Expanding the multivariate leptokurtic normal distribution to the high-dimensional clustering and factor analysis

Sohoon Youn

May 2023

Abstract

The finite mixture of multivariate leptokurtic-normal distributions is a flexible model based on mixtures of elliptical heavy-tailed distributions. However, the model's large number of parameters for estimation becomes computationally challenging, particularly in high-dimensional data. To address this issue, two procedures for the covariance structure are proposed. The first approach is based on the factor analyzers, while the second is based on a high-dimensional data clustering. Each methodology will be demonstrated with numerical experiments.

1 Introduction

Clustering is a widely used data analysis methodology employed to group data into several homogeneous groups. Its applications span various scientific fields (Yan and Shou-hua, 2012), where understanding and interpreting data patterns are vital. However, the challenge intensifies when dealing with high-dimensional data. The *curse of dimensionality* arises, leading to deteriorating performance as dimensions increases. Additionally, limited observations compared to the number of variables, further complicate the clustering task.

To address these challenges, it is crucial to find a balance between the number of parameters to estimate and the generality of the model. We propose multivariate leptokurtic normal (MLN) mixtures that involve the specific subspace in which each cluster is located. Regarding subspace clustering for the MLN mixtures, a factor analyzer (FA) and high-dimensional data clustering (HDC) have been suggested in this paper.

We estimate MLN mixture parameters using the Expectation-Maximization (EM) algorithm, as discussed by Dempster et al. (1977). The Bayesian Information Criterion (BIC) is used, as discussed by (Schwarz, 1978), to determine the number of factors in factor analysis and the intrinsic dimension of each cluster in HDC. These approaches give rise to a robust clustering method in high-dimensional spaces. Moreover, to further constrain the number of parameters, it can involve additional assumptions, such as parameter sharing across components.

This paper is organized as follows. Section 2 presents the brief overivew of background materials. Section 3 introduces the methodologies; i.e. factor analysis and high dimensional data clustering. Section 4 suggests the parameter estimation via EM algorithm. Analyses on simulations are reported in Section 5.

2 Background

2.1 Model-based clustering with high-dimensional data

Model-based clustering is a well-known and flexible methodology for its probabilistic principles. However, it tends to perform poorly when applied to high-dimensional data. With the increasing complexity in the data sets, model-based clustering methods have suffered from the *curse of dimens ionality* due to the over-parametrization of the model. There may exist subspaces with lower dimensions than the original space, *Empty space phenomenon*, which can result in disappointing performance in high dimensional data. To address this issue, dimension reduction methods, such as principal component analysis and factor analysis, as well as feature selection methods, are suggested as possible solutions. However, these methodologies may result in information loss, which can be critical in some cases. Several methods have been proposed to allow model-based methods to efficiently cluster high-dimensional data. They involve subspace clustering, constrained and parsimonious models (McNicholas and Murphy, 2008), regularization, and variable selection (Bouveyron et al., 2007). In this paper, we aim to group data into several homogeneous groups more precisely via subspace clustering, specifically factor analysis model (McNicholas and Murphy, 2008) and high-dimensional data clustering (Kim and Browne, 2019).

2.2 Finite mixture models

Finite mixture models are a mixture of probability distributions, each representing a cluster within the data. In this study, a data set of n observations $\{x_1, .., x_n\} \in \mathbb{R}^p$ is considered to be divided into G homogeneous group. To estimate the cluster membership of each observation, we introduce the concept of missing data, denoted as z_{ig}, where z_{ig} is 1 if observation i belongs to group g, and 0 otherwise.

The model-based clustering considers the overall population as a mixture of these groups, with each component represented by a conditional probability distribution. Thus, a G-component finite mixture model g($\mathbf{x}$) can be formulated as follows:

$$g(\mathbf{x}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}|\boldsymbol{\theta}_g); \quad \sum_{g=1}^{G} \pi_g = 1 \quad \text{for all } \pi_g,\ \pi_g > 0,$$

where π_g are the gth mixture proportion, and f_g are the conditional density function of gth mixture component with the parameter vector $\boldsymbol{\theta}_g$, for g = 1, ..., G. The finite mixture models have several advantages. They can effectively model complex data distributions by combining component densities with appropriate mixing proportions. Additionally, they can accommodate various types of distributions, allowing for flexibility in capturing diverse patterns present in the data.

In this paper, we assume that the component densities follow the MLN distribution. The identifiability of finite mixture of the MLN distributions is proved by Browne (2022).

2.3 Multivariate leptokurtic normal distribtuion

The MLN distribution, introduced by Bagnato et al. (2017), has the following joint probability density function

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta) = [1 + \beta g((\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}))]\phi(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where $\phi(\cdot)$ represents joint pdf of the multivariate normal distribution with mean $\boldsymbol{\mu}$, covariacne matrix $\boldsymbol{\Sigma}$, excess kurtosis β, and

$$g(r) = \frac{r^2 - 2(d+2)r + d(d+2)}{8d(d+2)},$$

$$\beta \in \left[0, \frac{4d(d+1)}{d+4}\right].$$

The MLN distribution is generated by applying a multivariate Gram-Charlier expansion to the multivariate normal distribution with one additional parameter for the kurtosis. With the excess kurtosis β, the MLN distribution makes a more general elliptical distribution with the desired amount of excess kurtosis. Browne(2022) gave a range for $\beta \in$ [0, 4d(d+2)/(d+4)] in which the MLN is distribution unimodal. It is desirable for the component densities to be unimodal for model based clustering. This distribution can be characterized by parameters directly corresponding to the moments of interest, particularly in each cluster.

2.4 Factor analysis

Factor analysis (for details see Bartholomew et al. (2011) and Spearman (1961)) is a method for modeling observed variables via a linear combination of unobservable latent factors. This approach offers several advantages, particularly in terms of data reduction and data interpretation. It can reduce a large number of variables into fewer numbers of factors, which leads to lower computational complexity. Additionally, it enables the explanation of the variability among important variables in terms of a reduced number of unobserved factors.

In factor analysis, McNicholas and Murphy (2008) considered parsimonious forms when each component of a mixture model had a factor model. A p-dimensional random vector $\mathbf{x}$ can be modelled through a q-dimensional vector of unobserved latent factors u, where q $\ll$ p. By default, the number of factors q must respect $(p-q)^2 >$ p+q, where p is the number of variables. Here, we take a random vector $\mathbf{x}$ as observed variables. Then, the model is $\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{U} + \boldsymbol{\epsilon}$ where $\boldsymbol{\Lambda}$ is a p×q matrix of factor loadings, the latent factors $\mathbf{U} \sim$ N(0, $\mathbf{I}_q$) and $\boldsymbol{\epsilon} \sim$ N(0, $\boldsymbol{\Psi}_q$), where $\boldsymbol{\Psi} = \text{diag}(\Psi_1, \Psi_2, ..., \Psi_p)$. With those orthogonal factor analysis model assumptions, the covariance among observed variables can be described as a linear combination of latent factors, $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$. It implies that the marginal distribution of $\mathbf{x}$ follows a distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$. The probability of an observation belonging to group g is represented by π_g. Therefore, the mixture of

factor analyzers model has density

$$f(\mathbf{x}_i) = \sum_{g=1}^{G} \frac{\pi_g}{(2\pi)^{p/2}|\mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \mathbf{\Psi}|^{1/2}} \times \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \mu_g)'(\mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \mathbf{\Psi})^{-1}(\mathbf{x}_i - \mu_g)\right\}.$$

2.5 High-dimensional data clustering

High-dimensional data clustering (HDC) combines the ideas of subspace clustering and parsimonious modeling, as discussed by Bouveyron et al. (2007) and Bouveyron and Brunet-Saumard (2014). Subspace clustering methods assume that each class within the data is situated on a distinct and unknown subspace. With this assumption, we can project the data points on to the distinct directions of orientation via eigen-decomposition,

$$\mathbf{\Sigma}_g = \mathbf{Q}_g\mathbf{D}_g\mathbf{Q}_g',$$

where $\boldsymbol{Q}_g$ is a p×p orthogonal matrix containing eigenvectors of $\mathbf{\Sigma}_g$ and $\mathbf{D}_g$ is a p×p diagonal matrix with p eigenvalues sorted in decreasing order, where the first q_g are distinct and the remaining (p-q_g) are the same. This approach enables us to find clusters in different subspaces within a data set. Morever, we obtain parsimonious models by introducing a parameterization that fixes certain parameters to be shared across components. With constraining parameters, it provides lower cost computations in high-dimensional data.

3 Methodology

The MLN distribution involves a covariance $\mathbf{\Sigma}$ which grows quadratically in size as the dimension increases. This growth can have a negative impact on clustering performance in high-dimensional spaces, especially when the number of variables is large. (Bouveyron et al., 2007) However, due to the *'Empty space phenomenon'*, we can assume that the majority of data points reside in lower-dimensional subspace. The original covariance $\mathbf{\Sigma}$ has $\frac{p(p+1)}{2}$ parameters but we consider two ways to reduce the number of parameter via factor analysis and HDC.

3.1 Factor analysis

We can develop a new class of MLN mixture models with parsimonious covariance structure by assuming a latent MLN for each population. With the orthogonal factor analysis model assumptions, the covariance among observed variables can be described as a linear combination of latent factors, $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$, where $\mathbf{\Lambda}$ is a p×q matrix of factor loadings and $\mathbf{\Psi} = \text{diag}(\psi_1, \psi_2, ..., \psi_p)$, where q $\ll$ p. Thus, we have $p \times q - q \times (q-1) + p$ parameters to estimate for $\mathbf{\Sigma}$ by applying factor analysis models to the covariance.

We provide the update for $\mathbf{\Lambda}_g$ and $\mathbf{\Psi}_g$ matrices as discussed by McNicholas and Murphy (2008). Table 1 represents a comprehensive range of constraints, resulting in eight distinct parsimonious covariance structures, as discussed by McNicholas and Murphy (2008). Specifically, the covariance structure UCU assumes the equal noise model, UUU assumes unequal noise, and UUC assumes

Table 1: Parsimonious covariance structures derived from the mixture of FA

Model	Loading matrix	Error variance	Isotropic	Covariance parameters
CCC	Constrained	Constrained	Constrained	$[pq - q(q-1)/2] + 1$
CCU	Constrained	Constrained	Unconstrained	$[pq - q(q-1)/2] + p$
CUC	Constrained	Unconstrained	Constrained	$[pq - q(q-1)/2] + G$
CUU	Constrained	Unconstrained	Unconstrained	$[pq - q(q-1)/2] + Gp$
UCC	Unconstrained	Constrained	Constrained	$G[pq - q(q-1)/2] + 1$
UCU	Unconstrained	Constrained	Unconstrained	$G[pq - q(q-1)/2] + p$
UUC	Unconstrained	Unconstrained	Constrained	$G[pq - q(q-1)/2] + G$
UUU	Unconstrained	Unconstrained	Unconstrained	$G[pq - q(q-1)/2] + Gp$

unequal but isotropic noise. The remaining five structures involve constraints on the loading matrices, which leads to a substantial reduction in the number of covariance parameters.

3.2 High-dimensional data clustering

We can apply the HDC methodology to the covariance parameter $\mathbf{\Sigma}_g$, g = 1,...,G, as discussed by Bouveyron et al. (2007) and Kim and Browne (2019). Since $\mathbf{\Sigma}_g$ is a positive definite $p \times p$ matrix, it is proposed to constraint the mixtures of MLN distribution through the eigen-decomposition of the covariance $\mathbf{\Sigma}_g$ of the gth group:

$$\mathbf{\Sigma}_g = [\mathbf{\Gamma}_g \mathbf{\Theta}_g] \mathbf{D}_g [\mathbf{\Gamma}_g \mathbf{\Theta}_g]',$$

where $\mathbf{D}_g$ is a $p \times p$ diagonal matrix containing p eigenvalues of $\mathbf{\Sigma}_g$ in descending order, and $[\mathbf{\Gamma}_g \mathbf{\Theta}_g]$ is a $p \times p$ orthogonal matrix whose columns are eigenvectors corresponding to the eigenvalues in $\mathbf{D}_g$. $\mathbf{\Gamma}_g$ is the matrix with the left q columns of $[\mathbf{\Gamma}_g \mathbf{\Theta}_g]$, and $\mathbf{\Theta}_g$ includes the remaining columns.

HDC relies on two assumptions. First, we assume that each cluster has an intrinsic dimension, denoted as q_g, where $g = 1$ ... G. The dimension q_g is significantly smaller than the total number of variables, p. Through this, we can project the data points onto the q_g-dimensional eigenspace of the covariance matrix, $\mathbf{\Sigma}_g$. This projection involves using the first q_g columns of the matrix $[\mathbf{\Gamma}_g \mathbf{\Theta}_g]$, which represent the distinct directions of orientation. Additionally, we can set the last $(p - q_g)$ eigenvalues to have the same value. This assumption indicates that the remaining directions are indistinguishable.

The following calculation on the Mahalanobis distance component in the density function of MLN illustrates the above idea with $r = x - \mu$,

$$\delta(\mathbf{x}|\mathbf{\Sigma}) = \mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{x} = \sum_{j=1}^{q} \lambda_j \|\boldsymbol{\gamma}_j \mathbf{x}\|^2 + \eta(\mathbf{x}'\mathbf{x} - \sum_{j=1}^{q} \|\boldsymbol{\gamma}_j \mathbf{x}\|^2) = \sum_{j=1}^{q} (\lambda_j - \eta)\|\boldsymbol{\gamma}_j \mathbf{x}\|^2 + \eta\, \mathbf{x}'\mathbf{x},$$

Table 2: Parsimonious covariance structures derived from HDC

Model	Eigenvectors	Eigenvalues	Remaining eigenvalues	Covariance parameters
VVV	Unconstrained	Unconstrained	Unconstrained	$[Gq(p-(q+1)/2)+Gq]+G$
VVE	Unconstrained	Unconstrained	Constrained	$[Gq(p-(q+1)/2)+Gq]+1$
VGV	Unconstrained	Equal within each g	Unconstrained	$[Gq(p-(q+1)/2)+G]+G$
VCV	Unconstrained	Mean across g	Unconstrained	$[Gq(p-(q+1)/2)+1]+G$
VGE	Unconstrained	Equal within each g	Constrained	$[Gq(p-(q+1)/2)+G]+1$
VEV	Unconstrained	Constrained across g	Unconstrained	$[Gq(p-(q+1)/2)+q]+G$
VEE	Unconstrained	Constrained across g	Constrained	$[Gq(p-(q+1)/2)+q]+1$
VCE	Unconstrained	Mean across g	Constrained	$[Gq(p-(q+1)/2)+1]+1$
EEE	Constrained	Constrained across g	Constrained	$[q(p-(q+1)/2)+q]+1$
EGV	Constrained	Equal within each g	Unconstrained	$[q(p-(q+1)/2)+G]+G$
EGE	Constrained	Equal within each g	Constrained	$[q(p-(q+1)/2)+G]+1$
ECV	Constrained	Mean across g	Unconstrained	$[q(p-(q+1)/2)+1]+G$
ECE	Constrained	Mean across g	Constrained	$[q(p-(q+1)/2)+1]+1$
EVV	Constrained	Unconstrained	Unconstrained	$[q(p-(q+1)/2)+Gq]+G$
EVE	Constrained	Unconstrained	Constrained	$[q(p-(q+1)/2)+Gq]+1$
EEV	Constrained	Constrained across g	Unconstrained	$[q(p-(q+1)/2)+q]+G$

where $\boldsymbol{\lambda}_j$: j = 1, ..., d are the first d eigenvalues of $\mathbf{D}$, $\boldsymbol{\gamma}_j$ is the first q eigenvectors of $\boldsymbol{\Sigma}_g$ or the first q eigenvectors of Γ_g, and η is the remaining $(p-q_g)$ eigenvalues. Therefore, we have $q+1+pq-\frac{q(q+1)}{2}$ parameters for $\boldsymbol{\Sigma}$ by HDC.

We examine the various models with different assumptions about the intrinsic dimension and orientation of the data set, as described in Table 2. This table lists out the full range of possible constraints provides a class of sixteen different parsimonious MLN mixture models. It shows that a model with constraints on eigenvectors and eigenvalues can reduce the number of parameters.

4 Parameter estimation

We use maximum likelihood for parameter estimation in MLN. Specifically, the Expectation-Maximization (EM) algorithm by Bouveyron and Brunet-Saumard (2014) and Browne (2022), which is a special case of the broader Majorize-Minimization algorithm by Hunter and Lange (2004), is applied here. The complete data set, denoted as $\{(x_i, z_i)\}_{i=1}^n$, plays a key role in this estimation process. Each pair (x_i, z_i) represents an observation x_i and its associated component

membership z_i, where $z_i = (z_{i1}, z_{i2}, ..., z_{iG})$. This complete data set captures all available information for the analysis. To quantify the likelihood of the observed data under the mixture model, we define the complete data likelihood as follows:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n}\prod_{g=1}^{G}\big[\pi_g f(\mathbf{x}_i|\boldsymbol{\mu}_g, \pi_g, \boldsymbol{\Sigma}_g, \beta_g)\big]^{z_{ig}},$$

where $\boldsymbol{\theta}$ represents the set of all parameters involved in the model. The expected value of the missing data given the observed data x_i and the current parameter values (at iteration j) is

$$\hat{w}_{ig}^{(j)} = E\big[z_{ig}|\mathbf{x}_i, \boldsymbol{\theta}_g^{(j)}\big] = \frac{\pi_g^{(j)} f\Big(\mathbf{x}_i|\boldsymbol{\mu}_g^{(j)}, \pi_g^{(j)}, \boldsymbol{\Sigma}_g^{(j)}, \beta_g^{(j)}\Big)}{\sum_{h=1}^{G}\pi_h^{(j)} f\Big(\mathbf{x}_i|\boldsymbol{\mu}_h^{(j)}, \pi_h^{(j)}, \boldsymbol{\Sigma}_h^{(j)}, \beta_h^{(j)}\Big)}.$$

Using these expected values we can construct the expected complete log-likelihood

$$\sum_{i=1}^{n}\sum_{g=1}^{G}\hat{w}_{ig}^{(j)}\big[\log\pi_g + \log f(\mathbf{x}_i|\boldsymbol{\mu_g}, \boldsymbol{\Sigma_g}, \beta_g)\big]. \tag{1}$$

Following Browne (2023) we have the fixed point approximation in (1),

$$\log f(\mathbf{x}|\theta) = -\frac{1}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}\Big[\big(\boldsymbol{\Sigma}_g^{(j)}\big)^{-1}\mathbf{A}\Big], \tag{2}$$

where

$$\mathbf{A} = \sum_{i=1}^{n} w_{ig}k_i\big(\boldsymbol{\mu}_g^{(j)}, \boldsymbol{\Sigma}_g^{(j)}, \beta_g^{(j)}\big)(\boldsymbol{x}_i - \boldsymbol{\mu}_g)(\boldsymbol{x}_i - \boldsymbol{\mu}_g)'.$$

To obtain updates, we maximize the expected complete log-likelihood. The update for π_g is

$$\pi_g^{(j+1)} = n_g^{(j)}/n, \tag{3}$$

where $n_g^{(j)} = \sum_{i=1}^{n} w_{ig}^{(j)}$. We can construct the surrogate function (2) using expected values as discussed by Browne (2022).

$$S\Big(\boldsymbol{\theta}_g^{(j)}\Big) = \sum_{g=1}^{G} n_g^{(j)}\log\pi_g^{(j)} - \frac{np}{2}\log(2\pi) - \sum_{g=1}^{G}\frac{n_g^{(j)}}{2}\log|\boldsymbol{\Sigma}_g^{(j)}| \tag{4}$$

$$-\frac{M}{2}\sum_{g=1}^{G}\mathrm{tr}\Big[\big(\boldsymbol{\Sigma}_g^{(j)}\big)^{-1}\sum_{i=1}^{n} w_{ig}^{(j)}k_i\big(\boldsymbol{\mu}_g^{(j)}, \boldsymbol{\Sigma}_g^{(j)}, \beta_g^{(j)}\big)\big(x_i - \boldsymbol{\mu}_g^{(j)}\big)\big(x_i - \boldsymbol{\mu}_g^{(j)}\big)'\Big],$$

$$\text{where}\quad k_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta) = 1 - 2\,\frac{\beta g'\big[(\boldsymbol{x_i} - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x_i} - \boldsymbol{\mu})\big]}{1 + \beta g\big[(\boldsymbol{x_i} - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x_i} - \boldsymbol{\mu})\big]}$$

with a boundary condition for $k_i(\cdot) \in [0,1]$. By taking the derivative and solving, it yields the following update for $\boldsymbol{\mu}_g$ and β_g. Then, the update for $\boldsymbol{\mu}_g$ is

$$\boldsymbol{\mu}_g^{(j+1)} = \frac{\sum_{i=1}^{n} w_{ig}^{(j)}k_i\Big(\boldsymbol{\mu}_g^{(j)}, \boldsymbol{\Sigma}_g^{(j)}, \beta_g^{(j)}\Big)\boldsymbol{x_i}}{\sum_{i=1}^{n} w_{ig}^{(j)}k_i\Big(\boldsymbol{\mu}_g^{(j)}, \boldsymbol{\Sigma}_g^{(j)}, \beta_g^{(j)}\Big)}. \tag{5}$$

For the parameter β_g, Browne (2022) showed that the log-likelihood in terms of β_g is a self-concordant function with respect to the parameter β_g. Self-concordant functions yield a bound on the number of Newton steps needed to minimize a function, as explained by Boyd and Vandenberghe (2004).

An unconstrained Newton step for β_g is

$$\beta_g^{(j+1)} = \beta_g^{(j)} + \frac{\sum_{i=1}^{n} w_{ig}\gamma_i\Big(\boldsymbol{\mu}_g^{(j)}, \boldsymbol{\Sigma}_g^{(j)}, \beta_g^{(j)}\Big)}{\sum_{i=1}^{n} w_{ig}\gamma_i\Big(\boldsymbol{\mu}_g^{(j)}, \boldsymbol{\Sigma}_g^{(j)}, \beta_g^{(j)}\Big)^2}, \tag{6}$$

where

$$\gamma_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta) = \frac{\beta g\big[(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\big]}{1 + \beta g\big[(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\big]},$$

with a boundary condition for β (Bagnato et al., 2017), $\beta \in$ [0,min(4d,4d(d+2)/5)]. We apply this update while maintaining that β_g stays within the range of unimodality. Then, for $\boldsymbol{\Sigma}_g$ we consider two different covariance structures; a high-dimensional data clustering and a factor analysis.

4.1 Factor analysis

We present the two procedures for estimating $\boldsymbol{\Sigma}_g$ with the structure given by high-dimensional data clustering and factor analysis. Note two methods have the same updates for π_g, μ_g and β_g given in equations (3), (5), and (6). In factor analysis the covariance parameters are $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$. At iteration q, with $\boldsymbol{\mu}_g = \boldsymbol{\mu}_g^{(j+1)}$, $\pi_g = \pi_g^{(j+1)}$, $w_{ig} = w_{ig}^{(j+1)}$ and $\beta_g = \beta_g^{(q+1)}$, the surrogate function for the component g can be defined as (1).

We apply the Majorize-Minimization algorithm (Hunter and Lange, 2004) to this objective function. The following function Q is a majorizing function for the objective function F. Appendix A shows that Q is a majorizing function. Here, a function $Q\big(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g | \boldsymbol{\Lambda}_g^{(j)}, \boldsymbol{\Psi}_g^{(j)}\big)$ is said to majorize the function $F(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g)$ at $\boldsymbol{\mu}_g, \pi_g, w_{ig}, \beta_g$ provided

$$Q\big(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g | \boldsymbol{\Lambda}_g^{(j)}, \boldsymbol{\Psi}_g^{(j)}\big) \geq F\big(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g\big) \quad \forall \boldsymbol{\Lambda}, \boldsymbol{\Psi},$$
$$Q\big(\boldsymbol{\Lambda}_g^{(j)}, \boldsymbol{\Psi}_g^{(j)} | \boldsymbol{\Lambda}_g^{(j)}, \boldsymbol{\Psi}_g^{(j)}\big) = F\big(\boldsymbol{\Lambda}_g^{(j)}, \boldsymbol{\Psi}_g^{(j)}\big)$$

In the Majorize-Minimization algorithm, we minimize the majorizing function $Q\big(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g | \boldsymbol{\Lambda}_g^{(j)}, \boldsymbol{\Psi}_g^{(j)}\big)$ rather than the actual function $F\big(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g\big)$. Then, the update for $\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g$ is estimated by

$$\big(\boldsymbol{\Lambda}_g^{(j+1)}, \boldsymbol{\Psi}_g^{(j+1)}\big) = \underset{\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g}{\operatorname{argmin}} Q(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g | \boldsymbol{\Lambda}_g^{(j)}, \boldsymbol{\Psi}_g^{(j)}),$$

when $Q\big(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g | \boldsymbol{\Lambda}_g^{(j)}, \boldsymbol{\Psi}_g^{(j)}\big)$ majorizes $F\big(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g\big)$. We iterate until $\boldsymbol{\Lambda}_g^{(j)}, \boldsymbol{\Psi}_g^{(j)}$ converges. Then, the resulting procedure is similar to the alternating expectation-conditional maximization (AECM) algorithm as discussed by Meng and Rubin (1993).

4.2 High-dimensional data clustering

In high-dimensional data clustering, the covariance parameters are ϕ_{gj}, γ_g, and η_g. The q_g distinct eigenvalues ϕ_{gj} and corresponding eigenvectors γ_g are obtained via eigen-decomposition of $\mathbf{A}_g$. The remaining $(p - q_g)$ eigenvalues are estimated by

$$\eta_g^{(j+1)} = \frac{1}{p - q_g^{(j+1)}} \left[\text{tr}\Big(\mathbf{A}_g^{(j+1)} - \sum_{k=1}^{q_g} \phi_{gk}^{(j+1)} \Big) \right],$$

where q_g is an intrinsic dimension, and k=1,2,3 ..., q_g, and g = 1,...,G.

4.3 Computational issues

Aitken's acceleration is used to measure the convergence of a model during the EM algorithm. The log-likelihood value of the model at each iteration is estimated asymptotically to assess the level of convergence, as discussed in Kim and Browne (2019) and McNicholas and Murphy (2008). At the k^{th} iteration, the asymptotic estimate of the log-likelihood, denoted as l_∞^k, is calculated using the formula:

$$l_\infty^{(k)} = l^{(k)} + \frac{1}{1 - a^{(k)}} \Big(l^{(k+1)} - l^{(k)} \Big),$$

where $l^{(k)}$ represents the log-likelihood at the k^{th} iteration, and $a^{(k)}$ is the ratio of the differences in log-likelihood between iterations k and k-1, and between iterations k+1 and k. The algorithm can be stopped when the difference between $l_\infty^{(k)}$ and $l^{(k)}$ is positive and less than a pre-defined threshold value $\epsilon > 0$. The latter criterion is used for all models, with $\epsilon = 10^{-4}$.

McNicholas and Murphy (2008) suggested the Bayesian information criterion (Schwarz, 1978) for selecting the appropriate $\mathbf{\Sigma}$. In the context of a model with parameters θ, the BIC is calculated as follows:

$$\text{BIC} = 2l(\hat{\theta}) - p\log(n),$$

where $l(\hat{\theta})$ is the maximized log-likelihood of a model, $\hat{\theta}$ is the maximum likelihood estimate of θ, p represents the number of free parameters in the model and n represents the number of observations. We test them using q ranging from 1 to $q-1$, where q represents the full dimension of the data set, in factor analysis and HDC, respectively. Then, the most appropriate $\mathbf{\Sigma}$ for each number of latent factors and intrinsic dimension will be chosen in each methodology.

5 Analyses on simulation

We conduct an evaluation of FA and HDC using simulated data sets. Both FA and HDC utilize the same randomly generated initial values. We generate a data set comprising 1000 observations based on a two-component ($g = 2$) MLN mixture model. The model parameters are as follows:

$$(\pi_1, \pi_2) = \Big(\frac{1}{2}, \frac{1}{2}\Big), \quad \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 = \{-2\mathbf{1}_d, +2\mathbf{1}_d\}, \quad \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}_d, \quad \text{and} \quad \beta_1 = \beta_2 = \frac{4d(d+2)}{(d+4)}.$$

Table 3: Number of times each parsimonious model was selected by BIC. Each column represents a summary of 100 data sets generated from a MLN mixture model using FA covariance structures.

Selected Model	$q = 1$	$q = 2$	$q = 3$	$q = 4$	$q = 5$	$q = 6$	$q = 7$	$q = 8$	$q = 9$
CUCC	16	18	14	26	16	19	14	9	17
CUCU	84	82	86	74	80	79	78	83	74
UCCC					2	2	1	1	1
UCCU					2		7	7	8
Total	100	100	100	100	100	100	100	100	100

Table 4: Number of times each parsimonious model was selected by BIC. Each column represents a summary of 100 data sets generated from a MLN mixture model using HDC covariance structures.

Selected Model	$q = 1$	$q = 2$	$q = 3$	$q = 4$	$q = 5$	$q = 6$	$q = 7$	$q = 8$	$q = 9$
EEEE	60	41	31	26	29	31	31	29	26
EVVE	26	46	55	66	60	67	64	67	66
EGVE	13	13	12	6	6				
EEVE	1			1	1		1		
EVVV			2	1	3	2	2	4	7
EEEV					1				
VVEE							2		1
Total	100	100	100	100	100	100	100	100	100

In our simulations, we fix the number of variables d to be 10, but vary the number of factors q and intrinsic dimension q to be 1, 2, 3, ..., $q-1$. EM algorithm is initialized using k-means clustering, and each experiment is replicated 100 times. During the iterative process, parameters such as $\boldsymbol{\mu}_g$, $\boldsymbol{\beta}_g$, and $\boldsymbol{\Sigma}_g$ for each component g are updated until convergence was achieved. The Aitken acceleration is utilized as a stopping criterion with a tolerance set to 10^{-4}.

As an initial comparison we count the number of times each covariance model is selected using the BIC across multiple replications. This analysis is conducted while varying the number of factors denoted as q and the intrinsic dimension q. The outcomes are presented in Tables 3 and 4, providing an overview of how each model was chosen under varying q values. Interesting, the covariance structure CUCU consistently emerged as the most frequently selected covariance model in FA across all q values. However, when q was equal to 1, the EEEE covariance exhibited dominance, while EVVE was chosen more often in all other cases of q.

A secondary analysis is performed to assess the behavior of FA and HDC when fitting finite mixture models. We consider different scenarios with varying numbers of components $G = 1, 2, 3$, and different numbers of factors, denoted as $q = 1, 2, 3, 4, 5$. In each case, both algorithms are iterated until convergence using identical initial values. Table 5 presents the mean computational time, log-likelihood at convergence, and iteration count for each combination of G and q. Notably, for $G = 1$, both FA and HDC yielded lower log-likelihood values, whereas $G = 2$ or 3 resulted in higher log-likelihoods. HDC consistently demonstrated faster convergence times than FA, regardless of the q value. However, FA consistently required the longest convergence time, especially for $q = 2$ or 3 across all G values. Furthermore, HDC exhibited fewer iterations to reach convergence compared to FA, regardless of G or q. These findings underscore HDC's efficiency in terms of both time and iterations across varying G and q scenarios, as compared to FA.

6 Conclusion

In this paper, we introduce subspace clustering methodologies for the finite mixture of multivariate leptokurtic-normal distributions. Additionally, we apply a parsimonious model by imposing constraints on the component covariance models and excess kurtoses. The covariance models are estimated using two different methodologies, involving FA and HDC, while the rest of the parameters are estimated using the fixed-point approximation. Two methodologies are assessed by mean computational time, log-likelihood at convergence, and iteration count for each combination of G and q and the number of times each covariance model is selected varying q. Moreover, the proposed methods can be compared in terms of computation time, the number of iterations, log-likelihood and BIC using real data.

Table 5: The average computational time, log-likelihood and number of iterations until convergence or reaching the maximum iterations of 5,000 for each algorithm. G represents the number of components.

	G	q	Time	Loglik	Iteration
FA	1	1	0.58	-19866.41	129.75
FA	1	2	13.10	-19859.78	3076.17
FA	1	3	22.60	-19885.78	4316.14
FA	1	4	4.81	-19888.42	929.24
FA	1	5	4.18	-19877.41	823.45
FA	2	1	52.66	-17733.12	3975.56
FA	2	2	53.31	-17695.47	4922.18
FA	2	3	52.08	-17727.92	4999
FA	2	4	21.60	-17741.53	2050.67
FA	2	5	0.83	-17738.79	72.73
FA	3	1	67.60	-17717.06	4562.31
FA	3	2	72.74	-17685.04	4968.53
FA	3	3	73.47	-17702.97	4999
FA	3	4	43.13	-17682.18	2928.09
FA	3	5	19.71	-17695.97	1320.17
HDC	1	1	0.13	-19894.18	18.37
HDC	1	2	0.12	-19874.65	18.59
HDC	1	3	0.12	-19865.44	18.74
HDC	1	4	0.13	-19882.92	18.78
HDC	1	5	0.13	-19879.75	18.98
HDC	2	1	0.21	-17713.69	15.46
HDC	2	2	0.22	-17747.96	15.73
HDC	2	3	0.23	-17741.05	15.86
HDC	2	4	0.23	-17753.05	16.11
HDC	2	5	0.25	-17738.16	16.16
HDC	3	1	22.78	-17733.35	1682.65
HDC	3	2	20.40	-17725.61	1697.37
HDC	3	3	27.87	-17733.62	1599.74
HDC	3	4	29.63	-17732.29	1678.08
HDC	3	5	31.42	-17708.75	1677.66

7 Appendix A

Setup for two surrogate functions

$$-\frac{1}{2}\log|\mathbf{\Lambda\Lambda'}+\mathbf{\Psi}|-\frac{1}{2}\mathrm{tr}[(\mathbf{\Lambda\Lambda'}+\mathbf{\Psi})^{-1}\mathbf{S}]$$

using the matrix determinant Lemma

$$\log|\mathbf{\Lambda\Lambda'}+\mathbf{\Psi}|=\log|\mathbf{\Psi}|+\log|\mathbf{I_q}+\mathbf{\Lambda'\Psi^{-1}\Lambda}|.$$

Then since the function $\log|\mathbf{X}|$ is convex we have

$$\log|\mathbf{X}|\geq\log|\mathbf{A}|+\mathrm{tr}[\mathbf{A^{-1}}(\mathbf{X}-\mathbf{A})]=\log|\mathbf{A}|+\mathrm{tr}[\mathbf{A^{-1}}X-I]$$

for any positive matrices $\mathbf{A}$ and $\mathbf{X}$. Letting $\mathbf{X}=\mathbf{I_q}+\mathbf{\Lambda'\Psi^{-1}\Lambda}$ and

$$\log|\mathbf{I_q}+\mathbf{\Lambda'\Psi^{-1}\Lambda}|\geq\log|\mathbf{A}|+\mathrm{tr}[\mathbf{A^{-1}}(\mathbf{I_q}+\mathbf{\Lambda'\Psi^{-1}\Lambda})-\mathbf{I_q}].$$

Now for $\mathrm{tr}[(\mathbf{\Lambda\Lambda'}+\mathbf{\Psi})^{-1}\mathbf{S}]$, consider the following for any $\boldsymbol{\beta}$, $\boldsymbol{\beta_0}\in M_{pq}$

$$\begin{aligned}\mathbf{0}&\preceq(\boldsymbol{\beta}-\boldsymbol{\beta_0})'(\mathbf{I}+\mathbf{\Lambda'\Psi\Lambda})(\boldsymbol{\beta}-\boldsymbol{\beta_0})\\&\preceq\boldsymbol{\beta}'(\mathbf{I}+\mathbf{\Lambda'\Psi\Lambda})\boldsymbol{\beta}-\boldsymbol{\beta}_0'(\mathbf{I}+\mathbf{\Lambda'\Psi\Lambda})\boldsymbol{\beta}-\boldsymbol{\beta}'(\mathbf{I}+\mathbf{\Lambda'\Psi\Lambda})\boldsymbol{\beta_0}+\boldsymbol{\beta}_0'(\mathbf{I}+\mathbf{\Lambda'\Psi\Lambda})\boldsymbol{\beta_0},\quad(1)\end{aligned}$$

where $\mathbf{A}\preceq\mathbf{B}$ means that $\mathbf{B}-\mathbf{A}$ is semi-positive. This matrix is semi-positive if $\boldsymbol{\beta}\neq\boldsymbol{\beta_0}$ and equal to zero when $\boldsymbol{\beta}=\boldsymbol{\beta_0}$.

Now if

$$\boldsymbol{\beta_0}=\mathbf{\Lambda'}(\mathbf{\Lambda\Lambda'}+\mathbf{\Psi})^{-1}=(\mathbf{I}+\mathbf{\Lambda'\Psi\Lambda})^{-1}\mathbf{\Lambda'\Psi^{-1}},$$

apply this to (1) to obtain

$$\mathbf{0}\preceq\boldsymbol{\beta}'(\mathbf{I}+\mathbf{\Lambda'\Psi\Lambda})\boldsymbol{\beta}-\mathbf{\Psi^{-1}\Lambda}\boldsymbol{\beta}-\boldsymbol{\beta}'\mathbf{\Lambda'\Psi^{-1}}+\mathbf{\Psi^{-1}\Lambda}(\mathbf{I}-\mathbf{\Lambda'\Psi\Lambda})^{-1}\mathbf{\Lambda'\Psi^{-1}}.$$

Then, adding and substracting $\mathbf{\Psi}$ using the woodbury identity yields

$$\begin{aligned}\mathbf{0}&\preceq\boldsymbol{\beta}'(\mathbf{I}+\mathbf{\Lambda'\Psi\Lambda})\boldsymbol{\beta}-\mathbf{\Psi^{-1}\Lambda}\boldsymbol{\beta}-\boldsymbol{\beta}'\mathbf{\Lambda'}\mathbf{\Psi}^{-1}+\mathbf{\Psi}^{-1}-[\mathbf{\Psi}^{-1}-\mathbf{\Lambda}(\mathbf{I}+\mathbf{\Lambda'\Psi\Lambda})^{-1}\mathbf{\Lambda'}\mathbf{\Psi}^{-1}]\\\mathbf{0}&\preceq\boldsymbol{\beta}'(\mathbf{I}+\mathbf{\Lambda'\Psi\Lambda})\boldsymbol{\beta}-\mathbf{\Psi^{-1}\Lambda}\boldsymbol{\beta}-\boldsymbol{\beta}'\mathbf{\Lambda'}\mathbf{\Psi}^{-1}+\mathbf{\Psi}^{-1}-(\mathbf{\Lambda\Lambda'}+\mathbf{\Psi})^{-1}\\\mathbf{\Lambda\Lambda'}+\mathbf{\Psi}&\preceq\boldsymbol{\beta}'(\mathbf{I}+\mathbf{\Lambda'\Psi\Lambda})\boldsymbol{\beta}-\mathbf{\Psi^{-1}\Lambda}\boldsymbol{\beta}-\boldsymbol{\beta}'\mathbf{\Lambda'}\mathbf{\Psi}^{-1}\end{aligned}$$

with equality when $\boldsymbol{\beta}=\boldsymbol{\beta_0}=\mathbf{\Lambda'}(\mathbf{\Lambda\Lambda'}+\mathbf{\Psi})^{-1}$. It implies that

$$\mathrm{tr}[(\mathbf{\Lambda\Lambda'}+\mathbf{\Psi})^{-1}\mathbf{S}]\leq\mathrm{tr}[\boldsymbol{\beta}'(\mathbf{I}+\mathbf{\Lambda'\Psi\Lambda})\boldsymbol{\beta}-\mathbf{\Psi^{-1}\Lambda}\boldsymbol{\beta}-\boldsymbol{\beta}'\mathbf{\Lambda'}\mathbf{\Psi}^{-1}\mathbf{S}]$$

with equality when $\boldsymbol{\beta}=\boldsymbol{\beta_0}=\mathbf{\Lambda'}(\mathbf{\Lambda\Lambda'}+\mathbf{\Psi})^{-1}$

References

Bagnato, L., Punzo, A., and Zoia, M. G. (2017). The multivariate leptokurtic-normal distribution and its application in model-based clustering. *Canadian Journal of Statistics*, 45(1):95–119.

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*, volume 904. John Wiley & Sons.

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

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

Boyd, S. P. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Browne, R. P. (2022). Revitalizing the multivariate elliptical leptokurtic-normal distribution and its application in model-based clustering. *Statistics & Probability Letters*, 190:109640.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Hunter, D. R. and Lange, K. (2004). A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37.

Kim, N.-H. and Browne, R. (2019). Subspace clustering for the finite mixture of generalized hyperbolic distributions. *Advances in Data Analysis and Classification*, 13:641–661.

McNicholas, P. D. and Murphy, T. B. (2008). Parsimonious gaussian mixture models. *Statistics and Computing*, 18:285–296.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80(2):267–278.

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

Spearman, C. (1961). The proof and measurement of association between two things.

Yan, Z. and Shou-hua, B. (2012). Research on optimizing recommend system for agriculture information personalization based on user clustering. In *2012 International Conference on Industrial Control and Electronics Engineering*, pages 1477–1480. IEEE.