389–1395.

Grau, V., Mewes, A., Alcaniz, M., Kikinis, R., and Warfield, S. K. (2004). Improved watershed transform for medical image segmentation using prior information. *Medical Imaging, IEEE Transactions on* **23,** 447–458.

Guillaume, B., Wang, C., Poh, J., Shen, M. J., Ong, M. L., Tan, P. F., Karnani, N., Meaney, M., and Qiu, A. (2018). Improving mass-univariate analysis of neuroimaging data by modelling important unknown covariates: Application to epigenome-wide association studies. *NeuroImage* **173,** 57–71.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2010). Pairwise variable selection for high-dimensional model-based clustering. *Biometrics* **66,** 793–804.

Hall, P., Müller, H.-G., Wang, J.-L., et al. (2006). Properties of principal component methods for functional and longitudinal data analysis. *The annals of statistics* **34,** 1493–1517.

Helmer, C., Damon, D., Letenneur, L., Fabrigoule, C., Barberger-Gateau, P., Lafont, S., Fuhrer, R., Antonucci, T., Commenges, D., Orgogozo, J., et al. (1999). Marital status and risk of alzheimer disease: a french population-based cohort study. *Neurology* **53,** 1953–1953.

Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research* **15,** 1351–1381.

Holzmann, H., Munk, A., and Gneiting, T. (2006). Identifiability of finite mixtures of elliptical distributions. *Scandinavian journal of statistics* **33,** 753–763.

Hornik, K. and Grün, B. (2014). movmf: an r package for fitting mixtures of von mises-fisher distributions. *Journal of Statistical Software* **58,** 1–31.

Huang, C., Shan, L., Charles, H. C., Wirth, W., Niethammer, M., and Zhu, H. (2015). Diseased region detection of longitudinal knee magnetic resonance imaging data. *IEEE transactions on medical imaging* **34,** 1914–1927.

Huang, C., Styner, M., and Zhu, H. (2015). Clustering high-dimensional landmark-based two-dimensional shape data. *Journal of the American Statistical Association* **110,** 946–961.

Huang, C., Thompson, P., Wang, Y., Yu, Y., Zhang, J., Kong, D., Colen, R. R., Knickmeyer, R. C., Zhu, H., Initiative, A. D. N., et al. (2017). Fgwas: Functional genome wide association analysis. *NeuroImage* **159,** 107–121.

Hubert, L. and Arabie, P. (1985a). Comparing partitions. *Journal of classification* **2,** 193–218.

Hubert, L. and Arabie, P. (1985b). Comparing partitions. *Journal of Classification* **2,** 193–218.

Hwang, J., Kim, C. M., Jeon, S., Lee, J. M., Hong, Y. J., Roh, J. H., Lee, J.-H., Koh, J.-Y., Na, D. L., Initiative, A. D. N., et al. (2016). Prediction of alzheimer's disease pathophysiology based on cortical thickness patterns. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* **2,** 58–67.

Inano, R., Oishi, N., Kunieda, T., Arakawa, Y., Yamao, Y., Shibata, S., Kikuchi, T., Fukuyama, H., and Miyamoto, S. (2014). Voxel-based clustered imaging by multiparameter diffusion tensor images for glioma grading. *NeuroImage: Clinical* **5,** 396–407.

Jaccard, P. (1901). *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* **8,** 118–127.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics* **29,** 295–327.

Jørgensen, D. R., Dam, E. B., and Lillholm, M. (2013). Predicting knee cartilage loss using adaptive partitioning of cartilage thickness maps. *Computers in biology and medicine* **43,** 1045–1052.

Karas, G., Scheltens, P., Rombouts, S., Van Schijndel, R., Klein, M., Jones, B., Van Der Flier, W., Vrenken, H., and Barkhof, F. (2007). Precuneus atrophy in early-onset alzheimer's disease: a morphometric structural mri study. *Neuroradiology* **49,** 967–976.

Kauermann, G., Müller, M., and Carroll, R. J. (1998). The efficiency of bias-corrected estimators for nonparametric kernel estimation based on local estimating equations. *Statistics & probability letters* **37,** 41–47.

Ke, Z. T., Fan, J., and Wu, Y. (2015). Homogeneity pursuit. *Journal of the American Statistical Association* **110,** 175–194.

Kellegren, J. and Lawrence, J. (1957). Radiological assessment of osteoarthritis. *Ann Rheum Dis* **16,** 494–501.

Kent, J. T. et al. (1983). Identifiability of finite mixtures for directional data. *The Annals of Statistics* **11,** 984–988.

Khalili, A. and Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association* **102,** 1025–1038.

Koo, S., Hargreaves, B. A., and Gold, G. E. (2014). Automatic segmentation of articular cartilage from mri. US Patent 8,706,188.

Kosorok, M. R. (2003). Bootstraps of sums of independent but not identically distributed stochastic processes. *Journal of Multivariate Analysis* **84,** 299–318.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference.* Springer.

Kovnatsky, A., Glashoff, K., and Bronstein, M. M. (2016). Madmm: a generic algorithm for non-smooth optimization on manifolds. In *European Conference on Computer Vision*, pages 680–696. Springer.

Kume, A. and Welling, M. (2010). Maximum likelihood estimation for the offset-normal shape distributions using em. *Journal of Computational and Graphical Statistics* **19,** 702–723.

Ledermann, W. (1937). On the rank of the reduced correlational matrix in multiple-factor analysis. *Psychometrika* **2,** 85–93.

Lee, S., Sun, W., Wright, F. A., and Zou, F. (2017). An improved and explicit surrogate variable analysis procedure by coefficient adjustment. *Biometrika* **104,** 303–316.

Lee, S., Zou, F., and Wright, F. A. (2014). Convergence of sample eigenvalues, eigenvectors, and principal component scores for ultra-high dimensional data. *Biometrika* **101,** 484–490.

Leek, J. T. (2011). Asymptotic conditional singular value decomposition for high-dimensional genomic data. *Biometrics* **67,** 344–352.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics* **3,** e161.

Leek, J. T. and Storey, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences* pages pnas–0808709105.

Li, S. Z. and Singh, S. (2009). *Markov random field modeling in image analysis.* Springer.

Lin, L., St. Thomas, B., Zhu, H., and Dunson, D. B. (2017). Extrinsic local regression on manifold-valued data. *Journal of the American Statistical Association* **112,** 1261–1273.

Lin, W. and Lv, J. (2013). High-dimensional sparse additive hazards regression. *Journal of the American Statistical Association* **108,** 247–264.

Listgarten, J., Kadie, C., Schadt, E. E., and Heckerman, D. (2010). Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences* **107,** 16465–16470.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis* **42,** 60–88.

Liu, R., Huang, C., Li, T., Yang, L., and Zhu, H. (2018). Statistical disease mapping for heterogeneous neuroimaging studies. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1415–1418. IEEE.

Liu, Y., Julkunen, V., Paajanen, T., Westman, E., Wahlund, L.-O., Aitken, A., Sobow, T., Mecocci, P., Tsolaki, M., Vellas, B., et al. (2012). Education increases reserve against alzheimer's disease-evidence from structural mri analysis. *Neuroradiology* **54,** 929–938.

Lv, J., Fan, Y., et al. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics* **37,** 3498–3528.

Lynch, J. A., Zaim, S., Zhao, J., Stork, A., Peterfy, C. G., and Genant, H. K. (2000). Cartilage segmentation of 3d mri scans of the osteoarthritic knee combining user knowledge and active contours. In *Medical Imaging 2000*, pages 925–935. International Society for Optics and Photonics.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. California, USA.

Madsen, S. K., Ho, A. J., Hua, X., Saharan, P. S., Toga, A. W., Jack Jr, C. R., Weiner, M. W., Thompson, P. M., Initiative, A. D. N., et al. (2010). 3d maps localize caudate nucleus atrophy in 400 alzheimer's disease, mild cognitive impairment, and healthy elderly subjects. *Neurobiology of aging* **31,** 1312–1325.

Mardia, K. V. and Jupp, P. E. (2009). *Directional statistics.* John Wiley & Sons.

Marroquin, J. L., Santana, E. A., and Botello, S. (2003). Hidden markov measure field models for image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **25,** 1380–1387.

Marroquín, J. L., Vemuri, B. C., Botello, S., Calderon, E., and Fernandez-Bouzas, A. (2002). An accurate and efficient bayesian method for automatic segmentation of brain mri. *Medical Imaging, IEEE Transactions on* **21,** 934–945.

Maruszak, A. and Thuret, S. (2014). Why looking at the whole hippocampus is not enough-a critical role for anteroposterior axis, subfield and activation analyses to enhance predictive value of hippocampal changes for alzheimer's disease diagnosis. *Frontiers in cellular neuroscience* **8,** 95.

McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019). Finite mixture models. *Annual review of statistics and its application* **6,** 355–378.

McLachlan, G. J. and Peel, D. (2000). Finite mixture models, volume 299 of probability and statistics–applied probability and statistics section.

McLachlan, G. J., Peel, D., and Bean, R. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis* **41,** 379–388.

McNicholas, P. D. (2016). Model-based clustering. *Journal of Classification* **33,** 331–373.

Mercer, J. (1909). Functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character* **209,** 415–446.

Miranda, M. F., Zhu, H., and Ibrahim, J. G. (2018). Tprm: Tensor partition regression models with applications in imaging biomarker detection. *The annals of applied statistics* **12,** 1422–1450.

Mirzaalian, H., Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., Grant, G., Marx, C., Morey, R. A., Flashman, L., et al. (2016). Inter-site and inter-scanner diffusion mri data harmonization. *NeuroImage* **135,** 311–323.

Moisio, K., Chang, A., Eckstein, F., Chmiel, J. S., Wirth, W., Almagor, O., Prasad, P., Cahue, S., Kothari, A., and Sharma, L. (2011). Varus–valgus alignment: reduced risk of subsequent cartilage loss in the less loaded compartment. *Arthritis & Rheumatism* **63,** 1002–1009.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W., and Beckett, L. (2005). The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics* **15,** 869–877.

Neal, R. M. (2011). Mcmc using hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, pages 133–182. Chapman and Hall/CRC.

Nie, J., Xue, Z., Liu, T., Young, G. S., Setayesh, K., Guo, L., and Wong, S. T. (2009). Automated brain tumor segmentation using spatial accuracy-weighted hidden markov random field. *Computerized Medical Imaging and Graphics* **33,** 431–441.

Nyúl, L. G., Udupa, J. K., and Zhang, X. (2000). New variants of a method of mri scale standardization. *IEEE transactions on medical imaging* **19,** 143–150.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics* **92,** 1004–1016.

Owen, A. B., Wang, J., et al. (2016). Bi-cross-validation for factor analysis. *Statistical Science* **31,** 119–139.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research* **8,** 1145–1164.

Pennec, X. (2006). Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision* **25,** 127–154.

Peterfy, C., Schneider, E., and Nevitt, M. (2008). The osteoarthritis initiative: report on the design rationale for the magnetic resonance imaging protocol for the knee. *Osteoarthritis and cartilage* **16,** 1433–1441.

Piplani, M. A., Disler, D. G., McCauley, T. R., Holmes, T. J., and Cousins, J. P. (1996). Articular cartilage volume in the knee: semiautomated determination from three-dimensional reformations of mr images. *Radiology* **198,** 855–859.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38,** 904.

Qian, W. and Titterington, D. (1991). Estimation of parameters in hidden markov models. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences* **337,** 407–428.

Rabl, U., Meyer, B. M., Diers, K., Bartova, L., Berger, A., Mandorfer, D., Popovic, A., Scharinger, C., Huemer, J., Kalcher, K., et al. (2014). Additive gene–environment effects on hippocampal structure in healthy humans. *Journal of Neuroscience* **34,** 9917–9926.

Rajapakse, J. C., Giedd, J. N., and Rapoport, J. L. (1997). Statistical approach to segmentation of single-channel cerebral mr images. *Medical Imaging, IEEE Transactions on* **16,** 176–186.

Rand, W. M. (1971a). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association* **66,** 846–850.

Rand, W. M. (1971b). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* **66,** 846–850.

Raynauld, J.-P. (2003). Quantitative magnetic resonance imaging of articular cartilage in knee osteoarthritis. *Current opinion in rheumatology* **15,** 647–650.

Rohlfing, T., Sullivan, E. V., and Pfefferbaum, A. (2007). Divergence-based framework for diffusion tensor clustering, interpolation, and regularization. In *Information Processing in Medical Imaging*, pages 507–518. Springer.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *The annals of statistics* **22,** 1346–1370.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics* **6,** 461–464.

Seo, S.-S., Kim, C.-W., and Jung, D.-W. (2011). Management of focal chondral lesion in the knee joint. *Knee surgery & related research* **23,** 185–196.

Shan, L., Charles, C., and Niethammer, M. (2012). Automatic multi-atlas-based cartilage segmentation from knee mr images. In *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*, pages 1028–1031. IEEE.

Shan, L., Zach, C., Charles, C., and Niethammer, M. (2014). Automatic atlas-based three-label cartilage segmentation from MR knee images. *Medical Image Analysis* **18,** 1233 – 1246.

Shen, X. and Huang, H.-C. (2010). Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association* **105,** 727–739.

Shi, F., Liu, B., Zhou, Y., Yu, C., and Jiang, T. (2009). Hippocampal volume and asymmetry in mild cognitive impairment and alzheimer's disease: Meta-analyses of mri studies. *Hippocampus* **19,** 1055–1064.

Shi, J., Thompson, P. M., Gutman, B., and Wang, Y. (2013). Surface fluid registration of conformal representation: Application to detect disease burden and genetic influence on hippocampus. *NeuroImage* **78,** 111–134.

Shi, Y., Wu, Y., Xu, D., and Jiao, Y. (2018). An admm with continuation algorithm for non-convex sica-penalized regression in high dimensions. *Journal of Statistical Computation and Simulation* **88,** 1826–1846.

Shi, Y., Zhou, Z., Jiao, Y., and Wang, J. (2019). A primal dual active set with continuation algorithm for high-dimensional nonconvex sica-penalized regression. *Journal of Statistical Computation and Simulation* **89,** 864–883.

Shinohara, R. T., Sweeney, E. M., Goldsmith, J., Shiee, N., Mateen, F. J., Calabresi, P. A., Jarso, S., Pham, D. L., Reich, D. S., Crainiceanu, C. M., et al. (2014). Statistical normalization techniques for magnetic resonance imaging. *NeuroImage: Clinical* **6,** 9–19.

Sommerlad, A., Ruegger, J., Singh-Manoux, A., Lewis, G., and Livingston, G. (2018). Marriage and risk of dementia: systematic review and meta-analysis of observational studies. *J Neurol Neurosurg Psychiatry* **89,** 231–238.

Soufi, M., Kamali-Asl, A., Geramifar, P., and Rahmim, A. (2017). A novel framework for automated segmentation and labeling of homogeneous versus heterogeneous lung tumors in [18 f] fdg-pet imaging. *Molecular Imaging and Biology* **19,** 456–468.

Srivastava, A., Joshi, S. H., Mio, W., and Liu, X. (2005). Statistical shape analysis: Clustering, learning, and testing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **27,** 590–602.

Städler, N., Bühlmann, P., and Van De Geer, S. (2010). L1-penalization for mixture regression models. *Test* **19,** 209–256.

Stahl, R., Jain, S. K., Lutz, J., Wyman, B. T., Le Graverand-Gastineau, M.-P. H., Vignon, E., Majumdar, S., and Link, T. M. (2011). Osteoarthritis of the knee at 3.0 t: comparison of a quantitative and a semi-quantitative score for the assessment of the extent of cartilage lesion and bone marrow edema pattern in a 24-month longitudinal study. *Skeletal radiology* **40,** 1315–1327.

Subbarao, R. and Meer, P. (2009). Nonlinear mean shift over riemannian manifolds. *International Journal of Computer Vision* **84,** 1–20.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine* **12,** e1001779.

Sun, Y., Zhang, N. R., Owen, A. B., et al. (2012). Multiple hypothesis testing adjusted for latent variables, with an application to the agemap gene expression data. *The Annals of Applied Statistics* **6,** 1664–1688.

Sundström, A., Westerlund, O., and Kotyrlo, E. (2016). Marital status and risk of dementia: a nationwide population-based prospective study from sweden. *BMJ open* **6,** e008565.

Sutton, C., McCallum, A., and Rohanimanesh, K. (2007). Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *Journal of Machine Learning Research* **8,** 693–723.

Tamez-Pena, J. G., Farber, J., Gonzalez, P. C., Schreyer, E., Schneider, E., and Totterman, S. (2012). Unsupervised segmentation and quantification of anatomical knee features: data from the osteoarthritis initiative. *Biomedical Engineering, IEEE Transactions on* **59,** 1177–1186.

Tang, L. and Song, P. X. (2016). Fused lasso approach in regression coefficients clustering: learning parameter heterogeneity in data integration. *The Journal of Machine Learning Research* **17,** 3915–3937.

Teicher, H. et al. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics* **34,** 1265–1269.

Tipping, M. and Bishop, C. (1999). Mixtures of probabilistic principal component analyzers. *Neural computation* **11,** 443–482.

Townsend, J., Koep, N., and Weichwald, S. (2016). Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. *Journal of Machine Learning Research* **17,** 1–5.

Vachet, C., Yvernault, B., Bhatt, K., Smith, R. G., Gerig, G., Hazlett, H. C., and Styner, M. (2012). Automatic corpus callosum segmentation using a deformable active fourier contour model. In *SPIE Medical Imaging*, volume 8317, pages 831707–831707–7. International Society for Optics and Photonics.

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.

Varol, E., Sotiras, A., Davatzikos, C., Initiative, A. D. N., et al. (2017). Hydra: revealing heterogeneity of imaging and genetic patterns through a multiple max-margin discriminative analysis framework. *Neuroimage* **145,** 346–364.

Vidal, R., Ma, Y., and Sastry, S. (2005). Generalized principal component analysis (gpca). *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **27,** 1945–1959.

Vincent, G., Wolstenholme, C., Scott, I., and Bowes, M. (2010). Fully automatic segmentation of the knee joint using active appearance models. *Medical Image Analysis for the Clinic: A Grand Challenge* pages 224–230.

Wang, D., Ding, C., and Li, T. (2009). K-subspace clustering. In *Machine learning and knowledge discovery in databases*, pages 506–521. Springer.

Wang, J., Zhao, Q., Hastie, T., Owen, A. B., et al. (2017). Confounder adjustment in multiple hypothesis testing. *The Annals of Statistics* **45,** 1863–1894.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application* **3,** 257–295.

Wang, S. and Zhu, J. (2008). Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics* **64,** 440–448.

Wang, Y. and Ji, Q. (2005). A dynamic conditional random field model for object segmentation in image sequences. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 264–270. IEEE.

Wang, Y., Loe, K.-F., and Wu, J.-K. (2006). A dynamic conditional random field model for foreground and shadow segmentation. *IEEE transactions on pattern analysis and machine intelligence* **28,** 279–289.

Wang, Y., Song, Y., Rajagopalan, P., T., A., Liu, K., Chou, Y. Y., Gutman, B., Toga, A. W., and Thompson, P. M. (2011). Surface-based tbm boosts power to detect disease effects on the brain: An n = 804 adni study. *NeuroImage* **56,** 1993–2010.

Wen, Z., Yang, C., Liu, X., and Marchesini, S. (2012). Alternating direction methods for classical and ptychographic phase retrieval. *Inverse Problems* **28,** 115010.

Whitwell, J. L., Dickson, D. W., Murray, M. E., Weigand, S. D., Tosakulwong, N., Senjem, M. L., Knopman, D. S., Boeve, B. F., Parisi, J. E., Petersen, R. C., et al. (2012). Neuroimaging correlates of pathologically defined subtypes of alzheimer's disease: a case-control study. *The Lancet Neurology* **11,** 868–877.

Wirth, W. and Eckstein, F. (2008). A technique for regional analysis of femorotibial cartilage thickness based on quantitative magnetic resonance imaging. *Medical Imaging, IEEE Transactions on* **27,** 737–744.

Witten, D. M. and Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association* **105,**.

Woolf, A. D. and Pfleger, B. (2003). Burden of major musculoskeletal conditions. *Bulletin of the World Health Organization* **81,** 646–656.

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *the Annals of Statistics* pages 1261–1295.

Xie, B., Pan, W., and Shen, X. (2008). Variable selection in penalized model-based clustering via regularization on grouped parameters. *Biometrics* **64,** 921–930.

Xie, B., Pan, W., and Shen, X. (2010). Penalized mixtures of factor analyzers with application to clustering high-dimensional microarray data. *Bioinformatics* **26,** 501–508.

Yakowitz, S. J., Spragins, J. D., et al. (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics* **39,** 209–214.

Yao, F. and Lee, T. (2006). Penalized spline models for functional principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68,** 3–25.

Yao, W. (2012). A bias corrected nonparametric regression estimator. *Statistics & Probability Letters* **82,** 274–282.

Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference* **82,** 171–196.

Yezzi, A. J. and Prince, J. L. (2003). An eulerian pde approach for computing tissue thickness. *Medical Imaging, IEEE Transactions on* **22,** 1332–1339.

Yin, J., Hu, D. H., and Yang, Q. (2009). Spatio-temporal event detection using dynamic conditional random fields. In *Proceedings of the 21st international jont conference on Artifical intelligence*, pages 1321–1326. Morgan Kaufmann Publishers Inc.

Yin, Y., Zhang, X., Williams, R., Wu, X., Anderson, D. D., and Sonka, M. (2010). LogismosÂ¡Âªlayered optimal graph image segmentation of multiple objects and surfaces: cartilage segmentation in the knee joint. *Medical Imaging, IEEE Transactions on* **29,** 2023–2037.

Yuan, Y., Failmezger, H., Rueda, O. M., Ali, H. R., Gräf, S., Chin, S.-F., Schwarz, R. F., Curtis, C., Dunning, M. J., Bardwell, H., et al. (2012). Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Science translational medicine* **4,** 157ra143–157ra143.

Zhang, J. and Chen, J. (2007). Statistical inference for functional data. *The Annals of Statistics* **35,** 1052–1079.

Zhang, J.-T. (2011). Statistical inferences for linear models with functional responses. *Statistica Sinica* **21,** 1431–1451.

Zhang, M. and Fletcher, P. T. (2013). Probabilistic principal geodesic analysis. In *Advances in Neural Information Processing Systems*, pages 1178–1186.

Zhang, T., Koutsouleris, N., Meisenzahl, E., and Davatzikos, C. (2014). Heterogeneity of structural brain changes in subtypes of schizophrenia revealed using magnetic resonance imaging pattern analysis. *Schizophrenia bulletin* **41,** 74–84.

Zhang, X., Mormino, E. C., Sun, N., Sperling, R. A., Sabuncu, M. R., Yeo, B. T., Weiner, M. W., Aisen, P., Weiner, M., Petersen, R., et al. (2016). Bayesian model reveals latent atrophy factors with dissociable cognitive trajectories in alzheimer's disease. *Proceedings of the National Academy of Sciences* **113,** E6535–E6544.

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *Medical Imaging, IEEE Transactions on* **20,** 45–57.

Zhou, H., Pan, W., and Shen, X. (2009). Penalized model-based clustering with unconstrained covariance matrices. *Electronic Journal of Statistics* **3,** 1473.

Zhu, H., Gu, M., and Peterson, B. (2007). Maximum likelihood from spatial random effects models via the stochastic approximation expectation maximization algorithm. *Statistics and computing* **17,** 163–177.

Zhu, H., Ibrahim, J. G., Tang, N., Rowe, D. B., Hao, X., Bansal, R., and Peterson, B. S. (2007). A statistical analysis of brain morphology using wild bootstrapping. *IEEE Transactions on Medical Imaging* **26,** 954–966.

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

Zhu, H., Li, R., and Kong, L. (2012). Multivariate varying coefficient model for functional responses. *The Annals of Statistics* **40,** 2634–2666.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association* **101,** 1418–1429.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics* **36,** 1509.