Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

*J Am Stat Assoc*. Author manuscript; available in PMC 2016 May 07.

Published in final edited form as:

*J Am Stat Assoc*. 2015 November 7; 110(115): 946–961. doi:10.1080/01621459.2015.1034802.

# Clustering High-Dimensional Landmark-based Two-dimensional Shape Data‡

**Chao Huang [Ph.d student under the supervision of Dr Zhu]**, **Martin Styner [Associate, Professor of Psychiatry and Computer Science]**, and **Hongtu Zhu [Professor of Biostatistics]**

Department of Biostatistics and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, NC 27599-7420, USA

Chao Huang: huangchao.seu@hotmail.com; Martin Styner: styner@cs.unc.edu; Hongtu Zhu: htzhu@email.unc.edu

## Abstract

An important goal in image analysis is to cluster and recognize objects of interest according to the shapes of their boundaries. Clustering such objects faces at least four major challenges including a curved shape space, a high-dimensional feature space, a complex spatial correlation structure, and shape variation associated with some covariates (e.g., age or gender). The aim of this paper is to develop a penalized model-based clustering framework to cluster landmark-based planar shape data, while explicitly addressing these challenges. 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. Our real data analysis has confirmed the excellent finite-sample performance of MOSFA in revealing meaningful clusters in the corpus callosum shape data obtained from the Attention Deficit Hyperactivity Disorder-200 (ADHD-200) study.

### Keywords

Alternating direction method of multipliers; Attention deficit hyperactivity disorder; Corpus callosum; Offset-normal shape distribution; Shape clustering

## 1 Introduction

Shape analysis has been an important research topic with various applications in computer vision, object recognition, and medical imaging for last several decades (Bookstein, 1991; Cootes et al., 1995; Dryden and Mardia, 1998; Younes, 2010; Srivastava et al., 2005). An important goal in shape analysis is to classify and recognize objects of interest according to the shapes of their boundaries. The majority of earlier work on shape analysis has focused

‡Address for correspondence and reprints: Hongtu Zhu, Ph.D., htzhu@email.unc.edu; Phone No: 919-966-7272.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

on landmark-based analysis, where shapes are represented by a coarse, discrete sampling of the object contours (Bookstein, 1991; Cootes et al., 1995; Dryden and Mardia, 1998; Rajpoot and Arif, 2008). As an illustration, Figure 1 shows the automatic corpus callosum (CC) segmentations of four randomly selected subjects by using the *CCSeg* package[1] (Vachet et al., 2012). The CC contour (red) is represented by 100 landmarks, spacing along the contour about 0.75 mm. These landmarks determine the important features of geometrical configuration represented by a matrix of coordinates (Small, 1996; Dryden and Mardia, 1998). Our motivating example is to use the CC shape data to unsupervisely cluster all 647 subjects from the ADHD-200 study into biologically meaningful subpopulations. Scientifically, we are interested in whether the CC shape information is a promising biomarker for the diagnosis of attention deficit hyperactivity disorder (ADHD) and may provide a clue to the topographical spread of ADHD disease.

Clustering landmark-based planar shape data raises four major challenges. First, planar shape data reside in a curved shape space, which is invariant under a similarity transformation including rigid rotation and translation, and non-rigid uniform scaling (Bookstein, 1991; Cootes et al., 1995; Dryden and Mardia, 1998; Younes, 2010). Therefore, most clustering methods (e.g., K-means) proposed for Euclidean data cannot be used to cluster data in the curved shape space (Srivastava et al., 2005; Amaral et al., 2010). Second, it is a standard high-dimensional-low-sample-size problem, since shape dimension, which is proportional to the number of landmark points, can be much larger than the sample size. Moreover, there may be significant amounts of noise in many of the landmark points, which is either associated with the complexity of the studied shapes or is caused by certain preprocessing steps such as image filtering and edge detection. Third, the landmark points along the boundaries of objects are inherently and spatially correlated with each other. Fourth, shape variation is commonly associated with some explanatory attributes (e.g., age, gender or disease status). As shown in simulations and real data analysis, ignoring such complex spatial correlation and explanatory attributes can introduce substantial errors in both clustering and classification results.

Little has been done on the development of methods for clustering high-dimensional landmark-based planar shape data. Most existing methods for shape data primarily extend standard clustering algorithms, such as K-means or mean-shift algorithm, by replacing the Euclidean metric by the metric of the curved shape space (Srivastava et al., 2005; Subbarao and Meer, 2009; Amaral et al., 2010). Furthermore, Kume and Welling (2010) developed a mixture model of offset-normal distributions, which explicitly models the spatial covariance matrix of all landmarks in each cluster. All these methods, however, do not address the noisy data in the high-dimensional feature space, the high-dimensional spatial correlation matrix, and the shape variation associated with explanatory attributes. When there are a large number of variables, they can mask underlying clustering structures and the spatial correlation matrix is not invertible. For instance, for the CC contours with 100 landmarks in Figure 1, there are about 4950 (100×99/2) unknown parameters in a single spatial covariance matrix.

[1] http://www.nitrc.org/projects/ccseg/

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The aim of this paper is to propose a mixture of offset-normal shape factor analyzers (MOSFA) model to address the four challenges. We use the offset-normal shape distribution (Dryden and Mardia, 1991; Kume and Welling, 2010) to characterize the variability of shape data in the curved shape space. To handle high dimensionality, we use a penalized clustering framework as an effective and powerful method to perform both variable selection and clustering (Pan and Shen, 2007; Guo et al., 2010). We integrate a latent factor analysis model to approximate the complex spatial correlation of shape data (McLachlan and Peel, 2004; Xie et al., 2010). We use a logistic regression model to build an association between mixing proportions and covariates of interest. We propose an expectation-maximization (EM) algorithm and establish its convergence property. We establish the asymptotic properties of penalized estimator obtained from the EM algorithm. Finally, we will develop companion software for MOSFA and release it to the public through http://www.nitrc.org/ and http://www.bias.unc.edu/.

In Section 2, we review the offset-normal shape distribution and introduce the MOSFA model. Moreover, we derive an EM algorithm to maximize the penalized likelihood function of the MOSFA model. The convergence properties of the proposed EM algorithm and the asymptotic properties of penalized estimator are also investigated. In Section 3, we use some simulations to examine the finite sample performance of our MOSFA model. In Section 4, we also apply the MOSFA model to the ADHD-200 CC shape data set. Our clustering results remarkably reveal an intrinsic subpopulation structure in the mixed population with controls and subjects with ADHD.

## 2 Methodology

### 2.1 Offset-normal shape distribution

We first review the Bookstein's shape variables of planar data. For a specific planar configuration $\boldsymbol{X}^{\dagger}$ with $k$ not-all-coincident landmarks, its coordinates can be written as a $k \times 2$ matrix as follows:

$$\boldsymbol{X}^{\dagger}=\begin{pmatrix} x_1^{\dagger} & x_2^{\dagger} & \cdots & x_k^{\dagger} \\ y_1^{\dagger} & y_2^{\dagger} & \cdots & y_k^{\dagger} \end{pmatrix}^T.$$

Let $\mathbf{1}_{k-1}$ and $\boldsymbol{I}_{k-1}$ be, respectively, a $(k-1)\times 1$ vector with all components being one and a $(k-1)\times(k-1)$ identity matrix. By left multiplying $\boldsymbol{X}^{\dagger}$ with a matrix $\boldsymbol{L}=(-\mathbf{1}_{k-1}, \boldsymbol{I}_{k-1})$, $\boldsymbol{X}^{\dagger}$ is translated such that the first landmark of $\boldsymbol{X}^{\dagger}$ is mapped to the origin $(0, 0)^T$. We call $\boldsymbol{X}=\boldsymbol{L}\boldsymbol{X}^{\dagger}$ as the preform of configuration $\boldsymbol{X}^{\dagger}$ and write it as

$$\boldsymbol{X}=\begin{pmatrix} x_2 & x_3 & \cdots & x_k \\ y_2 & y_3 & \cdots & y_k \end{pmatrix}^T.$$

Then, if $x_2^2+y_2^2>0$, then the rotation and scale information can be removed via

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\boldsymbol{X} \to \boldsymbol{X} \begin{pmatrix} x_2 & -y_2 \\ y_2 & x_2 \end{pmatrix} \frac{1}{x_2^2+y_2^2} = \begin{pmatrix} 1 & u_3 & \cdots & u_k \\ 0 & v_3 & \cdots & v_k \end{pmatrix}^T. \quad (1)$$

Since the landmarks in $\boldsymbol{X}^\dagger$ are not-all-coincident, we can choose another pair of landmarks instead if the first two landmarks are coincident. Since the first two landmarks are sent to $(0, 0)^T$ and $(1, 0)^T$, respectively, the shape coordinates in $\boldsymbol{u} = (u_3, \ldots, u_k, v_3, \ldots, v_k)^T$ are called the Bookstein's shape variables (Dryden and Mardia, 1998).

We introduce the offset-normal shape distribution of $\boldsymbol{u}$ as follows. It is assumed that the model for the landmarks in $\boldsymbol{X}^\dagger$ is $vec(\boldsymbol{X}^\dagger) \sim N_{2k}(vec(\boldsymbol{\mu}^\dagger), \boldsymbol{\Sigma}^\dagger)$, where $vec(\cdot)$ denotes the vectorization of a matrix. Since $\boldsymbol{X} = \boldsymbol{L}\boldsymbol{X}^\dagger$, $vec(\boldsymbol{X})$ is distributed as $N_p(vec(\boldsymbol{\mu}), \boldsymbol{\Sigma})$, where $p = 2k - 2$, $\boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{\mu}^\dagger$, and $\boldsymbol{\Sigma} = (\boldsymbol{I}_2 \otimes \boldsymbol{L})\boldsymbol{\Sigma}^\dagger(\boldsymbol{I}_2 \otimes \boldsymbol{L}^T)$, in which $\otimes$ denotes the matrix Kronecker product. We define

$$\boldsymbol{W} = \begin{pmatrix} 1 & u_3 & \cdots & u_k & 0 & v_3 & \cdots & v_k \\ 0 & -v_3 & \cdots & -v_k & 1 & u_3 & \cdots & u_k \end{pmatrix}^T \text{ and } \boldsymbol{h} = \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}, \quad (2)$$

and then we have $vec(\boldsymbol{X}) = \boldsymbol{W}\boldsymbol{h}$. Following Dryden and Mardia (1991), the distribution of shape variables $\boldsymbol{u}$ can be obtained by integrating out $\boldsymbol{h}$ from the distribution of $vec(\boldsymbol{X})$ and $\boldsymbol{u}$ follows the offset-normal shape probability density function given by

$$f_u(\boldsymbol{u};\boldsymbol{\mu},\sum) = \frac{|\boldsymbol{\Gamma}|^{\frac{1}{2}}\exp(-g/2)}{(2\pi)^{k-2}|\sum|^{\frac{1}{2}}} \sum_{i=0}^{k-2} \begin{pmatrix} k-2 \\ i \end{pmatrix} E(l_x^{2i}|\xi_x, \sigma_x^2) E(l_y^{2k-4-2i}|\xi_y, \sigma_y^2), \quad (3)$$

where $\boldsymbol{\Gamma} = (\boldsymbol{W}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{W})^{-1}$, $g = vec(\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}vec(\boldsymbol{\mu}) - \boldsymbol{\nu}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\nu}$, $\boldsymbol{\nu} = \boldsymbol{\Gamma}\boldsymbol{W}^T\boldsymbol{\Sigma}^{-1}vec(\boldsymbol{\mu})$, 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 $N(\xi, \sigma^2)$ and can be calculated based on the recursion relation (Willink, 2005), which is given by

$$E(l^{r+1}|\xi,\sigma^2) = \xi E(l^r|\xi,\sigma^2) + r\sigma^2 E(l^{r-1}|\xi,\sigma^2),\ r = 1, 2, \ldots. \quad (4)$$

## 2.2 Mixtures of offset-normal shape factor analyzers

We consider finite mixture models of offset-normal shape factor analyzers. It is assumed that $\boldsymbol{x}_1^\dagger, \ldots, \boldsymbol{x}_n^\dagger$ are independently and identically distributed (i.i.d.) random planar configurations. Equivalently, $vec(\boldsymbol{x}_i^\dagger)$, $i = 1, \ldots, n$ are independently generated from a mixture model of $M$ normal distributions given by $\sum_{m=1}^{M} \pi_m \phi(vec(\boldsymbol{x}^\dagger); vec(\boldsymbol{\mu}_m^\dagger), \sum_m^\dagger)$, where $\pi_m \geq 0$, $\sum_{m=1}^{M} \pi_m = 1$, and $\phi(\cdot)$ is the density function of multivariate normal distribution. Based on the results in Section 2.1, the induced probability density function of $vec(\boldsymbol{x})$ is given by

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$f_X(\boldsymbol{x};\boldsymbol{\theta})=\sum_{m=1}^{M}\pi_m\phi(vec(\boldsymbol{x});vec(\boldsymbol{\mu}_m),\sum\nolimits_m), \quad (5)$$

where $\boldsymbol{\mu}_m=\boldsymbol{L}\boldsymbol{\mu}_m^{\dagger}$, $\sum_m=(\boldsymbol{I}_2\otimes\boldsymbol{L})\sum_m^{\dagger}(\boldsymbol{I}_2\otimes\boldsymbol{L}^T)$, and $\boldsymbol{\theta}$ consists of all unknown parameters in $\{\pi_m,\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m\}_{m\le M}$.

We consider a factor analysis model of $\boldsymbol{\Sigma}_m$ in order to characterize the spatial correlation of high-dimensional shape data. Factor analysis is commonly used to uncover the latent structure (dimensions) of shape variables and allows us to extract a feature space from a high-dimensional shape space to a low-dimensional latent factor space. In the context of mixture modelling (McLachlan et al., 2003), it is assumed that $\boldsymbol{x}_i$ is modeled as follows:

$$vec(\boldsymbol{x}_i)=vec(\boldsymbol{\mu}_m)+\Lambda_m\boldsymbol{b}_{mi}+\boldsymbol{e}_{mi}\ \text{with prior probabilities}\ \pi_m\ (m=1,\ldots,M) \quad (6)$$

for $i = 1, \ldots, n$, where $\Lambda_m$ is a $p\times q$ factor loading matrix and $\boldsymbol{b}_{mi} \sim N_q(\boldsymbol{0}, \boldsymbol{I}_q)$ are independent of $\boldsymbol{e}_{mi} \sim N_p(\boldsymbol{0},\boldsymbol{\Omega})$, in which $\boldsymbol{\Omega}$ is a diagonal matrix. In this case, $\sum_m=\Lambda_m\Lambda_m^T+\boldsymbol{\Omega}$ and the number of parameters in $\boldsymbol{\Sigma}_m$ reduces from $p(p-1)/2$ to $p(q+1)$.

Let $w_{mi}^*$ be 1 or 0 according to whether $\boldsymbol{u}_i$ comes from the $m^{th}$ component or not. We consider a regression model of mixing proportions $\pi_{mi}=Pr(w_{mi}^*=1|\boldsymbol{z}_i)$. Since $w_{mi}^*$, $m = 1, \ldots, M$, indicate the group membership of $\boldsymbol{u}_i$, a good candidate is the widely used logistic regression model (Fokoué, 2005). Specifically, given the covariates in $\boldsymbol{z}_i \in \mathbb{R}^d$, the mixing proportions are defined through the logistic model given by

$$\log\left(\frac{\pi_{mi}(\boldsymbol{\beta})}{\pi_{Mi}(\boldsymbol{\beta})}\right)=\boldsymbol{z}_i^T\boldsymbol{\beta}_m\ \text{for}\ m=1,\ldots,M-1\ \text{and}\ i=1,\ldots,n, \quad (7)$$

in which $\boldsymbol{z}_i = (1, z_{i,1}, \ldots, z_{i,d-1})^T$, $\boldsymbol{\beta}_m = (\beta_{m,0}, \beta_{m,1}, \ldots, \beta_{m,d-1})^T$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{M-1})$, and $\boldsymbol{\beta}_M$ is set to $\boldsymbol{0}$ for identifiability. Under models (6) and (7), the shape variables $\boldsymbol{u}_i$ follow the MOSFA model given by

$$g_u(\boldsymbol{u}_i;\boldsymbol{z}_i,\boldsymbol{\theta})=\sum_{m=1}^{M}\pi_{mi}(\boldsymbol{\beta})f_u(\boldsymbol{u}_i;\boldsymbol{\mu}_m,\sum\nolimits_m=\Lambda_m\Lambda_m^T+\boldsymbol{\Omega}), \quad (8)$$

where $\boldsymbol{\theta}$ consists of the unknown elements of $\boldsymbol{\mu}_m$, $\Lambda_m$, $\boldsymbol{\Omega}$, and $\boldsymbol{\beta}$.

### 2.3 EM Algorithm for the MOSFA Model

Following Kume and Welling (2010), we first develop the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) to calculate the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$, denoted by $\tilde{\boldsymbol{\theta}}$, for low-dimensional shape 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 the MOSFA model, we

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

introduce $w^*_{mi}$, $\boldsymbol{b}_{mi}$, and $\boldsymbol{h}_i$ for $i = 1, \ldots, n$ and $m = 1, \ldots, M$ as missing data. Then, the complete-data log-likelihood function is given by

$$\log\tilde{L}(\boldsymbol{\theta})=\sum_{m=1}^{M}\sum_{i=1}^{n}w^*_{mi}\{\log\pi_{mi}+\log\left[\phi(vec(\boldsymbol{x}_i);vec(\boldsymbol{\mu}_m)+\Lambda_m\boldsymbol{b}_{mi},\boldsymbol{\Omega})\phi(\boldsymbol{b}_{mi};\boldsymbol{0},\boldsymbol{I}_q)\right]\}. \quad (9)$$

In the E-step, given $\tilde{\boldsymbol{\theta}}^{(r)}$ at the $r^{th}$ iteration, the Q-function is given by

$$\begin{aligned}Q(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}^{(r)})&=E\{\log\tilde{L}(\boldsymbol{\theta})|\{\boldsymbol{u}_i\}_{i\le n},\tilde{\boldsymbol{\theta}}^{(r)}\}\\ &=\sum_{m=1}^{M}\sum_{i=1}^{n}\tilde{\tau}^{(r)}_{mi}\{\log\pi_{mi}+E[\log\phi(vec(\boldsymbol{x}_i);vec(\boldsymbol{\mu}_m)+\Lambda_m\boldsymbol{b}_{mi},\boldsymbol{\Omega})|\{\boldsymbol{u}_i\}_{i\le n},\tilde{\boldsymbol{\theta}}^{(r)}]+E[\log\phi(\boldsymbol{b}_{mi};\boldsymbol{0},\boldsymbol{I}_q)|\{\boldsymbol{u}_i\}_{i\le n},\tilde{\boldsymbol{\theta}}^{(r)}]\},\end{aligned} \quad (10)$$

where $\tilde{\tau}^{(r)}_{mi}=\tilde{\pi}^{(r)}_{mi}f_u(\boldsymbol{u}_i;\tilde{\boldsymbol{\theta}}^{(r)}_m)/\sum_{m=1}^{M}\tilde{\pi}^{(r)}_{mi}f_u(\boldsymbol{u}_i;\tilde{\boldsymbol{\theta}}^{(r)}_m)$ and the calculation of expectations in (10) involves (i) the calculation of $E_B(\boldsymbol{b}_{mi}|\{\boldsymbol{u}_i\}_{i\le n},\boldsymbol{\theta})$ and $E_B(\boldsymbol{b}_{mi}\boldsymbol{b}^T_{mi}|\{\boldsymbol{u}_i\}_{i\le n},\boldsymbol{\theta})$ and (ii) the calculation of $E_x[vec(\boldsymbol{x}_i)|\{\boldsymbol{u}_i\}_{i\le n},\boldsymbol{\theta}]$ and $E_x[vec(\boldsymbol{x}_i)vec(\boldsymbol{x}_i)^T|\{\boldsymbol{u}_i\}_{i\le n},\boldsymbol{\theta}]$. The explicit expressions of these expectations are given in the supplementary document.

In the E-step, given the current estimate $\tilde{\boldsymbol{\theta}}^{(r)}$, we update $\tilde{\boldsymbol{\theta}}^{(r+1)}$ by maximizing the Q-function with respect to $\boldsymbol{\theta}$. For $\boldsymbol{\beta}$, we define an objective function $Q(\boldsymbol{\beta})$ given by

$$\sum_{m=1}^{M}\sum_{i=1}^{n}\tilde{\tau}^{(r)}_{mi}\log\pi_{mi}=\sum_{i=1}^{n}\left\{[\tilde{\boldsymbol{\tau}}^{(r)}_i]^T\boldsymbol{\Upsilon}_i\boldsymbol{\beta}-\log(1+\sum_{j=1}^{M-1}\exp(\boldsymbol{z}_i^T\boldsymbol{\beta}_m))\right\}, \quad (11)$$

where $\boldsymbol{\beta}=(\boldsymbol{\beta}_1^T,\cdots,\boldsymbol{\beta}_{M-1}^T)^T$, $\tilde{\boldsymbol{\tau}}^{(r)}_i=(\tilde{\tau}^{(r)}_{1i},\cdots,\tilde{\tau}^{(r)}_{(M-1)i})^T$, and $\boldsymbol{\Upsilon}_i=\boldsymbol{z}_i^T\otimes\boldsymbol{I}_{M-1}$. Then, we update $\tilde{\boldsymbol{\beta}}^{(r+1)}$ according to the Newton-Raphson algorithm. 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:

$$\tilde{\boldsymbol{\beta}}(s+1)^{(r+1)}=\tilde{\boldsymbol{\beta}}(s)^{(r+1)}=\left[\sum_{i=1}^{n}\boldsymbol{\Upsilon}_i^T\boldsymbol{C}_i(\tilde{\boldsymbol{\beta}}(s)^{(r+1)})\boldsymbol{\Upsilon}_i\right]^{-1}\sum_{i=1}^{n}\boldsymbol{\Upsilon}_i^T\{\tilde{\boldsymbol{\tau}}^{(r)}_i-\boldsymbol{\pi}_i(\tilde{\boldsymbol{\beta}}(s)^{(r+1)})\}, \quad (12)$$

where $\boldsymbol{\pi}_i(\boldsymbol{\beta})=(\pi_{1i}(\boldsymbol{\beta}),\ldots,\pi_{(M-1)i}(\boldsymbol{\beta}))^T$ and $\boldsymbol{C}_i(\boldsymbol{\beta})=\mathrm{diag}(\boldsymbol{\pi}_i(\boldsymbol{\beta}))-\boldsymbol{\pi}_i(\boldsymbol{\beta})\boldsymbol{\pi}_i(\boldsymbol{\beta})^T$. We update $\tilde{\boldsymbol{\beta}}(s+1)^{(r+1)}$ according to (12) until a pre-specified tolerance is reached and then set $\tilde{\boldsymbol{\beta}}(s+1)^{(r+1)}$ from the last iteration as $\tilde{\boldsymbol{\beta}}^{(r+1)}$.

We have much simpler formula to update $\boldsymbol{\mu}_m$, $\Lambda_m$, and $\boldsymbol{z}_m$ as follows. For $\boldsymbol{\mu}_m$, we have

$$vec(\boldsymbol{\mu}_m^{(r+1)})=\frac{1}{\sum_{i=1}^{n}\tilde{\tau}^{(r)}_{mi}}\sum_{i=1}^{n}\tilde{\tau}^{(r)}_{mi}E_x\left[vec(\boldsymbol{x}_i)-\tilde{\Lambda}^{(r)}_m E_B(\boldsymbol{b}_{mi}|\boldsymbol{x}_i,\tilde{\boldsymbol{\theta}}^{(r)}_m)|\{\boldsymbol{u}_i\}_{i\le n},\tilde{\boldsymbol{\theta}}^{(r)}_m\right]. \quad (13)$$

For $\Gamma_m$, we have

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\tilde{\Lambda}_m^{(r+1)} = \sum_{i=1}^{n} \tilde{\tau}_{mi}^{(r)} E_x\{vec(\boldsymbol{x}_i - \tilde{\boldsymbol{\mu}}_m^{(r)}) E_B(\boldsymbol{b}_{mi}^T | \boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_m^{(r)}) | \{\boldsymbol{u}_i\}_{i\le n}, \tilde{\boldsymbol{\theta}}_m^{(r)}\} \left( \sum_{i=1}^{n} \tilde{\tau}_{mi}^{(r)} E_x [E_B(\boldsymbol{b}_{mi}\boldsymbol{b}_{mi}^T | \boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_m^{(r)}) | \{\boldsymbol{u}_i\}_{i\le n}, \tilde{\boldsymbol{\theta}}_m^{(r)}] \right)^{-1}. \quad (14)$$

For $\boldsymbol{z}_m$, we have

$$\begin{aligned} \tilde{\boldsymbol{\Omega}}^{(r+1)} = \frac{1}{n} \mathrm{diag}\Big( & \sum_{m=1}^{M} \sum_{i=1}^{n} \tilde{\tau}_{mi}^{(r)} E_x[vec(\boldsymbol{x}_i - \tilde{\boldsymbol{\mu}}_m^{(r)}) vec(\boldsymbol{x}_i - \tilde{\boldsymbol{\mu}}_m^{(r)})^T | \{\boldsymbol{u}_i\}_{i\le n}, \tilde{\boldsymbol{\theta}}_m^{(r)}] \\ & - \sum_{m=1}^{M} \sum_{i=1}^{n} \tilde{\tau}_{mi}^{(r)} E_x [vec(\boldsymbol{x}_i - \tilde{\boldsymbol{\mu}}_m^{(r)}) E_B(\boldsymbol{b}_{mi}^T | \boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_m^{(r)}) | \{\boldsymbol{u}_i\}_{i\le n}, \tilde{\boldsymbol{\theta}}_m^{(r)}] \tilde{\Lambda}_m^{(r)^T} \\ & - \sum_{m=1}^{M} \sum_{i=1}^{n} \tilde{\tau}_{mi}^{(r)} \tilde{\Lambda}_m^{(r)} E_x [E_B(\boldsymbol{b}_{mi} | \boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_m^{(r)}) vec(\boldsymbol{x}_i - \tilde{\boldsymbol{\mu}}_m^{(r)})^T | \{\boldsymbol{u}_i\}_{i\le n}, \tilde{\boldsymbol{\theta}}_m^{(r)}] \\ & + \sum_{m=1}^{M} \sum_{i=1}^{n} \tilde{\tau}_{mi}^{(r)} \tilde{\Lambda}_m^{(r)} E_x [E_B(\boldsymbol{b}_{mi}\boldsymbol{b}_{mi}^T | \boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_m^{(r)}) | \{\boldsymbol{u}_i\}_{i\le n}, \tilde{\boldsymbol{\theta}}_m^{(r)}] \tilde{\Lambda}_m^{(r)^T} \Big). \end{aligned} \quad (15)$$

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 $10^{-4}$.

### 2.4 EM Algorithm for Penalized MOSFA Clustering

It can be very challenging to directly use $\tilde{\boldsymbol{\theta}}$ to cluster high-dimensional shape data in the presence of a large number of noisy landmarks, since these 'non-informative' variables can impede uncovering the underlying clustering structure of interest. Thus, it is critically important to remove such 'non-informative' variables to enhance interpretability. For high-dimensional Euclidean data, several authors (Pan and Shen, 2007; Zhou et al., 2009; Xie et al., 2010) have shown that it is necessary to perform variable selection to reduce such noisy variables during the clustering procedure. To achieve variable selection in MOSFA, we develop a penalized MOSFA clustering framework below.

To realize variable selection in MOSFA, we consider a penalized log-likelihood function, denoted as $\log L_p(\boldsymbol{\theta})$, which is given by

$$\sum_{i=1}^{n} \log\{\sum_{m=1}^{M} \pi_{mi} f_u(\boldsymbol{u}_i, \boldsymbol{z}_i, \boldsymbol{\theta}_m)\} - \lambda_1 \sum_{j=1}^{p} \sum_{1\le m,m'\le M} \kappa_{m,m'}^{(j)} |\mu_{mj} - \mu_{m'j}| - \lambda_2 \sum_{m=1}^{M} \sum_{j=1}^{p} \|\Lambda_{mj}\|_2, \quad (16)$$

where $\mu_{mj}$ is the $j^{th}$ element of $vec(\boldsymbol{\mu}m)$, $\kappa_{m,m'}^{(j)}$ are pre-specified weights, $\Gamma_{mj}$ is the $j^{th}$ row of the factor loading $\Gamma_m$, and $\|\cdot\|_2$ denotes the $L_2$ norm. In the second term of (16), the adaptive pairwise fusion Lasso penalization introduced on $\boldsymbol{\mu}m$ is to shrink the difference between every pair of cluster centers for each component $j$ (Guo et al., 2010). If $\hat{\mu}_{mj} = \hat{\mu}_{m'j}$, then the corresponding variable is considered to be non-informative for separating cluster $m$ from cluster $m'$. Furthermore, if all cluster means for one variable are shrunken to the same value, it indicates that such variable is non-informative for all clusters, and thus it can be removed from MOSFA. In the third term of (16), the $L_2$ penalty introduced on $\Gamma_m$ is to shrink small $\Gamma_{mj}$ to be exactly zero (Xie et al., 2010).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

We set the weights $\kappa_{m,m'}^{(j)}=\breve{\sigma}_j^{-1}|\tilde{\mu}_{mj}-\tilde{\mu}_{m'j}|^{-1}$, where $\breve{\sigma}_j^2$ is the estimate of the $j^{th}$ diagonal element of $\boldsymbol{\Omega}$ as $M = 1$ and $\Gamma_m = \mathbf{0}$, and $\mu_{mj}$ is the estimates of $\mu_{mj}$ in MOSFA without any penalization. By adding the weights $\kappa_{m,m'}^{(j)}$, we slightly penalize the difference between $\mu_{mj}$ and $\mu_{m'j}$ when the $j^{th}$ variable is informative for separating clusters $m$ and $m'$. Otherwise, we heavily penalize the difference between $\mu_{mj}$ and $\mu_{m'j}$ if the weight $\kappa_{m,m'}^{(j)}$ is large.

We also develop an EM algorithm to calculate the maximum penalized likelihood estimate (MPLE), which is denoted as $\hat{\boldsymbol{\theta}}$. For simplicity, we only highlight several key differences between the EM algorithm for MLE and that for MPLE. Similar to (10), the penalized Q-function, denoted by $Q_p(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(r)})$, is given by

$$Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(r)})-\lambda_1\sum_{j=1}^{p}\sum_{1\le m,m'\le M}\kappa_{m,m'}^{(j)}|\mu_{mj}-\mu_{m'j}|-\lambda_2\sum_{m=1}^{M}\sum_{j=1}^{p}\|\Lambda_{mj}\|_2. \quad (17)$$

Since the penalty functions do not depend on $\pi_{mi}$ and $\boldsymbol{\Omega}$, the updating formulas of $\tau_{mi}$, $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ are the same as those given in Section 2.3.

We update $\boldsymbol{\mu}_m$ and $\boldsymbol{\Gamma}_m$ by decomposing (17) into $p$ different functions as follows. Let $\boldsymbol{\mu}_{(j)} = (\mu_{1j}, \dots, \mu_{Mj})^T$, $\boldsymbol{K} = \mathrm{diag}(K_1, \cdots, K_M)$, $\boldsymbol{H} = \mathrm{diag}(H_1, \cdots, H_M)$, and $\boldsymbol{\Gamma}_{(j)} = (\Gamma_{1j}, \dots, \Gamma_{Mj})$. We solve a subproblem given by

$$\min_{\boldsymbol{\mu}_{(j)},\boldsymbol{\Lambda}_{(j)}} f_0(\boldsymbol{\mu}_{(j)},\boldsymbol{\Lambda}_{(j)})+g_0(\boldsymbol{A}_0^{(j)}\boldsymbol{\mu}_{(j)})+\sum_{m=1}^{M}g_m(\boldsymbol{A}_m^{(j)}\boldsymbol{\Lambda}_{(j)}^T), \quad (18)$$

where

$f_0(\boldsymbol{\mu}_{(j)},\boldsymbol{\Lambda}_{(j)})=\hat{\Omega}_j^{(r+1)^{-1}}\left(\frac{1}{2}\boldsymbol{\Lambda}_{(j)}\boldsymbol{K}\boldsymbol{\Lambda}_{(j)}^T+\frac{1}{2}\boldsymbol{\mu}_{(j)}^T\boldsymbol{\tau}\boldsymbol{\mu}_{(j)}-\boldsymbol{\alpha}_1^{(j)T}\boldsymbol{\Lambda}_{(j)}^T-\boldsymbol{\mu}_{(j)}^T\boldsymbol{\alpha}_2^{(j)}+\boldsymbol{\mu}_{(j)}^T\boldsymbol{H}\boldsymbol{\Lambda}_{(j)}^T\right)$, $g_0(\cdot) = \lambda_1\|\cdot\|_1$, and $g_m(\cdot) = \lambda_2\|\cdot\|_2$. The explicit expressions of all the matrices in (18) are given in the supplementary document.

To compute MPLE, we propose an efficient iterative algorithm based on the alternating direction method of multipliers (ADMM) (Glowinski and Marroco, 1975; Gabay and Mercier, 1976). ADMM is an algorithm that is intended to blend the decomposability of dual ascent with the superior convergence properties of the method of multipliers. Since all of the $M$+2 functions in (18) are convex, ADMM is ideal for solving subproblem (18). In the ADMM form, subproblem (18) can be written as

$$\begin{aligned}&\text{minimize} && f_0(\boldsymbol{\mu}_{(j)},\boldsymbol{\Lambda}_{(j)})+\sum_{m=0}^{M}g_m(\boldsymbol{v}_m)\\&\text{subject to} && \boldsymbol{A}_0^{(j)}\boldsymbol{\mu}_{(j)}-\boldsymbol{v}_0=\mathbf{0}\ \text{and}\ \boldsymbol{A}_m^{(j)}\boldsymbol{\Lambda}_{(j)}^T-\boldsymbol{v}_m^T-\mathbf{0}\ \text{for}\ m=1,\dots,M,\end{aligned} \quad (19)$$

where $\{\boldsymbol{v}_0, \cdots, \boldsymbol{v}_M\}$ is a set of augmented variables. The augmented Lagrangian can be written as

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$L_\rho(\breve{\boldsymbol{\theta}}_{(j)}, \boldsymbol{v}, \boldsymbol{y}) = f_0(\breve{\boldsymbol{\theta}}_{(j)}) + g(\boldsymbol{v}) + \boldsymbol{y}^T(\boldsymbol{A}^{(j)}\breve{\boldsymbol{\theta}}_{(j)} - \boldsymbol{v}) + \frac{\rho}{2}\|\boldsymbol{A}^{(j)}\breve{\boldsymbol{\theta}}_{(j)} - \boldsymbol{v}\|_2^2, \quad (20)$$

where $\breve{\boldsymbol{\theta}}_{(j)} = (\boldsymbol{\mu}_{(j)}^T, \boldsymbol{\Lambda}_{(j)})^T$, $\boldsymbol{v} = (\boldsymbol{v}_0^T, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_M)^T$, $g(\boldsymbol{v}) = \sum_{m=0}^{M} g_m(\boldsymbol{v}_m)$, $\boldsymbol{y}$ is the Lagrangian multiplier, $\rho > 0$ is called the step-size parameter, and $\boldsymbol{A}^{(j)}$ is a block diagonal matrix, i.e., $\boldsymbol{A}^{(j)} = \mathrm{diag}\{\boldsymbol{A}_0^{(j)}, \boldsymbol{I}_{Mq}\}$. Let $S^1_{\lambda_1/\rho}(\cdot)$ be the soft thresholding operator, which is interpreted elementwise, while $S^2_{\lambda_2/\rho}(\cdot)$ is a vector soft thresholding operator defined as

$$S^2_{\lambda_2/\rho}(\boldsymbol{\vartheta}) = \begin{cases} (1 - \lambda_2\rho^{-1}/\|\boldsymbol{\vartheta}\|_2) + \boldsymbol{\vartheta}, & \boldsymbol{\vartheta} \neq \boldsymbol{0} \\ \boldsymbol{0}, & \boldsymbol{\vartheta} = \boldsymbol{0}. \end{cases}$$

We obtain the ADMM algorithm for (19) as follows.

**Algorithm 1**

ADMM for solving subproblem (19)

**For** $t$ = 0, 1, …, **do**

- $\hat{\boldsymbol{\mu}}_{(j)}^{t+1} := [(\hat{\Omega}_j^{(r+1)})^{-1}\boldsymbol{\tau} + \rho(\boldsymbol{A}_0^{(j)})^T\boldsymbol{A}_0^{(j)}]^{-1}[(\hat{\Omega}_j^{(r+1)})^{-1}(\boldsymbol{\alpha}_2^{(j)} - \boldsymbol{H}(\hat{\boldsymbol{\Lambda}}_{(j)}^t)^T) + \rho(\boldsymbol{A}_0^{(j)})^T(\boldsymbol{v}_0^t - \boldsymbol{\varpi}_0^t)]$,
- $\boldsymbol{v}_0^{t+1} := S^1_{\lambda_1/\rho}(\boldsymbol{A}_0^{(j)}\hat{\boldsymbol{\mu}}_{(j)}^{t+1} + \boldsymbol{\varpi}_0^t)$,
- $\boldsymbol{\varpi}_0^{t+1} := \boldsymbol{\varpi}_0^t + \boldsymbol{A}_0^{(j)}\hat{\boldsymbol{\mu}}_{(j)}^{t+1} - \boldsymbol{v}_0^{t+1}$.

**For** $m$ = 1, …, $M$ **do**

- $\hat{\Lambda}_{mj}^{t+1} := [(\Omega_j^{(r+1)})^{-1}(\boldsymbol{\alpha}_{1,m}^{(j)} - \hat{\mu}_{mj}^{t+1}H_m) + \rho(\boldsymbol{v}_m^t - \boldsymbol{\varpi}_m^t)][(\hat{\Omega}_j^{(r+1)})^{-1}K_m + \rho\boldsymbol{I}_q]^{-1}$,
- $\boldsymbol{v}_m^{t+1} := S^2_{\lambda_1/\rho}(\hat{\Lambda}_{mj}^{t+1} + \boldsymbol{\varpi}_m^t)$,
- $\boldsymbol{\varpi}_m^{t+1} := \boldsymbol{\varpi}_m^t + \hat{\Lambda}_{mj}^{t+1} - \boldsymbol{v}_m^{t+1}$.

**End for**

**End for**

## 2.5 Convergence Properties and Asymptotic Properties

In this section, we first prove the convergence result of ADMM for the optimization problem (19). Technical conditions and proofs of Theorem 1 are provided in the supplementary document.

**Theorem 1 (ADMM Convergence Properties)—**Under Assumptions (A1) and (A2) in the supplementary document, we have the following results for Algorithm 1:

- *the residual* $\boldsymbol{r}^k = \boldsymbol{A}^{(j)}\breve{\boldsymbol{\theta}}_{(j)}^{(k)} - \boldsymbol{v}^k$ *converges to* $\boldsymbol{0}$ *as* $k \to \infty$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

- *the objective function* $f_0(\breve{\boldsymbol{\theta}}_{(j)}^{k})+g(\boldsymbol{v}^k)$ *converges to the optimal value* $p^*$ *as* $k \to \infty$, *where* $p^* = \inf_{\check{\boldsymbol{\theta}}_{(j)},\boldsymbol{v}}\{f_0(\check{\boldsymbol{\theta}}_{(j)}) + g(\boldsymbol{v})|\boldsymbol{A}^{(j)}\check{\boldsymbol{\theta}}_{(j)} - \boldsymbol{v} = \boldsymbol{0}\}$;
- *the dual variable* $\boldsymbol{y}^k \to \boldsymbol{y}^*$ *as* $k \to \infty$, *where* $\boldsymbol{y}^*$ *is the optimal value of the dual problem*

$$\max_{\boldsymbol{y}} \inf_{\breve{\boldsymbol{\theta}}_{(j)},\boldsymbol{v}} L_0(\breve{\boldsymbol{\theta}}_{(j)}, \boldsymbol{v}, \boldsymbol{y}).$$

Second, we prove the identifiability of MOSFA and the consistency of MPLE. A standard mixture model is not identified without any constraint, since different sets of parameters can parameterize the same distribution and therefore they are equivalent. The identifiability of finite mixture models has received a lot of attention in the literature (Yakowitz and Spragins, 1968; McLachlan and Peel, 2004; Holzmann et al., 2006; Teicher, 1963). For instance, Kent (1983) has systematically investigated the identifiability of finite mixtures of various directional distributions. By extending the existing results for identifiability, we are able to prove the identifiability property of MOSFA and present it in Proposition 1.

**Proposition 1 (Identifiability)—**Consider the proposed MOSFA model given by

$$g_u(\boldsymbol{u}_i;\boldsymbol{z}_i,\boldsymbol{\theta}) = \sum_{m=1}^{M} \pi_{mi}(\boldsymbol{\beta}) f_u(\boldsymbol{u}_i;\boldsymbol{\mu}_m, \Lambda_m\Lambda_m^T+\boldsymbol{\Omega}), \quad i=1,\ldots,n, \quad (21)$$

where $f_u(\cdot)$ is the offset-normal shape probability density function defined in (3). If the number of factors $q$ in (6) satisfies $q<B(p)=\{2p+1-(8p+1)^{\frac{1}{2}}\}/2$, where $B(p)$ is the Ledermann bound (Ledermann, 1937), and the design matrix $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ is full row rank, then $g_u(\boldsymbol{u}_i; \boldsymbol{z}_i, \boldsymbol{\theta})$ is generically identifiable in $\Theta$ up to a permutation of the components of MOSFA.

Based on the identifiability, we can further establish the convergence rate of MPLE for fixed $(q, M)$ (Khalili and Chen, 2007) when the number of parameters $\dim(\boldsymbol{\theta})$, denoted as $p_n$, tends to infinity as $n \to \infty$ (Fan et al., 2004; Städler et al., 2010). Proofs of Proposition 1 and Theorem 2 are also provided in the supplementary document.

**Theorem 2 (Consistency)—**Let $(\boldsymbol{z}_i, \boldsymbol{u}_i)$, $i = 1, 2, \ldots, n$, be a random sample drawn from $g_u(\boldsymbol{u}|\boldsymbol{z}, \boldsymbol{\theta}_0)p(\boldsymbol{z})$. If the penalty parameters satisfy $\lambda_j=O(n^{\frac{1}{2}})$ for $j = 1, 2$, $p_n^4/n=o(1)$, and the initial estimates $\check{\sigma}_j$, $\tilde{\mu}_{mj}$, and $\tilde{\mu}_{m'j}$ in the weights $\kappa_{m,m'}^{(j)}$ are $\sqrt{n}$-consistent, then there exists a local maximizer $\hat{\boldsymbol{\theta}}_n$ of the penalized log-likelihood function $\log L_p(\boldsymbol{\theta})$ in (16) such that

$$\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|_2=O_p(p_n^{\frac{1}{2}}n^{-\frac{1}{2}}), \quad (22)$$

where $\|\cdot\|_2$ represents the Euclidean norm.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

### 2.6 Model Selection & Computational Complexity

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 $M$, and the penalty parameters $\lambda_1$ and $\lambda_2$ 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 $(q, M, \lambda_1, \lambda_2)$, we estimate the MPLE $\hat{\boldsymbol{\theta}}$ 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 MPLE $\hat{\boldsymbol{\theta}}$ 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 $(\hat{q}, \hat{M}, \hat{\lambda}_1, \hat{\lambda}_2)$ is chosen based on the largest average predictive log-likelihood value.

Besides the tuning parameters in the proposed model, there is one more tuning parameter in Algorithm 1, the step-size parameter $\rho$. According to the results in Boyd et al. (2011) and Theorem 1 in this paper, our proposed method can be shown to converge for all values of the parameter $\rho$. In this paper, the parameter $\rho$ is fixed as 1.0 in both simulation studies and real data analysis. However, as discussed in Ghadimi et al. (2013), $\rho$ has a direct impact on the convergence factor of the algorithm, and inadequate tuning of this parameter can render the method slow. In this case, following the reviewers' comments, we have double checked the number of iterations in both simulation studies and real data analysis, it is found out that the iterates converge in all the cases, and the number ranges from 21 to 95, which is acceptable in terms of the high dimension settings and the big dataset.

We use the *random* EM algorithm to compute MPLE, since the EM algorithm is an iterative procedure and its performance strongly depends on its starting points. For MOSFA, a good initialization is crucial for calculating MPLE due to the presence of multiple local maxima of the penalized likelihood function. Specifically, for any given value of $(q, M, \lambda_1, \lambda_2)$, 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 $K$-means method is used for initializing mean parameter $\boldsymbol{\mu}_m$, while the principal component analysis method is used to initialize the factor loading matrices $\Lambda_m$ and the common covariance matrix $\boldsymbol{\Omega}$.

When generating the Bookstein shape coordinates, if the first two baseline landmarks are highly variable, then all the shape coordinates can be also noisy. Thus, it is critical to appropriately choose these baseline landmarks in order to cluster the shape data. To address this issue, we suggest to choose them near the midline of symmetrical or nearly symmetrical shapes, whereas we suggest to keep them far from the region with the greatest variation for non-symmetrical shapes (Bookstein, 1991). In this paper, we choose the two baseline landmarks based on the variability of all landmarks across all the subjects. Specifically, each landmark is initially set as the first baseline landmark and the related preform of configuration $\boldsymbol{X} = \boldsymbol{L}\boldsymbol{X}^{\dagger}$ is then calculated. The total variance across both the entries in the preform matrix $\boldsymbol{X}$ and the subjects is calculated as the loss function for such landmark. Finally, we choose the landmark with the smallest value of loss function as the first baseline

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

landmark and clockwisely reorder the rest of landmarks. Subsequently, given the first baseline landmark, we calculate the variance at each landmark in preform matrix $\boldsymbol{X}$ across all subjects and choose the second landmark that has the smallest variance among the rest of landmarks.

Finally, we analyze the computational complexity of the EM algorithm for both MOSFA and penalized MOSFA. The computational complexity is calculated per-iteration for two parts: E-step and M-step. First, in E-step, the EM algorithm for MOSFA has the same computational complexity, $O(Mn)$, as that for penalized MOSFA. In M-step, the computational complexity of updating $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ is $O(d(M-1) + np)$ for both models. For updating $\boldsymbol{\mu}_m$ and $\boldsymbol{\Lambda}_m$, however, its computational complexity is $O(Mp(q + 1))$ for MOSFA and $O(k\tilde{M}p(q + 1))$ for penalized MOSFA, where $k$ is the number of iterations in the ADMM algorithm. Then, the computational complexity of the whole procedure for MOSFA is $O(n_I n_M n_q r(\tilde{d}(M - 1) + np + nM + Mp(q + 1)))$, where $n_I$ is the number of initial values, $r$ is the number of iterations for the EM algorithm, and $n_M$ and $n_q$ are, respectively, the number of the alternative values for $M$ and that for $q$. Due to the additional variable selection procedure, the computational complexity of penalized MOSFA is $O(n_I n_\lambda^1 n_\lambda^2 \tilde{M}\tilde{q}\tilde{r}\{d(M-1)+np+nM+Mp(q+1)\})$, where $n_\lambda^1$ and $n_\lambda^2$ are, respectively, the number of the alternative values for $\lambda_1$ and that for $\lambda_2$. In the next section, we will show the running time of each simulation in seconds under different settings.

## 3 Simulation Studies

We conducted a set of Monte Carlo simulations to evaluate the finite sample performance MOSFA and compared it with the mixtures of offset-normal shape (MOS) model (Kume and Welling, 2010). We simulated CC shape data according to MOSFA as follows. We set $n$ = 100, $k$ = 50, and $M$ = 2. We randomly chose the CC contours of a normal control and a diseased patient from the ADHD-200 data set as the mean shapes of two different clusters. Figure 2 presents the CC contours and their corresponding landmarks from the two selected subjects.

In each cluster, the landmark configuration of each subject was set as the true value of the parameter $\boldsymbol{\mu}_m$ for $m$ = 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 spatial correlation structure, we considered three different cases as follows:

- Case 1: simple diagonal matrix: $\sum_m = \sigma_m^2 \boldsymbol{I}_{2(k-1)}$, $m$ = 1, 2;
- Case 2: cyclic Markov covariance:

$$\sum_m = \sigma_m^2 (\boldsymbol{I}_2 \otimes \boldsymbol{L}) \boldsymbol{I}_2 \otimes \boldsymbol{G}_m (\boldsymbol{I}_2 \otimes \boldsymbol{L}^T), m = 1, 2, \quad (23)$$

  where $(\boldsymbol{G}_m)_{j,j'} = [\gamma_m^{|j'-j|} + \gamma_m^{k-|j'-j|}]/(1-\gamma_m^k)$, $1 \le j', j \le k$ and $0 \le \gamma_m < 1$;
- Case 3: latent factor analysis model in (6): $\sum_m = \Lambda_m \Lambda_m^T + \boldsymbol{\Omega}$, $m$ = 1, 2.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The scale parameters $\sigma_m$, $m = 1, 2$ in Cases 1 and 2 were generated from $U(0.5, 0.6)$ and $U(0.8, 1)$, respectively. In Case 2, we set $\gamma_1 = 0.2$ and $\gamma_2 = 0.5$. In Case 3, the number of loading factors was set as $q = 2$. The latent variable $\boldsymbol{b}_{mi}$ was generated from $N(\mathbf{0}, \boldsymbol{I}_2)$, while the error terms $\boldsymbol{e}_{mi}$, independently of $\boldsymbol{b}_{mi}$, were generated from $N(\mathbf{0}, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega} = \text{diag}([\sigma_1, \ldots, \sigma_{2(k-1)}])$, in which we simulated $\sigma_l \sim U(1, 2)$ for all $l \leq 2(k-1)$. For the loading matrices $\Lambda_m$, $m = 1, 2$, the elements of the first $\ell_0$ 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. In each case, we simulated $N = 200$ data sets.

We fitted MOS, MOSFA, and penalized MOSFA to each simulated data set. For Cases 1 and 2, we considered two MOS models with the true correlation structure and an unspecified correlation structure. In Case 3, we only considered one MOS model with the unspecified correlation structure, while we considered two set-ups with different values of the parameters $\ell_0$, $c_1$, and $c_2$. For each data set, we randomly chose 10 sets of initial values in the *random* EM algorithm and then used the 2-fold cross predictive log-likelihood method to determine $M$ in MOS, $(q, M)$ in MOSFA, and $(q, M, \lambda_1, \lambda_2)$ in penalized MOSFA. For all the models, the baseline landmarks were chosen based on the method described in Section 2.6. The Rand index (RI) (Rand, 1971) and adjusted Rand index (aRI) (Hubert and Arabie, 1985) were used to compare the clustering results with the ground truth and to evaluate the finite sample performance of all the three models.

Table 1 presents the simulation results for Cases 1 and 2. All the four models show excellent clustering performance, while MOS with the true correlation structure and penalized MOSFA outperform the other two models. Moreover, although the correlation structure is misspecified for MOSFA and penalized MOSFA, penalized MOSFA performs as well as MOS with the true correlation structure. Therefore, our proposed model is robust to the misspecification of correlation structure.

Table 2 presents the simulation results corresponding to different values of $(\ell_0, c_1, c_2)$ for Case 3. Table 2 shows that MOSFA and penalized MOSFA outperform MOS. Furthermore, penalized MOSFA has the smallest Rand index and adjusted Rand index being larger than 0.85 for all values of $\ell_0$ in the two set-ups. In contrast, MOS performs well for $\ell_0 = 30$ and 60 in Set-up 1 with the indices larger than 0.9, whereas it performs very bad for $\ell_0 = 90$ in Set-up 1 and $\ell_0 \geq 60$ in Set-up 2.

We investigated the effect of baseline landmark selection on penalized MOSFA for the two set-ups in Case 3 and chose the baseline landmarks by either using the method in Section 2.6 or randomly choosing one. Table 3 presents the clustering results. When the loading matrix is sparse, penalized MOSFA is robust to the two choices of baseline landmarks. However, for more complex correlation structures, i.e, $\ell_0 = 90$ in Set-up 1 and $\ell_0 = 40, 60$ in Set-up 2, the penalized MOSFA model can perform poorly for the randomly selected baseline landmarks. These results may be caused by the fact that for more complex correlation structure, the variability in the baseline landmarks can be transferred to all the rest of landmarks, and thus all the shape coordinates can be very noisy. It may indicate that it is critical to choose baseline landmarks. Moreover, the results in Tables 1 and 2 indicate that

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the proposed method is effective and enhances the clustering performance of penalized MOSFA.

Table 4 includes the computation time of MOSFA and penalized MOSFA for different values of $n$ and numbers of landmarks $k$ in Cases 1–3. All computations were done in Matlab2013a on a server with a single core in a CPU, 4GB memory, and 2.93 GHz Intel processor. In Case 3, $m = 30$ and 40 are considered in Set-ups 1 and 2, respectively. For each model, its computation time is consistent with its computational complexity calculated in Section 2.6.

## 4 ADHD-200 Corpus Callosum Shape Data

Attention Deficit Hyperactivity Disorder (ADHD) is one of the most commonly diagnosed childhood behavioral disorders. It affects at least 5% of school-age children, causing them to be difficult to control their behaviors or focus their attentions. Despite a voluminous empirical literature, the scientific community remains without a comprehensive model of the pathophysiology of ADHD.

The corpus callosum (CC), the largest white matter structure in the brain, has been a structure of high interest in many neuroimaging studies of neuro-developmental pathology. It contains homotopic and heterotopic interhemispheric connections and is essential for communication between the two cerebral hemispheres. Individual differences in CC, and their possible implications regarding interhemispheric connectivity, have been investigated in last several decades (Witelson, 1989; Paul et al., 2007). There is a large body of work suggesting that CC plays an important role in attentional control in neurologically intact individuals and its integrity in clinical populations that suffer from ADHD is of particular interest (Lyoo et al., 1996; Hill et al., 2003; Hutchinson et al., 2007).

We consider the CC contour data of ADHD-200 Dataset[2]. The ADHD-200 sample contains both anatomical and resting-state functional MRI data of 776 labeled subjects across 8 independent imaging sites, 491 of which were obtained from typically developing individuals and 285 in children and adolescents with ADHD (ages: 7–21 years old). We processed the CC shape data for each subject in ADHD-200 Dataset as follows. We used *FreeSurfer* package[3] (Dale et al., 1999) to process each T1-weighted MRI, including motion correction, non-parametric non-uniform intensity normalization, affine transform to the MNI305 atlas, intensity normalization, skullstripping, and automatic subcortical segmentation. Some quality control procedures were done on each output image data. The intracranial volume (ICV) information was calculated from the output of *FreeSurfer* package. Subsequently, the midsagittal CC area was calculated in the *CCseg* package, which is measured by using subdivisions in Witelson (1989) motivated by neuro-histological studies (see Figure 3). Then, each T1-weighted MRI image and tissue segmentation calculated from *FreeSurfer* were used as the input files of *CCSeg* package to extract the planar CC shape data on the midsagittal slice, which contains 50 landmarks. The *CCseg* framework (Székely et al., 1996; Vachet et al., 2012) entails three main steps: (i) automatic

[2] http://fcon1000.projects.nitrc.org/indi/adhd200/
[3] http://surfer.nmr.mgh.harvard.edu/

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

initialization of the corpus callosum model, (ii) multi-step automatic (and potentially interactive) segmentation via constrained elastic deformation of a flexible Fourier contour model, and (iii) lobar area computation using a probabilistic subdivision model. After quality control, we obtained 647 CC shape data out of 776 subjects. The demographic information of the processed CC shape data set is presented in Table 5.

The area of CC is an important morphologic feature, which changes throughout infancy (Garel et al., 2011). Here, before we considered the CC shape information, we examined the association between the midsagittal CC area and some covariates of interest, such as gender or diagnosis status (Lyoo et al., 1996; Jäncke et al., 1997; Hill et al., 2003). We fitted a log-tranformed linear regression as follows:

$$\log y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + b_4 x_{i4} + b_5 \log x_{i5} + \varepsilon_i, \ \ \varepsilon_i \sim N(0,1), \ \ i = 1, \ldots, n, \quad (24)$$

where $y_i$ is the midsagittal CC area of the $i^{th}$ subject, $x_{i1}$=1(the $i^{th}$ subject is a diseased patient) is a dummy variable indicating the diagnosis status of the $i^{th}$ subject, $x_{i2}$ and $x_{i3}$ are gender and age, respectively, $x_{i4} = x_{i1}x_{i3}$ is the diagnosis by age interaction term, and $x_{i5}$ is the intracranial volume (ICV) of the $i^{th}$ subject. Moreover, we considered 5 different responses $y_i$: (i) the total midsagittal CC area; (ii) the area of prefrontal subdivision in CC; (iii) the area of frontal subdivision in CC; (iv) the area of parietal subdivision in CC; and (v) the area of occipito-temporal subdivision in CC. Table 6 presents the regression analysis results for all five responses. We observe that there is no significant gender difference in neither total CC area nor CC subareas, whereas the diagnosis by age interaction term is statistically significant, indicating that the midsagittal CC area and its subareas significantly change across groups as age varies.

Besides the CC volume data, the CC shape data is of great interest in many neuroimaging studies of neuro-developmental pathology. Many neurological studies indicate that the shape of CC for healthy young adults is associated with gender, age, cognitive performance, and neuro-degenerative diseases, among other factors (Farag et al., 2010; Joshi et al., 2013; Martín-Loeches et al., 2013). In Martín-Loeches et al. (2013), the CC shape variation was shown to be consistent and have significant correlations with attentional control, which is the core deficit in ADHD. Here we are more interested in whether the CC shape information is a promising biomarker for the diagnosis of ADHD and may provide a clue to the topographical spread of ADHD disease. We applied MOSFA to the CC shape data set to explore the relationship between CC shape data and ADHD diagnosis information. The 2-fold cross predictive log-likelihood method was adopted to select the tuning parameters and calculate the estimates. The cluster memberships of subjects in the testing data set were determined according to the fitted models.

We first applied the MOS model of Kume and Welling (2010), where the covariance matrix of each component is a diagonal matrix. The mean shape of all these shape data is presented in Figure 4(a). Meanwhile, the first two landmarks are also highlighted in Figure 4(a). Based on the 2-fold cross predictive log-likelihood method, the MOS model was fitted, but it selects only one cluster.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

As a comparison, the proposed penalized MOSFA model was used to cluster the training dataset. The first two landmarks were the same as those used in MOS. In MOSFA, we set $\boldsymbol{z}$ =(1, Gender, Age). We calculated MPLE by using the EM algorithm and then the final MOSFA model was able to detect 4 clusters with 239, 98, 64, and 246 subjects, respectively. The first three clusters contain 391 normal controls and 10 ADHD patients, whereas the fourth cluster includes 13 normal controls and 233 ADHD patients. Thus, the first three clusters contain almost all the normal controls, whereas most diseased subjects fall into the last cluster. The mean shape of the CC shape data in each cluster is presented in Figure 4(b). The mean shapes of the first three are similar to each other, whereas they are different from the mean shape of cluster four. We randomly chose 10 subjects from each cluster and presented their shape data in Figure 5.

To check the stability of our clustering results, we applied the leave-one-out method via removing one of these four clusters and reapplying penalized MOSFA to the rest subjects. We treated the original clustering result as the ground truth and calculated the Rand index and adjusted Rand index based on the clustering result for each reduced dataset. The stability performance is presented in Table 7. Based on the estimated number of clusters and the two indices, the clustering results show stable performance when each of the first three clusters is removed from the whole dataset. However, the clustering result becomes unstable when the last cluster is removed. It may indicate that the features in shape space among the first three clusters are not significantly different from each other. We will further investigate this issue below.

We are also interested in the estimated loading matrices for all the four clusters. The estimated number of factors in each cluster is 2. To extract the shape features of each cluster, we plotted each column in loading matrices for all the clusters in Figure 6. The columns of the loading matrices from the first three clusters have similar tendency, whereas they are different from those from the last cluster. It is consistent with the diagnosis information: most normal controls are in the first three clusters, whereas most ADHD patients are in the last cluster.

Then, we randomly chose subjects from each cluster and applied the *ClosedCurves2D3D* software[4] to compute a pair-wise geodesic path among the four clusters under the elastic Riemannian metric (Srivastava et al., 2011). For subjects in the same cluster and in different clusters, their shapes placed equidistant along the geodesic paths are plotted in Figure 7, and the geodesic distance between each pair of shapes is presented in Table 8. Figure 7 and Table 8 show that the geodesic distance between subjects in the same cluster is smaller than that between subjects in different clusters. Furthermore, the geodesic distance between subjects in the first three clusters is much smaller than the geodesic distance between subjects in the first three clusters and those in the fourth cluster.

We tested the mean shape difference among different clusters by using a bootstrap hypothesis testing approach (Amaral et al., 2007) and its related R package *shapes*[5]. The

[4] http://ssamg.stat.fsu.edu/software
[5] http://cran.r-project.org/web/packages/shapes/index.html

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

number of bootstrapping iterations was set as 500. The Hoteling $T^2$ test statistic and the related *p*-value (in brackets) are summarized in Table 9. It shows that the *p*-values of the two sample test for the first three clusters are not significant, while the *p*-values show a statistically significant mean shape difference between the first three clusters and the fourth one.

According to the data analysis results above, it seems that there is no statistical significant mean shape difference among the first three clusters. Therefore, we combined the first three clusters into group 1 and treated the fourth cluster as group 2. By using this information as the ground truth, we calculated the Rand index and adjusted Rand index to be $RI = 0.9311$ and $aRI = 0.8622$, respectively. This may indicate that the use of the planar CC shape leads to two meaningful and robust clusters in the ADHD-200 data set. Following the reviewers' comments, besides gender and age, we also took the disease status as the covariate, $z_{i,3} = 1$ for normal controls and 0 for ADHD patients. We reran our clustering method and 4 clusters were shown up. The Rand index and adjusted Rand index were also calculated to be $RI = 0.9347$ and $aRI = 0.8672$, respectively. It may indicate that there is only a small improvement by adding the disease status into MOSFA. Finally, based on the clustering results discussed above, the planar CC shape data may be a powerful biomarker for distinguishing ADHD patients from normal controls.

## 5 Conclusion

We have developed a penalized MOSFA clustering framework for clustering high-dimensional planar shape data. MOSFA is developed to specifically address four major challenges including the curved shape space, a high-dimensional feature space, complex spatial correlation, and shape variation associated with covariates (e.g., age or gender). We have developed an efficient EM algorithm coupled with the ADMM algorithm to calculate the MPLE of $\theta$. Our simulations have confirmed the excellent clustering performance of MOSFA in different scenarios. Our ADHD-200 data analysis has shown that penalized MOSFA can undercover meaningful clusters of the CC planar shape data.

Several important issues need to be addressed in future research. First, the shape space here are actually shapes of closed planar curves, and the two-dimensional (2D) shape data is modelled by the offset-normal shape distribution, which is the marginal distribution of the directional component of a multivariate normal distribution. It is meaningful to generalize the proposed model from a geometric and topological viewpoints for data from Rimennain symmetric space, which is of great importance for neuroimaging studies (Goh and Vidal, 2008; Shi et al., 2012; Joshi et al., 2013). Developing general methods for simultaneously performing variable selection and clustering on the data from Rimennain symmetric space faces up with many new challenges both computationally and theoretically. Second, the proposed MOSFA model can be treated as a landmark-based analysis, where shapes are represented by a coarse, discrete sampling of the object contours. However, this approach is limited in that automatic detection of landmarks is not straightforward and the ensuing shape analysis depends heavily on the choice of landmarks (Srivastava et al., 2005). An alternative approach of shape representation and analysis is the continuous framework, where a shape is represented by mappings (diffeomorphism) or functions (level sets), which are able to

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

handle topological changes such as merging and splitting of connected component (Krim and Yezzi, 2006). It is much more interesting to extend this landmark-based shape analysis work to the continuous shape framework, and more research is needed for formulating the continuous shape data clustering method. Finally, the methodology proposed here can be extended to shape classification by using 2D landmark-based shape data, which will be investigated in our future research.

## Acknowledgments

This work was partially supported by NIH grants MH086633, RR025747, 1UL1TR001111, and MH092335 and NSF grants SES-1357666 and DMS-1407655.

## References

Amaral GA, Dryden I, Wood ATA. Pivotal bootstrap methods for k-sample problems in directional statistics and shape analysis. Journal of the American Statistical Association. 2007; 102:695–707.

Amaral GJ, Dore LH, Lessa RP, Stosic B. k-Means Algorithm in Statistical Shape Analysis. Communications in Statistics-Simulation and Computation. 2010; 39:1016–1026.

Atienza N, Garcia-Heras J, Munoz-Pichardo J. A new condition for identifiability of finite mixture distributions. Metrika. 2006; 63:215–221.

Bekker PA, ten Berge JM. Generic global indentification in factor analysis. Linear Algebra and its Applications. 1997; 264:255–263.

Bertsekas, DP. Convex Optimization Theory. Athena Scientific; Cambridge: 2009.

Bookstein, FL. Morphometric Tools for Landmark Data. Cambridge University Press; 1991.

Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning. 2011; 3:1–122.

Cootes TF, Taylor CJ, Cooper DH, Graham J. Active shape models-their training and application. Computer Vision and Image Understanding. 1995; 61:38–59.

Dale AM, Fischl B, Sereno MI. Cortical surface-based analysis: I. Segmentation and surface reconstruction. Neuroimage. 1999; 9:179–194. [PubMed: 9931268]

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B (Methodological). 1977; 39:1–38.

Dryden, I.; Mardia, K. Statistical Analysis of Shape. Wiley; 1998.

Dryden I, Mardia KV. General shape distributions in a plane. Advances in Applied Probability. 1991; 23:259–276.

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

Farag, A.; Elhabian, S.; Abdelrahman, M.; Graham, J.; Chen, D.; Casanova, MF. Shape modeling of the corpus callosum. Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE; IEEE; 2010.

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

Gabay D, Mercier B. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Computers & Mathematics with Applications. 1976; 2:17–40.

Garel C, Cont I, Alberti C, Josserand E, Moutard M, le Pointe HD. Biometry of the corpus callosum in children: MR imaging reference data. American Journal of Neuroradiology. 2011; 32:1436–1443. [PubMed: 21799035]

Ghadimi E, Teixeira A, Shames I, Johansson M. Optimal parameter selection for the alternating direction method of multipliers (ADMM): Quadratic problems. 2013 arXiv preprint arXiv: 1306.2454.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Glowinski R, Marroco A. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires. ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique. 1975; 9:41–76.

Goh, A.; Vidal, R. Clustering and dimensionality reduction on Riemannian manifolds. Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on; IEEE; 2008.

Guo J, Levina E, Michailidis G, Zhu J. Pairwise variable selection for high-dimensional model-based clustering. Biometrics. 2010; 66:793–804. [PubMed: 19912170]

Hill DE, Yeo RA, Campbell RA, Hart B, Vigil J, Brooks W. Magnetic resonance imaging correlates of attention-deficit/hyperactivity disorder in children. Neuropsychology. 2003; 17:496–506. [PubMed: 12959515]

Holzmann H, Munk A, Gneiting T. Identifiability of finite mixtures of elliptical distributions. Scandinavian Journal of Statistics. 2006; 33:753–763.

Hubert L, Arabie P. Comparing partitions. Journal of Classification. 1985; 2:193–218.

Hutchinson M, Wilkes L, Vickers M, Jackson D. The development and validation of a bullying inventory for the nursing workplace. Nurse Researcher. 2007; 15:19–29. [PubMed: 18283759]

Jäncke L, Staiger J, Schlaug G, Huang Y, Steinmetz H. The relationship between corpus callosum size and forebrain volume. Cerebral Cortex. 1997; 7:48–56. [PubMed: 9023431]

Joshi SH, Narr KL, Philips OR, Nuechterlein KH, Asarnow RF, Toga AW, Woods RP. Statistical shape analysis of the corpus callosum in Schizophrenia. Neuroimage. 2013; 64:547–559. [PubMed: 23000788]

Kent JT. Identifiability of finite mixtures for directional data. The Annals of Statistics. 1983; 11:984–988.

Khalili A, Chen J. Variable selection in finite mixture of regression models. Journal of the american Statistical association. 2007; 102:1025–1038.

Krim, H.; Yezzi, AJ. Statistics and Analysis of Shapes. Springer; 2006.

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

Ledermann W. On the rank of the reduced correlational matrix in multiple-factor analysis. Psychometrika. 1937; 2:85–93.

Lyoo IK, Noam GG, Lee CK, Lee HK, Kennedy BP, Renshaw PF. The corpus callosum and lateral ventricles in children with attention-deficit hyperactivity disorder: a brain magnetic resonance imaging study. Biological Psychiatry. 1996; 40:1060–1063. [PubMed: 8915567]

Martín-Loeches M, Bruner E, De La Cuétara JM, Colom R. Correlation between corpus callosum shape and cognitive performance in healthy young adults. Brain Structure and Function. 2013; 218:721–731. [PubMed: 22581173]

McLachlan, G.; Peel, D. Finite Mixture Models. John Wiley & Sons; 2004.

McLachlan GJ, Peel D, Bean R. Modelling high-dimensional data by mixtures of factor analyzers. Computational Statistics & Data Analysis. 2003; 41:379–388.

Pan W, Shen X. Penalized model-based clustering with application to variable selection. The Journal of Machine Learning Research. 2007; 8:1145–1164.

Paul LK, Brown WS, Adolphs R, Tyszka JM, Richards LJ, Mukherjee P, Sherr EH. Agenesis of the corpus callosum: genetic, developmental and functional aspects of connectivity. Nature Reviews Neuroscience. 2007; 8:287–299. [PubMed: 17375041]

Rajpoot NM, Arif M. Unsupervised shape clustering using diffusion maps. The Annals of the BMVA. 2008; 2008:1–17.

Rand WM. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association. 1971; 66:846–850.

Shapiro A. Identifiability of factor analysis: Some results and open problems. Linear Algebra and its Applications. 1985; 70:1–7.

Shi X, Zhu H, Ibrahim JG, Liang F, Lieberman J, Styner M. Intrinsic regression models for medial representation of subcortical structures. Journal of the American Statistical Association. 2012; 107:12–23. [PubMed: 23794769]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Small, C. The Statistical Theory of Shape. Springer; 1996.

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

Srivastava A, Klassen E, Joshi SH, Jermyn IH. Shape analysis of elastic curves in euclidean spaces. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2011; 33:1415–1428.

Städler N, Bühlmann P, Van De Geer S. l1-penalization for mixture regression models. Test. 2010; 19:209–256.

Subbarao R, Meer P. Nonlinear mean shift over Riemannian manifolds. International Journal of Computer Vision. 2009; 84:1–20.

Székely G, Kelemen A, Brechbühler C, Gerig G. Segmentation of 2-D and 3-D objects from MRI volume data using constrained elastic deformations of flexible Fourier contour and surface models. Medical Image Analysis. 1996; 1:19–34. [PubMed: 9873919]

Teicher H. Identifiability of finite mixtures. The Annals of Mathematical Statistics. 1963; 34:1265–1269.

Vachet, C.; Yvernault, B.; Bhatt, K.; Smith, RG.; Gerig, G.; Hazlett, HC.; Styner, M. SPIE Medical Imaging. Vol. 8317. International Society for Optics and Photonics; 2012. Automatic corpus callosum segmentation using a deformable active Fourier contour model; p. 831707-7.

Willink R. Normal moments and Hermite polynomials. Statistics & Probability Letters. 2005; 73:271–275.

Witelson SF. Hand and sex differences in the isthmus and genu of the human corpus callosum a postmortem morphological study. Brain. 1989; 112:799–835. [PubMed: 2731030]

Xie B, Pan W, Shen X. Penalized mixtures of factor analyzers with application to clustering high-dimensional microarray data. Bioinformatics. 2010; 26:501–508. [PubMed: 20031967]

Yakowitz SJ, Spragins JD. On the identifiability of finite mixtures. The Annals of Mathematical Statistics. 1968; 39:209–214.

Younes, L. Shapes and Diffeomorphisms. Springer; 2010.

Zhou H, Pan W, Shen X. Penalized model-based clustering with unconstrained covariance matrices. Electronic Journal of Statistics. 2009; 3:1473. [PubMed: 20463857]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

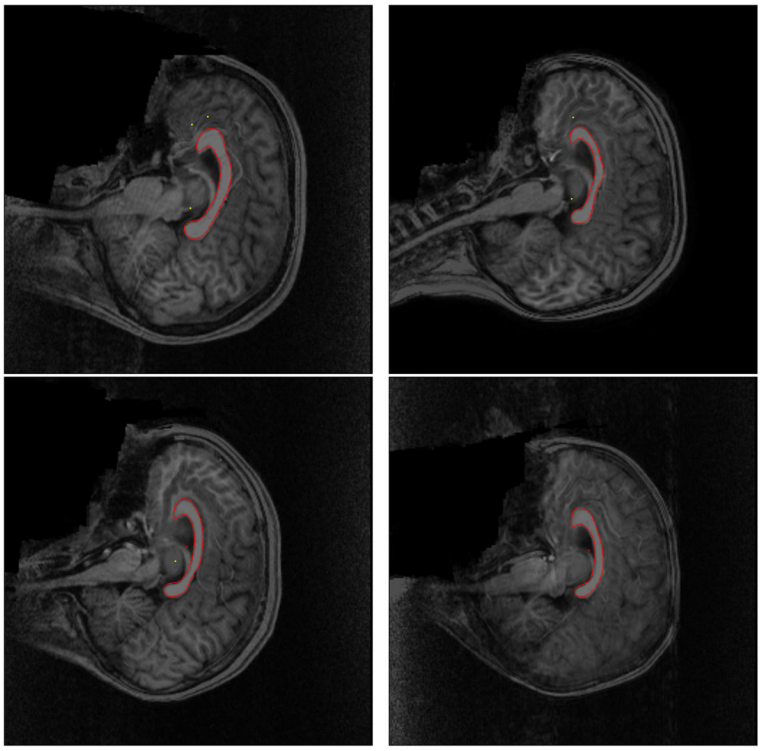

**Figure 1.**
Automatic corpus callosum segmentation of four randomly selected subjects from the ADHD-200 study.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

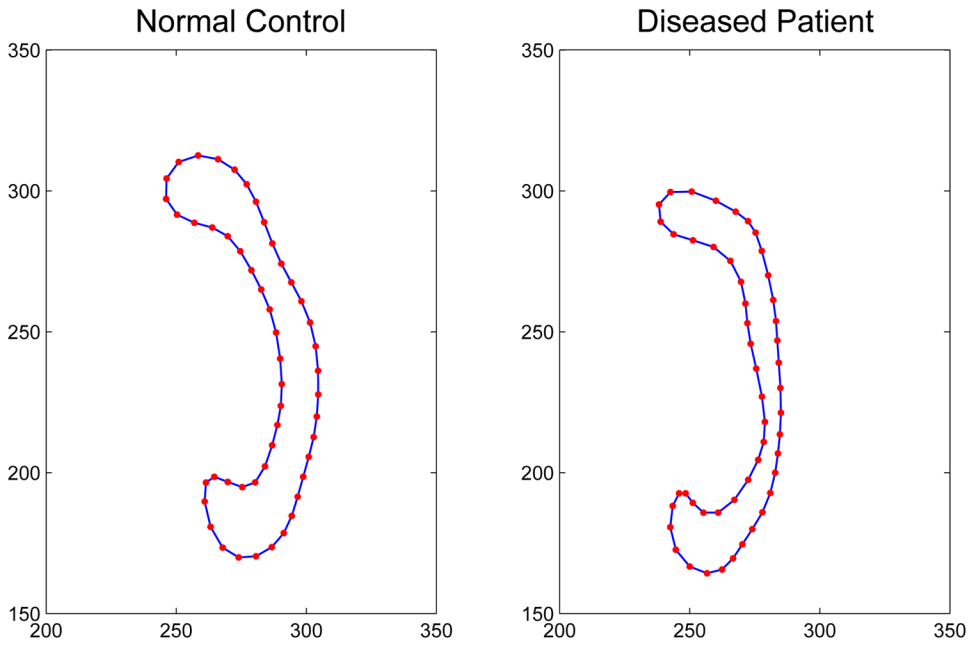


**Figure 2.**
Contours and landmarks of a normal control and an ADHD subject.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

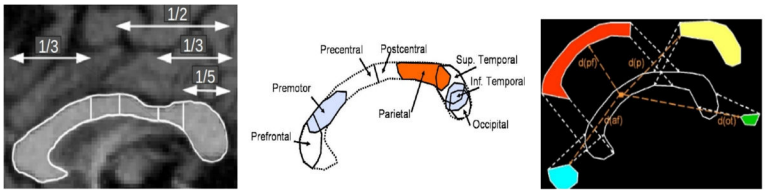


**Figure 3.**

Subdivisions of corpus callosum in Witelson (1989) approach (left), its neuro-histological motivation (Vachet et al. (2012), middle), and schematic visualization of the probability computation (Vachet et al. (2012), right): prefrontal subdivision (blue); frontal subdivision (red); parietal (yellow); and occipito-temporal subdivision (green).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

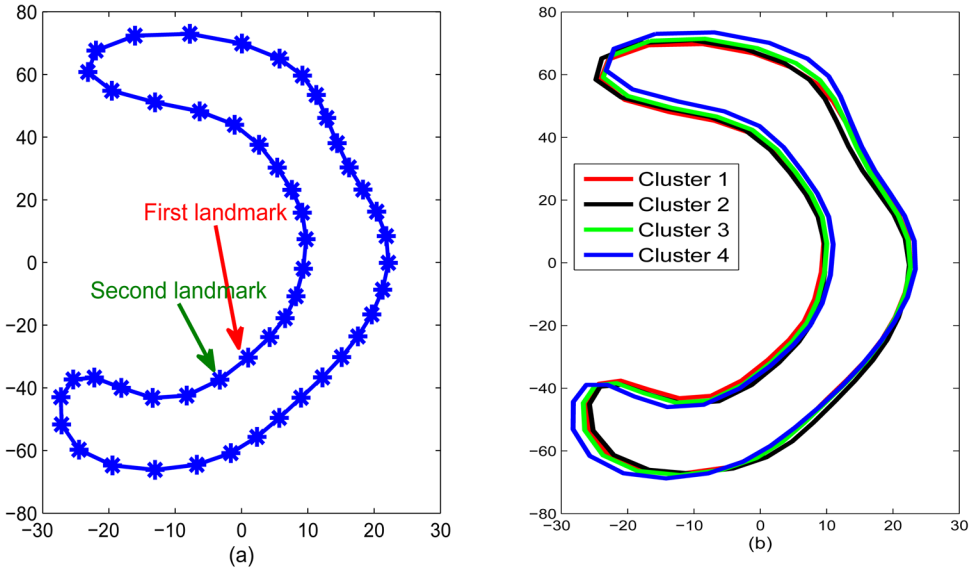


**Figure 4.**

ADHD-200 data analysis: (a) Mean shape of all CC shape data and the first two landmarks chosen based on the variation of landmarks; (b) Mean shape of CC shape data in four clusters.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

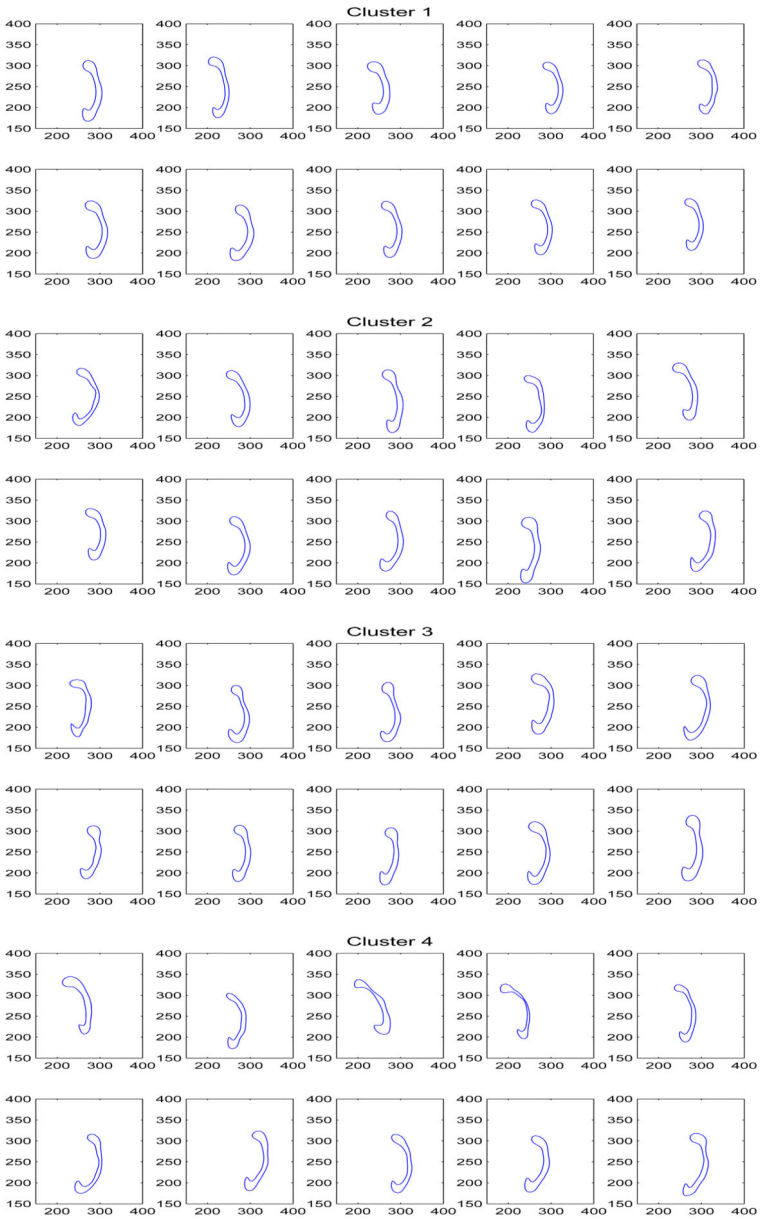


**Figure 5.**
ADHD-200 data analysis: landmarks of subjects in four clusters.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

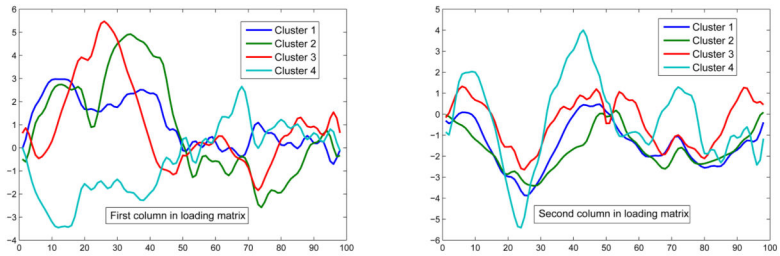


**Figure 6.**
ADHD-200 data analysis: columns in loading matrices from the four clusters.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

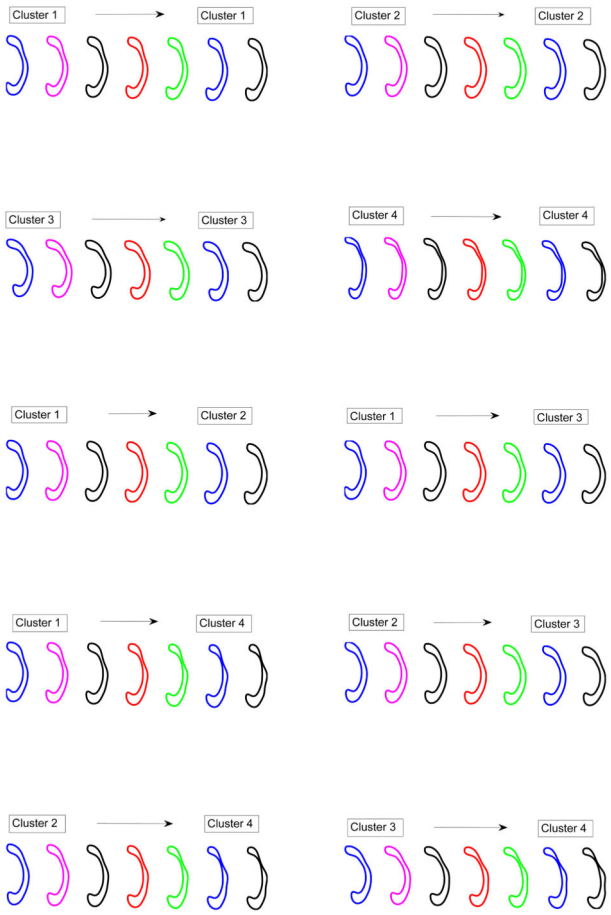


**Figure 7.**
ADHD-200 data analysis: shapes placed equidistant along the geodesic paths in four clusters.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

## Table 1

Performance of MOS with the true correlation structure (MOS-true), MOS with the general correlation structure (MOS-general), MOSFA and penalized MOSFA models in Case 1 and Case 2. RI and aRI denote the average of the Rand index and adjusted Rand index, respectively. For each case, 200 simulated data sets were used.

| Case 1 | | | | |
|---|---|---|---|---|
| **Cluster $\hat{M}$** | **MOS-true** | **MOS-general** | **MOSFA** | **penalized MOSFA** |
| 1 | 2 | 6 | 3 | 1 |
| 2 | 198 | 190 | 191 | 197 |
| 3 | 0 | 4 | 6 | 2 |
| RI(aRI) | 0.99(0.98) | 0.94(0.90) | 0.95(0.91) | 0.99(0.98) |

| Case 2 | | | | |
|---|---|---|---|---|
| **Cluster $\hat{M}$** | **MOS-true** | **MOS-general** | **MOSFA** | **penalized MOSFA** |
| 1 | 9 | 13 | 3 | 2 |
| 2 | 190 | 176 | 180 | 188 |
| 3 | 1 | 11 | 17 | 10 |
| RI(aRI) | 0.99(0.98) | 0.86(0.77) | 0.90(0.84) | 1(0.99) |

Author Manuscript Author Manuscript Author Manuscript Author Manuscript



**Table 2**

Comparison of MOS, MOSFA, and penalized MOSFA models for different values of ($\ell_0$, $c_1$, $c_2$) in Case 3. RI and aRI denote the average of the Rand index and adjusted Rand index, respectively. For each case, 200 simulated data sets were used.

| | | Set-up 1: $c_1 = 2$, $c_2 = 1$ | | | Set-up 2: $c_1 = 5$, $c_2 = 2$ | | |
|---|---|---|---|---|---|---|---|
| **Model** | **Cluster $\hat{M}$** | **$\ell_0 = 30$** | **$\ell_0 = 60$** | **$\ell_0 = 90$** | **$\ell_0 = 20$** | **$\ell_0 = 40$** | **$\ell_0 = 60$** |
| 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) |
| MOSFA | 1 | 0 | 3 | 13 | 2 | 3 | 11 |
| | 2 | 200 | 181 | 149 | 187 | 165 | 137 |
| | 3 | 0 | 17 | 38 | 11 | 32 | 52 |
| | RI(aRI) | 1(1) | 0.89(0.84) | 0.76(0.54) | 0.91(0.88) | 0.81(0.63) | 0.72(0.41) |
| 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) |

Author Manuscript Author Manuscript Author Manuscript Author Manuscript



**Table 3**

Performance of penalized MOSFA model in Case 3 when the baseline landmarks were chosen based on the method in Section 2.6 or randomly. RI and aRI denote the average of the Rand index and adjusted Rand index, respectively. For each case, 200 simulated data sets were used.

| | | Set-up 1: $c_1 = 2, c_2 = 1$ | | | Set-up 2: $c_1 = 5, c_2 = 2$ | | |
|---|---|---|---|---|---|---|---|
| **Mode** | **Cluster $\hat{M}$** | **$\ell_0 = 30$** | **$\ell_0 = 60$** | **$\ell_0 = 90$** | **$\ell_0 = 20$** | **$\ell_0 = 40$** | **$\ell_0 = 60$** |
| Method in Section 2.6 | 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) |
| Random | 1 | 0 | 1 | 9 | 1 | 14 | 16 |
| | 2 | 199 | 189 | 170 | 184 | 162 | 146 |
| | 3 | 1 | 10 | 21 | 15 | 24 | 38 |
| | RI(aRI) | 1(1) | 0.92(0.88) | 0.81(0.60) | 0.89(0.82) | 0.72(0.53) | 0.63(0.43) |

Author Manuscript Author Manuscript Author Manuscript Author Manuscript



**Table 4**

Computation time of MOSFA and penalized MOSFA models for different values of $k$ and $n$ in Cases 1–3. For each case, 200 simulated data sets were used.

| Model | $k$ | $n$ | Case 1 | Case 2 | Case 3 (1) | Case 3 (2) |
|---|---|---|---|---|---|---|
| | | 100 | 94s | 85s | 109s | 135s |
| MOSFA | 25 | 500 | 704s | 691s | 714s | 762s |
| | | 100 | 411s | 418s | 442s | 487s |
| MOSFA | 50 | 500 | 2631s | 2672s | 2680s | 2769s |
| | | 100 | 596s | 537s | 578s | 613s |
| penalized MOSFA | 25 | 500 | 4009s | 4047s | 4021s | 4147s |
| | | 100 | 2332s | 2471s | 2399s | 2483s |
| penalized MOSFA | 50 | 500 | 16391s | 16183s | 16547s | 17105s |

*J Am Stat Assoc*. Author manuscript; available in PMC 2016 May 07.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 5**

Demographic information about processed ADHD-200 CC shape dataset, including disease status, age, and gender.

| Disease status | NO. | Range of age in years (mean) | Gender (female/male) |
|---|---|---|---|
| Typically Developing Children | 404 | 7.09–21.83 (12.43) | 179/225 |
| ADHD-Combined | 150 | 7.17–20.15 (10.96) | 39/111 |
| ADHD-Hyperactive/Impulsive | 8 | 9.22–20.89 (14.69) | 1/7 |
| ADHD-Inattentive | 85 | 7.43–17.61 (12.23) | 18/67 |
| All data | 647 | 7.09–21.83(12.09) | 237/410 |

Author Manuscript Author Manuscript Author Manuscript Author Manuscript



**Table 6**

ADHD-200 data analysis: multivariate regression analysis of the midsagittal CC area.

| Parameter | $b_0$ | $b_1$ | $b_2$ | $b_3$ | $b_4$ | $b_5$ |
|---|---|---|---|---|---|---|
| Case 1: $y$ = total midsagittal CC area | | | | | | |
| Estimate | 6.416 | 4.102E-03 | 1.096E-02 | 2.319E-02 | −2.440E-02 | 1.546E-02 |
| Std. Error | 1.279E-02 | 1.493E-02 | 1.641E-02 | 8.6287E-03 | 1.540E-02 | 7.950E-03 |
| t-statistic | 5.017E+02 | 2.748E-01 | 6.681E-01 | 2.688 | −2.585 | 1.945 |
| p-value | <1E-05 | 7.836E-01 | 5.043E-01 | 7.392E-03 | 4.870E-03 | 5.222E-02 |
| Case 2: $y$ = area of prefrontal subdivision in CC | | | | | | |
| Estimate | 4.384 | 4.939E-03 | 1.638E-02 | 2.086E-02 | −3.675E-02 | 1.667E-02 |
| Std. Error | 1.308E-02 | 1.527E-02 | 1.678E-02 | 8.827E-03 | 1.575E-02 | 8.132E-03 |
| t-statistic | 3.351E+02 | 3.235E-01 | 9.759E-01 | 2.363 | −2.333 | 2.050 |
| p-value | <1E-05 | 7.464E-01 | 3.295E-01 | 1.844E-02 | 1.995E-02 | 4.079E-02 |
| Case 3: $y$ = area of frontal subdivision in CC | | | | | | |
| Estimate | 5.371 | 3.229E-03 | 9.294E-03 | 1.452E-02 | −2.640E-02 | 1.518E-02 |
| Std. Error | 1.316E-02 | 1.536E-02 | 1.688E-02 | 8.880E-03 | 1.584E-02 | 8.181E-03 |
| t-statistic | 4.081E+02 | 2.103E-01 | 5.505E-01 | 1.635 | −2.666 | 1.856 |
| p-value | <1E-05 | 8.335E-01 | 5.822E-01 | 1.026E-01 | 3.835E-03 | 6.396E-02 |
| Case 4: $y$ = area of parietal subdivision in CC | | | | | | |
| Estimate | 5.190 | −1.924E-03 | 3.560E-03 | 3.050E-02 | −2.813E-02 | 2.079E-02 |
| Std. Error | 1.245E-02 | 1.453E-02 | 1.597E-02 | 8.401E-03 | 1.499E-02 | 7.740E-03 |
| t-statistic | 4.168E+02 | −1.324E-01 | 2.229E-01 | 3.631 | −2.876 | 2.686 |
| p-value | <1E-05 | 8.947E-01 | 8.237E-01 | 3.061E-04 | 2.012E-03 | 7.430E-03 |
| Case 5: $y$ = area of occipito-temporal subdivision in CC | | | | | | |
| Estimate | 4.929 | −1.532E-03 | 5.180E-03 | 2.964E-02 | −2.809E-02 | 2.032E-02 |
| Std. Error | 1.237E-02 | 1.444E-02 | 1.587E-02 | 8.348E-03 | 1.489E-02 | 7.691E-03 |
| t-statistic | 3.984E+02 | −1.061E-01 | 3.264E-01 | 3.550 | −2.886 | 2.642 |

*J Am Stat Assoc*. Author manuscript; available in PMC 2016 May 07.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

| Parameter | $b_0$ | $b_1$ | $b_2$ | $b_3$ | $b_4$ | $b_5$ |
|---|---|---|---|---|---|---|
| p-value | <1E-05 | 9.156E-01 | 7.443E-01 | 4.137E-04 | 1.950E-03 | 8.459E-03 |

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 7**

ADHD-200 data analysis: stability analysis of clustering results by leave-one-out method.

| | Cluster ID removed | | | |
|---|---|---|---|---|
| | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** |
| *M●* | 3 | 3 | 3 | 2 |
| RI | 0.8213 | 0.8827 | 0.8789 | 0.7163 |
| aRI | 0.7164 | 0.8091 | 0.8002 | 0.5035 |

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 8**

ADHD-200 data analysis: geodesic distance between each pair of shapes.

| Distance | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---|---|---|---|---|
| Cluster 1 | 0.0801 | 0.0824 | 0.0843 | 0.1168 |
| Cluster 2 | - | 0.0418 | 0.0812 | 0.1295 |
| Cluster 3 | - | - | 0.0510 | 0.1496 |
| Cluster 4 | - | - | - | 0.0989 |

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 9**

DHD-200 data analysis: mean shape difference test statistics and their associated *p*–values in the parentheses among different clusters.

| | **Cluster 2** | **Cluster 3** | **Cluster 4** |
|---|---|---|---|
| Cluster 1 | 0.0552 (0.3284) | 0.0881 (0.1343) | 0.6431 (0.0050) |
| Cluster 2 | - | 0.0044 (0.5808) | 0.0474 (0.0040) |
| Cluster 3 | - | - | 0.0218 (0.0349) |