IISE Transactions

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uiie21

A novel hierarchically-structured factor mixture model for cluster discovery from multi-modality data

Bing Si , Todd J. Schwedt , Catherine D. Chong , Teresa Wu & Jing Li

To cite this article: Bing Si , Todd J. Schwedt , Catherine D. Chong , Teresa Wu & Jing Li (2020): A novel hierarchically-structured factor mixture model for cluster discovery from multi-modality data, IISE Transactions, DOI: 10.1080/24725854.2020.1800149

To link to this article: https://doi.org/10.1080/24725854.2020.1800149

Published online: 30 Sep 2020.

IISE TRANSACTIONS
https://doi.org/10.1080/24725854.2020.1800149

A novel hierarchically-structured factor mixture model for cluster discovery from multi-modality data

Bing Si[a], Todd J. Schwedt[b], Catherine D. Chong[b], Teresa Wu[a], and Jing Li[a]

[a]Industrial Engineering, Arizona State University, Tempe, AZ, USA; [b]Department of Neurology, Mayo Clinic Arizona, Scottsdale, AZ, USA

ABSTRACT

Advances in sensing technology have generated multi-modality datasets with complementary information in various domains. In health care, it is common to acquire images of different types/modalities for the same patient to facilitate clinical decision making. We propose a clustering method called hierarchically-structured Factor Mixture Model (hierFMM) that enables cluster discovery from multi-modality datasets to exploit their joint strength. HierFMM employs a novel double-L_{21}-penalized likelihood formulation to achieve hierarchical selection of modalities and features that are nested within the modalities. This formulation is proven to satisfy a Quadratic Majorization condition that allows for an efficient Group-wise Majorization Descent algorithm to be developed for model estimation. Simulation studies show significantly better performance of hierFMM than competing methods. HierFMM is applied to an application of identifying clusters/subgroups of migraine patients based on brain cortical area, thickness, and volume datasets extracted from Magnetic Resonance Imaging. Two subgroups are found, whose patients significantly differ in clinical characteristics. This finding shows the promise of using multi-modality imaging data to help patient stratification and develop optimal treatment for different subgroups with migraine.

ARTICLE HISTORY
Received 19 January 2019
Accepted 8 July 2020

KEYWORDS
Factor model; clustering; sparse learning; health care

1. Introduction

Clustering analysis is a classic research area in statistical modeling and machine learning. Conventional approaches have focused on data from a single modality. Multi-modality datasets are becoming increasingly common in various domains. These datasets contain complementary information that potentially leads to better cluster identification. For example, in health care, neuroimaging datasets of different kinds have been used to characterize the brain from complementary aspects to facilitate clinical decision making for neurological diseases. One important clinical decision is to identify subgroups of patients so that treatment can be optimized for each subgroup. This task, by nature, needs a clustering method for multi-modality datasets.

There are multifold challenges in developing a clustering method for multi-modality datasets:

1. There can be quite a few modalities of data used in a study, and some of them may not contribute to the differentiation of clusters. These non-informative modalities should be automatically selected out, so that they will not mask the underlying cluster structure. For instance, in the aforementioned neuroimaging example, there could be a variety of different imaging modalities for characterizing the brain, such as cortical thickness, cortical area, cortical volume, white matter integrity, functional connectivity, metabolism, etc. When focusing on a particular type of disease, it is likely that cluster structure only exists within a subset of the imaging modalities.
2. Within each modality, it is commonplace that the modality includes many features and some of the features may not contribute to the differentiation of clusters. These non-informative features should be automatically selected out. Using the neuroimaging example again, because features are extracted from the image of the whole brain, the number of features is typically large. However, when focusing on a particular type of disease, it is likely that not all the features are relevant for differentiation of clusters.

To address these challenges, we propose a novel clustering method called hierarchically-structured Factor Mixture Model (hierFMM) that enables an automatic, hierarchical selection of modalities and features. A unique characteristic of multi-modality datasets is that features are nested within modalities. To accommodate this unique data structure, hierFMM selects features in a hierarchical way. That is, if a modality does not contribute to the differentiation of clusters, hierFMM will unselect all the features included in this modality, i.e., eliminate this modality in its entirety. For each remaining modality, hierFMM will select features within the modality that contribute to cluster differentiation. The contributions of this research are summarized as follows:

CONTACT Jing Li jinglz@asu.edu
Bing Si is now at the Department of Systems Science & Industrial Engineering, State University of New York at Binghamton, Binghamton, NY, USA.
Jing Li is now at H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Georgia, GA, USA.

1. Contribution to methodological development: HierFMM intersects with two research areas in statistical modeling and machine learning: sparse learning and factor models. However, no existing methods in these areas aim to cluster multi-modality datasets with hierarchical modality and feature selection. HierFMM uses a novel double-L_{21}-penalized likelihood formulation to achieve hierarchical modality and feature selection. Furthermore, we integrate the Expectation-Maximization (EM) framework and an efficient Group-wise Majorization Descent (GMD) algorithm for parameter estimation of hierFMM.
2. Contribution to the application domain: We apply hierFMM to identify clusters/subgroups of migraine patients using their brain cortical area, thickness, and volume datasets extracted from structural Magnetic Resonance Imaging (MRI). HierFMM finds two well-separated clusters, indicating that subjects in the clusters have distinct imaging phenotypes. The imaging features selected by hierFMM are valid because they have been reported in previous migraine studies. Interestingly, we find that the two clusters also significantly differ in terms of clinical characteristics, with one cluster having more allodynia symptoms during migraine attacks, more migraine-related disability, and a greater number of years with migraine. In essence, this study contributes to understanding of migraine heterogeneity from an imaging perspective and shows promise of using multi-modality imaging data to help stratify patients toward more refined and ultimately personalized management of migraine.

The rest of this article is organized as follows: Section 2 provides a literature review. Section 3 presents the development of hierFMM. Section 4 shows simulation experiments. Section 5 presents the application case study. Conclusions are drawn in Section 6.

2. Literature review

The proposed method mainly intersects with two research areas: sparse learning and factor models. Next, we will review each area and point out gaps that stress the need for new methodological development.

Sparse learning (SL): LASSO-type models are among the best-known SL models, which design different penalty functions to enable variable selection from a high-dimensional feature set (Tibshirani, 1996). Hierarchical variable selection has been investigated by some researchers. For example, Yuan *et al.* (2009) proposed a non-negative garrote method to incorporate the hierarchical or structural relationship among predictors by imposing corresponding constraints on coefficients. Jenatton *et al.* (2011) extended LASSO and group LASSO by allowing the subgroups of parameters to overlap and introduced structured sparsity-inducing norms. Zhao *et al.* (2009) proposed a composite absolute penalty for grouped and hierarchical variable selection based on l_r-norms. A few other variations of group LASSO have been discussed in the context of hierarchical sparse modeling to implement desired hierarchical sparsity relations among parameters (Yan and Bien, 2017). However, all these models are supervised learning methods.

SL has also been integrated with unsupervised learning, primarily with model-based clustering methods. The basic idea is to maximize the likelihood function of a mixture distribution subject to a sparsity-inducing penalty on the features. Pan and Shen (2007) proposed to impose a L_1-penalty on the cluster-wise mean vectors. Xie *et al.* (2008) proposed to penalize cluster-wise variances together with the means. Raftery and Dean (2006) proposed a stepwise Model-Based Clustering (sMBC) method, which recast feature selection as a model selection problem by comparing models containing nested subsets of features. Bouveyron *et al.* (2007) proposed a High Dimensional Data Clustering (HDDC) method with a new parametrization of the Gaussian mixture model that limited the number of parameters to estimate while providing flexibility in modeling the data. However, all these methods target "flat" features, i.e., there is no hierarchy of the features organized in different modalities.

Less research has been performed on clustering with multi-modality data. Khalidov *et al.* (2011) proposed conjugate mixture models for identifying objects in a scene (e.g., human subjects participating in a meeting) by fusing auditory and visual sensing data. The word "clustering" was used in their paper to refer to finding the grouping of features corresponding to each object. Multi-modality sensor fusion for object identification is a popular research area, but its objective is different from ours or what a typical clustering algorithm aims to achieve — grouping the subjects. Another related research area is two-level clustering. For example, Wang *et al.* (2015) proposed to have feature selection using sparse coding nested within the optimization that minimizes the loss function of clustering, which was referred to as bi-level. Wang (2019) extended this framework to integrate with deep learning for improving scalability and efficiency. However, these frameworks are only applicable to flat features. In summary, these related research areas have a different objective from ours that focuses on clustering by coupling features in different modalities with feature and modality selection capabilities.

Factor models: For clustering of high-dimensional datasets, one approach is to adopt a sparsity-inducing penalty for feature selection, an alternative approach assumes that the observed high-dimensional features lie on a low-dimensional latent space, which is the idea of factor models. Factor models are more appropriate in applications where the features typically embrace a complex correlation structure, suggesting the existence of latent factors. For example, in the migraine application presented in Section 5 of this article, the observed features in each modality correspond to 64 anatomically-defined Regions of Interest of the brain. These features are naturally correlated as a result of the spatial proximity and/or functional similarity of the regions.

Factor mixture models (FMM) is an extension of the classic factor analysis (FA). FA assumes that the sample observations are from a single distribution. In contrast,

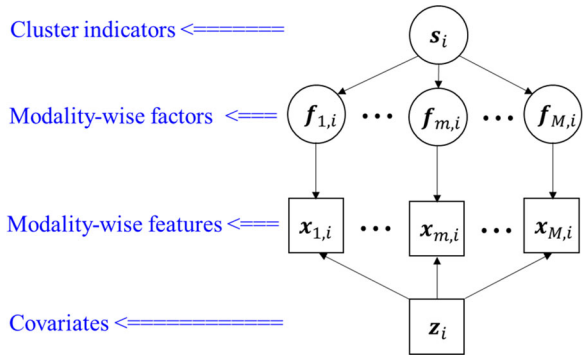

Figure 1. Graphical representation of the hierFMM model (circles represent latent variable; rectangles represent observed variables).

FMM assumes that the factors are distributed as a mixture model, and thus being a clustering approach. Different FMM models have been developed based on different assumptions on the mixture distribution (Muthén and Asparouhov, 2006; Lubke and Muthén, 2005; Baek *et al.*, 2010; Montanari and Viroli, 2010): Some assume different means for the components in the mixture distribution; some additionally assume that the covariance matrices for the components are also different.

However, the existing FMM models are essentially a single-modality approach, i.e., they force all features to share the same latent factors even when the features are indeed from distinct modalities. As different modalities usually have different physical meanings, it is not valid to assume that their respective features have the same underlying constructs (i.e., the latent factors). Also, blindly using FMM on multi-modality data destroys the inherent data structure within each modality, which may lead to poor clustering performance. Additionally, there is a lack of capacity for hierarchical modality and feature selection.

3. Development of HierFMM

3.1. Preliminaries

Since the proposed hierFMM model is an extension of the FMM model, we first introduce the FMM model. Suppose there are N subjects to be clustered. For each subject i, let $\boldsymbol{x}_i$ denote a P-dimensional feature vector, $i = 1, ..., N,$. In FMM, the observed feature vector $\boldsymbol{x}_i$ is first linked with R-dimensional latent factors, $\boldsymbol{f}_i$, $R \ll P$:

$$\boldsymbol{x}_i = \mathbf{H}\boldsymbol{f}_i + \boldsymbol{\varepsilon}_i,$$

where $\mathbf{H}$ is a $P \times R$ coefficient/loading matrix and $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Psi})$ contains random errors. Furthermore, $\boldsymbol{f}_i$ is linked with cluster indicator variables $\boldsymbol{s}_i$ by

$$\boldsymbol{f}_i = \mathbf{A}\boldsymbol{s}_i + \boldsymbol{\xi}_i,$$

where $\boldsymbol{s}_i = (s_{1,i}, \ldots, s_{K,i})^T$. $s_{k,i} = 1$ if subject i belongs to the kth cluster and 0 otherwise, $k = 1, ..., K$. K is the number of cluster. $\boldsymbol{s}_i$ is assumed to follow a multinomial distribution with parameters $\boldsymbol{w} = (w_1, ..., w_K)^T$ that correspond to the probabilities of different clusters. $\mathbf{A}$ is a $R \times K$ loading matrix and $\boldsymbol{\xi}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ contains random errors.

3.2. HierFMM model formulation

In hierFMM, we consider that the features are nested within M different modalities. To accommodate this change, we use $\boldsymbol{x}_{m,i}$ to denote a P_m-dimensional feature vector within the mth modality, $m = 1, ..., M$. Also, the two major equations in FMM are changed to

$$\boldsymbol{x}_{m,i} = \mathbf{H}_m\boldsymbol{f}_{m,i} + \mathbf{B}_m\boldsymbol{z}_i + \boldsymbol{\varepsilon}_{m,i}, \text{ and} \tag{1}$$

$$\boldsymbol{f}_{m,i} = \mathbf{A}_m\boldsymbol{s}_i + \boldsymbol{\xi}_{m,i}. \tag{2}$$

In Equation (1), we consider that the features in each modality, $\boldsymbol{x}_{m,i}$, are linked with R_m-dimensional latent factors specific to that modality, $\boldsymbol{f}_{m,i}$. Also, we add L covariates in $\boldsymbol{z}_i$ to make the model more general. In Equation (2), we consider that the modality-wise factor vectors, $\boldsymbol{f}_{m,i}, m = 1, ..., M$, share the same clustering structure. Figure 1 provides a graphical illustration of this model. $\mathbf{H}_m$, $\mathbf{B}_m$, and $\mathbf{A}_m$ are loading matrices of size $P_m \times R_m$, $P_m \times L$, $R_m \times K$, respectively. $\boldsymbol{\varepsilon}_{m,i}$ and $\boldsymbol{\xi}_{m,i}$ follow zero-mean Gaussian distributions with covariance matrices $\boldsymbol{\Psi}_m$ and $\boldsymbol{\Sigma}_m$ of size $P_m \times P_m$ and $R_m \times R_m$, respectively.

Next, we discuss how to estimate the parameters in the proposed hierFMM. Put all the parameters in a set $\boldsymbol{\Theta}$, i.e., $\boldsymbol{\Theta} = \{\{\boldsymbol{\Theta}_m\}_{m=1}^M, \boldsymbol{w}\}$, where $\boldsymbol{\Theta}_m = \{\mathbf{H}_m, \mathbf{B}_m, \mathbf{A}_m, \boldsymbol{\Psi}_m, \boldsymbol{\Sigma}_m\}$. We can write the complete log-likelihood function as:

$$l(\boldsymbol{\Theta}) = \sum_{i=1}^{N} \left\{ \sum_{m=1}^{M} \log\left(f\left(\boldsymbol{x}_{m,i} | \boldsymbol{f}_{m,i}, \boldsymbol{z}_i; \boldsymbol{\Theta}\right)\right) + \sum_{m=1}^{M} \log\left(f\left(\boldsymbol{f}_{m,\ i} | \boldsymbol{s}_i; \boldsymbol{\Theta}\right)\right) + \log\left(f(\boldsymbol{s}_i; \boldsymbol{\Theta})\right) \right\}. \tag{3}$$

As Equation (3) involves latent variables, we could use an EM algorithm to estimate the parameters. However, a regular EM does not consider that some modalities or some features within a modality may not contribute to the differentiation of clusters. To see this more clearly, let us insert Equation (2) into Equation (1) and obtain the distribution of $\boldsymbol{x}_{m,i}$ in the kth cluster:

$$\boldsymbol{x}_{m,i} | s_{k,i} = 1 \sim N\left(\mathbf{H}_m\mathbf{a}_{m,k} + \mathbf{B}_m\boldsymbol{z}_i,\ \mathbf{H}_m\boldsymbol{\Sigma}_m\mathbf{H}_m^T + \boldsymbol{\Psi}_m\right), \tag{4}$$

where $\mathbf{a}_{m,k}$ is the kth column of $\mathbf{A}_m$. If $\mathbf{A}_m = 0$, then Equation (4) becomes

$$\boldsymbol{x}_{m,i} | s_{k,i} = 1 \sim N\left(\mathbf{B}_m\boldsymbol{z}_i,\ \mathbf{H}_m\boldsymbol{\Sigma}_m\mathbf{H}_m^T + \boldsymbol{\Psi}_m\right). \tag{5}$$

It is easy to note that the distribution in Equation (5) does not include cluster membership "k". This means that the mth modality (including all the features in this modality) does not have a cluster structure, and thus does not contribute to cluster differentiation. Furthermore, suppose $\mathbf{A}_m \neq \mathbf{0}$ but $\mathbf{h}_m^j = \mathbf{0}$, where $\mathbf{h}_m^j$ is the jth row of $\mathbf{H}_m$ corresponding to the jth feature. Then, the distribution of the jth feature in the mth modality, $x_{m,i}^j$, is

$$x_{m,i}^j | s_k = 1 \sim N\left(\left(\mathbf{b}_m^j\right)^T\boldsymbol{z}_i, \psi_m^{jj}\right), \tag{6}$$

where $\mathbf{b}_m^j$ is the jth row of $\mathbf{B}_m$ corresponding to the jth feature and ψ_m^{jj} is the jth diagonal element of $\boldsymbol{\Psi}_m$. Equation (6)

indicates that the jth feature in the mth modality, i.e., $x^j_{m,i}$, does not contribute to cluster differentiation.

The challenge, however, is that we do not know which $\mathbf{A}_m$, $m = 1, \ldots, M$, and which $\mathbf{h}^j_m$, $j = 1, \ldots, P_m$, are zero because they are part of the parameter set to be estimated. To automatically zero out the $\mathbf{A}_m$'s and $\mathbf{h}^j_m$'s that do not contribute to the differentiation of clusters, we propose to add two L_{21}-penalties to the complete log-likelihood function in Equation (3), which results in the following optimization problem:

$$\min_{\mathbf{\Theta}} g(\mathbf{\Theta}) \triangleq \min_{\mathbf{\Theta}} \left\{ -l(\mathbf{\Theta}) + \lambda_1 \sum_{m=1}^{M} \sum_{j=1}^{P_m} \|\mathbf{h}^j_m\|_2 + \lambda_2 \sum_{m=1}^{M} \|\mathbf{A}_m\|_2 \right\}. \tag{7}$$

$\|\cdot\|_2$ is the L_2-norm of a vector or matrix. λ_1 and λ_2 are penalty parameters. It is well-known that an L_{21}-penalty can zero out all the coefficients within the $\|\cdot\|_2$ as a group. Due to this property, the optimization in Equation (7) can eliminate features not contributing to cluster differentiation in a *hierarchical* manner. That is, $\sum_{m=1}^{M} \|\mathbf{A}_m\|_2$ helps eliminate modalities that do not contribute to cluster differentiation. Furthermore, $\sum_{m=1}^{M} \sum_{j=1}^{P_m} \|\mathbf{h}^j_m\|_2$ helps eliminate features not contributing to cluster differentiation within a modality.

A final note regarding the optimization in Equation (7) is how to address the identifiability issue of the parameters. This issue can be seen from Equation (1), where the observed features $\boldsymbol{x}_{m,i}$ is linked with a product of $\mathbf{H}_m$ and $\boldsymbol{f}_{m,i}$, both being unobserved. Therefore, $\mathbf{H}_m$ and $\boldsymbol{f}_{m,i}$ are not uniquely identifiable given $\boldsymbol{x}_{m,i}$. The identifiability issue also exists in classic FMM. The strategy for ensuring model identifiability in FMM is to add constraints on the mean values and covariance matrix of the latent factors. Following a similar idea, we add mean and covariance constraints to the latent factors of each modality in our model:

$E(\boldsymbol{f}_{m,i}) = 0$, and $Cov(\boldsymbol{f}_{m,i}) = \mathbf{I}$, where $\mathbf{I}$ is an identity matrix of an appropriate size. The mean constraint can be easily satisfied by standardizing $\boldsymbol{x}_{m,i}$. The covariance constraint can be satisfied using a mathematical trick as follows: Assume the original covariance, $Cov(\boldsymbol{f}^*_{m,i})$, is not an identity matrix and let $\mathbf{H}^*_m$ be the original loading matrix. Using Cholesky Decomposition we can decompose $Cov(\boldsymbol{f}^*_{m,i}) = \boldsymbol{L}_{m,i}\boldsymbol{L}^T_{m,i}$ and update the loading matrix by $\mathbf{H}_m = \mathbf{H}^*_m \boldsymbol{L}_{m,i}$. As $Cov(\mathbf{H}^*_m \boldsymbol{f}^*_{m,i}) = \mathbf{H}^*_m\ Cov(\boldsymbol{f}^*_{m,i})\ \mathbf{H}^*_m T = \mathbf{H}^*_m\ \boldsymbol{L}_{m,i}\boldsymbol{L}^T_{m,i}\ \mathbf{H}^*_m T =$ $\mathbf{H}_m\mathbf{H}^T_m$, the updated $\mathbf{H}_m$ automatically sets the covariance of the latent factors to be an identify matrix, i.e., the covariance constraint is satisfied.

3.3. HierFMM model estimation by integrating EM and an efficient GMD algorithm

3.3.1. The EM framework

As hierFMM involves latent variables, we adopt the EM framework for model estimation. Let $\{\mathbf{X}_m\}_{m=1}^M$ be the observed data, i.e., the features of M modalities. Let $\{\mathbf{F}_m\}_{m=1}^M$ and $\mathbf{S}$ be the missing data, i.e., the latent factors and cluster indicators, respectively. In the E-step, we will need to derive the expectation of the $g(\mathbf{\Theta})$ in Equation (7) with respect to the conditional distribution of $\{\mathbf{F}_m\}_{m=1}^M, \mathbf{S}$ given $\{\mathbf{X}_m\}_{m=1}^M$ and the current estimate $\tilde{\mathbf{\Theta}}$, i.e., $E_{\{\mathbf{F}_m\}_{m=1}^M, \mathbf{S}\ |\{\mathbf{X}_m\}_{m=1}^M; \tilde{\mathbf{\Theta}}}\ (g(\mathbf{\Theta}))$. We find that this expectation can be written into a sum of three functions that involve non-overlapping parameters in $\mathbf{\Theta}$, that is

$$E_{\{\mathbf{F}_m\}_{m=1}^M, \mathbf{S}|\{\mathbf{X}_m\}_{m=1}^M; \tilde{\mathbf{\Theta}}}\ (g(\mathbf{\Theta})) \triangleq \varphi\left(\{\mathbf{H}_m, \mathbf{B}_m, \mathbf{\Psi}_m\}_{m=1}^M\right) + \varphi\left(\{\mathbf{A}_m,\ \mathbf{\Sigma}_m\}_{m=1}^M\right) + \varphi(\boldsymbol{w}),$$

where the $\varphi(\cdot)'s$ take the following forms:

$$\begin{aligned}
\varphi\left(\{\mathbf{H}_m, \mathbf{B}_m, \mathbf{\Psi}_m\}_{m=1}^M\right) &= \sum_{i=1}^{N}\sum_{m=1}^{M} E_{f_{m,i}|x_{m,i};\tilde{\mathbf{\Theta}}}\left(-\log f\left(\boldsymbol{x}_{m,i}|\boldsymbol{f}_{m,i}, \boldsymbol{z}_i; \{\mathbf{H}_m, \mathbf{B}_m, \mathbf{\Psi}_m\}_{m=1}^M\right)\right) + \lambda_1 \sum_{m=1}^{M}\sum_{j=1}^{P_m} \|\mathbf{h}^j_m\|_2, \\
\varphi\left(\{\mathbf{A}_m, \mathbf{\Sigma}_m\}_{m=1}^M\right) &= \sum_{i=1}^{N}\sum_{m=1}^{M} E_{f_{m,i}, s_i|x_{1,i}, \ldots, x_{M,i};\tilde{\mathbf{\Theta}}}\left(-\log f\left(\boldsymbol{f}_{m,i}|\boldsymbol{s}_i; \{\mathbf{A}_m, \mathbf{\Sigma}_m\}_{m=1}^M\right)\right) + \lambda_2 \sum_{m=1}^{M} \|\mathbf{A}_m\|_2, \\
\varphi(\boldsymbol{w}) &= \sum_{i=1}^{N} E_{s_i|x_{1,i}, \ldots, x_{M,i};\tilde{\mathbf{\Theta}}}\left(-\log f(\boldsymbol{s}_i; \boldsymbol{w})\right).
\end{aligned} \tag{8}$$

Due to this property of the expectation, the optimization in the M-step becomes solving three separate optimization problems corresponding to the three $\varphi(\cdot)'s$ in the expectation, respectively, that is

$$\mathbf{\Theta}^* = \left\{ \begin{array}{l} \underset{\{\mathbf{H}_m, \mathbf{B}_m, \mathbf{\Psi}_m\}_{m=1}^M}{\operatorname{argmin}}\ \varphi\left(\{\mathbf{H}_m, \mathbf{B}_m, \mathbf{\Psi}_m\}_{m=1}^M\right), \\ \underset{\{\mathbf{A}_m, \mathbf{\Sigma}_m\}_{m=1}^M}{\operatorname{argmin}}\ \varphi\left(\{\mathbf{A}_m,\ \mathbf{\Sigma}_m\}_{m=1}^M\right),\ \ \underset{\boldsymbol{w}}{\operatorname{argmin}}\ \varphi(\boldsymbol{w}) \end{array} \right\}. \tag{9}$$

Solving the three optimizations yields the following solutions:

$$\begin{aligned}
\mathbf{H}^*_m, \mathbf{B}^*_m = &\underset{\mathbf{H}_m, \mathbf{B}_m}{\operatorname{argmin}} \sum_{i=1}^{N} E_{f_{m,i}|x_{m,i};\tilde{\mathbf{\Theta}}}\left(-\log f\left(\boldsymbol{x}_{m,i}|\boldsymbol{f}_{m,i}, \boldsymbol{z}_i; \{\mathbf{H}_m, \mathbf{B}_m\}\right)\right) \\
&+ \lambda_1 \sum_{j=1}^{P_m} \|\mathbf{h}^j_m\|_2,
\end{aligned} \tag{10}$$

$$\left\{\mathbf{A}_m^*\right\}_{m=1}^M = \underset{\{\mathbf{A}_m\}_{m=1}^M}{\operatorname{argmin}} \sum_{i=1}^N \sum_{m=1}^M E_{f_{m,i},\boldsymbol{s_i}|\boldsymbol{x}_{1,i},\ldots,\boldsymbol{x}_{M,i};\tilde{\boldsymbol{\Theta}}} \left(-\log f\left(\boldsymbol{f}_{m,i}|\boldsymbol{s}_i;\{\mathbf{A}_m\}_{m=1}^M\right)\right) + \lambda_2 \sum_{m=1}^M \|\mathbf{A}_m\|_2, \tag{11}$$

$$w_k^* = \frac{\sum_{i=1}^N f\left(s_{k,i}=1|\boldsymbol{x}_{1,i},\ldots,\boldsymbol{x}_{M,i};\tilde{\boldsymbol{\Theta}}\right)}{\sum_{i=1}^N \sum_{k=1}^K f\left(s_{k,i}=1|\boldsymbol{x}_{1,i},\ldots,\boldsymbol{x}_{M,i};\tilde{\boldsymbol{\Theta}}\right)}, k=1,\ldots,K,$$

$$\boldsymbol{\Psi}_m^* = diag\left(\frac{1}{N}\left(\sum_{i=1}^N \left(\boldsymbol{x}_{m,i}-\mathbf{B}_m^*\boldsymbol{z}_i\right)\left(\boldsymbol{x}_{m,i}-\mathbf{B}_m^*\boldsymbol{z}_i\right)^T - \left(\sum_{i=1}^N \left(\boldsymbol{x}_{m,i}-\mathbf{B}_m^*\boldsymbol{z}_i\right) E\left(\left(\boldsymbol{f}_{m,i}\right)^T|\boldsymbol{x}_{m,i};\tilde{\boldsymbol{\Theta}}\right)\right)\left(\mathbf{H}_m^*\right)^T\right)\right),$$

$$\boldsymbol{\Sigma}_m^* = \frac{\sum_{k=1}^K \sum_{i=1}^N f\left(s_{k,i}=1|\boldsymbol{x}_{1,i},\ldots,\boldsymbol{x}_{M,i};\tilde{\boldsymbol{\Theta}}\right)\left(E\left(\boldsymbol{f}_{m,i}\left(\boldsymbol{f}_{m,i}\right)^T\boldsymbol{x}_{1,i},\ldots,\boldsymbol{x}_{M,i},s_{k,i}=1;\tilde{\boldsymbol{\Theta}}\right)-\mathbf{a}_{m,k}^*\left(\mathbf{a}_{m,k}^*\right)^T\right)}{\sum_{i=1}^N f\left(s_{k,i}=1|\boldsymbol{x}_{1,i},\ldots,\boldsymbol{x}_{M,i};\tilde{\boldsymbol{\Theta}}\right)},$$

for $m=1,\ldots,M$. Except for (10) and (11), all other parameters can be estimated analytically. Therefore, the key to speeding up the EM iterations is to develop an efficient algorithm to solve the optimization problems in (10) and (11).

3.3.2. The GMD algorithm

The original GMD algorithm was proposed by Yang and Zou (2015), as an efficient algorithm to solve L_{21}-penalized optimization problems. An L_{21}-penalized optimization can be written in a general form as:

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, L(\boldsymbol{\beta}|\mathbf{D}) + \lambda \sum_{j=1}^J \|\boldsymbol{\beta}^{(j)}\|_2, \tag{12}$$

where $\mathbf{D}$ denotes a dataset; $\boldsymbol{\beta}$ contains p-dimensional parameters to be estimated, which can be partitioned into J groups, $\boldsymbol{\beta}^{(1)},\ldots,\boldsymbol{\beta}^{(J)}$. $L(\boldsymbol{\beta}|\mathbf{D})$ is an empirical loss function. Compared with classic algorithms such as block-wise descent (Yuan and Lin, 2006), block coordinate gradient descent (Meier *et al.*, 2008), and Nesterov's method (Liu *et al.*, 2009), GMD is 5–10 times faster. To apply GMD, the optimization problem must satisfy a Quadratic Majorization (QM) condition.

Definition 1: QM condition (Yang and Zou, 2015): The L_{21}-penalized optimization in (12) satisfies the QM condition if and only if the following two assumptions hold:

(i) $L(\boldsymbol{\beta}|\mathbf{D})$ is differentiable as a function of $\boldsymbol{\beta}$, i.e., $\nabla L(\boldsymbol{\beta}|\mathbf{D})$ exists everywhere.

(ii) There exists a $p\times p$ matrix $\boldsymbol{\Lambda}$, which may only depend on the data $\mathbf{D}$, such that for all $\boldsymbol{\beta}$, $\boldsymbol{\beta}^*$,

$$L(\boldsymbol{\beta}|\mathbf{D}) \le L(\boldsymbol{\beta}^*|\mathbf{D}) + (\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\nabla L(\boldsymbol{\beta}^*|\mathbf{D}) + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\boldsymbol{\Lambda}(\boldsymbol{\beta}-\boldsymbol{\beta}). \tag{13}$$

Given that the QM condition is satisfied, the optimization in Equation (12) can be solved using the GMD algorithm in the following way: At step-$(\omega+1)$ of the algorithm, one wants to update the jth group in $\boldsymbol{\beta}^{(\omega)}$ while keeping the other groups unchanged, that is

$$\boldsymbol{\beta}^{(\omega+1)} - \boldsymbol{\beta}^{(\omega)} = \left(0,\ \ldots 0,\ \underbrace{\left(\boldsymbol{\beta}^{(j)(\omega+1)} - \boldsymbol{\beta}^{(j)(\omega)}\right)^T}_{j-\text{th group}},\ 0,\ldots 0\right)^T.$$

According to the QM condition in (ii), one can get following inequality:

$$L\left(\boldsymbol{\beta}^{(\omega+1)}|\mathbf{D}\right) \le L\left(\boldsymbol{\beta}^{(\omega)}|\mathbf{D}\right) + \left(\boldsymbol{\beta}^{(j)(\omega+1)} - \boldsymbol{\beta}^{(j)(\omega)}\right)^T\nabla L^{(j)} + \frac{1}{2}\left(\boldsymbol{\beta}^{(j)(\omega+1)} - \boldsymbol{\beta}^{(j)(\omega)}\right)^T\boldsymbol{\Lambda}^{(j)}\left(\boldsymbol{\beta}^{(j)(\omega+1)} - \boldsymbol{\beta}^{(j)(\omega)}\right), \tag{14}$$

where $\nabla L^{(j)}$ and $\boldsymbol{\Lambda}^{(j)}$ are sub-matrices of $\nabla L\left(\boldsymbol{\beta}^{(\omega)}|\mathbf{D}\right)$ and $\boldsymbol{\Lambda}$ only including the rows and columns corresponding to the jth group. Furthermore, let τ_j be the largest eigenvalue of $\boldsymbol{\Lambda}^{(j)}$ and set $\rho_j = (1+10^{-6})\tau_j$. Then, Equation (14) can be further relaxed as

$$L\left(\boldsymbol{\beta}^{(\omega+1)}|\mathbf{D}\right) \le L\left(\boldsymbol{\beta}^{(\omega)}|\mathbf{D}\right) + \left(\boldsymbol{\beta}^{(j)(\omega+1)} - \boldsymbol{\beta}^{(j)(\omega)}\right)^T\nabla L^{(j)} + \frac{1}{2}\rho_j\left(\boldsymbol{\beta}^{(j)(\omega+1)} - \boldsymbol{\beta}^{(j)(\omega)}\right)^T\left(\boldsymbol{\beta}^{(j)(\omega+1)} - \boldsymbol{\beta}^{(j)(\omega)}\right), \tag{15}$$

where the inequality strictly holds unless $\boldsymbol{\beta}^{(j)(\omega+1)} = \boldsymbol{\beta}^{(j)(\omega)}$. Using Equation (15), one can solve the optimization in (12) by solving:

$$\underset{\boldsymbol{\beta}^{(j)(\omega+1)}}{\operatorname{argmin}}\ L\left(\boldsymbol{\beta}^{(\omega)}|\mathbf{D}\right) + \left(\boldsymbol{\beta}^{(j)(\omega+1)} - \boldsymbol{\beta}^{(j)(\omega)}\right)^T\nabla L^{(j)} + \frac{1}{2}\rho_j\left(\boldsymbol{\beta}^{(j)(\omega+1)} - \boldsymbol{\beta}^{(j)(\omega)}\right)^T\left(\boldsymbol{\beta}^{(j)(\omega+1)} - \boldsymbol{\beta}^{(j)(\omega)}\right) + \lambda\|\boldsymbol{\beta}^{(j)(\omega)}\|_2, \tag{16}$$

which has an analytical solution:

$$\boldsymbol{\beta}^{(j)*(\omega+1)} = \frac{1}{\rho_j}\left(-\nabla L^{(j)} + \rho_j\boldsymbol{\beta}^{(j)(\omega)}\right)\left(1 - \frac{\lambda}{\|-\nabla L^{(j)} + \rho_j\boldsymbol{\beta}^{(j)(\omega)}\|_2}\right)_+. \tag{17}$$

Due to the analytic form in Equation (17), GMD is computationally efficient.

We propose to use GMD to solve the optimizations in Equations (10) and (11) of our hierFMM model. To use GMD, we first prove that these optimizations satisfy the QM condition.

Proposition 1: *The optimization problems in (10) and (11) of the hierFMM model satisfy the QM condition* (please see the proof in Appendix B).

Given Proposition 1, we can further customize the general GMD algorithm to solve Equations (10) and (11). The key is to write Equations (10) and (11) into the general form of Equation (12). First compare Equation (10) with Equation (12): if we make $\mathbf{h}_m^j = \boldsymbol{\beta}^{(j)}$ and $\{\boldsymbol{x}_{m,i}\}_{i=1}^N = \mathbf{D}$, Equation (10) becomes Equation (12). Next compare Equation (11) with Equation (12): if we make $\mathbf{A}_m = \boldsymbol{\beta}^{(j)}$, with a slight adjustment that the summation in the L_{21}-norm is over m instead of j, and $\{\boldsymbol{x}_{1,i}, ..., \boldsymbol{x}_{M,i}\}_{i=1}^N = \mathbf{D}$, then Equation (11) becomes Equation (12). After such transformations, the optimizations in Equations (10) and (11) can be solved by GMD. Finally, we can show that the GMD algorithms used to solve Equations (10) and (11) are guaranteed to converge to a local optimal solution.

Proposition 2: *The GMD algorithms for solving the optimizations in (10) and (11) are guaranteed to converge* (please see the proof in Appendix C).

3.3.3. Model selection

The numbers of clusters, the number of factors, and the penalty parameters, i.e., λ_1 and λ_2, can be selected according to a model selection criterion that balances the model fit and complexity. The former is measured by the log-likelihood of the observed data, $\tilde{l}$. The latter is given by the degree of freedom, df, that counts the number of non-zeros in the estimated parameters. There are various criteria to combine the fit and complexity in the literature such as Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC). For example, BIC takes the form of $-2\tilde{l} + \log(N) \times df$. A common approach is to exhaustively search for all the combinations of tuning parameters on a four-dimensional (4-D) grid of parameters (number of clusters, number of factors, λ_1, λ_2) and choose the set of parameters that minimizes the BIC criterion. Note that the first two tuning parameters can take only integer values, resulting in a limited number of options to be tuned. The main computational cost is on the search for the last two parameters, i.e., λ_1 and λ_2. To ease the computational burden, parallelized programs can be adopted for model fitting based on different combinations of the tuning parameters. In our experiments in Sections 4 and 5, we adopt the aforementioned 4-D grid search strategy using parallel computing resources and find that the computational time is acceptable. More efficient computational algorithms based on optimization heuristics such as genetic algorithms and simulated annealing can be used to further improve the computational efficiency. Finally, once a model that minimizes the BIC has been identified, each sample i is classified to a cluster for which the posterior probability of belonging to that cluster, i.e., $P(s_{k,i} = 1 | \boldsymbol{x}_{1,i}, ..., \boldsymbol{x}_{M,i}; \boldsymbol{\Theta}^*)$, is maximized.

Table 1. Factor means of each cluster within each modality.

	Clusters	
Modalities	$k = 1$	$k = 2$
$m = 1$	$\mathbf{a}_{1,1} = (1, 1)^T$	$\mathbf{a}_{1,2} = (-1, -1)^T$
$m = 2$	$\mathbf{a}_{2,1} = (0.75, 0.75)^T$	$\mathbf{a}_{2,2} = (-0.75, -0.75)^T$
$m = 3$	$\mathbf{a}_{3,1} = (0, 0)^T$	$\mathbf{a}_{3,2} = (0, 0)^T$
$m = 4$	$\mathbf{a}_{4,1} = (0, 0)^T$	$\mathbf{a}_{4,2} = (0, 0)^T$

4. Simulation studies

4.1. Simulation setup

We generate simulation data to assess the performance of hierFMM. Consider 120 samples each belonging to one of two clusters. The prior probability of each cluster is $w_1 = 0.4$ and $w_2 = 0.6$. Furthermore, suppose there are four modalities with two factors in each modality. To demonstrate the capability of hierFMM for modality selection, we assume that the first two modalities have a two-cluster structure whereas the other two modalities do not, thus they do not contribute to cluster differentiation. To accomplish this, we set the cluster-wise factor means of each modality according to Table 1. The factor means differ between two clusters in modalities 1 and 2, but not in modalities 3 and 4. We set the covariance matrices of the four modalities according to Equation (18), which do not differ between two clusters:

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}_3 = \mathbf{I}, \quad \text{and } \boldsymbol{\Sigma}_4 = \mathbf{I}. \tag{18}$$

Also, we consider three covariates that are sampled from $N(0, 1)$. Once the data for the factors and covariates are generated from the aforementioned distributions, we proceed to generate the features using Equation (1). The number of features is set to be 40 in each modality. In order to demonstrate the capability of hierFMM for feature selection, we assume that 10 features have a two-cluster structure whereas the other 30 do not. To accomplish this, we set the coefficient matrices $\mathbf{H}_m$ in Equation (1) to be

$$\mathbf{H}_m = \begin{pmatrix} \tilde{\mathbf{H}}_{10\times 2} \\ \mathbf{0}_{30\times 2} \end{pmatrix},$$

with each element of $\tilde{\mathbf{H}}$ generated as follows: Generate a number from *Uniform* $[1, 1.5]$, which is used as the magnitude of the element. To decide the sign of the element, generate another number from $N(0, 1)$. If this number is greater than 0.5, create a positive sign; otherwise, create a negative sign. Finally, we sample the random errors $\boldsymbol{\varepsilon}_m$ in Equation (1) from $N(\mathbf{0}, 0.1 \times \mathbf{I})$.

4.2. HierFMM application and results

HierFMM is applied to the simulation data. The experiment is repeated 100 times, which allows us to compute the average performance of the model. We compare two model selection criteria, BIC and AIC, for selecting the number of clusters, the number of factors, and the penalty parameters. Among the 100 repetitions, BIC selects the correct modalities for 82 times, i.e., BIC has a modality selection accuracy of 82%, whereas AIC has an 98% accuracy.

Furthermore, we evaluate the feature selection accuracy of the model. Since our model is hierarchical, the accuracy of feature selection should consider errors made at both the modality level and the feature level. Thus, the sensitivity is defined as the percentage of selected features among the features that truly differentiate the clusters. According to our simulation setup, modalities 1 and 2 are cluster-specific and within each of the two modalities, the first 10 features contribute to the cluster differentiation. These are the features used in the denominator of our sensitivity calculation. Furthermore, the specificity is defined as the percentage of unselected features among the features that do not contribute to cluster differentiation (i.e., features in modalities 3 and 4, as well as the last 30 features in modalities 1 and 2).

Table 2 shows the average and standard deviation of sensitivity and specificity under the BIC and AIC criteria. In general, the average sensitivity and specificity are fairly high. BIC has better specificity, whereas AIC has better sensitivity. This is consistent with the literature that BIC is known to select a sparser (less complex) model. The standard deviation of the sensitivity, especially under BIC is high because of the hierarchical nature of feature selection, i.e., if a cluster-informative modality is not selected by our model, then all the features in that modality are considered missed, which results in low sensitivity in some of the experimental runs. On the other hand, the fact that the average sensitivity over all the experimental runs is fairly good means that a large majority of the runs does not miss any informative modality. This is also backed up by the good modality selection accuracy of 82% reported earlier.

Table 2. Sensitivity and specificity of feature selection by hierFMM (average ± standard deviation over all experiments).

Model selection criterion	*Sensitivity* (%)	*Specificity* (%)
BIC	82± 38.6	95.8±4.6
AIC	98 ± 14.1	85.1 ± 9.5

4.3. Comparison with competing methods

HierFMM is a model-based clustering method. Within the category of model-based clustering methods, one sub-category of methods performs clustering on the observed features whereas the other sub-category assumes observed features are generated from latent factors that follow a mixture distribution. The latter sub-category consists of FMM models. We choose a relatively recent FMM model (Montanari and Viroli, 2010) to compare with hierFMM. Furthermore, we include two models from the first sub-category with a sparsity consideration to allow for feature selection, i.e., sMBC (Raftery and Dean, 2006) and HDDC (Bouveyron *et al.*, 2007).

Since the three competing methods can only be used to cluster data of a single modality, we adopt two strategies to make them work for multi-modality datasets, so that they can be compared with hierFMM. One strategy is to apply a competing method on one modality at a time; the other strategy is to apply the competing method on pooled features from all the modalities. As the ultimate performance metric for a clustering algorithm is its accuracy in discovering the true clustering structure, we compare the clustering accuracy of the competing methods with hierFMM.

Figure 2 shows the clustering accuracy of each method, defined as the percentage of samples correctly classified to their ground-truth clusters. For the first strategy of applying the competing methods, we only show the results of modalities 1 and 2, as the accuracies of modalities 3 and 4 (i.e., the modalities not contributing to cluster differentiation) are poor. HierFMM achieves an accuracy of 89.5%±12.2%,

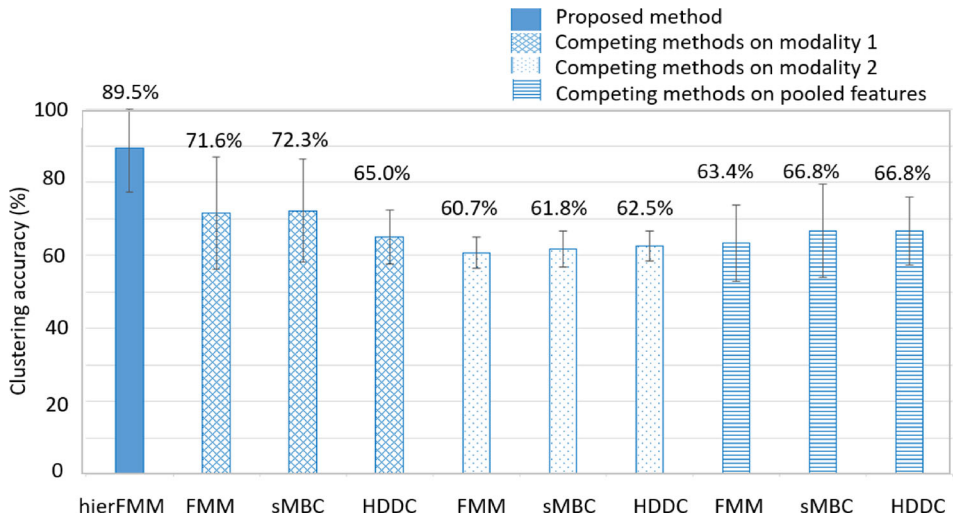

Figure 2. Clustering accuracy of hierFMM in comparison with three competing methods under two strategies (i.e., applying a competing method on a single modality (modality 1 and 2 individually) and applying a competing method on pooled features from all the modalities).

which is significantly higher than all three competing methods under the two strategies.

5. Application in migraine patient clustering using multi-modality imaging data

5.1. Data collection and image processing

The data used for this application were obtained from Mayo Clinic Arizona through our clinical collaborators. A total of 120 subjects were included in this analysis. Migraine was diagnosed in accordance with the diagnostic criteria defined by the International Classification of Headache Disorders. Data collected from all subjects included demographics such as age and sex, and clinical characteristics and symptoms measured through a number of instruments such as Beck Depression Inventory (BDI), State-Trait Anxiety Inventory (STAI), Allodynia Symptom Checklist 12 (ASC-12), Migraine Disability Assessment (MIDAS), Hyperacusis Questionnaire, Photosensitivity Assessment Questionnaire, together with a few individually measured key symptom variables such as headache frequency, number of years with migraine, and aura status.

Structural MRI data were obtained from two Siemens 3T MRI machines. Details of the MRI acquisition were described in prior publications (Schwedt *et al.,* 2015a, b; Schwedt *et al.,* 2017). Using a cortical reconstruction and segmentation program in the FreeSurfer image analysis suite (version 5.3, http://www.surfer.nmr.mgh.harvard.edu/), cortical area, thickness, and volume measurements of 68 Regions of Interest (ROIs) were extracted. In our study, cortical area, cortical thickness, and volume are treated as three modalities. Within each modality, 34 features correspond to ROIs at the right brain hemisphere, while the other 34 correspond to same-name ROIs at the left hemisphere. Our analysis found no difference in the clustering structure between using 68 features in each modality and using 34 by averaging the features at the left and right hemispheres corresponding to the same-name ROI. Therefore, we will only present the result for the latter situation in this article.

5.2. Data augmentation with nuisance modalities and features

A challenge in applying any clustering method to real data is that the ground-truth clustering structure is unknown. This prohibits rigorous performance assessment for the clustering result using accuracy metrics, as can be done in a simulation study. On the other hand, this is what makes a clustering method appealing, because it can lead to new discovery to extend the boundary of the existing knowledge in a domain. In this article, we design our study in a way that allows for performance assessment. Specifically, we add artificially generated nuisance modalities and features to the real modalities and features, apply hierFMM to the combined data, and examine if hierFMM is able to identify the real modalities and features. In a prior study of ours, Schwedt *et al.* (2017) applied a non-penalized version of hierFMM to the same dataset and found all three modalities and 34 features within each modality to be relevant to a two-cluster structure. This result was validated with the medical knowledge of our clinical collaborators and the existing literature on migraine studies. Therefore, the three modalities are treated as real modalities and 34 features as real features in the present study.

To add nuisance modalities and features, we employ the following steps: First, we add 68 nuisance features $\tilde{\boldsymbol{x}}_m$ to each real modality, which are generated by

$$\tilde{\boldsymbol{x}}_m = \mathbf{0}_{68\times 1}\boldsymbol{f}_m + \tilde{\mathbf{B}}_m\boldsymbol{z} + \tilde{\boldsymbol{\varepsilon}}_m, \quad m = 1, 2, 3. \tag{19}$$

Equation (19) takes the same form as Equation (1) with $\mathbf{H}_m = \mathbf{0}_{68\times 1}$, as nuisance features are not supposed to have any clustering structure. To make the distribution of nuisance data as close as possible to the real data, we do not give the coefficient matrix $\tilde{\mathbf{B}}_m$ arbitrarily, but rather sample each row of $\tilde{\mathbf{B}}_m$ with replacement from the rows of $\mathbf{B}_m$. Although the true $\mathbf{B}_m$ is unknown, we have a reliable estimate $\widehat{\mathbf{B}}_m$ from a previous study that applied a non-penalized version of hierFMM to the real dataset (Schwedt *et al.,* 2017). Similarly, we sample $\tilde{\boldsymbol{\varepsilon}}_m$ from $N(\mathbf{0},\ \widehat{\boldsymbol{\Psi}}_m)$, where $\widehat{\boldsymbol{\Psi}}_m$ is from the previous study. Furthermore, we add three nuisance modalities each having $34 + 68 = 102$ features that match the size of augmented real modalities. The features in a nuisance modality are generated by

$$\tilde{\tilde{\boldsymbol{x}}}_m = \tilde{\tilde{\mathbf{H}}}_m\tilde{\tilde{\boldsymbol{f}}}_m + \tilde{\tilde{\mathbf{B}}}_m\boldsymbol{z} + \tilde{\tilde{\boldsymbol{\varepsilon}}}_m, \quad m = 1, 2, 3, \tag{20}$$

where each row of $\tilde{\tilde{\mathbf{B}}}_m$ is sampled with replacement from the rows of $\widehat{\mathbf{B}}_m$. Each of the first 34 rows of $\tilde{\tilde{\mathbf{H}}}_m$ are sampled from the rows of $\widehat{\mathbf{H}}_m$ while the remaining 68 rows are all zeros. $\tilde{\tilde{\boldsymbol{\varepsilon}}}_m$ is sampled from $N(\mathbf{0},\ \widehat{\boldsymbol{\Psi}}_m)$. We sample $\tilde{\tilde{\boldsymbol{f}}}_m$ from $N(0,\ 1)$ because nuisance modalities are not supposed to have any clustering structure. Through this procedure, we create an augmented dataset of six modalities (three real and three nuisance modalities) and 102 features within each modality (34 real features in each real modality).

5.3. Results from the application of HierFMM

We apply hierFMM to the augmented data together with two patient-specific covariates, sex and age. All three real modalities are correctly selected, resulting in 100% accuracy. Within the two real modalities, the sensitivity and specificity of selecting out the real features is 100% and 96.6%, respectively. HierFMM found two clusters among the 120 subjects, each consisting of 53 and 67 subjects, respectively. Call these clusters A and B hereafter. A total of seven factors (two, three and two, from cortical area, cortical thickness, and volume modalities, respectively) are found to differentiate clusters A and B. The correspondence between these factors and the imaging features is encoded in the estimated loading matrix $\widehat{\mathbf{H}}_m$ and is shown in Figure 3. A clear pattern is that the within each modality, loadings that reflect the contributions to the imaging features from one factor (i.e., bars of one color) are different from another factor (i.e., bars of

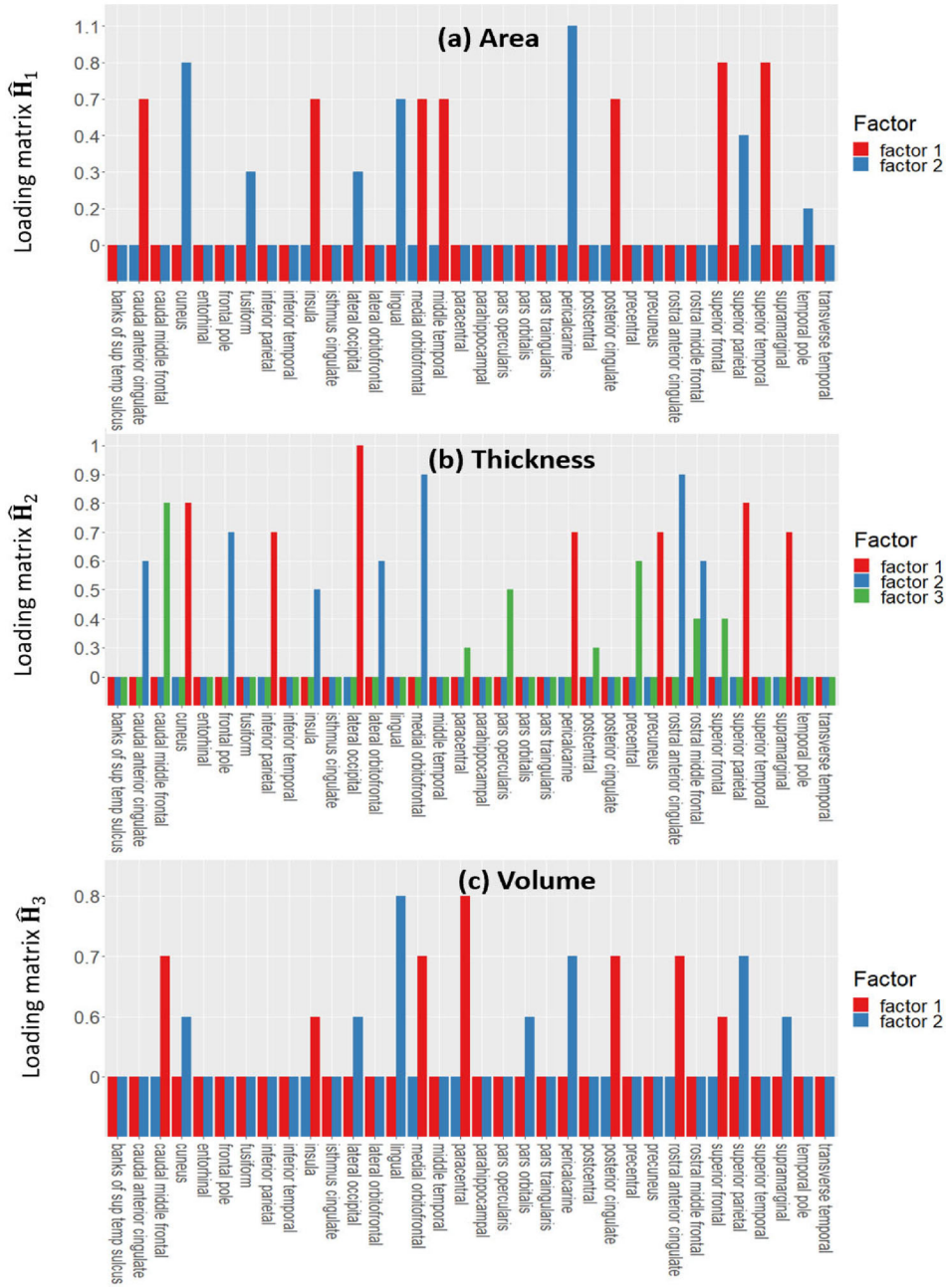

Figure 3. Estimated loadings (*y*-axis) that show the contribution of factors to original features (*x*-axis) for: (a) area, (b) thickness, and (c) volume. Loadings whose magnitudes are less than the 80th percentile of all loadings are suppressed and represented by short bars for better visualization.

another color). This indicates that there may be more than one biological underpinning underlying the observed imaging features, and thus supporting the validity of multiple factors found in each modality. Furthermore, we highlight the ROIs in each modality whose measurements mostly contribute to differentiation of the two clusters in Figure 4. These ROIs are those whose loading magnitudes are greater than the 80th percentile of all the loadings estimated by hierFMM.

Next, we would like to see how well subjects belonging to clusters A and B are separated in terms of the seven factors. Since it is impossible to visualize the separation on a seven-dimensional space, we choose to visualize it modality by modality. Figure 5(a)-(c) plots the subjects in terms of the two, three, and two factors within area, thickness, and volume modalities, respectively. Figure 5(d) plots the posterior probability of each subject belonging to cluster B, which reflects the joint effect of the seven factors in separating subjects belonging to the two clusters. These results demonstrate that all the factors in each modality and all three modalities contribute to the cluster separation. Also, the two clusters are separated very well, as the vast majority of the subjects in each cluster have a high posterior probability of being in the cluster they are assigned to, as shown in Figure 5(d).

Finally, we would like to see how the two clusters/subgroups of patients derived from MRI differ in clinical characteristics and symptoms. We focus on a panel of variables including the number of headache days per month, number of years with migraine, aura status, MIDAS score, STAI

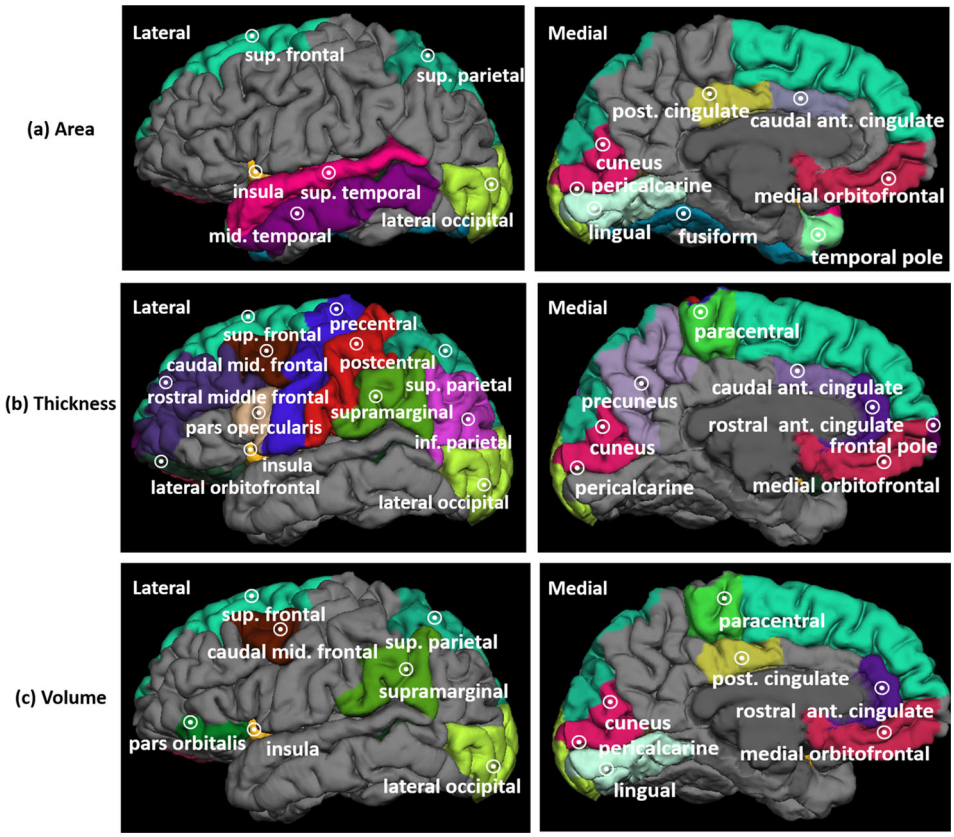

Figure 4. The ROIs whose (a) area, (b) thickness, or (c) volume measurements mostly contribute to differentiation of the two clusters are color-highlighted on a 3-D rending of the brain.

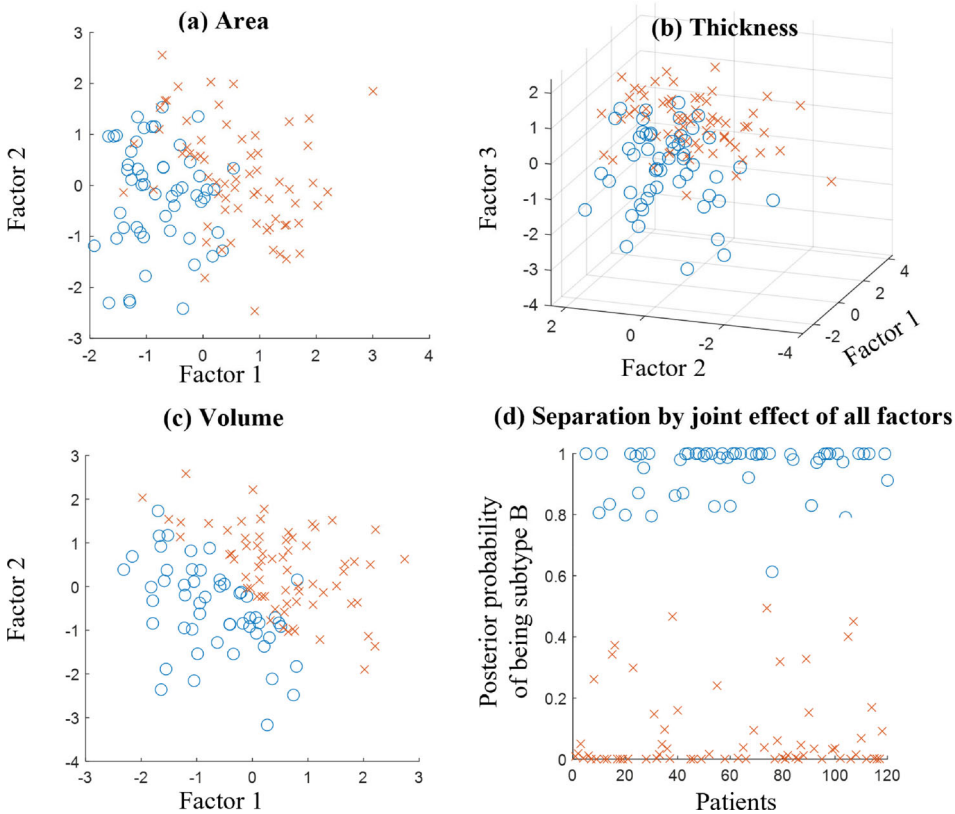

Figure 5. Separation of subjects belonging to cluster A (red cross) and B (blue circle) in terms of the factors in each modality and the posterior probability of cluster membership.

score, BDI score, allodynia during and between migraine attacks, hyperacusis, and photophobia. We perform hypothesis testing to compare the patient subgroups/clusters A and B in terms of each variable. Three variables are found to have statistically significant between-cluster difference: migraine subjects belonging to cluster A have a greater number of years with migraine (p-value = 0.01), more migraine-related disability as measured by the MIDAS score (p-value = 0.04), and greater symptoms of allodynia during migraine attacks (p-value = 0.03).

5.4. Discussion on medical implications

The main finding of this application was identification of two clusters/subgroups of patients within the study cohort based on structural MRI measurements of brain cortical area, cortical thickness, and volume. The two clusters significantly differ in a number of clinical characteristics including the number of years with migraine, allodynia during migraine attacks, and migraine-related disability. These clinical variables have been previously reported to relate to brain imaging findings in migraine. For example, a number of studies have shown that the number of years with migraine is associated with brain structure and function (Chong and Schwedt, 2015; Chong *et al.*, 2017). In general, the longer a person has had migraine and the more attacks they have had, the greater the brain differences. Allodynia symptom severity was measured using the ASC-12, a questionnaire that collects information about cutaneous allodynia – the sensation of pain to normally non-noxious stimulation of the skin (Lipton *et al.,* 2008). Several imaging studies have demonstrated associations between brain structure and function with symptoms of allodynia (Moulton *et al.*, 2008; Chong *et al.*, 2017). Disability could be a marker for the severity of migraine symptoms as well as the person's ability to cope with their migraine symptoms (Ford *et al.*, 2008). Migraine severity and coping mechanisms could both associate with measures of brain structure and function. The structural measurements that differentiated the two clusters in this study were of brain regions that have previously been shown to be aberrant in migraine and/or in individuals who have allodynia (Schwedt *et al.,* 2015a, b; Russo *et al.*, 2017; Schwedt *et al.*, 2017).

In essence, this study shed some light on the potential of using multi-modality imaging data to stratify migraine patients into subgroups/clusters, which can help with delineating patient heterogeneity and optimizing treatment for each subgroup.

6. Conclusion

In this article, we proposed a new method, hierFMM, for clustering multi-modality datasets. HierFMM employed a double-L_{21}-penalized likelihood formulation to enable hierarchical modality and feature selection. We developed an efficient GMD algorithm embedded in the EM framework to estimate the model parameters. We performed simulation experiments to compare hierFMM with competing methods. HierFMM performed significantly better in terms of modality selection accuracy, feature selection accuracy, and clustering accuracy. We applied hierFMM to identify subgroups of migraine patients based on brain cortical area, thickness, and volume measurements extracted from MRI. Two clusters were found and well separated using a total of seven factors. Subjects in the two clusters had significant different clinical characteristics. There are a number of extensions for the current study, such as extending hierFMM to include mixed-type features, and applying the model to other application areas with multi-modality high-dimensional datasets.

Funding

This work was partially supported by NIH K23NS070891, NSF CAREER 1149602, and NSF DMS 1903135.

Notes on contributors

Bing Si, PhD, is an assistant professor in the Department of Systems Science & Industrial Engineering at State University of New York at Binghamton. She received her PhD in industrial engineering from Arizona State University. Her research focuses on developing data analytics and statistical learning methodologies to support health care decisions in diagnosis, prognosis, treatment, and care delivery. She is a member of IISE, INFORMS, and IEEE.

Todd J. Schwedt, MD, is a professor of neurology at Mayo Clinic Arizona. His research investigates the mechanisms, classification, and treatment of migraine, post-traumatic headache, and other headaches. A core research goal is to use advanced magnetic resonance imaging (MRI) techniques to identify biomarkers that will help with the diagnosis and treatment of headache. He has published over 90 manuscripts, lectures nationally and internationally, and serves on the Board of Directors for the American Headache Society and the Board of Trustees for the International Headache Society.

Catherine D. Chong, PhD, is an assistant professor in the Department of Neurology at Mayo Clinic Arizona. She completed her PhD in neuroscience at the University of Utah in 2002. Her research interests focus on using structural and functional neuroimaging techniques for delineating the neuropathology associated with migraine.

Teresa Wu, PhD, is professor in industrial engineering at Arizona State University. She received her PhD from the University of Iowa in 2007. Her research interests are health informatics and distributed decision supports. She is a recipient of an NSF CAREER award. She is a member of IISE and INFORMS.

Jing Li, PhD, is professor in industrial engineering at Arizona State University. She received her PhD in industrial and operations engineering from the University of Michigan in 2007. Her research interests are statistical modeling and machine learning for health care applications. She is a recipient of an NSF CAREER award. She is a member of IISE, INFORMS, and IEEE.

References

Baek, J., Mclachlan, G.J. and Flack, L.K. (2010) Mixtures of factor analyzers with common factor loadings: Applications to the clustering and visualization of high-dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**(7), 1298–1309.

Bouveyron, C., Girard, S. and Schmid, C. (2007) High-dimensional data clustering. *Computational Statistics & Data Analysis*, **52**(1), 502–519.

Chong, C.D., Plasencia, J.D., Frakes, D.H. and Schwedt, T.J. (2017) Structural alterations of the brainstem in migraine. *NeuroImage: Clinical*, **13**, 223–227.
Chong, C.D. and Schwedt, T.J. (2015) Migraine affects white-matter tract integrity: A diffusion-tensor imaging study. *Cephalalgia*, **35**(13), 1162–1171.
Ford, S., Calhoun, A., Kahn, K., Mann, J. and Finkel, A. (2008) Predictors of disability in migraineurs referred to a tertiary clinic: Neck pain, headache characteristics, and coping behaviors. *Headache: The Journal of Head and Face Pain*, **48**(4), 523–528.
Jenatton, R., Audibert, J.Y. and Bach, F. (2011) Structured variable selection with sparsity-inducing norms. *The Journal of Machine Learning Research*, **12**(Oct), 2777–2824.
Khalidov, V., Forbes, F. and Horaud, R. (2011) Conjugate mixture models for clustering multimodal data. *Neural Computation*, **23**(2), 517–557.
Lipton, R.B., Bigal, M.E., Ashina, S., Burstein, R., Silberstein, S., Reed, M.L., Serrano, D., Stewart, W.F. and American Migraine Prevalence Prevention Advisory Group. (2008) Cutaneous allodynia in the migraine population. *Annals of Neurology*, **63**(2), 148–158.
Liu, J., Ji, S. and Ye, J. (2009) *SLEP: Sparse learning with efficient projections*. Arizona State University, **6**(491), 7.
Lubke, G.H. and Muthén, B. (2005) Investigating population heterogeneity with factor mixture models. *Psychological Methods*, **10**(1), 21–39.
Meier, L., Geer, S.V. and Bühlmann, P. (2008) The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(1), 53–71.
Montanari, A. and Viroli, C. (2010) Heteroscedastic factor mixture analysis. *Statistical Modelling: An International Journal*, **10**(4), 441–460.
Moulton, E.A., Burstein, R., Tully, S., Hargreaves, R., Becerra, L. and Borsook, D. (2008) Interictal dysfunction of a brainstem descending modulatory center in migraine patients. *PLoS ONE*, **3**(11), e3799.
Muthén, B. and Asparouhov, T. (2006) Item response mixture modeling: Application to tobacco dependence criteria. *Addictive Behaviors*, **31**(6), 1050–1066.
Pan, W. and Shen, X. (2007) Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, **8**, 1145–1164.
Raftery, A.E. and Dean, N. (2006) Variable selection for model-based clustering. *Journal of the American Statistical Association*, **101**(473), 168–178.
Russo, A., Esposito, F., Conte, F., Fratello, M., Caiazzo, G., Marcuccio, L., Giordano, A., Tedeschi, G. and Tessitore, A. (2017) Functional interictal changes of pain processing in migraine with ictal cutaneous allodynia. *Cephalalgia*, **37**(4), 305–314.
Schwedt, T.J., Chiang, C., Chong, C.D. and Dodick, D.W. (2015a) Functional MRI of migraine. *The Lancet Neurology*, **14**(1), 81–91.
Schwedt, T.J., Chong, C.D., Wu, T., Gaw, N., Fu, Y. and Li, J. (2015b) Accurate classification of chronic migraine via brain magnetic resonance imaging. *Headache: The Journal of Head and Face Pain*, **55**(6), 762–777.
Schwedt, T.J., Si, B., Li, J., Wu, T. and Chong, C.D. (2017) Migraine subclassification via a data-driven automated approach using multimodality factor mixture modeling of brain structure measurements. *Headache: The Journal of Head and Face Pain*, **57**(7), 1051–1064.
Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**(1), 267–288.
Wang, Z. (2019) From bi-level sparse clustering to deep clustering, in *Deep Learning Through Sparse and Low-Rank Modeling*, Academic Press, Cambridge, Massachusetts, pp. 87–119.
Wang, Z., Yang, Y., Chang, S., Li, J., Fong, S. and Huang, T.S. (2015) A joint optimization framework of sparse coding and discriminative clustering, in *Proceedings of the Twenty-fourth International Joint Conference on Artificial Intelligence*.
Xie, B., Pan, W. and Shen, X. (2008) Variable selection in penalized model-based clustering via regularization on grouped parameters. *Biometrics*, **64**(3), 921–930.
Yan, X. and Bien, J. (2017) Hierarchical sparse modeling: A choice of two group lasso formulations. *Statistical Science*, **32**(4), 531–560.
Yang, Y. and Zou, H. (2015) A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, **25**(6), 1129–1141.
Yuan, M., Joseph, V.R. and Zou, H. (2009) Structured variable selection and estimation. *The Annals of Applied Statistics*, **3**(4), 1738–1757.
Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**(1), 49–67.
Zhao, P., Rocha, G. and Yu, B. (2009) The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, **37**(6A), 3468–3497.

Appendices

Appendix A: Deriving the expectations in Equation (6)

(I.1) Deriving $E_{f_{m,i},s_i|x_{1,i},\ldots,x_{M,i};\tilde{\Theta}}\left(-\log f\left(f_{m,i}|s_i;\Theta\right)\right)$.

According to Equation (3), the distribution of $f_{m,i}|s_{k,i}=1;\Theta$ is $N(\mathbf{a}_{m,k},\ \Sigma_m)$. Inserting the probability density function of this distribution into the above expectation and ignoring constants, we can get

$$\sum_{k=1}^{K}\left\{\begin{array}{c}\frac{1}{2}\log|\Sigma_m|+\\ \frac{1}{2}\left(E\left(f_{m,i}|x_{m,i},s_{k,i}=1;\tilde{\Theta}\right)-\mathbf{a}_{m,k}\right)^T\\ \Sigma_m^{-1}\left(E\left(f_{m,i}|x_{m,i},s_{k,i}=1;\tilde{\Theta}\right)-\mathbf{a}_{m,k}\right)+\\ \frac{1}{2}tr\left(\Sigma_m^{-1}E\left(f_{m,}f_{m.i}^T|x_{m,i},s_{k,i}=1;\tilde{\Theta}\right)\right)\end{array}\right\}f\left(s_{k,i}=1|x_{1,i},\ldots,x_{M,i,};\tilde{\Theta}\right). \tag{S-1}$$

The distribution of $f_{m,i}|x_{m,i},s_{k,i}=1;\tilde{\Theta}$ is $N(\tilde{\rho}_{mk}(x_{m,i}),\tilde{\Upsilon}_{mk})$, where $\tilde{\rho}_{mk}(x_{m,i})=\tilde{\Upsilon}_{mk}(\tilde{\mathbf{H}}_m^T\tilde{\Psi}_m^{-1}x_{m,i}+\tilde{\Sigma}_m^{-1}\tilde{\mathbf{a}}_{m,k})$ and $\tilde{\Upsilon}_{mk}=(\tilde{\mathbf{H}}_m^T\tilde{\Psi}_m^{-1}\tilde{\mathbf{H}}_m+\tilde{\Sigma}_m^{-1})^{-1}$. Therefore,

$$E\left(f_{m,i}|x_{m,i},\ s_{k,i}=1;\tilde{\Theta}\right)=\tilde{\rho}_{mk}(x_{m,i})$$

and

$$E\left(f_{m,i}f_{m.i}^T|x_{m,i},\ s_{k,i}=1;\tilde{\Theta}\right)=\tilde{\Upsilon}_{mk}-\tilde{\rho}_{mk}(x_{m,i})^T\tilde{\rho}_{mk}(x_{m,i})$$

in (S-1). Finally, to derive the $f\left(s_{k,i}=1|x_{1,i},\ldots,x_{M,i,};\tilde{\Theta}\right)$ in (S-1), we can use the Bayes' theorem and get

$$f\left(s_{k,i}=1|x_{1,i},\ldots,x_{M,i,};\tilde{\Theta}\right)=\frac{\tilde{w}_k\prod_{m=1}^{M}f\left(x_{m,i}|s_{k,i}=1;\tilde{\Theta}\right)}{\sum_{k=1}^{K}\tilde{w}_k\prod_{m=1}^{M}f\left(x_{m,i}|s_{k,i}=1;\tilde{\Theta}\right)},$$

where $x_{m,i}|s_{k,i}=1;\tilde{\Theta}\sim N\left(\tilde{\mathbf{H}}_m\tilde{\mathbf{a}}_{m,k}+\tilde{\mathbf{B}}_m z_i,\tilde{\mathbf{H}}_m\tilde{\Sigma}_m\tilde{\mathbf{H}}_m^T+\tilde{\Psi}_m\right)$.

(I.2) Deriving $E_{f_{m,i}|x_{m,i};\tilde{\Theta}}\left(-\log f\left(x_{m,i}|f_{m,i},z_i;\Theta\right)\right)$.

According to Equation (1), the distribution of $x_{m,i}|f_{m,i},z_i;\Theta$ is $N\left(\mathbf{H}_m f_{m,i}+\mathbf{B}_m z_i,\Psi_m\right)$. Inserting the probability density function of this distribution into the above expectation and ignoring constants, we can get

$$\begin{array}{l}\frac{1}{2}\log|\Psi_m|\\ +\frac{1}{2}\left(x_{m,i}-\mathbf{H}_m E\left(f_{m,i}|x_{m,i};\tilde{\Theta}\right)-\mathbf{B}_m z_i\right)^T\\ \Psi_m^{-1}\left(x_{m,i}-\mathbf{H}_m E\left(f_{m,i}|x_{m,i};\tilde{\Theta}\right)-\mathbf{B}_m z_i\right)\\ +\frac{1}{2}tr\left(\mathbf{H}_m^T\Psi_m^{-1}\mathbf{H}_m\left(E\left(f_{m,}f_{m.i}^T x_{m,i};\tilde{\Theta}\right)\right.\right.\\ \left.\left.-E\left(f_{m,i}x_{m,i};\tilde{\Theta}\right)E\left(f_{m,i}x_{m,i};\tilde{\Theta}\right)^T\right)\right)\end{array} \tag{S-2}$$

Using the result in (I.1), we can get

$$E\left(\boldsymbol{f}_{m,i}\boldsymbol{x}_{m,i};\tilde{\boldsymbol{\Theta}}\right)=\sum_{k=1}^{K}\tilde{\boldsymbol{\rho}}_{mk}(\boldsymbol{x}_{m,i})f\left(s_{k,i}=1|\boldsymbol{x}_{m,i};\tilde{\boldsymbol{\Theta}}\right),$$

and

$$E\left(\boldsymbol{f}_{m,i}\boldsymbol{f}_{m,i}^{T}\boldsymbol{x}_{m,i};\tilde{\boldsymbol{\Theta}}\right)=\sum_{k=1}^{K}\left(\tilde{\Upsilon}_{mk}-\tilde{\boldsymbol{\rho}}_{mk}(\boldsymbol{x}_{m,i})^{T}\tilde{\boldsymbol{\rho}}_{mk}(\boldsymbol{x}_{m,i})\right)f\left(s_{k,i}=1|\boldsymbol{x}_{m,i};\boldsymbol{\Theta}^{(\omega)}\right)$$

in (S-2).

(I.3) Deriving $E_{\boldsymbol{s}_i|\boldsymbol{x}_{1,i},\ldots,\boldsymbol{x}_{M,i};\tilde{\boldsymbol{\Theta}}}\left(-\log f(\boldsymbol{s}_i;\boldsymbol{\Theta})\right)$.

According to Equation (2), $\log f(\boldsymbol{s}_i;\boldsymbol{\Theta})=\sum_{k=1}^{K}s_{k,i}\log(w_k)$. Inserting it into the above expectation, we can get

$$E_{\boldsymbol{s}_i|\boldsymbol{x}_{1,i},\ldots,\boldsymbol{x}_{M,i};\tilde{\boldsymbol{\Theta}}}\left(-\log f(\boldsymbol{s}_i;\boldsymbol{\Theta})\right)=-\sum_{k=1}^{K}\log(w_k)f\left(s_{k,i}=1|\boldsymbol{x}_{1,i},\ldots,\boldsymbol{x}_{M,i};\tilde{\boldsymbol{\Theta}}\right).$$

□

Appendix B: Proof of Proposition 1

Here we show the proof that the optimization in Equation (10) satisfies the QM condition. The proof for Equation (11) follows similar ideas and is thus omitted. We first need to write the minimization problem in Equation (10) into the form of Equation (12). This can be achieved by making $J=P_m$, $\boldsymbol{\beta}^{(j)}=\mathbf{h}_m^j$, and

$$L(\boldsymbol{\beta}|\mathbf{D})=\sum_{i=1}^{N}E_{\boldsymbol{f}_{m,i}|\boldsymbol{x}_{m,i};\tilde{\boldsymbol{\Theta}}}\left(-\log f\left(\boldsymbol{x}_{m,i}|\boldsymbol{f}_{m,i},\boldsymbol{z}_i;\boldsymbol{\Theta}\right)\right).\tag{S-3}$$

Since the QM condition requires $L(\boldsymbol{\beta}|\mathbf{D})$ satisfying two assumptions, we will need to write $L(\boldsymbol{\beta}|\mathbf{D})$ into a format that facilitates checking of the assumptions. Through some derivation and dropping the terms not involving $\boldsymbol{\beta}$, we can write $L(\boldsymbol{\beta}|\mathbf{D})$ as

$$L(\boldsymbol{\beta}|\mathbf{D})=\boldsymbol{\beta}^{T}\left(\sum_{i=1}^{N}\mathbf{C}_{mi}\right)\boldsymbol{\beta}-2\left(\sum_{i=1}^{N}\boldsymbol{b}_{m,i}^{T}\right)\boldsymbol{\beta}\tag{S-4}$$

where

$$\begin{aligned}\mathbf{C}_{mi}=\frac{1}{2}tr\left(\boldsymbol{\Psi}_m^{-1}\right)\Bigg[&\left(\mathbf{1}_{P_M\times P_M}\otimes\left(E_{\boldsymbol{f}_{m,i}|\boldsymbol{x}_{m,i};\boldsymbol{\Theta}^{(\omega)}}\left(\boldsymbol{f}_{m,i}\right)\right)^{T}\right)^{T}\\&\left(\mathbf{1}_{P_M\times P_M}\otimes\left(E_{\boldsymbol{f}_{m,i}|\boldsymbol{x}_{m,i};\boldsymbol{\Theta}^{(\omega)}}\left(\boldsymbol{f}_{m,i}\right)\right)^{T}\right)\\&+\left(\mathbf{1}_{P_m}^{T}\otimes\mathbf{1}_{r_m}^{T}\right)\left(\mathbf{1}_{P_m}^{T}\otimes\mathbf{1}_{r_m}^{T}\right)^{T}tr\left(E\left(\boldsymbol{f}_{m,i}\left(\boldsymbol{f}_{m,i}\right)^{T}\boldsymbol{x}_{m,i};\boldsymbol{\Theta}^{(\omega)}\right)\right)\\&-E\left(\boldsymbol{f}_{m,i}\boldsymbol{x}_{m,i};\boldsymbol{\Theta}^{(\omega)}\right)E\left(\boldsymbol{f}_{m,i}\boldsymbol{x}_{m,i};\boldsymbol{\Theta}^{(\omega)}\right)^{T}\Bigg)\Bigg]\end{aligned}$$

and

$$\boldsymbol{b}_{m,i}^{T}=\frac{1}{2}tr\left(\boldsymbol{\Psi}_m^{-1}\right)\boldsymbol{x}_{m,i}^{T}\left(\mathbf{1}_{P_M\times P_M}\otimes\left(E_{\boldsymbol{f}_{m,i}|\boldsymbol{x}_{m,i};\boldsymbol{\Theta}^{(\omega)}}\left(\boldsymbol{f}_{m,i}\right)\right)^{T}\right).$$

Next, we will prove (S-4) satisfy the two assumptions required by the QM condition:

(i) It is straightforward to get $\nabla L(\boldsymbol{\beta}|\mathbf{D})=2(\sum_{i=1}^{N}\mathbf{C}_{mi}\boldsymbol{\beta}-\sum_{i=1}^{N}\boldsymbol{b}_{m,i}^{T})$, which exist everywhere.

(ii) To prove this assumption, we define a function $l(t)=L(\boldsymbol{\beta}^{*}+t(\boldsymbol{\beta}-\boldsymbol{\beta}^{*})|\mathbf{D})$. By the mean value theorem, there exists $a\in(0,1)$ such that:

$$l(1)=l(0)+l'(a)=l(0)+l'(0)+\left(l'(a)-l'(0)\right).\tag{S-5}$$

Using the $L(\boldsymbol{\beta}|\mathbf{D})$ in (S-4), we can get

$$l'(0)=2(\boldsymbol{\beta}-\boldsymbol{\beta}^{*})^{T}\left(\sum_{i=1}^{N}\mathbf{C}_{mi}\boldsymbol{\beta}^{*}-\sum_{i=1}^{N}\boldsymbol{b}_{m,i}^{T}\right)=(\boldsymbol{\beta}-\boldsymbol{\beta}^{*})^{T}\nabla L(\boldsymbol{\beta}^{*}|\mathbf{D}),\tag{S-6}$$

And

$$\begin{aligned}l'(a)-l'(0)&=2a(\boldsymbol{\beta}-\boldsymbol{\beta}^{*})^{T}\sum_{i=1}^{N}\mathbf{C}_{mi}(\boldsymbol{\beta}-\boldsymbol{\beta}^{*})\\&=\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}^{*})^{T}\left(4a\cdot\sum_{i=1}^{N}\mathbf{C}_{mi}\right)(\boldsymbol{\beta}-\boldsymbol{\beta}^{*})\\&\le(\boldsymbol{\beta}-\boldsymbol{\beta}^{*})^{T}\left(4\cdot\sum_{i=1}^{N}\mathbf{C}_{mi}\right)(\boldsymbol{\beta}-\boldsymbol{\beta}^{*}).\end{aligned}\tag{S-7}$$

Substituting (S-6) and (S-7) into (S-5), we have:

$$l(1)\le l(0)+(\boldsymbol{\beta}-\boldsymbol{\beta}^{*})^{T}\nabla L(\boldsymbol{\beta}^{*}|\mathbf{D})+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}^{*})^{T}\left(4\cdot\sum_{i=1}^{N}\mathbf{C}_{mi}\right)(\boldsymbol{\beta}-\boldsymbol{\beta}^{*}).$$

Noting that $l(1)=L(\boldsymbol{\beta}|\mathbf{D})$, $l(0)=L(\boldsymbol{\beta}^{*}|\mathbf{D})$, and let $\boldsymbol{\Lambda}=4\cdot\sum_{i=1}^{N}\mathbf{C}_{mi}$, we proved the second assumption of the QM condition.

Appendix C: Proof of Proposition 2

The convergence of the GMD used to solve the optimizations in our model follows from the convergence of the general GMD algorithm (Yang and Zou, 2015). First, we can add

$$\lambda\|\boldsymbol{\beta}^{(j)(\omega+1)}\|_2$$

to both sides of Equation (14), which give us the following inequality:

$$\begin{aligned}L\left(\boldsymbol{\beta}^{(\omega+1)}|\mathbf{D}\right)+\lambda\|\boldsymbol{\beta}^{(j)(\omega+1)}\|_2\le L\left(\boldsymbol{\beta}^{(\omega)}|\mathbf{D}\right)+\left(\boldsymbol{\beta}^{(j)(\omega+1)}-\boldsymbol{\beta}^{(j)(\omega)}\right)^{T}\nabla L^{(j)}\\+\frac{1}{2}\left(\boldsymbol{\beta}^{(j)(\omega+1)}-\boldsymbol{\beta}^{(j)(\omega)}\right)^{T}\boldsymbol{\Lambda}^{(j)}\left(\boldsymbol{\beta}^{(j)(\omega+1)}-\boldsymbol{\beta}^{(j)(\omega)}\right)+\lambda\left\|\boldsymbol{\beta}^{(j)(\omega+1)}\right\|_2.\end{aligned}\tag{S-8}$$

Then we can substitute $\boldsymbol{\beta}^{(\omega+1)}$ with $\boldsymbol{\beta}^{*(\omega+1)}$ and have

$$\begin{aligned}L\left(\boldsymbol{\beta}^{*(\omega+1)}|\mathbf{D}\right)+\lambda\|\boldsymbol{\beta}^{(j)*(\omega+1)}\|_2\le L\left(\boldsymbol{\beta}^{(\omega)}|\mathbf{D}\right)+\left(\boldsymbol{\beta}^{(j)*(\omega+1)}-\boldsymbol{\beta}^{(j)(\omega)}\right)^{T}\nabla L^{(j)}\\+\frac{1}{2}\left(\boldsymbol{\beta}^{(j)*(\omega+1)}-\boldsymbol{\beta}^{(j)(\omega)}\right)^{T}\boldsymbol{\Lambda}^{(j)}\left(\boldsymbol{\beta}^{(j)*(\omega+1)}-\boldsymbol{\beta}^{(j)(\omega)}\right)+\lambda\|\boldsymbol{\beta}^{(j)*(\omega+1)}\|_2.\end{aligned}\tag{S-9}$$

On the right-hand side of (S-9), we can update $\boldsymbol{\beta}^{(j)*(\omega+1)}$ using (17), and then (S-9) becomes

$$\begin{aligned}L\left(\boldsymbol{\beta}^{*(\omega+1)}|\mathbf{D}\right)+\lambda\|\boldsymbol{\beta}^{(j)*(\omega+1)}\|_2&\le L\left(\boldsymbol{\beta}^{(\omega)}|\mathbf{D}\right)+\left(\boldsymbol{\beta}^{(j)(\omega)}-\boldsymbol{\beta}^{(j)(\omega)}\right)^{T}\nabla L^{(j)}\\&+\frac{1}{2}\left(\boldsymbol{\beta}^{(j)(\omega)}-\boldsymbol{\beta}^{(j)(\omega)}\right)^{T}\boldsymbol{\Lambda}^{(j)}\left(\boldsymbol{\beta}^{(j)(\omega)}-\boldsymbol{\beta}^{(j)(\omega)}\right)+\lambda\|\boldsymbol{\beta}^{(j)(\omega)}\|_2\\&=L\left(\boldsymbol{\beta}^{(\omega)}|\mathbf{D}\right)+\lambda\|\boldsymbol{\beta}^{(j)(\omega)}\|_2.\end{aligned}\tag{S-10}$$

Inequality (S-10) shows that the objective function in (12) strictly decreases at each iteration unless $\boldsymbol{\beta}^{*(\omega+1)}=\boldsymbol{\beta}^{(\omega)}$. Furthermore, if there exists a $\boldsymbol{\beta}^{*}$, in which

$$\boldsymbol{\beta}^{(j)*}=\boldsymbol{\beta}^{(j)(\omega)}\ \text{ for } j=1,\ldots,J,$$

we can prove that $\boldsymbol{\beta}^{*}$ satisfy the Karush-Kuhn-Tucker (KKT) conditions, which implies that the GMD algorithm has converged to the optimal solution. Specifically, Equation (17) can be rewritten as

$$\boldsymbol{\beta}^{(j)*}=\begin{cases}\frac{1}{\rho_j}\left(-\nabla L^{(j)}+\rho_j\boldsymbol{\beta}^{(j)(\omega)}\right)\left(1-\frac{\lambda}{\|-\nabla L^{(j)}+\rho_j\boldsymbol{\beta}^{(j)(\omega)}\|_2}\right) & \text{if } \|-\nabla L^{(j)}+\rho_j\boldsymbol{\beta}^{(j)(\omega)}\|_2>\lambda\\ 0 & \text{if } \|-\nabla L^{(j)}+\rho_j\boldsymbol{\beta}^{(j)(\omega)}\|_2\le\lambda\end{cases},$$

from which we can derive KKT conditions for the optimization problem in (12). Therefore, if the objective function in (12) stays unchanged after a cycle, the GMD algorithm can be proved to converge to the optimal solution. □