IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 June 2023, accepted 12 July 2023, date of publication 19 July 2023, date of current version 27 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3296790

RESEARCH ARTICLE

Model-Based Clustering of Mixed Data With Sparse Dependence

YOUNG-GEUN CHOI[1], (Member, IEEE), SOOHYUN AHN[2], AND JAYOUN KIM[3]

[1]Department of Mathematics Education, Sungkyunkwan University, Seoul 03063, Republic of Korea
[2]Department of Mathematics, Ajou University, Suwon, Gyeonggi-do 16499, Republic of Korea
[3]Medical Research Collaborating Center, Seoul National University Hospital, Seoul 03080, Republic of Korea

Corresponding author: Jayoun Kim (nunadli03@snu.ac.kr)

The work of Young-Geun Choi was supported by the National Research Foundation of Korea under Grant RS-2023-00252026. The work of Soohyun Ahn was supported by the National Research Foundation of Korea under Grant 2019R1F1A1056779. The work of Jayoun Kim was supported by the Ministry of Health and Welfare, Republic of Korea, under Grant HL19C0026.

ABSTRACT Mixed data refers to a mixture of continuous and categorical variables. The clustering problem with mixed data is a long-standing statistical problem. The latent Gaussian mixture model, a model-based approach for such a problem, has received attention owing to its simplicity and interpretability. However, these approaches are prone to dimensionality problems. Specifically, parameters must be estimated for each group, and the number of covariance parameters is quadratic in the number of variables. To address this, we propose ''regClustMD,'' a novel model-based clustering method that can address sparse dependence among variables. We consider a sparse latent Gaussian mixture model, assuming that the precision matrix between variables has sparse nonzero elements. We propose maximizing a penalized complete log-likelihood using the Monte Carlo expectation-maximization (MCEM) algorithm. Our numerical experiments and real data analyses demonstrated that our method outperformed a counterpart algorithm in both accuracy and failure rate under the correlated data structure.

INDEX TERMS Latent Gaussian mixture model, maximum likelihood, model-based clustering, Monte Carlo expectation-maximization algorithm.

I. INTRODUCTION

In clustering problems, the observations are clustered into groups that share similar features. These features are commonly observed in continuous and categorical mixed data types (ordinal, nominal, or binary). Mixed-type datasets are prevalent in many applications, for example, finance, marketing, medicine, and healthcare sciences [1]. Although there is a wealth of clustering approaches, they often face challenges in correctly and simultaneously explaining the correlation structure in large datasets that consist of mixed types of variables. One of the main issues is the choice of the most appropriate distance or model to simultaneously deal with both data types. When an acceptable or reasonable model for cluster structure in the data can be found, model-based clustering provides persuasive results [2]. It is a popular toolkit and a principled statistical approach for clustering. In particular, the Gaussian mixture model is popular for model-based clustering of continuous data [3], [4].

Recently, various model-based approaches have been proposed for categorical or mixed data analysis. The Gaussian mixture models in the presence of categorical variables were proposed in [5] and [6] for binary data, [7] for ordinal data, [8] for the combination of binary and continuous data. In addition, [9] and [10] considered a mixture of latent trait and factor models, respectively. Reference [11] then proposed the Gaussian mixture model in the ClustMD latent variable framework. ClustMD considers six specific covariance structures for latent variables. However, these structures assume uncorrelated variables, which may limit their practical use. The adaptation of the uncorrelatedness assumption may be due to the high dimensionality of the parameters when an arbitrary covariance structure is assumed.

In this study, we developed a novel model-based clustering algorithm called ''regClustMD'' (regularized model-based clustering for mixed data), that can model sparse but arbitrary

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Abdur Razzaque.

dependence structures for mixed data while addressing the dimensionality problem. We consider the latent Gaussian mixture model for the observed data and maximize the ℓ_1-penalized version of the complete log-likelihood. By imposing the ℓ_1 penalty on the inverse covariance matrix term, we expect to explain the strong partial correlation among variables while reducing the number of active parameters. We propose a Monte Carlo expectation maximization (MCEM) algorithm to maximize the complete log-likelihood.

One key advantage of regClustMD is its ability to naturally incorporate the dependence structure between variables into the algorithm, even when it varies across different groups. As a result, compared to ClustMD, the proposed regClustMD can better estimate the within-covariance structure of each cluster. The improvement in estimating cluster-specific covariance can enhance the quality of clustering, as we will demonstrate in both our simulation experiments and real data examples.

The remainder of this paper is organized as follows. In Section III, we introduce our probability model of the Gaussian mixture model with latent continuous variables for mixed data. In Section IV, we describe the objective function and the detailed procedure of the regClustMD algorithm. In Section V, we propose a model selection procedure based on BIC. In Sections VI and VII, we demonstrate the usefulness of the proposed method through simulations and real data examples. Section VIII summarizes this paper.

The codes used for the simulation and real data analysis section is available at:
`http://github.com/shahn63/regclustMD`.

II. RELATED WORK

A. MODEL-BASED CLUSTERING ALGORITHMS FOR MIXED DATA

Since the work of McParland and Gormley [11] proposed a clustering algorithm for a mixed type of continuous, binary, ordinal, or nominal variables, it may be the most relevant to our method. It employs a latent variable model where observed categorical variables are considered as a discretization of the continuous latent variables. Then, the latent variables are assumed to follow a mixture of Gaussian distributions. To reduce the number of parameters, it introduces a diagonal covariance structure for the latent Gaussian variables. A major drawback of this approach is not being able to model the dependence structure between variables.

Several other papers have proposed clustering methods for a narrower scope of a mixture of specific types of data. Reference [8] proposed a clustering algorithm for mixed binary and continuous variables, where each binary attribute is generated by a latent continuous variable that is dichotomized with a suitable threshold value. Reference [5] introduced a latent variable model for binary data with heterogeneity accounted for replacing the traditional assumption of Gaussian distributed factors with a finite mixture of multivariate Gaussian. Reference [6] proposed a mixture of latent trait models with common slope parameters for model-based clustering of high-dimensional binary data. Reference [7] developed a mixture model for ordinal data using a pairwise likelihood approach. They considered the observed categorical variables as a discretization of an underlying finite mixture of Gaussian estimated within the EM framework. Reference [9] assumed a model for the categorical response variables that depends on both a categorical latent class and a continuous latent trait variable. The discrete latent class accommodated group structure and the continuous latent trait held dependence within these groups. Last but not least, [10] presented a latent variable based-algorithm for a mixture of binary, ordinal, and nominal response data.

B. REGULARIZED MODEL-BASED CLUSTERING ALGORITHMS FOR CONTINUOUS DATA

For continuous-type datasets, the Gaussian mixture model and its variants have been popular for model-based clustering algorithms. We refer to [3] and [4] for a comprehensive review. Regularized mixture models have been proposed to deal with the dimensionality problem when the data are only of the continuous type. References [12] and [13] proposed penalizing the inverse covariance matrix in terms of the log-likelihood of the Gaussian mixture model. Some variants can be made; for example, the group membership variable can be extended from a binary variable to a continuous variable that sums up to one [14] and [15]. In addition, the distribution of the data can extend to multivariate t-mixtures [16]. We refer the reader to [17] for a more comprehensive and extensive review of model selection strategies for the Gaussian mixture model-based clustering problem. However, as mentioned previously, these methods are limited to continuous data.

C. NON-MODEL-BASED CLUSTERING ALGORITHMS FOR MIXED DATA

The extensive literature on non-model-based clustering algorithms encompasses a wide variety of methods, such as K-means, hierarchical clustering, and density-based algorithms among others [18], [19], [20], [21]. Some of these methods rely on a measure of distance between data values. A common approach for non-model-based clustering of mixed data involves calculating a 'generalized distance', a weighted combination of the distance between continuous variables, and a dissimilarity measure for categorical variables such as Gower distance. One can run hierarchical clustering algorithms for mixed data by employing this generalized distance instead of the conventional distance measure. Reference [22] proposed the K-prototype algorithm, a variant of K-means, that iteratively calculates modes of temporary clusters. However, one major drawback of these approaches is the difficulty in balancing the weight between distances for categorical and continuous variables. These are controlled by a weighting factor which, if not set appropriately, may cause one type of

variable to dominate the other, which may lead to suboptimal clustering.

III. MODEL

We denote the observed record of the i-th subject as $y_i = (y_{i1}, \ldots, y_{ip})^T \in \mathbb{R}^p$, $i = 1, \ldots, N$. We denote hidden cluster group memberships by the random variable $g_i \in \{1, \ldots, G\}$. In addition, we define $l_i = (l_{i1}, \ldots, l_{iG})^T$ as a one-hot representation of group membership, that is, $l_{ig} = I(g = g_i)$, $g = 1, \ldots, G$.

A. MODELING LATENT VARIABLES

We employ the setting described by [11] for the latent variables. For completeness, we define the latent variables as follows. Assume that each j-th variable of the i-th subject, y_{ij}, has a latent random variable associated with it, say, z_{ij}. We assume that z_{ij} is always continuous, whereas y_{ij} can be continuous, binary, ordinal, or nominal. For each case, we postulate the relationship between y_{ij} and z_{ij} as described below.

1) CONTINUOUS y_{ij}

We assume that the observed y_{ij} exactly matches the latent variable; that is,

$$y_{ij} = z_{ij}.$$

2) ORDINAL y_{ij}

It is assumed that the observed category y_{ij} is a split of univariate z_{ij}. Let us say that the j-th variable consists of ordinal categories $1, \ldots, K_j$. We assume that there exist $K_j - 1$ threshold values. We include $\pm\infty$ as the threshold for notational convenience. Let partitions $-\infty = c_{j,0} < c_{j,1} < \cdots < c_{j,K_j} = \infty$ split support of z_{ij}. Then, we suppose that

$$y_{ij} = k \iff z_{ij} \in [c_{j,k-1}, c_{j,k}), \quad k = 1, \ldots, K_j,$$

in other words, $y_{ij} = k \cdot I(c_{j,k-1} \le z_{ij} < c_{j,k})$. It is assumed that the mean and variance parameters of z_{ij} are unknown and free to change. Thus, for identifiability issue, the threshold values $c_{j,0}, c_{j,1}, \ldots, c_{j,K_j}$ are used as predetermined values. Following the literature [10], [11], we fix $c_{j,k} = \Phi^{-1}(b_k)$, where $b_k = \sum_{i=1}^N I(y_{ij} \le k)/N$ is the frequency of subject i satisfying $y_{ij} \le k$ and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

3) NOMINAL y_{ij}

Because nominal categories are unordered, a nominal variable requires a multivariate representation. With slight notational abuse, we propose that a nominal variable y_{ij} with K_j categories be represented by $z_{ij} := (z_{ij}^1, \ldots, z_{ij}^{K_j-1})^T \in \mathbb{R}^{K_j-1}$ as follows:

$$y_{ij} = \begin{cases} k, & \text{if } z_{ij}^k \ge 0 \text{ and } z_{ij}^k = \max_{l \in [K_j-1]} z_{ij}^l; \\ K_j, & \text{if } z_{ij}^l < 0 \text{ for all } l \in [K_j - 1]. \end{cases} \tag{1}$$

In other words, if at least one z_{ij}^l ($l = 1, \ldots, K_j - 1$) is non-negative then y_{ij} is the index l that maximizes z_{ij}^l otherwise, y_{ij} takes K_j.

4) BINARY y_{ij}

Binary y_{ij} can be viewed as a special case of both ordinal and nominal variables when $K_j = 2$. The formulation of an ordinal variable with $K_j = 2$ is the same as that of a nominal variable.

B. PROBABILITY MODEL AND OBJECTIVE FUNCTION

We impose a normal mixture assumption on the latent space of z_i. Suppose that

$$\begin{aligned} &\mathbb{P}(g_i = g) = \pi_g \\ &\text{(equivalently } l_i \sim \text{Multinom}\big(1, (\pi_1, \ldots, \pi_G)\big)\text{)}, \\ &z_i | g_i = g \sim \mathcal{N}_D(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \end{aligned}$$

where $\sum_{i=1}^G \pi_g = 1, \pi_g \ge 0, \boldsymbol{\mu}_g \in \mathbb{R}^D$, $\boldsymbol{\Sigma}_g$ is a p by p positive definite matrix for $g = 1, \ldots, G$ and $i = 1, \ldots, N$. Because z_i may have a larger dimension than y_i, we have $D \ge p$. We let $\boldsymbol{\Omega}_g = \boldsymbol{\Sigma}_g^{-1}$ denote the precision matrix for the group g. Then, the log-likelihood for complete data $\{(l_i, z_i)\}_{i=1}^N$ is written as

$$\begin{aligned} \log \mathcal{L}_C = \sum_{i=1}^N \sum_{g=1}^G \Big[& l_{ig} \log \pi_g + \text{const} + \frac{l_{ig}}{2} \log \det(\boldsymbol{\Omega}_g) \\ & - \frac{l_{ig}}{2} z_i^T \boldsymbol{\Omega}_g z_i + l_{ig} \boldsymbol{\mu}_g^T \boldsymbol{\Omega}_g z_i - \frac{l_{ig}}{2} \boldsymbol{\mu}_g^T \boldsymbol{\Omega}_g \boldsymbol{\mu}_g \Big]. \end{aligned} \tag{2}$$

The unknown parameters to be estimated are $\{\pi_g, \boldsymbol{\mu}_g, \boldsymbol{\Omega}_g : g = 1, \ldots, G\}$. In [11], $\boldsymbol{\Omega}_g$ is strictly restricted to a class of diagonal matrices. We impose only the sparsity of $\boldsymbol{\Omega}_g$, that is, $\boldsymbol{\Omega}_g$ is assumed to have sparse nonzero off-diagonal elements. This assumption relaxes the diagonal assumption and is advantageous because non-zero off-diagonal elements can address the dependency between variables. To reflect the sparsity assumption in the estimation procedure, we consider penalizing $\boldsymbol{\Omega}_g$, $g = 1, \ldots, G$, when maximizing $\log \mathcal{L}_C$. If complete data were available, the penalized log-likelihood would be written as

$$\underset{\pi_g, \boldsymbol{\mu}_g, \boldsymbol{\Omega}_g : g = 1, \ldots, G}{\text{argmax}} \left[\log \mathcal{L}_C - \sum_{g=1}^G P_g(\boldsymbol{\Omega}_g) \right],$$

where $P_g(\boldsymbol{\Omega}_g)$ is a sparse penalty function for $\boldsymbol{\Omega}_g$; for example, the vectorized ℓ_1 norm of $\boldsymbol{\Omega}_g$ multipled by a tuning parameter. Because the observed data $\{y_i\}_{i=1}^N$ are incomplete, one may want to maximize the expected value of the objective function given $\{y_i\}_{i=1}^N$.

IV. regClustMD: PROPOSED ALGORITHM

We now describe our proposed algorithm (regClustMD). The complete algorithm is described in Algorithm 1. It maximizes the penalized log-likelihood using the Monte Carlo expectation-maximization (MCEM) algorithm. We iterate the following E- and M-steps until the convergence of the

Algorithm 1 Proposed Algorithm (regClustMD)

Input: Observed data $\{y_i\}_{i=1}^N$, the number of groups G.
Initialize $\pi_g^{(1)}$, $\boldsymbol{\mu}_g^{(1)}$, $\boldsymbol{\Omega}_g^{(1)}$ for $g = 1, \ldots, G$.
while until convergence **do**
  (E-step)
  for $i = 1, \ldots, N$ and $g = 1, \ldots, G$ **do**
    Calculate τ_{ig} as in (3).
    Calculate m_{ig} as in (4).
    Calculate V_{ig} as in (5).
  end for
  (M-step)
  for $g = 1, \ldots, G$ **do**
    Update $\pi_g^{(t+1)}$ as in (7).
    Update $\boldsymbol{\mu}_g^{(t+1)}$ as in (8).
    Calculate $\boldsymbol{\Sigma}_g^{(t+1)}$ as in (10).
    Update $\boldsymbol{\Omega}_g^{(t+1)}$ as in (9).
  end for
end while

objective function or estimated parameters. In what follows, we explain the detail of each line in the algorithm.

A. E-STEP

This subsection describes the calculation of $\mathbb{E}(\log \mathcal{L}_C|\mathcal{D}, \boldsymbol{\Theta}^{(t)})$, where $\mathcal{D} = \{y_i\}_{i=1}^N$ is the observed dataset, $\boldsymbol{\Theta} = \{\boldsymbol{\mu}_g, \boldsymbol{\Omega}_g, \pi_g\}_{g=1}^N$ is the collection of all parameters and $\boldsymbol{\Theta}^{(t)} = \{\boldsymbol{\mu}_g^{(t)}, \boldsymbol{\Omega}_g^{(t)}, \pi_g^{(t)}\}_{g=1}^G$ is the t-th update of the loop of the MCEM algorithm. For simplicity, we let $\boldsymbol{\Theta} = \boldsymbol{\Theta}^{(t)}$ if there is no confusion.

To conveniently refer to the continuous and categorical parts without loss of generality, we write y_i as $y_i^T = (y_i^{\alpha T}, y_i^{\beta T})$, where $y_i^\alpha \in \mathcal{R}^C$ and $y_i^\beta \in \mathcal{R}^{p-C}$ are the continuous and categorical variables, respectively. For the continuous part, the designation of latent space yields $y_i^\alpha = z_i^\alpha$. For the categorical part, let k_s $(s = 1, \ldots, q)$ be all possible values that y^β can take and let $\mathcal{I}_s \subseteq \mathbb{R}^{D-C}$ be the set of all possible values of $z^\beta \in \mathbb{R}^{D-C}$ that generates an outcome k_s. We partition other notations accordingly, i.e., $z_i^T = (z_i^{\alpha T}, z_i^{\beta T})$, $\boldsymbol{\mu}_g^T = (\boldsymbol{\mu}_g^{\alpha T}, \boldsymbol{\mu}_g^{\beta T})$, and $\boldsymbol{\Sigma}_g = \begin{bmatrix} \boldsymbol{\Sigma}_g^{\alpha\alpha} & \boldsymbol{\Sigma}_g^{\alpha\beta} \\ \boldsymbol{\Sigma}_g^{\beta\alpha} & \boldsymbol{\Sigma}_g^{\beta\beta} \end{bmatrix}$. From (2), the expected values for the calculation are $\mathbb{E}(l_{ig}|\mathcal{D}, \boldsymbol{\Theta})$, $\mathbb{E}(l_{ig}z_i|\mathcal{D}, \boldsymbol{\Theta})$, and $\mathbb{E}(l_{ig}z_i z_i^T|\mathcal{D}, \boldsymbol{\Theta})$. In addition, we define $\tau_{ig} = \mathbb{E}(l_{ig}|\mathcal{D}, \boldsymbol{\Theta})$. With a slight abuse of notation, let $\mathcal{N}(z|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the density function of the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Assuming that the observed value of y_i^β is k_s, according to Bayes' rule,

$$\tau_{ig} = \mathbb{P}(g_i = g|\mathcal{D}, \boldsymbol{\Theta}) = \frac{\pi_g \mathcal{N}(z_i^\alpha|\boldsymbol{\mu}_g^\alpha, \boldsymbol{\Sigma}_g^{\alpha\alpha}) \int_{\mathcal{I}_s} \mathcal{N}(z_i^\beta|\boldsymbol{\mu}_g^{\beta|\alpha}, \boldsymbol{\Sigma}_g^{\beta|\alpha}) dz_i^\beta}{\sum_{g=1}^G \pi_g \mathcal{N}(z_i^\alpha|\boldsymbol{\mu}_g^\alpha, \boldsymbol{\Sigma}_g^{\alpha\alpha}) \int_{\mathcal{I}_s} \mathcal{N}(z_i^\beta|\boldsymbol{\mu}_g^{\beta|\alpha}, \boldsymbol{\Sigma}_g^{\beta|\alpha}) dz_i^\beta}, \tag{3}$$

where $\boldsymbol{\mu}_g^{\beta|\alpha} = \boldsymbol{\mu}_g^\beta + \boldsymbol{\Sigma}_g^{\beta\alpha}(\boldsymbol{\Sigma}_g^{\alpha\alpha})^{-1}(y_i^\alpha - \boldsymbol{\mu}_g^\alpha)$ and $\boldsymbol{\Sigma}_g^{\beta|\alpha} = \boldsymbol{\Sigma}_g^{\beta\beta} - \boldsymbol{\Sigma}_g^{\beta\alpha}(\boldsymbol{\Sigma}_g^{\beta\beta})^{-1}\boldsymbol{\Sigma}_g^{\alpha\beta}$. Evaluations of the truncated integrals of the multivariate normal densities are numerically conducted; for example, the minimax tilting Gibbs sampling method proposed by [23] implemented in the `R` package `mvNcdf`.

Once τ_{ig} is evaluated, $\mathbb{E}(l_{ig}z_i^\beta|\mathcal{D}, \boldsymbol{\Theta})$ and $\mathbb{E}(l_{ig}z_i^\beta z_i^{\beta T}|\mathcal{D}, \boldsymbol{\Theta})$ can be calculated as

$$\begin{aligned} &\mathbb{E}(l_{ig}z_i^\beta|\mathcal{D}, \boldsymbol{\Theta}) = \\ &\mathbb{P}(g_i = g|\mathcal{D}, \boldsymbol{\Theta}) \int_{\mathcal{I}_s} z_i^\beta \mathcal{N}(z_i^\beta|\boldsymbol{\mu}_g^{\beta|\alpha}, \boldsymbol{\Sigma}_g^{\beta|\alpha}) dz_i^\beta =: \tau_{ig} m_{ig}, \\ &\mathbb{E}(l_{ig}z_i^\beta z_i^{\beta T}|\mathcal{D}, \boldsymbol{\Theta}) = \\ &\mathbb{P}(g_i = g|\mathcal{D}, \boldsymbol{\Theta}) \int_{\mathcal{I}_s} z_i^\beta z_i^{\beta T} \mathcal{N}(z_i^\beta|\boldsymbol{\mu}_g^{\beta|\alpha}, \boldsymbol{\Sigma}_g^{\beta|\alpha}) dz_i^\beta =: \tau_{ig} V_{ig}. \end{aligned} \tag{4, 5}$$

Note that m_{ig} and V_{ig} are the first and second moments of the truncated multivariate normal distribution, respectively. Their calculations were implemented using the `mtmvtnorm` of the `R` package `tmvtnorm`. We refer the reader to [24] for the closed-form formula for moments.

B. M-STEP

Let $\tau_{ig}^{(t)}$, $m_{ig}^{(t)}$, and $V_{ig}^{(t)}$ be the resulting values from (3), (4), and (5), respectively, when the given parameter is $\boldsymbol{\Theta}^{(t)}$. Let $N_g^{(t)} = \sum_{i=1}^N \mathbb{E}(l_{ig}|\mathcal{D}, \boldsymbol{\Theta}^{(t)}) = \sum_{i=1}^N \tau_{ig}^{(t)}$ for $g = 1, \ldots, G$. For each M-step, we propose to encourage sparsity on $\boldsymbol{\Omega}_g$s by maximizing ℓ_1-penalized expected complete log-likelihood, that is,

$$\begin{aligned} Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}) &:= \mathbb{E}(\log \mathcal{L}_C|\mathcal{D}, \boldsymbol{\Theta}^{(t)}) - \sum_{g=1}^G P_g(\boldsymbol{\Omega}_g) \\ &\sum_{g=1}^G \Bigg[\frac{1}{2} \sum_{i=1}^N \Big\{ 2\mathbb{E}(l_{ig}|\mathcal{D}, \boldsymbol{\Theta}^{(t)}) \log \pi_g + \mathbb{E}(l_{ig}|\mathcal{D}, \boldsymbol{\Theta}^{(t)}) \log \det(\boldsymbol{\Omega}_g) \\ &- \mathrm{tr}\left(\boldsymbol{\Omega}_g \mathbb{E}\left(l_{ig}(z_i - \boldsymbol{\mu}_g)(z_i - \boldsymbol{\mu}_g)^T|\mathcal{D}, \boldsymbol{\Theta}^{(t)} \right) \right) \Big\} - \lambda N_g^{(t)} |\boldsymbol{\Omega}_g| \Bigg] \\ &\sum_{g=1}^G N_g^{(t)} \Bigg[\frac{1}{2} \Big\{ 2 \log \pi_g + \log \det(\boldsymbol{\Omega}_g) \\ &- \mathrm{tr}\Big(\boldsymbol{\Omega}_g \frac{\sum_{i=1}^N \mathbb{E}\left(l_{ig}(z_i - \boldsymbol{\mu}_g)(z_i - \boldsymbol{\mu}_g)^T|\mathcal{D}, \boldsymbol{\Theta}^{(t)} \right)}{N_g} \Big) \Big\} - \lambda |\boldsymbol{\Omega}_g| \Bigg]. \end{aligned} \tag{6}$$

Here, $|A| = \sum_{i \geq j} |a_{ij}|$ is the vectorized ℓ_1 norm of matrix $A = (a_{ij})$ and $\lambda \geq 0$ is a tuning parameter. For each group g, we have weighed the penalty $\lambda|\boldsymbol{\Omega}_g|$ by $N_g^{(t)}$ to ensure that the parameters across different groups were penalized in the same amount.

We now derive $\boldsymbol{\Theta}^{(t+1)} := \mathrm{argmax}_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)})$. It is noteworthy that (6) is separable over g. For each g, maximization over π_g and $\boldsymbol{\mu}_g$ is independent of maximization over $\boldsymbol{\Omega}_g$. It is

straightforward to derive the following equation:

$$\pi_g^{(t+1)} = \frac{N_g^{(t)}}{\sum_{g=1}^{G} N_g^{(t)}}, \tag{7}$$

and

$$\begin{aligned}\boldsymbol{\mu}_g^{(t+1)} &= \frac{1}{N_g^{(t)}} \sum_{i=1}^{N} \mathbb{E}\left(l_{ig}^{(t)} z_i | \mathcal{D}, \boldsymbol{\Theta}^{(t)}\right) \\ &= \frac{1}{N_g^{(t)}} \sum_{i=1}^{N} \tau_{ig}^{(t)} \begin{bmatrix} y_i^{\alpha} \\ m_{ig}^{(t)} \end{bmatrix}. \end{aligned} \tag{8}$$

To update $\boldsymbol{\Sigma}_g$, the maximization of (6) is a weighted version of the graphical lasso problem. To be precise,

$$\boldsymbol{\Omega}_g^{(t+1)} = \underset{\boldsymbol{\Omega}}{\operatorname{argmax}} \left\{ \log\det(\boldsymbol{\Omega}) - \operatorname{tr}\left(\boldsymbol{\Omega}\boldsymbol{\Sigma}_g^{(t+1)}\right) - \lambda|\boldsymbol{\Omega}| \right\}, \tag{9}$$

where

$$\begin{aligned}\boldsymbol{\Sigma}_g^{(t+1)} &:= \mathbb{E}\left(l_{ig} z_i z_i^T | \mathcal{D}, \boldsymbol{\Theta}^{(t)}\right) - \boldsymbol{\mu}_g^{(t+1)} \left(\boldsymbol{\mu}_g^{(t+1)}\right)^T \\ &= \frac{1}{N_g^{(t)}} \sum_{i=1}^{N} \tau_{ig}^{(t)} \begin{bmatrix} y_i^{\alpha} (y_i^{\alpha})^T & y_i^{\alpha} m_{ig}^{(t)T} \\ m_{ig}^{(t)} (y_i^{\alpha})^T & V_{ig}^{(t)} \end{bmatrix} \\ &\quad - \boldsymbol{\mu}_g^{(t+1)} \left(\boldsymbol{\mu}_g^{(t+1)}\right)^T. \end{aligned} \tag{10}$$

To solve (9), we can use off-the-shelf statistical software, for example, the graphical lasso [25] or QUIC [26].

Remark. To encourage sparsity, one may consider other penalty functions, such as the smoothly clipped absolute deviation (SCAD) or the minimax concave penalty (MCP). An advantage of the choice of ℓ_1-penalty is that each M-step becomes a concave maximization problem, which facilitates the scalability of our proposed algorithm.

Remark. The time complexity of the algorithm for each iteration is dominated by the combination of the number of Monte Carlo samples in the E-step and the precision matrix estimation in the M-step. Thus, it is $O(NM(D - C) + \min(GD^3, GND^2))$. Compared to ClustMD, a benchmark algorithm in the Simulation and Real Data Analysis Section, our algorithm has an additional cost of $O(\min \hat{a}(GD^3, GND^2))$ due to the estimation of the precision matrix.

V. MODEL SELECTION

The regClustMD procedure requires tuning G and λ, which determines the number of clusters and sparsity of the estimated precision matrix.

We consider BIC-based model selection, widely employed in model-based clustering [3], [11], [13], [27]. Let $\widehat{\boldsymbol{\Theta}} = \widehat{\boldsymbol{\Theta}}(G, \lambda)$ be the estimate of $\boldsymbol{\Theta}$ given G and λ. In EM algorithm-based procedures, the BIC value is the expected negative complete likelihood added by $DF \cdot \log(N)$. Similarly, we propose the BIC value as

$$BIC(G, \lambda) := -\mathbb{E}(\log \mathcal{L}_C | \widehat{\boldsymbol{\Theta}}(G, \lambda)) + DF \cdot \log(N),$$

where DF is regarded as the number of non-zero elements of the estimated parameters in the sparse estimation literature [28], that is, letting $\widehat{\boldsymbol{\Theta}} = \{\widehat{\pi}_g, \widehat{\boldsymbol{\mu}}_g, \widehat{\boldsymbol{\Omega}}_g\}_{g=1}^{G}$,

$$DF = N + NG + \sum_{g=1}^{G} \sum_{i \le j} I([\widehat{\boldsymbol{\Omega}}_g]_{ij} \neq 0),$$

where $[\widehat{\boldsymbol{\Omega}}_g]_{ij}$ is the (i, j)-th element of $\widehat{\boldsymbol{\Omega}}_g$. One advantage of our method is that it directly maximizes the expected complete likelihood, which does not require an additional approximation procedure, as in [11]. Finally, we select G and λ to minimize $BIC(G, \lambda)$.

VI. NUMERICAL STUDY

A. EXPERIMENTAL SETTINGS

To assess the performance of the proposed method, we compared the clustering accuracy rate and the number of failures. Our model was compared with ClustMD [11], a latent Gaussian mixture model approach that does not allow for an association between variables. For the clustering error rate, although BIC is known to identify the true model consistently across a range of applications, as in Section V, the number of clusters was assumed to be known as two ($G = 2$), so that the selection of cluster numbers between two methods does not confound the performance evaluation. With $G = 2$, the clusters produced by each method are re-labeled to be the most consistent with the real membership.

We consider 100 replications consisting of $N = 100$ subjects with $p = 10$ variables from a 2-cluster model with a mixing probability $\pi_1 = \pi_2 = 0.5$. Seven variables were continuous, two were ordinal with two and three levels, and one was nominal with three levels. The last three categorical variables were obtained by categorizing a part of z_{ij}. As in Section III, we consider the median and the first and third quantiles as threshold values for binary and ordinal data with three levels, respectively. For nominal data, we generate $K_j = 3$ levels using a $K_j - 1 = 2$-dimensional latent continuous variable and a threshold 0 using (1).

Fixing $G = 2$ and $\pi_1 = 0.5$, we generate 50 subjects for the first cluster from a $MVN(\boldsymbol{\mu}_1, \boldsymbol{\Omega}_1^{-1})$ with a mean vector $\boldsymbol{\mu}_1$ and a covariance matrix $\boldsymbol{\Omega}_1^{-1}$, and similarly 50 subjects for the second cluster from $MVN(\boldsymbol{\mu}_2, \boldsymbol{\Omega}_2^{-1})$. For simplicity, we fix the mean vector of each mixture component as $\boldsymbol{\mu}_1 = 0_p$ and $\boldsymbol{\mu}_2 \in \{1_p/2, 1_p\}$, where 1_p denotes a p-dimensional vector of ones. Here, the norm of $\boldsymbol{\mu}_2$ determines the separability of two clusters. Finally, we consider two sparse covariance structures:

- AR(1) model: the two precision matrices for both clusters follow AR(1) structure, respectively:

$$\boldsymbol{\Omega}_g(i, j) = \begin{cases} 1, & \text{if } i = j; \\ \rho, & \text{if } |i - j| = 1; \\ 0, & \text{otherwise,} \end{cases}$$

TABLE 1. Clustering accuracy summarized by mean (Mean) and standard error (S.E.) $\times 10^3$, and the number of simulation failures (# failures) over 100 replications with $\mu_2 = 1_p/2$.

ρ	Accuracy	AR(1) model		Random model	
		ClustMD	regClustMD	ClustMD	regClustMD
0	Mean	0.641	0.625	0.641	0.625
	S.E.$\times 10^3$	7.25	7.38	7.25	7.38
	# failures	-	5	-	5
0.2	Mean	0.583	0.579	0.604	0.623
	S.E.$\times 10^3$	5.63	6.11	6.40	7.01
	# failures	-	6	3	-
0.4	Mean	0.563	0.572	0.568	0.652
	S.E.$\times 10^3$	4.90	5.27	5.87	9.71
	# failures	-	-	4	-
0.6	Mean	0.555	0.568	0.569	0.674
	S.E.$\times 10^3$	4.33	5.08	5.78	13.71
	# failures	-	-	12	-
0.8	Mean	0.553	0.575	0.581	0.670
	S.E.$\times 10^3$	3.72	7.09	7.09	12.90
	# failures	-	-	23	-

TABLE 2. Clustering accuracy summarized by mean (Mean) and standard error (S.E.) $\times 10^3$, and the number of simulation failures (# failures) over 100 replications with $\mu_2 = 1_p$.

ρ	Accuracy	AR(1) model		Random model	
		ClustMD	regClustMD	ClustMD	regClustMD
0	Mean	0.909	0.890	0.909	0.890
	S.E.$\times 10^3$	3.73	4.60	3.73	4.60
	# failures	-	-	-	-
0.2	Mean	0.939	0.942	0.814	0.823
	S.E.$\times 10^3$	3.45	3.29	8.23	10.98
	# failures	-	-	4	-
0.4	Mean	0.952	0.966	0.746	0.862
	S.E.$\times 10^3$	5.14	6.06	7.85	13.11
	# failures	-	-	7	-
0.6	Mean	0.944	0.984	0.718	0.893
	S.E.$\times 10^3$	9.15	6.34	9.02	12.11
	# failures	-	-	23	-
0.8	Mean	0.919	0.993	0.689	0.839
	S.E.$\times 10^3$	11.52	3.68	9.08	13.95
	# failures	1	-	43	-

- Random model: the two precision matrices are generated by adding random noise ϵ to off-diagonal elements:

$$\boldsymbol{\Omega}_g(i,j) = \boldsymbol{\Omega}_g(j,i) = \begin{cases} 1, & \text{if } i = j; \\ \epsilon_{ij} \cdot I(|\epsilon_{ij}| < \rho), & \text{if } i \neq j, \end{cases}$$

where $g = 1, 2$, $\epsilon_{ij} \in [-1, 1]$ is uniformly and independently generated but fixed over replications, $I(\cdot)$ is the indicator function, and $\rho \in \{0, 0.2, 0.4, 0.6, 0.8\}$ denotes covariance structure parameters controlling the dependency and sparsity of the covariance matrix. A larger ρ indicates stronger dependence in both AR(1) and Random models. We scaled the covariance matrices for both models with diagonal elements of 1. All analyses were conducted using R version 4.0.2 [29].

B. RESULTS

Tables 1 and 2 show the results across different association levels by ρ under both AR(1) and Random models, with $\boldsymbol{\mu}_2 = 1_p/2$ and 1_p denoting medium and large separability, respectively. We compared the two methods in terms of the averaged clustering accuracy and the simulation failure rate over the 100 replications. In terms of clustering accuracy, regClustMD tended to have increased accuracy as the dependency between variables gets stronger (larger ρ). On the other hand, as we expected, ClustMD tended to have decreased accuracy for most settings as ρ increased. We hypothesize that the inefficiency of ClustMD in larger ρ may be because ClustMD assumes an independent covariance structure and does not account for dependency. When we compared the two methods for each setting, the proposed regClustMD was more accurate than ClustMD in most cases, except for the independent covariance structure case ($\rho = 0$) that is favorable to ClustMD. In terms of the simulation failure rate, the proposed regClustMD outperformed ClustMD which does not allow estimation of the correlated data structure (e.g., with $\rho = 0.8$ under the Random model, ClustMD's failure rates are 23% in Table 1 and 43% in Table 2).

TABLE 3. Unequal mixing probability π_1: clustering accuracy summarized by mean (Mean) and standard error (S.E.) $\times 10^3$, and the number of simulation failures (# failures) over 100 replications with $\mu_2 = 1_p$.

ρ	Accuracy	$\pi_1 = 0.75, \pi_2 = 0.25$			
		AR(1) model		Random model	
		clustMD	regclustMD	clustMD	regclustMD
0.0	Mean	0.920	0.903	0.920	0.903
	S.E.$\times 10^3$	0.034	0.048	0.034	0.048
	# failures	0	0	0	0
0.2	Mean	0.936	0.921	0.660	0.650
	S.E.$\times 10^3$	0.055	0.078	0.100	0.141
	# failures	1	1	7	0
0.4	Mean	0.886	0.904	0.619	0.585
	S.E.$\times 10^3$	0.128	0.145	0.083	0.061
	# failures	2	0	12	0
0.6	Mean	0.778	0.959	0.591	0.614
	S.E.$\times 10^3$	0.171	0.117	0.088	0.090
	# failures	0	0	22	0
0.8	Mean	0.750	0.929	0.609	0.628
	S.E.$\times 10^3$	0.170	0.150	0.098	0.117
	# failures	0	0	38	0

C. SENSITIVITY ANALYSIS AGAINST THE CHOICE OF EXPERIMENTAL PARAMETERS

We evaluated our proposed method across different parameters: mixing probability π_1, the number of groups G, sample size n, and the number of variables p. To illustrate this, we varied a parameter once at a time while fixing other parameters in Table 2.

1) MIXING PROBABILITY

Table 3 shows the results with different mixing probability $\pi_1 = 0.75, \pi_2 = 0.25$ compared to Table 2 with $\pi_1 = \pi_2 = 0.5$. In terms of both clustering accuracy and failure rate, the proposed regClustMD consistently outperformed ClustMD while the accuracy level was lower than those for each case in Table 2.

2) NUMBER OF GROUPS

Tables 4 present the results with different numbers of groups $G = 3, 4$ compared to Table 2 with $G = 2$. We again fixed the mean vectors as $\boldsymbol{\mu}_1 = 0_p$ and $\boldsymbol{\mu}_2 = 1_p$ and consider $\boldsymbol{\mu}_3 = (1_{p-q}, -1_q)$ and $\boldsymbol{\mu}_4 = -1_p$ where q denotes the

TABLE 4. Different number of groups G: clustering accuracy summarized by mean (Mean) and standard error (S.E.) $\times 10^3$, and the number of simulation failures (# failures) over 100 replications with $\mu_1 = 0_p$, $\mu_2 = 1_p$, $\mu_3 = (1_{p-q}, -1_q)$ and $\mu_4 = -1_p$.

(a)

ρ	Accuracy	$G = 3$			
		AR(1) model		Random model	
		clustMD	regclustMD	clustMD	regclustMD
0.0	Mean	0.719	0.682	0.719	0.682
	S.E. $\times 10^3$	0.115	0.107	0.115	0.107
	# failures	3	5	3	5
0.2	Mean	0.728	0.721	0.642	0.679
	S.E. $\times 10^3$	0.124	0.111	0.073	0.105
	# failures	4	7	3	2
0.4	Mean	0.692	0.746	0.636	0.692
	S.E. $\times 10^3$	0.107	0.115	0.097	0.128
	# failures	2	2	41	5
0.6	Mean	0.673	0.847	0.619	0.652
	S.E. $\times 10^3$	0.094	0.125	0.091	0.121
	# failures	1	1	36	1
0.8	Mean	0.650	0.891	0.567	0.632
	S.E. $\times 10^3$	0.094	0.104	0.072	0.117
	# failures	1	1	42	0

(b)

ρ	Accuracy	$G = 4$			
		AR(1) model		Random model	
		clustMD	regclustMD	clustMD	regclustMD
0.0	Mean	0.708	0.672	0.708	0.672
	S.E. $\times 10^3$	0.093	0.073	0.093	0.073
	# failures	48	14	48	14
0.2	Mean	0.721	0.706	0.628	0.704
	S.E. $\times 10^3$	0.080	0.082	0.074	0.085
	# failures	46	19	63	19
0.4	Mean	0.696	0.718	0.592	0.678
	S.E. $\times 10^3$	0.106	0.097	0.089	0.089
	# failures	39	15	83	15
0.6	Mean	0.675	0.742	0.705	0.706
	S.E. $\times 10^3$	0.102	0.122	0.101	0.081
	# failures	38	10	94	11
0.8	Mean	0.695	0.796	0.591	0.694
	S.E. $\times 10^3$	0.093	0.120	0.050	0.088
	# failures	39	11	85	10

TABLE 5. Different sample size n: clustering accuracy summarized by mean (Mean) and standard error (S.E.) $\times 10^3$, and the number of simulation failures (# failures) over 100 replications with $\mu_2 = 1_p$.

(a)

ρ	Accuracy	$n = 50$			
		AR(1) model		Random model	
		clustMD	regclustMD	clustMD	regclustMD
0.0	Mean	0.894	0.869	0.894	0.869
	S.E. $\times 10^3$	0.061	0.072	0.061	0.072
	# failures	2	3	2	3
0.2	Mean	0.912	0.907	0.817	0.799
	S.E. $\times 10^3$	0.061	0.087	0.090	0.117
	# failures	1	4	9	9
0.4	Mean	0.917	0.926	0.776	0.778
	S.E. $\times 10^3$	0.083	0.092	0.110	0.117
	# failures	2	5	15	14
0.6	Mean	0.899	0.953	0.754	0.775
	S.E. $\times 10^3$	0.114	0.087	0.107	0.120
	# failures	1	7	32	12
0.8	Mean	0.884	0.951	0.741	0.775
	S.E. $\times 10^3$	0.132	0.113	0.124	0.109
	# failures	0	6	37	17

(b)

ρ	Accuracy	$n = 200$			
		AR(1) model		Random model	
		clustMD	regclustMD	clustMD	regclustMD
0.0	Mean	0.924	0.915	0.924	0.915
	S.E. $\times 10^3$	0.023	0.026	0.023	0.026
	# failures	0	0	0	0
0.2	Mean	0.953	0.955	0.810	0.922
	S.E. $\times 10^3$	0.019	0.019	0.059	0.078
	# failures	0	0	0	0
0.4	Mean	0.978	0.989	0.749	0.950
	S.E. $\times 10^3$	0.016	0.008	0.057	0.088
	# failures	0	0	3	0
0.6	Mean	0.989	0.995	0.704	0.961
	S.E. $\times 10^3$	0.013	0.026	0.087	0.091
	# failures	0	0	16	0
0.8	Mean	0.986	0.999	0.698	0.913
	S.E. $\times 10^3$	0.033	0.014	0.093	0.114
	# failures	0	0	28	0

number of categorical variables. As the number of groups increased with a fixed sample size $n = 100$, both methods tended to have decreased accuracy and increased failure rate. Especially, clustMD failed most of the cases with $G = 4$ but regClustMD consistently outperformed clustMD with a small failure rate for both $G = 3, 4$.

3) SAMPLE SIZE

Tables 5 illustrate the results with different sample sizes $n = 50, 200$ compared to Table 2 with $n = 100$. As the sample size become larger, accuracy also increased except for clustMD under the Random model. The proposed method consistently outperformed clustMD and the difference in accuracy between the two methods tended to increase as n increased under the Random model which required the estimation of the more sophisticated correlated data structure.

4) NUMBER OF VARIABLES

Tables 6 illustrate the results with different sample size $p = 5, 20$ compared to Table 2 with $p = 10$. The accuracy positively correlated with the number of variables p. The results were consistent in accuracy and the differences between two methods were negligible with large $p = 20$ in which the performances of both methods were almost perfect.

VII. DATA EXAMPLE

A. PROSTATE CANCER DATA

Firstly, we considered the prostate cancer data introduced by [30]. The data consist of 12 mixed-type variables. To be specific, eight variables (Age, Diastolic blood pressure, Index of tumor stage and histologic grade, Serum hemoglobin, Serum prostatic acid phosphatase, Size of the primary tumor, Systolic blood pressure, Weight) are continuous, three variables (Bone metastases, Cardiovascular disease history, and performance rating) are ordinal, and only one variable (Electrocardiogram code) is nominal. We preprocessed 12 variables, as in [11]. Some exploration of the correlation structure showed dependency across variables. For example, the absolute values of the correlation coefficients ranged from 0.011 to 0.797 in the correlation matrix of the data. In addition, the range of the inverse covariance matrix of the prostate cancer data is from 0.000 to 1.461.

As in Section VI, we fitted the proposed method and ClustMD with a BIC-based model selection. A line plot of the BIC values estimated using our proposed method (the

TABLE 6. Different number of variables p: clustering accuracy summarized by mean (Mean) and standard error (S.E.) $\times 10^3$, and the number of simulation failures (# failures) over 100 replications with $\mu_2 = 1_p$.

(a)

ρ	Accuracy	$p = 5$			
		AR(1) model		Random model	
		clustMD	regclustMD	clustMD	regclustMD
0.0	Mean	0.777	0.760	0.777	0.760
	S.E.$\times 10^3$	0.065	0.082	0.065	0.082
	# failures	0	0	0	0
0.2	Mean	0.740	0.748	0.713	0.716
	S.E.$\times 10^3$	0.097	0.097	0.081	0.100
	# failures	0	0	3	0
0.4	Mean	0.696	0.674	0.667	0.697
	S.E.$\times 10^3$	0.096	0.120	0.081	0.109
	# failures	2	0	1	0
0.6	Mean	0.648	0.633	0.702	0.698
	S.E.$\times 10^3$	0.085	0.112	0.048	0.070
	# failures	0	0	4	0
0.8	Mean	0.621	0.626	0.693	0.733
	S.E.$\times 10^3$	0.078	0.103	0.066	0.081
	# failures	2	0	8	0

(b)

ρ	Accuracy	$p = 20$			
		AR(1) model		Random model	
		clustMD	regclustMD	clustMD	regclustMD
0.0	Mean	0.978	0.977	0.978	0.977
	S.E.$\times 10^3$	0.016	0.017	0.016	0.017
	# failures	0	0	0	0
0.2	Mean	0.992	0.994	0.953	0.988
	S.E.$\times 10^3$	0.009	0.008	0.055	0.033
	# failures	0	0	3	0
0.4	Mean	0.999	0.999	0.897	0.980
	S.E.$\times 10^3$	0.004	0.002	0.089	0.049
	# failures	0	0	1	0
0.6	Mean	1.000	1.000	0.889	0.965
	S.E.$\times 10^3$	0.000	0.000	0.096	0.067
	# failures	0	0	3	1
0.8	Mean	1.000	1.000	0.878	0.955
	S.E.$\times 10^3$	0.000	0.000	0.111	0.087
	# failures	0	0	3	0

left panel of Figure 1) reveals that the two-cluster model ($G = 2$) leads to the minimum BIC value. The right panel of Figure 1 shows the estimated group means for the ClustMD with $G = 2$. The patients in group 2 have a larger primary tumor size, higher serum prostatic acid phosphatase levels, higher tumor stage and histologic grade index, and lower serum hemoglobin levels than those of group 1.

In addition, we estimated the existing validation indices that are widely used in the cluster analysis introduced by [31]. In other words, we calculated eight validation indices to measure the quality of clusters (*C, Dunn, Gamma, Gplus, Mcclain, Ptbiserial, Silhouette, Tau*) in the ClustMD method with three clusters (selected by the BIC) and regClustMD method with two clusters to compare their results. Table 7 summarizes the cluster results for each method in terms of eight cluster validity indices. Six of the eight performance measures represent that our proposed method with two clusters fits better than ClustMD with three clusters.

B. AUSTRALIAN INSTITUTE OF SPORTS DATA

We also considered the Australian Institute of Sports (AIS) data [32]. The data contains 202 observations with 13 mixed-type variables. The variables consist of nine continuous

TABLE 7. The comparison of cluster validity indices calculated from the clustering results on the prostate cancer data.

Index	ClustMD ($G = 3$)	regClustMD ($G = 2$)
C	**0.2568**	0.2782
Dunn	**0.0091**	0.0086
Gamma	0.4043	**0.4649**
Gplus	0.1425	**0.1280**
Mcclain	0.6943	**0.6751**
Ptbiserial	0.2976	**0.3458**
Silhouette	0.1927	**0.2634**
Tau	0.2796	**0.3216**

* Note: The optimal criterion of three indices (*C, Gplus, Mcclain*) is minimum value. The rest of the eight indices is the maximum value. The bold numbers represent better performances compared with the other method.

TABLE 8. The comparison of cluster validity indices calculated from the clustering results on the AIS data.

Index	ClustMD ($G = 3$)	regClustMD ($G = 2$)
C	**0.1479**	0.1619
Dunn	0.0221	**0.0480**
Gamma	0.6404	**0.6460**
Gplus	**0.0862**	0.0885
Mcclain	**0.4718**	0.4831
Ptbiserial	0.4659	**0.4946**
Silhouette	0.3670	**0.4974**
Tau	0.4435	**0.4569**

* Note: The optimal criterion of three indices (*C, Gplus, Mcclain*) is minimum value. The rest of the eight indices is the maximum value. The bold numbers represent better performances compared with the other method.

(Red blood cell count, White blood cell count, Hematocrit, Hemoglobin concentration, Plasma ferritins, Body mass index, Sum of skin folds, Percent body fat, Lean body mass), one binary (Sex), and one nominal variable (Sport), respectively. We recategorized the 'Sport' variable into three categories for computational efficiency. That is, we merged basketball, Netball, and Tennis as the first group, waterpolo and Rowing as the second group, and the rest of the nine sports as the third group, respectively. A total of 13 variables revealed pairwise dependency showing that absolute values of the correlation coefficients ranged from 0.080 to 0.964.

Two cluster model in our proposed method with a minimum BIC value was selected (Figure 2). The right panel of Figure 2 represents that the patients in group 2 have higher lean body mass, higher body mass index, higher plasma ferritins, higher hemoglobin concentration, higher hematocrit, higher white blood cell count, higher red blood cell count, lower percent body fat, lower sum of skin folds, and are more likely to be females than those of group 1, respectively.

We measured the eight validation indices that measure the quality of clustering to compare the ClustMD result with three clusters which were also selected by the BIC value and that of regClustMD with two clusters. As summarized in Table 8, five of the eight indices illustrate that our proposed regClustMD with two clusters outperforms the clustMD with two clusters.

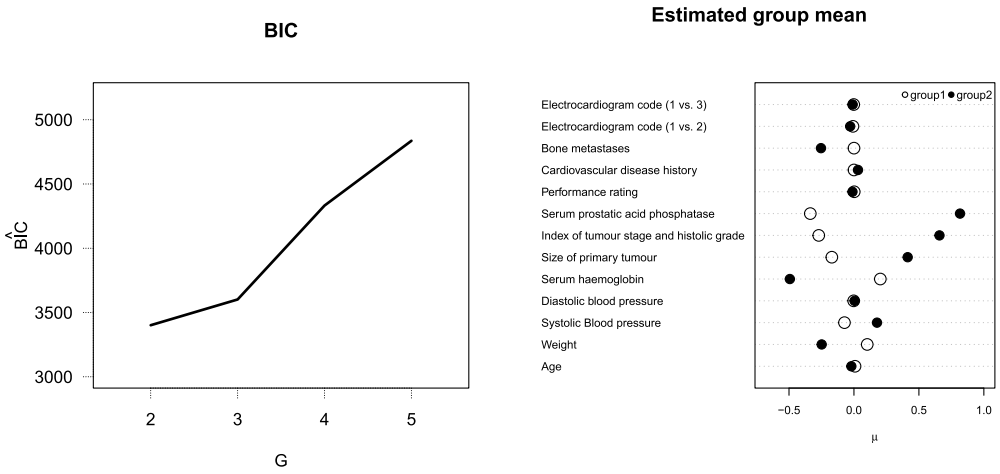


FIGURE 1. Left panel, Line plot of the number of groups (G) versus BIC. For each group, only the result for best λ is reported; Right panel, the estimated group mean ($\widehat{\mu}_1, \widehat{\mu}_2$) in the proposed regClustMD method.

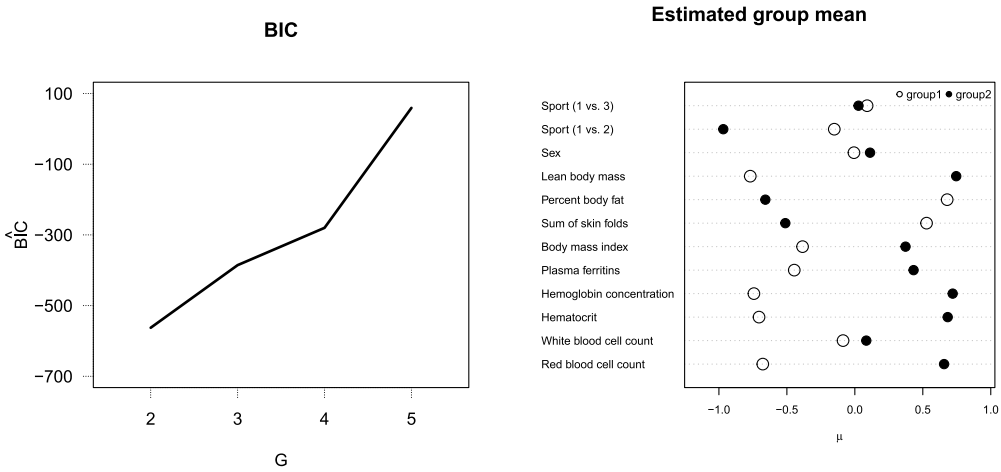


FIGURE 2. Left panel, Line plot of the number of groups (G) versus BIC. For each group, only the result for best λ is reported; Right panel, the estimated group mean ($\widehat{\mu}_1, \widehat{\mu}_2$) in the proposed regClustMD method.

VIII. CONCLUDING REMARKS

It is well known that modeling the correlation structure between variables improves clustering results. Most of the data were correlated regardless of the data type. In this study, we proposed the regClustMD algorithm, which is a model-based clustering method for mixed data, that can address sparse dependence between variables. Our probability model postulates that categorical variables are generated from latent continuous variables before categorization. We then considered a sparse latent Gaussian mixture model. Through simulation studies and prostate cancer data analysis, we showed that regClustMD outperformed existing approaches that do not address dependence.

Our study considered the l_1-penalty in regularization for implemental simplicity. However, the l_1-penalty generally produces biased estimates, and other nonconvex penalties such as SCAD and MCP may be beneficial for mitigating the bias. Another future direction is to accelerate the proposed algorithm by parallelization because many computation routines in our algorithm can be conducted in parallel. Since our method relies on the assumption of a Gaussian distribution, it may be susceptible to outliers. One systematic approach to handling outliers is by developing a latent t-mixture model suitable for mixed data. An example of such a model is presented by [16] where a multivariate t-mixture model is proposed for clustering continuous-type data.

ACKNOWLEDGMENT

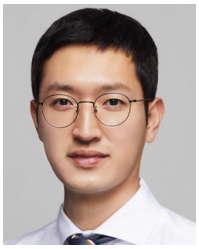
(Young-Geun Choi and Soohyun Ahn are co-first authors.)

REFERENCES

[1] A. Ahmad and S. S. Khan, "Survey of state-of-the-art mixed data clustering algorithms," *IEEE Access*, vol. 7, pp. 31883–31902, 2019.

[2] G. Preud'homme, K. Duarte, K. Dalleau, C. Lacomblez, E. Bresso, M. Smaïl-Tabbone, M. Couceiro, M.-D. Devignes, M. Kobayashi, O. Huttin, J. P. Ferreira, F. Zannad, P. Rossignol, and N. Girerd, "Head-to-head comparison of clustering methods for heterogeneous data: A simulation-driven benchmark," *Sci. Rep.*, vol. 11, no. 1, p. 4202, Feb. 2021.

[3] C. Fraley, "How many clusters? Which clustering method? Answers via model-based cluster analysis," *Comput. J.*, vol. 41, no. 8, pp. 578–588, Aug. 1998.
[4] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *J. Amer. Stat. Assoc.*, vol. 97, no. 458, pp. 611–631, Jun. 2002.
[5] S. Cagnone and C. Viroli, "A factor mixture analysis model for multivariate binary data," *Stat. Model.*, vol. 12, no. 3, pp. 257–277, Jun. 2012.
[6] Y. Tang, R. P. Browne, and P. D. McNicholas, "Model based clustering of high-dimensional binary data," *Comput. Statist. Data Anal.*, vol. 87, pp. 84–101, Jul. 2015.
[7] M. Ranalli and R. Rocci, "Mixture models for ordinal data: A pairwise likelihood approach," *Statist. Comput.*, vol. 26, nos. 1–2, pp. 529–547, Jan. 2016.
[8] I. Morlini, "A latent variables approach for clustering mixed binary and continuous variables within a Gaussian mixture model," *Adv. Data Anal. Classification*, vol. 6, no. 1, pp. 5–28, Apr. 2012.
[9] I. Gollini and T. B. Murphy, "Mixture of latent trait analyzers for model-based clustering of categorical data," *Statist. Comput.*, vol. 24, no. 4, pp. 569–588, Jul. 2014.
[10] D. McParland, I. C. Gormley, T. H. McCormick, S. J. Clark, C. W. Kabudula, and M. A. Collinson, "Clustering South African households based on their asset status using latent variable models," *Ann. Appl. Statist.*, vol. 8, no. 2, pp. 747–776, Jun. 2014.
[11] D. McParland and I. C. Gormley, "Model based clustering for mixed data: clustMD," *Adv. Data Anal. Classification*, vol. 10, no. 2, pp. 155–169, Jun. 2016.
[12] H. Zhou, W. Pan, and X. Shen, "Penalized model-based clustering with unconstrained covariance matrices," *Electron. J. Statist.*, vol. 3, pp. 1473–1496, Jan. 2009.
[13] L. Ruan, M. Yuan, and H. Zou, "Regularized parameter estimation in high-dimensional Gaussian mixture models," *Neural Comput.*, vol. 23, no. 6, pp. 1605–1622, Jun. 2011.
[14] M.-S. Yang, S.-J. Chang-Chien, and Y. Nataliani, "Unsupervised fuzzy model-based Gaussian clustering," *Inf. Sci.*, vol. 481, pp. 1–23, May 2019.
[15] M.-S. Yang and W. Ali, "Fuzzy Gaussian lasso clustering with application to cancer data," *Math. Biosci. Eng.*, vol. 17, no. 1, pp. 250–265, 2020.
[16] W. Ali, M.-S. Yang, M. Ali, and S. Ud-Din, "Fuzzy model-based sparse clustering with multivariate t-mixtures," *Appl. Artif. Intell.*, vol. 37, no. 1, Dec. 2023, Art. no. 2169299.
[17] M. Fop and T. B. Murphy, "Variable selection methods for model-based clustering," *Statist. Surv.*, vol. 12, pp. 18–65, Jan. 2018.
[18] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 6th ed. London, U.K.: Pearson, 2007.
[19] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
[20] K. P. Sinaga, I. Hussain, and M.-S. Yang, "Entropy k-means clustering with feature reduction under unknown number of clusters," *IEEE Access*, vol. 9, pp. 67736–67751, 2021.
[21] M.-S. Yang and I. Hussain, "Unsupervised multi-view k-means clustering algorithm," *IEEE Access*, vol. 11, pp. 13574–13593, 2023.
[22] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining Knowl. Discovery*, vol. 2, no. 3, pp. 283–304, 1998.
[23] Z. I. Botev, "The normal law under linear restrictions: Simulation and estimation via minimax tilting," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 79, no. 1, pp. 125–148, Jan. 2017.
[24] B. G. Manjunath and S. Wilhelm, "Moments calculation for the doubly truncated multivariate normal density," 2012, *arXiv:1206.5387*.
[25] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, Jul. 2008.
[26] C.-J. Hsieh, "QUIC: Quadratic approximation for sparse inverse covariance estimation," *J. Mach. Learn. Res.*, vol. 15, no. 83, pp. 2911–2947, 2014.
[27] S. Ahn, H. Choi, J. Lim, and K. E. Lee, "Self-semi-supervised clustering for large scale data with massive null group," *J. Korean Stat. Soc.*, vol. 49, no. 1, pp. 161–176, Mar. 2020.
[28] H. Zou, T. Hastie, and R. Tibshirani, "On the 'degrees of freedom' of the lasso," *Ann. Statist.*, vol. 35, no. 5, pp. 2173–2192, 2007.
[29] (2022). *R: A language and Environment for Statistical Computing. R Foundation for Statistical Computing*. Vienna, Austria: R Core Team. [Online]. Available: https://www.R-project.org/
[30] D. P. Byar and S. B. Green, "The choice of treatment for cancer patients based on covariate information," *Bull. du Cancer*, vol. 67, no. 4, pp. 477–490, 1980.
[31] M. Halkidi, M. Vazirgiannis, and C. Hennig, "Method-independent indices for cluster validation and estimating the number of clusters," in *Handbook of Cluster Analysis*. London, U.K.: Chapman & Hall, 2015, pp. 595–618.
[32] R. D. Telford and R. B. Cunningham, "Sex, sport, and body-size dependency of hematology in highly trained athletes," *Med. Sci. Sports Exerc.*, vol. 23, no. 7, pp. 788–794, 1991.



YOUNG-GEUN CHOI (Member, IEEE) received the B.S. degree in mathematics education and the Ph.D. degree in statistics from Seoul National University, Seoul, Republic of Korea, in 2010 and 2015, respectively.

From 2016 to 2018, he was a Postdoctoral Fellow with the Public Health Division, Fred Hutchinson Cancer Research Center. From 2018 to 2019, he was a Research Scientist with Data Labs, SK Telecom. From 2019 to 2023, he was an Assistant Professor with the Department of Statistics, Sookmyung Women's University. Since 2023, he has been an Assistant Professor with the Mathematics Education Department, Sungkyunkwan University, Seoul. His research interests include developing statistical and machine learning methods for multivariate data analysis, causal inference, multi-armed bandits, and dynamic pricing.

SOOHYUN AHN received the Ph.D. degree in statistics from Seoul National University, Seoul, Republic of Korea, in 2016.

From 2016 to 2018, she was a Senior Researcher with the Statistics and Data Center, Samsung Medical Center. From 2018 to 2021, she was an Assistant Professor with the Department of Mathematics, Ajou University, Suwon, Republic of Korea. Since 2022, she has been an Associate Professor with the Department of Mathematics, Ajou University. Her research interests include developing statistical methods for multivariate data analysis, large-scale mass-spectrometry data analysis, and order related statistical inference and analyzing medical data.

JAYOUN KIM received the Ph.D. degree in statistics from the Department of Applied Statistics, Yonsei University, in 2011. She is currently a Research Professor with the Medical Research Collaborating Center, Seoul National University Hospital, Seoul, Republic of Korea. Her research interests include survival analysis with rare events, multi-state models, extended joint models, and statistical methods in medical fields.