Journal of Classification (2023) 40:588–613
https://doi.org/10.1007/s00357-023-09449-9

Clustered Sparse Structural Equation Modeling for Heterogeneous Data

Ippei Takasawa[1] · Kensuke Tanioka[2] · Hiroshi Yadohisa[3]

Accepted: 15 August 2023 / Published online: 30 November 2023

Abstract
Joint analysis with clustering and structural equation modeling is one of the most popular approaches to analyzing heterogeneous data. The methods involved in this approach estimate a path diagram of the same shape for each cluster and interpret the clusters according to the magnitude of the coefficients. However, these methods have problems with difficulty in interpreting the coefficients when the number of clusters and/or paths increases and are unable to deal with any situation where the path diagram for each cluster is different. To tackle these problems, we propose two methods for simplifying the path structure and facilitating interpretation by estimating a different form of path diagram for each cluster using sparse estimation. The proposed methods and related methods are compared using numerical simulation and real data examples. The proposed methods are superior to the existing methods in terms of both fitting and interpretation.

Keywords Clustering · Factor analysis · Path diagram · SEM

1 Introduction

In recent years, various types of multivariate data have become available for analysis. Structural Equation Modeling (SEM;Bollen 1989; Heck and Thomas 2015; Smid et al. 2020) is a method that can reveal the relationships between observed variables and latent variables in the data. SEM is an effective method in terms of visualization as it can be visualized using a path diagram and is used in many fields, such as, marketing, psychology, education, and so on, because it estimates relationships among variables and a path structure (causal structure) behind the data. Confirmatory factor analysis would be formulated as an SEM model because it estimates the size of each arrow in a path diagram after assuming a causal structure based on the prior knowledge and experience of analysts and/or researchers. There are two approaches in SEM, namely, covariance-based and component-based approaches (Hair Jr. et al., 2017;

✉ Kensuke Tanioka
ktanioka@mail.doshisha.ac.jp

[1] Graduate School of Culture and Information Science, Doshisha University, Kyoto, Japan

[2] Department of Biomedical Sciences and Informatics, Doshisha University, Kyoto, Japan

[3] Department of Culture and Information Science, Doshisha University, Kyoto, Japan

Reinartz et al., 2009). The covariance-based approach is also called covariance structure analysis and estimates parameters in a model such that the sample covariance matrix of the data and the covariance matrix based on the model are closer. However, when applying the component-based approach, it is assumed that the components comprise observed variables, such as principal component analysis (PCA), which estimates parameters by minimizing the error between observed variables and the component; the typical example of this is the partial least square SEM (PLSSEM;Chin 1998).

When SEM is applied to multivariate data, the individuals are obtained not only from the same population, but also from some clusters that potentially exist. The data obtained in this case are called heterogeneous data. For example, consumer belief structures are assumed to be different for different market segments by the expectancy-value model (Bagozzi, 1982), and the decision-making process for brand choice may differ among different groups of consumers (Kamakura et al., 1996). When SEM is applied to the heterogeneous data, there may be a risk of bias in the estimation because a single-path structure is estimated even when different path structures potentially exist for each cluster. To overcome this problem, one approach is to consider the cluster structure behind the data and simultaneously perform clustering and estimate the size of each arrow in SEM (Fordellone & Vichi, 2020; Hwang et al., 2007; Jedidi et al., 1997a, b). We refer to the methods based on this approach as clustered SEM.Clustered SEM considers the heterogeneity of the data, even when the cluster structure from the data is not known beforehand. Moreover, they have the benefit that cluster features are grasped visually by the path structures. In clustered SEM on a covariance-based approach, mixture SEM (MSEM;Jedidi et al., 1997a, b; Muthén and Shedden 1999), which assumes a Gaussian mixture distribution for SEM, is proposed. In that case, the component-based approach, Fuzzy clusterwise GSCA (FCGSCA;Hwang et al., 2007), which combines Generalized structured component analysis (GSCA) and is an extension of PLSSEM and fuzzy clustering, and a method combining PLSSEM and Kmeans clustering (Fordellone and Vichi, 2020), is proposed. However, there are two problems in the existing clustered SEM. First, the shape of the path diagram for each cluster is assumed to be the same, and the clusters are interpreted according to the size of the path coefficients only, which makes it difficult to interpret and compare clusters as the number of clusters increases. Second, the existing methods estimate the coefficients in the same way for all clusters, even if the shape of a reasonable path diagram is different for each cluster. Nevertheless, owing to the characteristics of applying SEM, it is not feasible to assume different path diagrams for each cluster in advance.

Therefore, in this study, we propose a clustered sparse SEM to grasp different path structures for each cluster and interpret the path diagram for each cluster more easily using sparse estimation. Concretely, there are three advantages to clustered sparse SEM. First, even if the number of clusters and that of variables are larger, it is easy to interpret the estimated coefficients compared with non-sparse clustered SEM because some coefficients are estimated as zero. Second, in clustered sparse SEM, it is easy to interpret the distinctive feature of each cluster because each path diagram can be estimated differently by each cluster, while non-sparse clustered SEM cannot provide such path structures. Generally, it is natural that these clusters have different path structures. From that, it can be considered that the assumption of sparse SEM is relaxed from that of non-sparse SEM. Third, sparse estimation also eliminates relationships among variables that do not affect the cluster structure; thus, only the relationships that affect clustering remain and clustering accuracy is improved (Pan & Shen, 2007; Xie et al., 2010). The purpose of this study is to propose an extension of the existing methods via two approaches by comparing and reviewing them. The first approach is the extension of sparse SEM (SSEM) to heterogeneous data. In Huang (2018) and Liang and

Jacobucci (2020), sparse SEM for multi-group datasets, which have labels dividing individuals into groups, multi-group sparse SEM (MGSSEM), was proposed. We propose partitioning sparse SEM (PS-SEM), which can be used to estimate cluster-specific path structures, by incorporating Kmeans clustering into MGSSEM in the known group,

.In MGSSEM, the causal structure specific to each group and the causal structure common to all groups can be expressed. Similarly, in PS-SEM, the causal structure specific to each cluster and the causal structure common to all clusters can be expressed. The second approach is the extension of MSEM by sparse estimation. Mixture models using sparse estimation have been proposed, for example, Galimberti et al. (2009) and Xie et al. (2010) in the framework of factor analysis and Fop et al. (2019); Pan and Shen (2007) and Zhou et al. (2009) in the framework of clustering. In this paper, we propose a mixture sparse SEM (MS-SEM) as an extension of MSEM based on sparse estimation. In addition, MSEM combines the EM algorithm with the conjugate gradient method (Powell, 1977), which slows down the convergence speed, while MS-SEM does not use the conjugate gradient method, which has the advantage of faster convergence.

This paper is organized as follows: in Sect. 2, the proposed methods, PS-SEM and MS-SEM, are described; in Sect. 3, numerical simulations and their results are relayed to compare and investigate the performance of the proposed and existing methods; in Sect. 4, the evaluation and comparison of the proposed methods are relayed, while applying real data and some interpretation examples are provided; and in Sect. 5, we summarize and discuss this paper.

2 Proposed Method

In this section, we describe the proposed PS-SEM and MS-SEM methods. Both can perform sparse estimation of SEM and clustering, simultaneously. As a result, they obtain a different causal structure for each cluster and interpretation of each cluster becomes easier. PS-SEM is an extension of MGSSEM and can group without the group labels. MS-SEM is a mixture SEM with sparse estimation. The model is the same for both methods, and their parameters are estimated using the EM algorithm. In parameter estimation, iterative conditional fitting is used to estimate the covariance matrix of variables.

Before describing the proposed methods, we introduce some notations. Given a vector $\boldsymbol{x} \in \mathbb{R}^{P\times 1}$ and matrix $\boldsymbol{Y} \in \mathbb{R}^{N\times P}$, the notations are defined in Table 1.

2.1 The Model

Let $\boldsymbol{V} \in \mathbb{R}^{N\times P}$ denote the data matrix and G denote the number of clusters. Then, the model for PS-SEM and MS-SEM can be defined as

$$\boldsymbol{\eta}_g = \boldsymbol{\alpha}_g + \boldsymbol{B}_g\boldsymbol{\eta}_g + \boldsymbol{\zeta}_g, \tag{1}$$

where

$$\boldsymbol{\eta}_g = \begin{pmatrix} \boldsymbol{v}_g \\ \boldsymbol{f}_g \end{pmatrix}, \boldsymbol{\alpha}_g = \begin{pmatrix} \boldsymbol{\alpha}_g^{(v)} \\ \boldsymbol{\alpha}_g^{(f)} \end{pmatrix}, \ \boldsymbol{B}_g = \begin{pmatrix} \boldsymbol{B}_g^{(vv)} & \boldsymbol{B}_g^{(vf)} \\ \boldsymbol{B}_g^{(fv)} & \boldsymbol{B}_g^{(ff)} \end{pmatrix}, \ \boldsymbol{\zeta}_g = \begin{pmatrix} \boldsymbol{\zeta}_g^{(v)} \\ \boldsymbol{\zeta}_g^{(f)} \end{pmatrix}.$$

$\boldsymbol{\eta}_g \in \mathbb{R}^{(P+M)\times 1}$ is a random vector composed of the observed variable vector $\boldsymbol{v}_g \in \mathbb{R}^{P\times 1}$ and the latent variable vector $\boldsymbol{f}_g \in \mathbb{R}^{M\times 1}$ for cluster $g = 1, 2, \ldots, G$. $\boldsymbol{v}_g$ has different meanings for PS-SEM and MS-SEM. In PS-SEM, $\boldsymbol{v}_g$ denotes the observed variables for

Table 1 Notation table

Notation	Description
$x_j,\ \boldsymbol{x}[j]$	j th element of $\boldsymbol{x}$
$\boldsymbol{x}[-j]$	Vector with j th elements deleted from $\boldsymbol{x}$, $\boldsymbol{x}[-j] \in \mathbb{R}^{(P-1)\times 1}$
$y_{jk},\ Y[j,\ k]$	$(j,\ k)$ element of $\boldsymbol{Y}$
$\boldsymbol{Y}[j,\ ]$	j th row of $\boldsymbol{Y}$, $\boldsymbol{Y}[j,\ ] \in \mathbb{R}^{1\times P}$
$\boldsymbol{Y}[,\ k]$	k th column of $\boldsymbol{Y}$, $\boldsymbol{Y}[,\ k] \in \mathbb{R}^{N\times 1}$
$\boldsymbol{Y}[-j,\ ]$	Submatrix with j th row deleted from $\boldsymbol{Y}$, $\boldsymbol{Y}[-j,\ ] \in \mathbb{R}^{(N-1)\times P}$
$\boldsymbol{Y}[,\ -k]$	Submatrix with k th column deleted from $\boldsymbol{Y}$, $\boldsymbol{Y}[,\ -k] \in \mathbb{R}^{N\times (P-1)}$
$\boldsymbol{Y}[-j,\ -k]$	Submatrix with j th row and k th column deleted from $\boldsymbol{Y}$, $\boldsymbol{Y}[-j,\ -k] \in \mathbb{R}^{(N-1)\times (P-1)}$
$\boldsymbol{Y}[j,\ -k]$	Subvector with k th element deleted from $\boldsymbol{Y}[j,\ ]$, $\boldsymbol{Y}[j,\ -k] \in \mathbb{R}^{1\times (P-1)}$
$\boldsymbol{Y}[-j,\ k]$	Subvector with j th element deleted from $\boldsymbol{Y}[,\ k]$, $\boldsymbol{Y}[-j,\ k] \in \mathbb{R}^{(N-1)\times 1}$

individuals belonging to cluster g, whereas in MS-SEM, $\boldsymbol{v}_g$ is assumed to be distributed from a normal distribution corresponding to the cluster g. $\boldsymbol{\alpha}_g \in \mathbb{R}^{(P+M)\times 1}$ is the intercept vector comprising an intercept vector of observed variables $\boldsymbol{\alpha}_g^{(v)} \in \mathbb{R}^{P\times 1}$ and that of latent variables $\boldsymbol{\alpha}_g^{(f)} \in \mathbb{R}^{M\times 1}$. $\boldsymbol{B}_g \in \mathbb{R}^{(P+M)\times(P+M)}$ is a coefficient matrix that is partitioned into four coefficient matrices; $\boldsymbol{B}_g^{(vv)} \in \mathbb{R}^{P\times P}$ is a coefficient matrix that describes the relations from the observed variables to observed variables, $\boldsymbol{B}_g^{(vf)} \in \mathbb{R}^{P\times M}$ from latent variables to observed variables, $\boldsymbol{B}_g^{(fv)} \in \mathbb{R}^{M\times P}$ from observed variables to latent variables, and $\boldsymbol{B}_g^{(ff)} \in \mathbb{R}^{M\times M}$ is from latent variables to latent variables. diag$(\boldsymbol{B}_g)$ is set to $\boldsymbol{0}_{P+M}$ because the relations from the observed (latent) variable to itself are not considered in almost SEM applications. $\boldsymbol{\zeta}_g \in \mathbb{R}^{(P+M)\times 1}$ is also the residual vector comprising $\boldsymbol{\zeta}_g^{(v)} \in \mathbb{R}^{P\times 1}$ and $\boldsymbol{\zeta}_g^{(f)} \in \mathbb{R}^{M\times 1}$.

It is assumed that $\boldsymbol{\zeta}_g$ is distributed from the multivariate normal distribution $\mathcal{N}(0_{P+M}, \Phi_g)$, in which $\boldsymbol{\zeta}_g^{(v)}$ and $\boldsymbol{\zeta}_g^{(f)}$ are independent. That is, $\boldsymbol{\Phi}_g \in \mathbb{R}^{(P+M)\times(P+M)}$ is the covariance matrix of $\boldsymbol{\zeta}_g$ and can be written as

$$\boldsymbol{\Phi}_g = \begin{pmatrix} \boldsymbol{\Phi}_g^{(vv)} & \boldsymbol{\Phi}_g^{(vf)} \\ \boldsymbol{\Phi}_g^{(fv)} & \boldsymbol{\Phi}_g^{(ff)} \end{pmatrix},$$

where $\boldsymbol{\Phi}^{(vv)} = \mathrm{Cov}(\boldsymbol{\zeta}^{(v)},\ \boldsymbol{\zeta}^{(v)})$, $\boldsymbol{\Phi}^{(vf)} = \mathrm{Cov}(\boldsymbol{\zeta}^{(v)},\ \boldsymbol{\zeta}^{(f)}) = \boldsymbol{\Phi}^{(fv)\top}$, $\boldsymbol{\Phi}^{(ff)} = \mathrm{Cov}(\boldsymbol{\zeta}^{(f)},\ \boldsymbol{\zeta}^{(f)})$, and $\boldsymbol{\Phi}_g^{(vf)} = \boldsymbol{O}_{P\times M} = \boldsymbol{\Phi}_g^{(fv)\top}$.

Under the assumptions, expectation and covariance matrices of $\boldsymbol{\eta}_g$ can be written as

$$\boldsymbol{\mu}_g^{(\eta)} = (\boldsymbol{I}_{P+M} - \boldsymbol{B}_g)^{-1}\boldsymbol{\alpha}_g, \quad \text{and} \quad \boldsymbol{\Sigma}_g^{(\eta\eta)} = (\boldsymbol{I}_{P+M} - \boldsymbol{B}_g)^{-1}\boldsymbol{\Phi}_g\{(\boldsymbol{I}_{P+M} - \boldsymbol{B}_{\boldsymbol{g}})^{-1}\}^\top, \tag{2}$$

respectively. As $\boldsymbol{v}_g = (\boldsymbol{I}_P,\ \boldsymbol{O}_{P\times M})\boldsymbol{\eta}_g$, the expectation and covariance matrix of $\boldsymbol{v}_g$ and covariance matrix btween $\boldsymbol{v}_g$ and $\boldsymbol{\eta}_g$ can also be written as

$$\begin{aligned} &\boldsymbol{\mu}_g^{(v)} = (\boldsymbol{I}_P,\ \boldsymbol{O}_{P\times M})\boldsymbol{\mu}_g^{(\eta)}, \quad \boldsymbol{\Sigma}_g^{(vv)} = (\boldsymbol{I}_P,\ \boldsymbol{O}_{P\times M})\boldsymbol{\Sigma}_g^{(\eta\eta)}(\boldsymbol{I}_P,\ \boldsymbol{O}_{P\times M})^\top, \\ &\boldsymbol{\Sigma}_g^{(v\eta)} = \boldsymbol{\Sigma}_g^{(\eta v)\top} = (\boldsymbol{I}_P,\ \boldsymbol{O}_{P\times M})\boldsymbol{\Sigma}_g^{(\eta\eta)}. \end{aligned} \tag{3}$$

2.2 Partitioning Sparse SEM

PS-SEM can consider parameters for both common effects for all clusters and cluster-specific effects for each cluster. Nevertheless, MGSSEM (Huang, 2018; Lindstrøm and Dahl, 2020) can consider them for both group effects for all groups and group-specific effects for each group. However, in many practical situations, the group structure is unknown. Therefore, we extended the MGSSEM to address this situation. The parameter estimation of PS-SEM can be performed in the same way as MGSSEM.

Let $\boldsymbol{\Theta}_{\text{KMSSEM}}$ denote the parameter space of the PS-SEM. Then, the parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\text{KMSSEM}}$ is defined as

$$\boldsymbol{\theta} = \left(\underline{\boldsymbol{\theta}}^\top,\ \underline{\boldsymbol{\theta}}_1^\top,\ \underline{\boldsymbol{\theta}}_2^\top,\ \ldots,\ \underline{\boldsymbol{\theta}}_G^\top\right)^\top \in \mathbb{R}^{(G+1)Q\times 1},$$

where $\underline{\boldsymbol{\theta}} = (\underline{\theta}_1,\ \underline{\theta}_2,\ \ldots,\ \underline{\theta}_Q)^\top \in \mathbb{R}^{Q\times 1}$ is the parameter for effects for all clusters. Specifically, $\left(\underline{\boldsymbol{\alpha}}^\top,\ \mathrm{vec}(\underline{\boldsymbol{B}})^\top,\ \mathrm{vec}(\underline{\boldsymbol{\Phi}})^\top\right)^\top$. $\underline{\boldsymbol{\theta}}_g = (\underline{\theta}_{g1},\ \underline{\theta}_{g2},\ \ldots,\ \underline{\theta}_{gQ})^\top \in \mathbb{R}^{Q\times 1}$ is the parameter for cluster-specific effects. Specifically, $\left(\underline{\boldsymbol{\alpha}}_g^\top,\ \mathrm{vec}(\underline{\boldsymbol{B}}_g)^\top,\ \mathrm{vec}(\underline{\boldsymbol{\Phi}}_g)^\top\right)^\top$. We call each effect a common effect and cluster effect. Then, we assume that $\boldsymbol{\theta}_g$, which includes each element of $\boldsymbol{\alpha}_g,\ \boldsymbol{B}_g$, and $\boldsymbol{\Phi}_g$, is described as $\boldsymbol{\theta}_g = \underline{\boldsymbol{\theta}} + \underline{\boldsymbol{\theta}}_g$.

The objective function of PS-SEM is defined as

$$\mathcal{U}_{\text{KMSSEM}}(\boldsymbol{\theta},\ \lambda) = \mathcal{L}_{\text{KMSSEM}}(\boldsymbol{\theta}) - \mathcal{R}_{\text{KMSSEM}}(\boldsymbol{\theta}, \lambda), \tag{4}$$

where $\mathcal{L}_{\text{KMSSEM}}(\boldsymbol{\theta})$ is the log-likelihood function, which is similar to the log-likelihood function of the multi-group SEM (Jöreskog, 1971),

$$\mathcal{L}_{\text{KMSSEM}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{g=1}^{G}\sum_{n=1}^{N_g}\left\{-\frac{1}{2}\log|\boldsymbol{\Sigma}_g^{(vv)}| - \frac{1}{2}\left(\boldsymbol{m}_g - \boldsymbol{\mu}_g^{(v)}\right)^\top \left(\boldsymbol{\Sigma}_g^{(vv)}\right)^{-1}\left(\boldsymbol{m}_g - \boldsymbol{\mu}_g^{(v)}\right)\right\}, \tag{5}$$

and $\mathcal{R}_{\text{KMSSEM}}(\boldsymbol{\theta}, \lambda)$ is the penalization term, which is similar to the penalization term of the multi-group sparse SEM (Huang, 2018),

$$\mathcal{R}_{\text{KMSSEM}}(\boldsymbol{\theta}, \lambda) = \sum_{q=1}^{Q} c_{\underline{\theta}_q}\lambda|\underline{\theta}_q| + \sum_{g=1}^{G}\sum_{q=1}^{Q} c_{\underline{\theta}_{gq}}\lambda|\underline{\theta}_{gq}|. \tag{6}$$

N_g denotes the number of individuals belonging to cluster g, and $\boldsymbol{v}_{gn}$ denotes the n th observed variable vector of $\boldsymbol{v}_g$. The first term of $\mathcal{R}_{\text{KMSSEM}}(\boldsymbol{\theta}, \lambda)$ is the penalization term for $\underline{\boldsymbol{\theta}}$, and the second term, the penalization term for $\underline{\boldsymbol{\theta}}_g$ $(g = 1,\ 2,\ \ldots,\ G)$. Here, λ is a regularization parameter. $c_{\underline{\theta}_q}$ and $c_{\underline{\theta}_{gq}}$ are the regularization indicators for $\underline{\theta}_q$ and $\underline{\theta}_{gq}$, respectively, which takes 0 or 1. Using these indicators, we can determine if the parameter is sparsely estimated in advance. If the indicator is 1, the parameter is estimated with penalization, whereas if the indicator is 0, the parameter is estimated without penalization. There are two penalization terms in $\mathcal{R}_{\text{KMSSEM}}(\boldsymbol{\theta}, \lambda)$, that is, a zero or non-zero pattern in θ_{gq}, which means the q th elements of $\boldsymbol{\theta}_g$. θ_{gq} is expressed as follows:

$$\theta_{gq} = \begin{cases} \underline{\theta}_q + \underline{\theta}_{gq} & (\underline{\theta}_q \neq 0,\ \underline{\theta}_{gq} \neq 0) \\ \underline{\theta}_{gq} & (\underline{\theta}_q = 0,\ \underline{\theta}_{gq} \neq 0) \\ \underline{\theta}_q & (\underline{\theta}_q \neq 0,\ \underline{\theta}_{gq} = 0) \\ 0 & (\underline{\theta}_q = 0,\ \underline{\theta}_{gq} = 0) \end{cases}.$$

The first case is that there are both common and cluster effects. The path diagram in this case shows that the coefficients are different for each cluster, although the shape of the path diagram is the same. The second case is that there is only a cluster effect. In this case, it is different for each cluster if the path diagram has an arrow, that is, if each cluster has a differently shaped path diagram. The third case is that there is only a common effect. In this case, both the shape and coefficients are the same for each cluster. The fourth case is that there is neither a common effect nor a cluster effect. In this case, the path diagram of each cluster does not have arrows for the parameter. This means that there was no relationship between these variables. In particularly, our goal that the proposed method should obtain a path diagram of a different shape for each cluster is achieved in the second case.

Log-likelihood Eq. 5 can be transformed as follows:

$$\mathcal{L}_{\mathrm{KMSSEM}}(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{g=1}^{G} w_g \left\{\log|\boldsymbol{\Sigma}^{(vv)}| + \mathrm{tr}(\boldsymbol{\Sigma}^{(vv)}\boldsymbol{S}_g)\right\} - \frac{1}{2}\sum_{g=1}^{G} w_g (\boldsymbol{v}_{gn} - \boldsymbol{\mu}_g^{(v)})^\top (\boldsymbol{\Sigma}_g^{(vv)})^{-1} (\boldsymbol{v}_{gn} - \boldsymbol{\mu}_g^{(v)}), \tag{7}$$

where $w_g = N_g/N$, $\boldsymbol{m}_g = \frac{1}{N_g}\sum_{n=1}^{N_g} \boldsymbol{v}_{gn}$ and $\boldsymbol{S}_g = \frac{1}{N_g}\sum_{n=1}^{N_g}(\boldsymbol{v}_{gn} - \boldsymbol{m}_g)(\boldsymbol{v}_{gn} - \boldsymbol{m}_g)^\top$.

To maximize the objective function, there are parameter estimation and clustering phases. The parameter estimation phase is performed by the EM algorithm (Rubin and Thayer, 1982) and updates the parameters for SEM, especially the M-step. The clustering phase determines the cluster in which the individual belongs.

E-step

By Eq. 4, the completely penalized log-likelihood of the PS-SEM is written as

$$\mathcal{U}^{C}_{\mathrm{KMSSEM}}(\boldsymbol{\theta},\ \lambda) = \frac{1}{N}\sum_{g=1}^{G}\sum_{n=1}^{N_g}\left\{-\frac{1}{2}\log|\boldsymbol{\Phi}_g| - \frac{1}{2}\boldsymbol{\zeta}_{gn}^\top \boldsymbol{\Phi}_g \boldsymbol{\zeta}_{gn}\right\} - \left(\sum_{q=1}^{Q} c_{\underline{\theta}_q}\lambda|\underline{\theta}_q| + \sum_{g=1}^{G}\sum_{q=1}^{Q} c_{\underline{\theta}_{gq}}\lambda|\underline{\theta}_{gq}|\right) \tag{8}$$

Note that Eq. 8 is not a completely penalized likelihood of $\boldsymbol{\eta}_{gn}$ but $\boldsymbol{\zeta}_{gn}$ $(= \boldsymbol{\eta}_g - \boldsymbol{\alpha}_{gn} - \boldsymbol{B}_g\boldsymbol{\eta}_{gn})$, because $\boldsymbol{\eta}_{gn}$ has a determinant of the estimated parameter, which makes it difficult to estimate the parameter. The expectation of $\mathcal{U}^{C}_{\mathrm{KMSSEM}}(\boldsymbol{\theta},\ \lambda)$ is derived from Eq. 8 as follows:

$$\begin{aligned}
&\mathcal{M}_{\mathrm{KMSSEM}}(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)}) \\
=&\mathbb{E}\left[\mathcal{U}^{C}_{\mathrm{KMSSEM}}(\boldsymbol{\theta},\ \lambda) \mid \boldsymbol{V},\ \hat{\boldsymbol{\theta}}^{(t)},\ \hat{w}_g^{(t)}\right] \\
=&-\frac{1}{2}\sum_{g=1}^{G}\frac{\hat{w}_g^{(t)}}{N_g}\sum_{n=1}^{N_g}\left\{\log|\hat{\boldsymbol{\Phi}}_g^{(t)}| + \mathrm{tr}\Big((\boldsymbol{I}_{P+M} - \hat{\boldsymbol{B}}_g^{(t)})^\top (\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}(\boldsymbol{I}_{P+M} - \hat{\boldsymbol{B}}_g^{(t)})\mathbb{E}[\boldsymbol{\eta}_{gn}\boldsymbol{\eta}_{gn}^\top \mid \boldsymbol{V},\ \hat{\boldsymbol{\theta}}^{(t)},\ \hat{w}_g^{(t)}]\Big)\right. \\
&\left. - 2\hat{\boldsymbol{\alpha}}_g^{(t)\top}(\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}(\boldsymbol{I}_{P+M} - \hat{\boldsymbol{B}}_g^{(t)})\mathbb{E}[\boldsymbol{\eta}_{gn} \mid \boldsymbol{V},\ \hat{\boldsymbol{\theta}}^{(t)},\ \hat{w}_g^{(t)}]) + \hat{\boldsymbol{\alpha}}_g^{(t)\top}(\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}\hat{\boldsymbol{\alpha}}_g^{(t)}\right\} \\
&- \Big(\sum_{q=1}^{Q} c_{\underline{\theta}_q}\lambda|\underline{\theta}_q| + \sum_{g=1}^{G}\sum_{q=1}^{Q} c_{\underline{\theta}_{gq}}\lambda|\underline{\theta}_{gq}|\Big),
\end{aligned} \tag{9}$$

where $\hat{\boldsymbol{\theta}}^{(t)}$ is the parameter updated by the t th time update in the parameter estimation phase, and $\hat{w}_g^{(t)}$ is obtained by updating the t th time in the clustering phase.

Now, we define $\boldsymbol{\gamma}_g^{(t+1)} = \mathbb{E}\left[\frac{1}{N_g}\sum_{n=1}^{N_g} \boldsymbol{\eta}_{gn} \mid \boldsymbol{V}, \hat{\boldsymbol{\theta}}^{(t)}\right]$, $\boldsymbol{\Gamma}_g^{(t+1)} = \mathbb{E}\left[\frac{1}{N_g}\sum_{n=1}^{N_g} \boldsymbol{\eta}_{gn}\boldsymbol{\eta}_{gn}^\top \mid \boldsymbol{V}, \hat{\boldsymbol{\theta}}^{(t)}\right]$, and $\boldsymbol{\Psi}_g^{(t+1)} = \mathbb{E}\left[\frac{1}{N_g}\sum_{n=1}^{N_g} \boldsymbol{\zeta}_{gn}\boldsymbol{\zeta}_{gn}^\top \mid \boldsymbol{V}, \hat{\boldsymbol{\theta}}^{(t)}\right]$. Then, they are described as

$$\boldsymbol{\gamma}_g^{(t+1)} = \boldsymbol{c}_g + \boldsymbol{C}_g \boldsymbol{m}_g \tag{10}$$

$$\begin{aligned}\boldsymbol{\Gamma}_g^{(t+1)} =& \hat{\boldsymbol{\Sigma}}_g^{(\eta\eta)(t)} - \left(\hat{\boldsymbol{\Sigma}}_g^{(\eta v)(t)}\right)^{-1} \hat{\boldsymbol{\Sigma}}_g^{(vv)(t)} \hat{\boldsymbol{\Sigma}}_g^{(v\eta)(t)} \\ &+ \boldsymbol{c}_g\boldsymbol{c}_g^\top + \boldsymbol{c}_g\boldsymbol{m}_g^\top\boldsymbol{C}_g^\top + \boldsymbol{C}_g\boldsymbol{m}_g\boldsymbol{c}_g^\top + \boldsymbol{C}_g(\boldsymbol{S}_g + \boldsymbol{m}_g\boldsymbol{m}_g^\top)\boldsymbol{C}_g^\top\end{aligned} \tag{11}$$

$$\begin{aligned}\boldsymbol{\Psi}_g^{(t+1)} =& (\boldsymbol{I}_{P+M} - \hat{\boldsymbol{B}}_g^{(t)})\boldsymbol{\Gamma}_g^{(t+1)}(\boldsymbol{I}_{P+M} - \hat{\boldsymbol{B}}_g^{(t)})^\top - (\boldsymbol{I}_{P+M} - \hat{\boldsymbol{B}}_g^{(t)})\boldsymbol{\gamma}_g^{(t+1)}\hat{\boldsymbol{\alpha}}_g^{(t)\top} \\ &- \hat{\boldsymbol{\alpha}}_g^{(t)}\boldsymbol{\gamma}_g^{(t+1)\top}(\boldsymbol{I}_{P+M} - \hat{\boldsymbol{B}}_g^{(t)})^\top + \hat{\boldsymbol{\alpha}}_g^{(t)}\hat{\boldsymbol{\alpha}}_g^{(t)\top}\end{aligned} \tag{12}$$

where $\boldsymbol{c}_g = \hat{\boldsymbol{\mu}}_g^{(\eta)(t)} - \hat{\boldsymbol{\Sigma}}_g^{(\eta v)(t)}(\hat{\boldsymbol{\Sigma}}_g^{(vv)(t)})^{-1}$ and $\boldsymbol{C}_g = \hat{\boldsymbol{\Sigma}}_g^{(\eta v)(t)}(\hat{\boldsymbol{\Sigma}}_g^{(vv)(t)})^{-1}\hat{\boldsymbol{\mu}}_g^{(v)(t)} + \hat{\boldsymbol{\Sigma}}_g^{(\eta v)(t)}(\hat{\boldsymbol{\Sigma}}_g^{(vv)(t)})^{-1}$.

The E-step is performed by computing Eqs. 10, 11, and 12.

M-step

Updating parameters are achieved by the following:

$$\hat{\underline{\theta}}_q^{(t+1)}(\lambda) = \mathrm{S}\left(\hat{\underline{\theta}}_q^{(t+1)}(0),\ W_{\underline{\theta}_q}^{(t+1)} c_{\underline{\theta}_q}\lambda\right),\quad \hat{\underline{\theta}}_{gq}^{(t+1)}(\lambda) = \mathrm{S}\left(\hat{\underline{\theta}}_{gq}^{(t+1)}(0),\ W_{\underline{\theta}_{gq}}^{(t+1)} c_{\underline{\theta}_{gq}}\lambda\right), \tag{13}$$

where $\mathrm{S}(x,\ y) = \mathrm{sign}(x)\max(|x| - y,\ 0)$, and $W_{\underline{\theta}_q}^{(t+1)}$ and $W_{\underline{\theta}_{gq}}^{(t+1)}$ are weights. $\hat{\underline{\theta}}_q^{(t+1)}(\lambda)$ and $\hat{\underline{\theta}}_{gq}^{(t+1)}(\lambda)$ Stand for the updated estimate on regularization parameter λ. When $\lambda = 0$, they become $\hat{\underline{\theta}}_q^{(t+1)}(0)$ and $\hat{\underline{\theta}}_{gq}^{(t+1)}(0)$, respectively. According to Eq. 13, $\hat{\underline{\theta}}_q^{(t+1)}(0)$, $\hat{\underline{\theta}}_{gq}^{(t+1)}(0)$, $W_{\underline{\theta}_q}^{(t+1)}$, and $W_{\underline{\theta}_{gq}}^{(t+1)}$ need to be updated. Hence, the M-step is done by computing them. They are derived as follows.

Updating formula on $\boldsymbol{B}_g$

In relation to $\boldsymbol{B}_g$, the updating formulas of $\underline{\boldsymbol{B}}$ and $\underline{\boldsymbol{B}}_g$ are given as follows:

$$\begin{aligned}\hat{\underline{b}}_{jk}^{(t+1)}(0) =& W_{\underline{b}_{jk}}^{(t+1)} \sum_{g=1}^{G} \hat{w}_g^{(t)} \Big(\boldsymbol{\Gamma}_g^{(t+1)}[k,\ ](\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}[\ ,\ j] - \boldsymbol{\Gamma}_g^{(t+1)}[k,\ k]\hat{\underline{b}}_{gjk}^{(t)}\hat{\phi}_g^{jj(t)} \\ &- \boldsymbol{\Gamma}_g^{(t+1)}[k,\ k]\hat{\boldsymbol{B}}_g^{(t)}[-j,\ k]^\top(\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}[-j,\ j] - \boldsymbol{\Gamma}_g^{(t+1)}[k,\ -k]\hat{\boldsymbol{B}}_g^{(t)}[\ ,\ -k]^\top(\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}[\ ,\ j] \\ &- \boldsymbol{\gamma}_g^{(t)}[k]\hat{\boldsymbol{\alpha}}_g^{(t)\top}(\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}[\ ,\ j]\Big),\end{aligned} \tag{14}$$

where $W_{\underline{b}_{jk}}^{(t+1)} = 1/\sum_{g=1}^{G} \hat{w}_g^{(t)}\boldsymbol{\Gamma}_g^{(t+1)}[k, k]\hat{\phi}_g^{jj(t)}$, and

$$\begin{aligned}\hat{\underline{b}}_{gjk}^{(t+1)}(0) =& W_{\underline{b}_{gjk}}^{(t+1)} \Big(\boldsymbol{\Gamma}_g^{(t+1)}[k,\ ](\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}[\ ,\ j] - \boldsymbol{\Gamma}_g^{(t+1)}[k,\ k]\hat{\underline{b}}_{jk}^{(t)}\hat{\phi}_g^{jj(t)} \\ &- \boldsymbol{\Gamma}_g^{(t+1)}[k,\ k]\hat{\boldsymbol{B}}_g^{(t)}[-j,\ k]^\top(\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}[-j,\ j] - \boldsymbol{\Gamma}_g^{(t+1)}[k,\ -k]\hat{\boldsymbol{B}}_g^{(t)}[\ ,\ -k]^\top(\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}[\ ,\ j] \\ &- \boldsymbol{\gamma}_g^{(t)}[k]\hat{\boldsymbol{\alpha}}_g^{(t)\top}(\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}[\ ,\ j]\Big),\end{aligned} \tag{15}$$

where $W_{\underline{b}_{gjk}}^{(t+1)} = 1/\boldsymbol{\Gamma}_g^{(t+1)}[k, k]\hat{\phi}_g^{jj(t)}$.

Updating formula on $\boldsymbol{\alpha}_g$

In relation to $\boldsymbol{\alpha}_g$, the updating formulas of $\underline{\boldsymbol{\alpha}}$ and $\underline{\boldsymbol{\alpha}}_g$ are given as follows:

$$\begin{aligned}\hat{\underline{\alpha}}_j^{(t+1)}(0) =& W_{\underline{\alpha}_j}^{(t+1)} \sum_{g=1}^{G} \hat{w}_g^{(t)}\Big((\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}[j,\ ]\boldsymbol{\gamma}_g^{(t)} - (\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}[j,\ ]\hat{\boldsymbol{B}}_g^{(t)}\boldsymbol{\gamma}_g^{(t)} - \hat{\phi}^{jj(t)}\underline{\alpha}_{gj}^{(t)} \\ &- \hat{\boldsymbol{\Phi}}_g^{(t)}[j,\ -j]\hat{\alpha}_g^{(t)}[-j]\Big), \end{aligned} \tag{16}$$

where $W_{\underline{\alpha}_j}^{(t+1)} = 1/\sum_{g=1}^{G} \hat{w}_g^{(t)}\hat{\phi}_g^{jj(t)}$, and

$$\hat{\underline{\alpha}}_{gj}^{(t+1)}(0) =W_{\underline{\alpha}_{gj}}^{(t+1)}\Big((\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}[j,\ ]\boldsymbol{\gamma}_g^{(t)} - (\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}[j,\ ]\hat{\boldsymbol{B}}_g^{(t)}\boldsymbol{\gamma}_g^{(t)} - \hat{\phi}^{jj(t)}\underline{\alpha}_j^{(t)} - \hat{\boldsymbol{\Phi}}_g^{(t)}[j,\ -j]\hat{\alpha}_g^{(t)}[-j]\Big), \tag{17}$$

where $W_{\underline{\alpha}_{gj}}^{(t+1)} = 1/\hat{\phi}_g^{jj(t)}$.

Updating formula on $\boldsymbol{\Phi}_g$

In relation to $\boldsymbol{\Phi}_g$, the updating formulas of non-diagonal elements in $\underline{\boldsymbol{\Phi}}$ and $\underline{\boldsymbol{\Phi}}_g$ are given as the following. Here, we define

$$\hat{\varphi}_{gj}^{(t)} = \hat{\phi}_{gjj}^{(t)} - \hat{\boldsymbol{\Phi}}_g^{(t)}[j,\ -j]\left(\hat{\boldsymbol{\Phi}}_g^{(t)}[-j,\ -j]\right)^{-1}\hat{\boldsymbol{\Phi}}_g^{(t)}[-j,\ -j] \tag{18}$$

and

$$\begin{aligned}\widetilde{\boldsymbol{\Psi}}_{g(j)}^{(t+1)} &= \left(\hat{\boldsymbol{\Phi}}_g^{(t)}[-j, -j]\right)^{-1}\boldsymbol{\Psi}_g^{(t+1)}[-j,\ ], \\ \overset{\approx}{\boldsymbol{\Psi}}_{g(j)}^{(t+1)} &= \left(\hat{\boldsymbol{\Phi}}_g^{(t)}[-j, -j]\right)^{-1}\boldsymbol{\Psi}_g^{(t+1)}[-j,\ -j]\left\{\left(\hat{\boldsymbol{\Phi}}_g^{(t)}[-j,\ -j]\right)^{-1}\right\}^{\top}.\end{aligned}$$

Then, we obtain

$$\hat{\underline{\phi}}_{jk}^{(t+1)}(0) = W_{\underline{\phi}_{jk}}^{(t+1)}\sum_{g=1}^{G}\frac{\hat{w}_g^{(t)}}{\hat{\varphi}_{gj}^{(t)}}\Big(\widetilde{\boldsymbol{\Psi}}_{g(j)}^{(t+1)}[\ell_k,\ j] - \hat{\underline{\phi}}_{gjk}^{(t)}\overset{\approx}{\boldsymbol{\Psi}}_{g(j)}^{(t+1)}[\ell_k,\ \ell_k] - \hat{\Phi}_g^{(t)}[j,\ -(j,k)]\overset{\approx}{\boldsymbol{\Psi}}_{g(j)}^{(t+1)}[-\ell_k,\ \ell_k]\Big), \tag{19}$$

where $W_{\underline{\phi}_{jk}}^{(t+1)} = 1/\sum_{g=1}^{G}\frac{\hat{w}_g^{(t)}}{\hat{\varphi}_{gj}^{(t)}}\overset{\approx}{\boldsymbol{\Psi}}_{g(j)}^{(t+1)}[\ell_k,\ \ell_k]$, and

$$\hat{\underline{\phi}}_{gjk}^{(t+1)}(0) = W_{\underline{\phi}_{gjk}}^{(t+1)}\Big(\widetilde{\boldsymbol{\Psi}}_{g(j)}^{(t+1)}[\ell_k,\ j] - \hat{\underline{\phi}}_{jk}^{(t)}\overset{\approx}{\boldsymbol{\Psi}}_{g(j)}^{(t+1)}[\ell_k,\ \ell_k] - \hat{\Phi}_g^{(t)}[j,\ -(j,k)]\overset{\approx}{\boldsymbol{\Psi}}_{g(j)}^{(t+1)}[-\ell_k,\ \ell_k]\Big), \tag{20}$$

where $W_{\underline{\phi}_{gjk}}^{(t+1)} = 1/\overset{\approx}{\boldsymbol{\Psi}}_{g(j)}^{(t+1)}[\ell_k,\ \ell_k]$. ℓ_k denotes the column index of $\hat{\phi}_{gjk}^{(t)}$ in $\hat{\boldsymbol{\Phi}}_g^{(t)}[j,\ -j]$. Meanwhile, diagonal elements of them are given as follows:

$$\hat{\underline{\phi}}_{gjj}^{(t+1)}(0) = \hat{\underline{\varphi}}_{gj}^{(t+1)}(0) + \hat{\boldsymbol{\Phi}}_g^{(t)}[j,\ -j]\left(\hat{\boldsymbol{\Phi}}_g^{(t)}[-j,\ -j]\right)^{-1}\hat{\boldsymbol{\Phi}}_g^{(t)}[-j,\ j], \tag{21}$$

where

$$\underline{\hat{\varphi}}_{gj}^{(t+1)}(0) = \boldsymbol{\Psi}_g^{(t+1)}[j,\ j] - 2\hat{\boldsymbol{\Phi}}_g^{(t)}[j,\ -j]\widetilde{\boldsymbol{\Psi}}_{g(j)}^{(t+1)}[\ ,\ j] + \hat{\boldsymbol{\Phi}}_g^{(t)}[j,\ -j]\overset{\approx}{\boldsymbol{\Psi}}_{g(j)}^{(t+1)}\hat{\boldsymbol{\Phi}}_g^{(t)}[-j,\ j]. \tag{22}$$

Estimating the covariance matrix of $\boldsymbol{\zeta}_g$ is achieved by iterative conditional fitting (ICF) (Chaudhuri et al., 2007). In ICF, the diagonal and non-diagonal elements of the covariance matrix are updated differently. The diagonal elements of $\boldsymbol{\Phi}_g$ are updated by updating $\boldsymbol{\varphi}_g$ in Eq. 18. As they are variances of residuals variable, they are not regularized, that is, $\lambda = 0$. Therefore, we need not consider the common effect in the variance of residuals variable but only the cluster effect. In other words, we can set $\underline{\phi}_{jj} = 0$, and then $\phi_{gjj} = \underline{\phi}_{gjj}$. In the clustering of PS-SEM, individuals are assigned to either of the clusters where the likelihood function becomes maximum. To compute the likelihood function, Eq. 5 is transformed into

$$\mathcal{L}_{\mathrm{KMSSEM}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{g=1}^{G}\sum_{n=1}^{N_g}\log \mathrm{p}(\boldsymbol{v}_{gn};\ \boldsymbol{\mu}_g^{(v)},\ \boldsymbol{\Sigma}_g^{(vv)}) = \frac{1}{N}\sum_{n=1}^{N}\sum_{g=1}^{G} u_{ng}\log \mathrm{p}(\boldsymbol{v}_{gn};\ \boldsymbol{\mu}_g^{(v)},\ \boldsymbol{\Sigma}_g^{(vv)}),$$

where $\mathrm{p}(\boldsymbol{v}_{gn};\ \boldsymbol{\mu}_g^{(v)},\ \boldsymbol{\Sigma}_g^{(vv)})$ denotes the probability density function of normal distribution. u_{ng} is an element of a cluster assignment matrix $\boldsymbol{U} \in \mathbb{R}^{N\times G}$ and values 1 or 0. $u_{ng} = 1$ if the n th individual belongs to the g th cluster, otherwise $u_{ng} = 0$. Then, u_{ng} is updated by the following:

$$\hat{u}_{ng}^{(t+1)} = \begin{cases} 1 & \left(g = \underset{k\in\{1,\ \ldots,\ G\}}{\operatorname{argmax}}\ \log \mathrm{p}\left(\boldsymbol{v}_{kn}\ ;\ \boldsymbol{\mu}_k^{(v)(t+1)},\ \boldsymbol{\Sigma}_k^{(vv)(t+1)}\right)\right) \\ 0 & (\text{others}) \end{cases}. \tag{23}$$

Using $\hat{u}_{ng}^{(t+1)}$, w_g is updated by the following:

$$\hat{w}_g^{(t+1)} = \frac{\sum_{n=1}^{N}\hat{u}_{ng}^{(t+1)}}{N}. \tag{24}$$

From the above updating formulas, the PS-SEM algorithm is described as Algorithm 1.

2.3 Mixture Sparse SEM

In MS-SEM, it is assumed that the observed variable $\boldsymbol{v} \in \mathbb{R}^{P\times 1}$ follows a Gaussian mixture distribution. Then, the probability density function is described as

$$\mathrm{p}(\boldsymbol{v}) = \sum_{g=1}^{G}\pi_g \mathrm{p}(\boldsymbol{v};\ \boldsymbol{\mu}_g^{(v)},\ \boldsymbol{\Sigma}_g^{(vv)}), \tag{25}$$

where π_g are the mixing proportions such that $0 < \pi_g \leq 1$ and $\sum_{g=1}^{G}\pi_g = 1$.

Let $\boldsymbol{\Theta}_{\mathrm{MSSEM}}$ denote the parameter space of MS-SEM, then the parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\mathrm{MSSEM}}$ is defined as

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top,\ \boldsymbol{\theta}_2^\top,\ \ldots,\ \boldsymbol{\theta}_G^\top)^\top \in \mathbb{R}^{G(Q+1)\times 1},$$

where $\boldsymbol{\theta}_g = (\theta_{g1},\ \theta_{g2}, \ldots,\ \theta_{gQ+1})^\top \in \mathbb{R}^{(Q+1)\times 1}$, specially $\left(\boldsymbol{\alpha}_g^\top, \mathrm{vec}(\boldsymbol{B}_g)^\top, \mathrm{vec}(\boldsymbol{\Phi}_g)^\top,\ \pi_g\right)^\top$.

Algorithm 1 The algorithm of PS-SEM

```
Input: $V,\ G,\ \lambda,\ \epsilon$
Output: $\hat{\boldsymbol{\theta}},\ \widehat{\boldsymbol{U}}$
  Set $t = 0$ and initialize $\hat{\boldsymbol{\theta}}^{(t)} \in \boldsymbol{\Theta}_{\text{KMSSEM}}$ and $\widehat{\boldsymbol{U}}^{(t)} = (\hat{\boldsymbol{u}}_{ng}^{(t)})$.
  Compute $\hat{w}_g^{(t+1)}$ from Eq. 24
  repeat
    E-step
    for $g = 1,\ \ldots,\ G$ do
      Compute $\hat{\boldsymbol{\gamma}}_g^{(t+1)}$, $\hat{\boldsymbol{\Gamma}}_g^{(t+1)}$ and $\hat{\boldsymbol{\Psi}}_g^{(t+1)}$ by Eqs. 10, 11, and 12
    end for

    M-step
    for $q = 1,\ \ldots,\ Q$ do
      Compute $\underline{\hat{b}}_{jk}^{(t)}$, $\underline{\hat{\alpha}}_j^{(t)}$ and $\underline{\hat{\phi}}_{jk}^{(t)}$ according to Eqs. 14, 16, and 19.
      for $g = 1,\ \ldots,\ G$ do
        Compute $\underline{\hat{b}}_{gjk}^{(t+1)}$, $\underline{\hat{\alpha}}_{gj}^{(t+1)}$, $\underline{\hat{\phi}}_{gjk}^{(t+1)}$, and $\underline{\hat{\phi}}_{gjj}^{(t+1)}$ according to Eq. 15, 17, 20, and 21.
      end for
    end for

    Clustering-step
    Compute $\widehat{\boldsymbol{U}}^{(t+1)} = (\hat{\boldsymbol{u}}_{ng}^{(t+1)})$ by Eq. 23.
    Compute $\hat{w}_g^{(t+1)}$ from Eq. 24

    Let $t = t + 1$
  until $|\mathcal{U}_{\text{KMSSEM}}(\hat{\boldsymbol{\theta}}^{(t)}) - \mathcal{U}_{\text{KMSSEM}}(\hat{\boldsymbol{\theta}}^{(t-1)})| < \epsilon$
  Let $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^{(t)}$ and $\widehat{\boldsymbol{U}} = \widehat{\boldsymbol{U}}^{(t)}$ return $\hat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{U}}$
```

The objective function of MS-SEM is defined as

$$\mathcal{U}_{\text{MSSEM}}(\boldsymbol{\theta},\ \lambda) = \mathcal{L}_{\text{MSSEM}}(\boldsymbol{\theta}) - \mathcal{R}_{\text{MSSEM}}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log \sum_{g=1}^{G} \pi_g \mathrm{p}(\boldsymbol{v}_n;\ \boldsymbol{\mu}_g^{(v)},\ \boldsymbol{\Sigma}_g^{(vv)}) - \sum_{g=1}^{G}\sum_{q=1}^{Q} c_{\theta_{gq}} \lambda |\theta_{gq}|, \tag{26}$$

where $c_{\theta_{gq}}$ is a regularization indicator. $\mathcal{R}_{\text{MSSEM}}(\boldsymbol{\theta})$ has one penalization term because the parameters of MS-SEM are not divided, unlike those of PS-SEM.

Maximization of the objective function is conducted only by the EM algorithm, which differs from the PS-SEM, because the E-step in MS-SEM computes the conditional expectation of z_{ng} contribute clustering. In addition, the M-step in MS-SEM updates the mixing proportions π_g, although it is not a parameter for SEM.

E-step

Considering another latent variable $\boldsymbol{z} = (z_1,\ z_2,\ \ldots,\ z_G)^\top \in \mathbb{R}^{G\times 1}$ which represents assignment individuals such that $z_g \in \{0,\ 1\}$ and $\sum_{g=1}^{G} z_g = 1$, the completely penalized log-likelihood is written as

$$\mathcal{U}^{C}_{\mathrm{MSSEM}}(\boldsymbol{\theta},\ \lambda) = \sum_{n=1}^{N}\sum_{g=1}^{G} z_{ng}\Big(\log \pi_g - \frac{1}{2}|\boldsymbol{\Phi}_g| - \frac{1}{2}\boldsymbol{\zeta}_{gn}^{\top}\boldsymbol{\Phi}_g^{-1}\boldsymbol{\zeta}_{gn}\Big) - \sum_{g=1}^{G}\sum_{q=1}^{Q} c_{\theta_{gq}}\lambda|\theta_{gq}|. \tag{27}$$

Moreover, the expectation of $\mathcal{U}^{C}_{\mathrm{MSSEM}}(\boldsymbol{\theta},\ \lambda)$ is given as follows:

$$\begin{aligned}
&\mathcal{M}_{\mathrm{MSSEM}}(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)})\\
=&\mathbb{E}\Big[\mathcal{U}^{C}_{\mathrm{MSSEM}}(\boldsymbol{\theta},\ \lambda) \mid \boldsymbol{V},\ \hat{\boldsymbol{\theta}}^{(t)}\Big]\\
=&\sum_{n=1}^{N}\sum_{g=1}^{G}\mathbb{E}[z_{ng} \mid \boldsymbol{V},\ \hat{\boldsymbol{\theta}}^{(t)}]\Big[\log \hat{\pi}_g^{(t)} - \frac{1}{2}\log|\hat{\boldsymbol{\Phi}}_g^{(t)}|\\
&-\frac{1}{2}\mathrm{tr}\Big((\boldsymbol{I}_{P+M} - \hat{\boldsymbol{B}}_g^{(t)})^{\top}(\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}(\boldsymbol{I}_{P+M} - \hat{\boldsymbol{B}}_g^{(t)})\mathbb{E}[\boldsymbol{\eta}_{gn}\boldsymbol{\eta}_{gn}^{\top} \mid \boldsymbol{V},\ \hat{\boldsymbol{\theta}}^{(t)}]\Big)\\
&+\hat{\boldsymbol{\alpha}}_g^{(t)\top}(\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}(\boldsymbol{I}_{P+M} - \hat{\boldsymbol{B}}_g^{(t)})\mathbb{E}[\boldsymbol{\eta}_{gn} \mid \boldsymbol{V},\ \hat{\boldsymbol{\theta}}^{(t)}]) - \frac{1}{2}\hat{\boldsymbol{\alpha}}_g^{(t)\top}(\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}\hat{\boldsymbol{\alpha}}_g^{(t)}\Big] - \sum_{g=1}^{G}\sum_{q=1}^{Q} c_{\theta_{gq}}\lambda|\theta_{gq}|.
\end{aligned} \tag{28}$$

Let $\hat{r}_{ng}^{(t+1)} = \mathbb{E}[z_{ng} \mid \boldsymbol{V},\ \hat{\boldsymbol{\theta}}^{(t)}]$, $\boldsymbol{\delta}_{gn}^{(t+1)} = \mathbb{E}[\boldsymbol{\eta}_{gn} \mid \boldsymbol{V},\ \hat{\boldsymbol{\theta}}^{(t)}]$, $\boldsymbol{\Delta}_{gn}^{(t+1)} = \mathbb{E}[\boldsymbol{\eta}_{gn}\boldsymbol{\eta}_{gn}^{\top} \mid \boldsymbol{V},\ \hat{\boldsymbol{\theta}}^{(t)}]$ and $\boldsymbol{\Omega}_{gn}^{(t+1)} = \mathbb{E}[\boldsymbol{\zeta}_{gn}\boldsymbol{\zeta}_{gn}^{\top} \mid \boldsymbol{V},\ \hat{\boldsymbol{\theta}}^{(t)}]$, then, they are described as

$$\hat{r}_{ng}^{(t+1)} = \frac{\hat{\pi}_g^{(t)}\mathrm{p}(\boldsymbol{v}_n;\ \hat{\boldsymbol{\mu}}_g^{(v)(t)},\ \hat{\boldsymbol{\Sigma}}_g^{(vv)(t)})}{\sum_{j=1}^{G}\hat{\pi}_j^{(t)}\mathrm{p}_j(\boldsymbol{v}_n;\ \hat{\boldsymbol{\mu}}_j^{(v)(t)},\ \hat{\boldsymbol{\Sigma}}_j^{(vv)(t)})} \tag{29}$$

$$\boldsymbol{\delta}_{gn}^{(t+1)} = \hat{\boldsymbol{\mu}}_g^{(\eta)(t)} - \hat{\boldsymbol{\Sigma}}_g^{(\eta v)(t)}(\hat{\boldsymbol{\Sigma}}_g^{(vv)(t)})^{-1}\left(\boldsymbol{v}_n - \hat{\boldsymbol{\mu}}_g^{(v)(t)}\right) \tag{30}$$

$$\boldsymbol{\Delta}_{gn}^{(t+1)} = \hat{\boldsymbol{\Sigma}}_g^{(\eta\eta)(t)} - \left(\hat{\boldsymbol{\Sigma}}_g^{(\eta v)(t)}\right)^{-1}\hat{\boldsymbol{\Sigma}}_g^{(vv)(t)}\hat{\boldsymbol{\Sigma}}_g^{(v\eta)(t)} + \boldsymbol{\delta}_{gn}^{(t+1)}\boldsymbol{\delta}_{gn}^{(t+1)\top} \tag{31}$$

$$\begin{aligned}
\boldsymbol{\Omega}_{gn}^{(t+1)} =&(\boldsymbol{I}_{P+M} - \hat{\boldsymbol{B}}_g^{(t)})\boldsymbol{\Delta}_{gn}^{(t+1)}(\boldsymbol{I}_{P+M} - \hat{\boldsymbol{B}}_g^{(t)})^{\top} - (\boldsymbol{I}_{P+M} - \hat{\boldsymbol{B}}_g^{(t)})\boldsymbol{\delta}_{gn}^{(t+1)}\hat{\boldsymbol{\alpha}}_g^{(t)\top}\\
&-\hat{\boldsymbol{\alpha}}_g^{(t)}\boldsymbol{\delta}_{gn}^{(t+1)\top}(\boldsymbol{I}_{P+M} - \boldsymbol{B}_g)^{\top} + \hat{\boldsymbol{\alpha}}_g^{(t)}\hat{\boldsymbol{\alpha}}_g^{(t)\top}.
\end{aligned} \tag{32}$$

E-step is performed by computing Eqs. 29, 30, 31, and 32. $\hat{r}_{ng}^{(t+1)}$ is called responsibility.

M-step

The M-step was performed in the same way as PS-SEM without π_g. Here, π_g, $\boldsymbol{B}_g$, $\boldsymbol{\alpha}_g$, and $\boldsymbol{\Phi}_g$ are estimated, iteratively, Consequently, parameters are updated as follows.

Updating Formula on π_g

The π_g is given as follows with responsibility.

$$\hat{\pi}_g^{(t+1)} = \frac{\sum_{n=1}^{N}\hat{r}_{ng}^{(t+1)}}{N} \tag{33}$$

Updating formula on $\boldsymbol{B}_g$

The updating formula on $\boldsymbol{B}_g$ is given as follows:

$$\begin{aligned}\hat{b}_{gjk}^{(t+1)}(0) =& W_{b_{gjk}}^{(t+1)} \sum_{n=1}^{N} \hat{r}_{ng}^{(t+1)} \Big(\boldsymbol{\Delta}_{gn}^{(t+1)}[k, \,](\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}[\, , \, j] - \boldsymbol{\Delta}_{gn}^{(t+1)}[k, k] \hat{\boldsymbol{B}}_g^{(t)}[-j, k]^{\top} (\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}[-j, j] \\ &- \boldsymbol{\Delta}_{gn}^{(t+1)}[k, \, -k] \hat{\boldsymbol{B}}_g^{(t)}[\, , -k]^{\top} (\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}[\, , \, j] - \boldsymbol{\delta}_{gn}^{(t)}[k] \hat{\boldsymbol{\alpha}}_g^{(t)\top} (\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}[\, , \, j] \Big), \end{aligned} \tag{34}$$

where $W_{b_{gjk}}^{(t+1)} = 1/\sum_{n=1}^{N} \hat{r}_{ng}^{(t+1)} \boldsymbol{\Delta}_{gn}^{(t+1)}[k, k] \hat{\phi}_g^{jj(t)}$.

Updating Formula on $\boldsymbol{\alpha}_g$

The updating formula on $\boldsymbol{\alpha}_g$ is given as follows:

$$\hat{\alpha}_{gj}^{(t+1)}(0) = W_{\alpha_{gj}}^{(t+1)} \sum_{n=1}^{N} \hat{r}_{ng}^{(t+1)} \Big((\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}[j, \,] \boldsymbol{\delta}_{gn}^{(t)} - (\hat{\boldsymbol{\Phi}}_g^{(t)})^{-1}[j, \,] \hat{\boldsymbol{B}}_g^{(t)} \boldsymbol{\delta}_{gn}^{(t)} - \hat{\boldsymbol{\Phi}}_g^{(t)}[j, -j] \hat{\alpha}_g^{(t)}[-j] \Big), \tag{35}$$

where $W_{\alpha_{gj}}^{(t+1)} = 1/\sum_{n=1}^{N} \hat{r}_{ng}^{(t+1)} \hat{\phi}_g^{jj(t)}$.

Updating Formula on$\boldsymbol{\Phi}_g$

The updating formula on the non-diagonal elements of $\boldsymbol{\Phi}_g$ is given as the following. Here, we define

$$\begin{aligned} \widetilde{\boldsymbol{\Omega}}_{gn(j)}^{(t+1)} &= \left(\hat{\boldsymbol{\Phi}}_g^{(t)}[-j, \, -j] \right)^{-1} \boldsymbol{\Omega}_{gn}^{(t+1)}[-j, \,], \\ \approx{\boldsymbol{\Omega}}_{gn(j)}^{(t+1)} &= \left(\hat{\boldsymbol{\Phi}}_g^{(t)}[-j, \, -j] \right)^{-1} \boldsymbol{\Omega}_{gn}^{(t+1)}[-j, \, -j] \left\{ \left(\hat{\boldsymbol{\Phi}}_g^{(t)}[-j, \, -j] \right)^{-1} \right\}^{\top}. \end{aligned}$$

Then, we obtain

$$\hat{\phi}_{gjk}^{(t+1)}(0) = W_{\phi_{gjk}}^{(t+1)} \sum_{n=1}^{N} \hat{r}_{ng}^{(t+1)} \Big(\widetilde{\boldsymbol{\Omega}}_{gn(j)}^{(t+1)}[\ell_k, \, j] - \hat{\boldsymbol{\Phi}}_g^{(t)}[j, \, -(j, \, k)] \overset{\approx}{\boldsymbol{\Omega}}_{gn(j)}^{(t+1)}[-\ell_k, \, \ell_k] \Big), \tag{36}$$

where $W_{\phi_{gjk}}^{(t+1)} = 1/\sum_{n=1}^{N} \hat{r}_{ng}^{(t+1)} \overset{\approx}{\boldsymbol{\Omega}}_{gn(j)}^{(t+1)}[\ell_k, \, \ell_k]$.

The updating formula on the diagonal elements of$\boldsymbol{\Phi}_g$ is given as follows:

$$\hat{\phi}_{gjj}^{(t+1)}(0) = \hat{\varphi}_{gj}^{(t+1)}(0) + \hat{\boldsymbol{\Phi}}_g^{(t)}[j, \, -j] \left(\hat{\boldsymbol{\Phi}}_g^{(t)}[-j, \, -j] \right)^{-1} \hat{\boldsymbol{\Phi}}_g^{(t)}[-j, \, j], \tag{37}$$

where

$$\begin{aligned} \hat{\varphi}_{gj}^{(t+1)}(0) =& \frac{1}{\sum_{n=1}^{N} \hat{r}_{ng}^{(t+1)}} \sum_{n=1}^{N} \hat{r}_{ng}^{(t+1)} \Big(\boldsymbol{\Omega}_{gn}^{(t+1)}[j, \, j] - 2 \hat{\boldsymbol{\Phi}}_g^{(t)}[j, \, -j] \widetilde{\boldsymbol{\Omega}}_{gn(j)}^{(t+1)}[\, , \, j] \\ &+ \hat{\boldsymbol{\Phi}}_g^{(t)}[j, \, -j] \overset{\approx}{\boldsymbol{\Omega}}_{gn(j)}^{(t+1)} \hat{\boldsymbol{\Phi}}_g^{(t)}[-j, \, j] \Big). \end{aligned} \tag{38}$$

Clustering in MS-SEM is the same as model-based clustering, that is, $\hat{r}_{ng}$ denotes the degree to which the individual n belongs to the cluster g. When performing hard clustering in MS-SEM, the individuals are assigned to either of clusters where $\hat{r}_{ng}$ becomes maximum.

According to the above updating formulas, the algorithm of MS-SEM is described as Algorithm 2

Algorithm 2 The algorithm of MS-SEM.

```
Input: $\boldsymbol{V}$, $G$, $\lambda$, $\epsilon$
Output: $\hat{\boldsymbol{\theta}}$, $\widehat{\boldsymbol{U}}$
  Set $t = 0$ and initialize $\hat{\boldsymbol{\theta}}^{(t)} \in \boldsymbol{\Theta}_{\text{MSSEM}}$
  repeat
    E-step
    for $n = 1, \ldots, N$ do
      for $g = 1, \ldots, G$ do
        Compute $\hat{r}_{ng}^{(t+1)}$, $\hat{\boldsymbol{\delta}}_{gn}^{(t+1)}$, $\hat{\boldsymbol{\Delta}}_{gn}^{(t+1)}$ and $\hat{\boldsymbol{\Omega}}_{gn}^{(t+1)}$ by Eqs. 29, 30, 31 and 32.
      end for
    end for
    M-step
    for $q = 1, \ldots, Q + 1$ do
      for $g = 1, \ldots, G$ do
        Compute $\hat{\pi}_{g}^{(t+1)}$, $\hat{b}_{gjk}^{(t+1)}$, $\hat{\alpha}_{gj}^{(t+1)}$, $\hat{\phi}_{gjk}^{(t+1)}$, and $\hat{\phi}_{gjj}^{(t+1)}$ according to Eqs. 33, 34, 35, 36, and 37.
      end for
    end for
    Let $t = t + 1$.
  until $|\mathcal{U}_{\text{MSSEM}}(\hat{\boldsymbol{\theta}}^{(t)}) - \mathcal{U}_{\text{MSSEM}}(\hat{\boldsymbol{\theta}}^{(t-1)})| < \epsilon$
  Let $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^{(t)}$ and $\hat{r}_{ng} = \hat{r}_{ng}^{(t+1)}$
  Let $\widehat{\boldsymbol{U}} = (\hat{r}_{ng})$
  return $\hat{\boldsymbol{\theta}}$, $\widehat{\boldsymbol{U}}$
```

3 Simulation Study

In this section, we conducted a simulation study to evaluate the performance of the proposed method. The simulation design and simulation results are described below.

3.1 Simulation Design

Both PS-SEM and MS-SEM estimate path diagrams and the size of the arrows in them and assume that there is a latent cluster structure that has a different path diagram for each. Therefore, we generated artificial data from different path diagrams and compared the proposed methods with existing methods in the simulation study. Concretely, we let $P = 8$, $M = 2$, and $G = 3$, the path diagram of Cluster 1, as shown in Fig. 1, Cluster 2 as Fig. 2, and Cluster 3 as Fig. 3 The dashed lines in Figs. 2 and 3 represent the fact that the existence of the relations is assumed beforehand, although these relations do not exist in the true path diagram. According to them, if the proposed methods can be estimated correctly, their corresponding parameters are estimated as 0 by sparse estimation.

True coefficient matrices that express the path diagrams in Figs. 1, 2, and 3 are described as

$$\boldsymbol{B}_1 = \begin{pmatrix} \boldsymbol{O}_{P\times P} & \boldsymbol{B}_1^{(vf)} \\ \boldsymbol{O}_{M\times P} & \boldsymbol{B}_1^{(ff)} \end{pmatrix}, \ \boldsymbol{B}_2 = \begin{pmatrix} \boldsymbol{O}_{P\times P} & \boldsymbol{B}_2^{(vf)} \\ \boldsymbol{O}_{M\times P} & \boldsymbol{B}_2^{(ff)} \end{pmatrix}, \text{ and } \boldsymbol{B}_3 = \begin{pmatrix} \boldsymbol{O}_{P\times P} & \boldsymbol{B}_3^{(vf)} \\ \boldsymbol{O}_{M\times P} & \boldsymbol{B}_3^{(ff)}, \end{pmatrix}$$

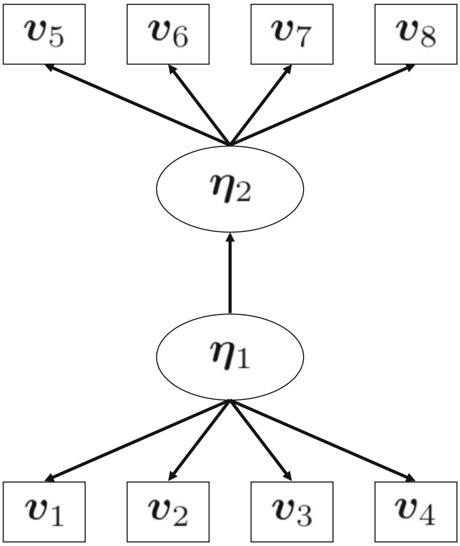


Fig. 1 Cluster 1

where

$$\boldsymbol{B}_1^{(vf)} = \begin{pmatrix} 1.0 & 0.0 \\ 1.1 & 0.0 \\ 1.2 & 0.0 \\ 1.3 & 0.0 \\ 0.0 & 1.0 \\ 0.0 & 1.1 \\ 0.0 & 1.2 \\ 0.0 & 1.3 \end{pmatrix}, \ \boldsymbol{B}_2^{(vf)} = \begin{pmatrix} 1.0 & 0.0 \\ 1.1 & 0.0 \\ 1.2 & 0.0 \\ 1.3 & 0.0 \\ 0.0 & 1.0 \\ 0.0 & \mathbf{0.0} \\ 0.0 & \mathbf{0.0} \\ 0.0 & \mathbf{0.0} \end{pmatrix}, \text{ and } \boldsymbol{B}_3^{(vf)} = \begin{pmatrix} 1.0 & 0.0 \\ \mathbf{0.0} & 0.0 \\ \mathbf{0.0} & 0.0 \\ \mathbf{0.0} & 0.0 \\ 0.0 & 1.0 \\ 0.0 & 1.1 \\ 0.0 & 1.2 \\ 0.0 & 1.3 \end{pmatrix}.$$

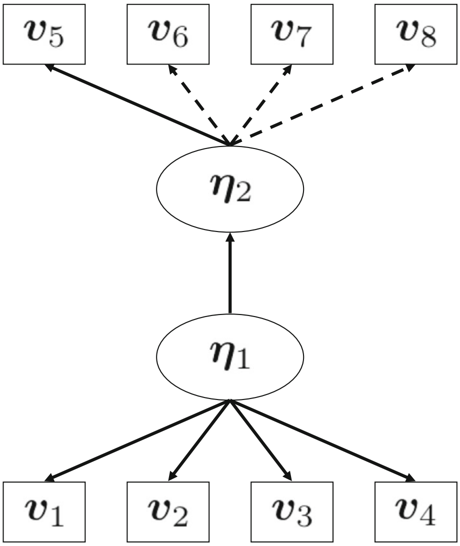
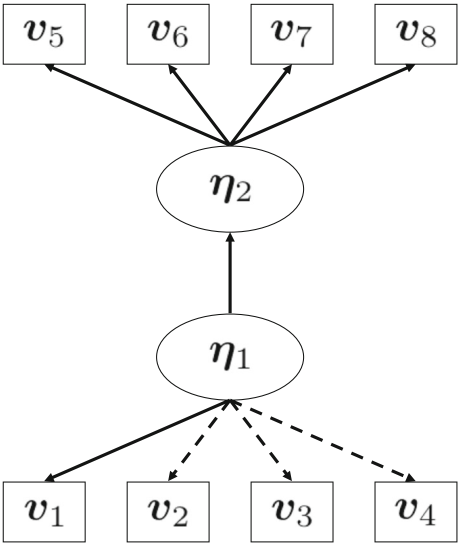


Fig. 2 Cluster 2

Fig. 3 Cluster 3



These bold types are part of coefficient matrices corresponding with dashed lines. Moreover, to maintain identification we set other parameters as follows.

$$\boldsymbol{B}_g^{(ff)} = \begin{pmatrix} 0 & 0 \\ 5 & 0 \end{pmatrix}, \ \boldsymbol{\alpha}_g = \boldsymbol{0}_{P+M}, \ \boldsymbol{\Phi}_g = \boldsymbol{I}_{P+M} \quad (g = 1,\ 2,\ 3)$$
$$\underline{\boldsymbol{B}} = \ \boldsymbol{O}_{(P+M)\times(P+M)}, \ \underline{\boldsymbol{\alpha}} = \boldsymbol{0}_{P+M}, \ \underline{\boldsymbol{\Phi}} = \boldsymbol{O}_{(P+M)\times(P+M)}$$

By fixing these parameters, $\boldsymbol{B}_1^{(vf)}$, $\boldsymbol{B}_2^{(vf)}$, and $\boldsymbol{B}_3^{(vf)}$ are estimated. Computing $\boldsymbol{\mu}_g^{(v)}$ and $\boldsymbol{\Sigma}_g^{(vv)}$ based on the above, N_g observed data in cluster g are generated from $\boldsymbol{v}_{gn} \sim \mathcal{N}(\boldsymbol{\mu}_g^{(v)},\ \boldsymbol{\Sigma}_g^{(vv)})$, where $N_g = N/G$. Then, the data are $\boldsymbol{V} = (\boldsymbol{v}_{11},\ \boldsymbol{v}_{12},\ \ldots,\ \boldsymbol{v}_{1N_1},\ \boldsymbol{v}_{21},\ \boldsymbol{v}_{22},\ \ldots,\ \boldsymbol{v}_{2N_2},\ \boldsymbol{v}_{31},\ \boldsymbol{v}_{32},\ \ldots,\ \boldsymbol{v}_{3N_3})^\top$. We apply each method to the data 100 times in the different sample sizes, $N = 90$, $N = 150$, and $N = 300$. b_{gjk}, which is an element of $\boldsymbol{B}_g$, is initialized by $b_{gjk} \sim \mathrm{U}(0,\ 3)$, where $\mathrm{U}(x,\ y)$ denotes a uniform distribution. In KM-SEM, the initial value of the indicator matrix $\boldsymbol{U}$ is required. In each row of $\boldsymbol{U}$, we randomly set one of the elements of the column as 1 and all the others as 0.

In MS-SEM the initial value of π_g is set as $\pi_g = \rho_g / \sum_{j=1}^{G} \rho_j$ $(g = 1,\ 2,\ 3)$, where $\rho_g \sim \mathrm{U}(0.9,\ 1)$. Due to this, π_g satisfies $\sum_g^G \pi_g = 1$. A random start is conducted 30 times because algorithms depend on initial value.

We applied four methods, PS-SEM, MS-SEM, MSEM, and KM+MGSSEM, where KM+MGSSEM is a tandem analysis that applies MGSSEM after clustering using Kmeans. Three methods without MSEM are needed to set the regularization parameter λ. λ is chosen based on BIC from Λ, which is the grid search range, because in Huang (2018) the selection of λ was based on BIC. Let $\hat{\boldsymbol{\theta}}(\lambda)$ be the optimized parameter of λ, then, the BIC of PS-SEM and MS-SEM are

$$\mathrm{BIC}_{\mathrm{KMSSEM}}(\lambda) = -2\mathcal{L}_{\mathrm{KMSSEM}}(\hat{\boldsymbol{\theta}}(\lambda)) + \tfrac{\log N}{N}\mathrm{d}(\lambda), \text{ and}$$
$$\mathrm{BIC}_{\mathrm{MSSEM}}(\lambda) = -2\mathcal{L}_{\mathrm{MSSEM}}(\hat{\boldsymbol{\theta}}(\lambda)) + \log N(\mathrm{d}(\lambda) - 1),$$

respectively, where d(λ) is the number of nonzero parameters. AS MS-SEM has the constraint $\sum_g^G \pi_g = 1$, d(λ) − 1 appears in the BIC of the MS-SEM. Moreover, $\Lambda_{\text{KMSSEM}} = \{0.01,\ 0.02,\ \ldots,\ 0.1\}$, and $\Lambda_{\text{MSSEM}} = \{1.1,\ 1.2,\ \ldots,\ 2\}$. Λ in KM+MGSSEM is the same as Λ_{KMSSEM}. In this simulation, to choose the tuning parameter, grid search is used and selected based on BIC.

We evaluated each method using four indices: adjusted rand index (ARI) (Hubert and Arabie, 1985), Accuracy for each cluster (ACC), true positive rate (TPR), false positive rate (FPR), and root mean square error (RMSE). ARI is calculated between the true clustering structure and the estimated clustering structure. If the ARI is close to 1, the estimated clustering structure is considered as good, otherwise, the estimated clustering result is considered as not good. ACC is calculated by each cluster and calculated as follows:

$$\text{ACC cls}_g = \frac{\sum_{n=1}^N \hat{u}_{ng} u_{ng}}{\sum_{n=1}^N u_{ng}},$$

for $g = 1, 2, \ldots, G$, where $\hat{u}_{ng}$ and $\hat{u}_{ng}$ are elements of the true indicator matrix and the estimated indicator matrix, respectively. The estimated label with the largest number of labels corresponding to each true label is determined to be the corresponding label. ACC is sensitivity for true cluster labels. Accuracy refers to clustering accuracy. TPR refers to the rate of relations that do not exist in the true structure (the dashed lines in Figs. 2 and 3) that are estimated to be zero by sparse estimation. On the other hand, FPR is the rate of relations that exist in the true structure and are incorrectly estimated to be 0 by sparse estimation. Therefore, the closer the TPR is to 1, the better, while the closer the FPR is to 0, the better. Finally, RMSE is the estimation error. We set the true parameter vector and the estimated parameter vector as $\boldsymbol{\theta} = (\theta_1,\ \theta_2,\ \ldots,\ \theta_R)$ and $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1,\ \hat{\theta}_2,\ \ldots,\ \hat{\theta}_R)$, respectively, as follows:

$$\text{RMSE} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_r - \theta_r)^2}.$$

The reason RMSE is added to the evaluation of the numerical simulation is to evaluate the magnitude level of the error variance.

3.2 Simulation Result

The simulation results are shown in Tables 2, 3, and 4. These are the means and standard deviations of the evaluation indices by 30 simulations for $N = 90$, $N = 150$, and $N = 300$ and the bold type in the table shows the largest mean for each indicator. As MSEM did not perform sparse estimation, TPR and FPR were not calculated in MSEM. In the other methods, TPR and FPR were also calculated for each cluster. For example, TPR cls2 meant TPR in Cluster 2, and TPR cls1 were not calculated because the truth path structure of Cluster 1 did not require sparse estimation.

First, in terms of ARI, PS-SEM demonstrates the best accuracy for all cases of $N = 90,\ 150$, and 300. From the results of ACC, it can be confirmed that the clustering results of PS-SEM is stable. As the clustering of the PS-SEM is hard clustering, it is more accurate than that of the MS-SEM and MSEM, which use soft clustering. However, the clustering KM+MGSSEM has the lowest accuracy, although it is clustered by Kmeans method, which is a hard clustering method. Accordingly, because the tandem method does not perform

Table 2 Result in $N = 90$

N=90	PS-SEM		MS-SEM		MSEM		KM+MGSSEM	
Indices	Mean	sd	Mean	SD	Mean	SD	Mean	SD
ARI	**0.772**	**0.038**	0.751	0.047	0.757	0.049	0.529	0.042
ACC cls1	0.651	**0.083**	**0.710**	0.126	0.639	0.179	0.353	0.198
ACC cls2	**0.962**	**0.043**	0.940	0.062	0.939	0.054	0.932	0.169
ACC cls3	**0.703**	**0.093**	0.602	0.132	0.692	0.176	0.301	0.195
TPR	**0.517**	0.202	0.344	0.185	-	-	0.071	0.113
TPR cls2	**0.378**	0.312	0.167	0.191	-	-	0.048	0.159
TPR cls3	**0.656**	0.254	0.522	0.335	-	-	0.095	0.187
FPR	0.000	0.000	0.000	0.000	-	-	0.000	0.000
FPR cls1	0.000	0.000	0.000	0.000	-	-	0.000	0.000
FPR cls2	0.000	0.000	0.000	0.000	-	-	0.000	0.000
FPR cls3	0.000	0.000	0.000	0.000	-	-	0.000	0.000
RMSE	0.050	0.014	**0.039**	**0.011**	0.041	**0.011**	0.115	0.082

clustering based on the path structure estimated by SEM, it is not suitable for clustering when there is a causal structure behind the data. Another reason the Kmeans method does not perform clustering correctly is that the expectation of the true distribution used to generate the data for this simulation is equal for all clusters. The clustering accuracies of MS-SEM and MSEM were not significantly different in all situations. The reason for this, as will be discussed later in the comparison of TPR, is that the MS-SEM does not sufficiently prune the arrows of the path diagram by sparse estimation and does not significantly differ from MSEM.

Table 3 Result in $N = 150$

N=150	PS-SEM		MS-SEM		MSEM		KM+MGSSEM	
Indices	Mean	SD	Mean	SD	Mean	SD	Mean	SD
ARI	**0.764**	0.038	0.763	0.029	0.757	0.040	0.520	0.022
ACC cls1	0.645	**0.080**	0.657	0.131	**0.702**	0.140	0.387	0.222
ACC cls2	0.951	0.038	**0.953**	**0.030**	0.945	0.044	0.950	0.116
ACC cls3	**0.695**	**0.095**	0.679	0.138	0.624	0.160	0.228	0.231
TPR	**0.667**	0.196	0.267	0.178	-	-	0.000	0.000
TPR cls2	**0.611**	0.324	0.133	0.225	-	-	0.000	0.000
TPR cls3	**0.822**	0.243	0.400	0.296	-	-	0.000	0.000
FPR	0.003	0.015	**0.000**	0.000	-	-	**0.000**	0.000
FPR cls1	0.006	0.030	**0.000**	0.000	-	-	**0.000**	0.000
FPR cls2	0.000	0.000	0.000	0.000	-	-	0.000	0.000
FPR cls3	0.000	0.000	0.000	0.000	-	-	0.000	0.000
RMSE	0.040	0.017	**0.029**	**0.007**	0.032	0.008	0.127	0.053

Table 4 Result in $N = 300$

N=300	PS-SEM		MS-SEM		MSEM		KM+MGSSEM	
Indices	Mean	SD	Mean	SD	Mean	SD	Mean	SD
ARI	**0.770**	0.023	0.765	0.028	0.765	0.025	0.541	0.018
ACC cls1	0.634	**0.057**	0.659	0.116	**0.666**	0.134	0.366	0.201
ACC cls2	0.958	0.022	0.946	0.022	0.945	0.021	**0.989**	**0.015**
ACC cls3	**0.718**	**0.054**	0.689	0.124	0.682	0.141	0.256	0.211
TPR	**0.700**	0.177	0.250	0.147	-	-	0.000	0.000
TPR cls2	**0.522**	0.324	0.155	0.212	-	-	0.000	0.000
TPR cls3	**0.878**	0.185	0.345	0.212	-	-	0.000	0.000
FPR	0.000	0.000	0.000	0.000	-	-	0.000	0.000
FPR cls1	0.000	0.000	0.000	0.000	-	-	0.000	0.000
FPR cls2	0.000	0.000	0.000	0.000	-	-	0.000	0.000
FPR cls3	0.000	0.000	0.000	0.000	-	-	0.000	0.000
RMSE	0.029	0.007	**0.022**	0.006	0.024	**0.005**	0.097	0.029

In terms of TPR, PS-SEM is also the best method, it performs the most accurate pruning. In PS-SEM, the TPR value improves as the sample size increases. In particular, the TPR in Cluster 3 is better than that in Cluster 2, which suggests that there are parts of the path structure that can be easily sparsely estimated. Nevertheless, MS-SEM did not show any difference in TPR depending on the sample size and did not show sufficient pruning in any situation. Lastly, in KM+MGSSEM, TPR was 0 or close to 0 in all situations, even though sparse estimation was performed after clustering. The reason for this is that, as described the above, Kmeans does not perform clustering correctly, and the data generated from the path structure, which both require and do not require pruning, are mixed in the same cluster.

The FPR was almost 0 for all the methods. This indicates that the PS-SEM and MS-SEM methods do not erroneously perform sparse estimation and that the pruned parts are unnecessary structures.

Lastly, in terms of RMSE, MS-SEM showed the best value for all sample sizes. The estimation error of MS-SEM and MSEM based on the mixture normal distribution was better than that of PS-SEM. It is considered that the estimation error is smaller than that of MSEM owing to the sparse estimation because the unpruned coefficients, which should be pruned, are reduced by sparse estimation in MS-SEM.

Owing to calculating the RMSE, sparse estimation was performed only for the measurement model in this simulation to ensure model identification. However, it is noted that sparse estimation is possible for other parts of path diagrams as well.

4 Real Data Illustration

In this section, we apply four methods, PS-SEM, MS-SEM, MSEM, and KM+MGSSEM, which are discussed in Sect. 3 to the real data. We also compare the results and provide an interpretation example for each of the four methods. For the evaluation index, we adopted the goodness-of-fit index (GFI) following Huang (2018) and Steiger (1998). The reason GFI

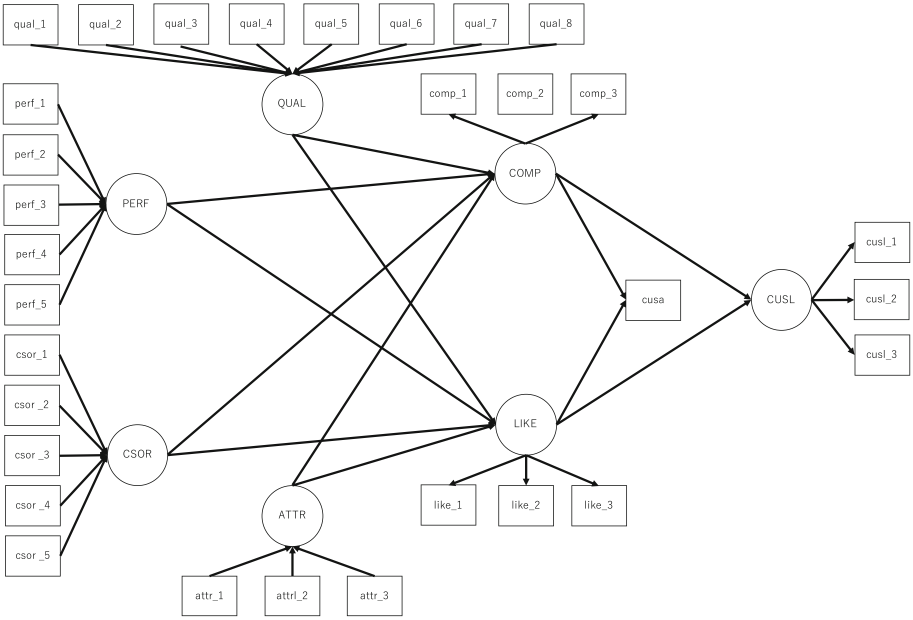


Fig. 4 Corporate reputation model

is used is to provide the GFI values referenced from previous studies. The data used are from a survey on satisfaction with and perception of cell phone service providers in the German cell phone market (Sarstedt & Mooi, 2014) and are considered to include a heterogeneous structure. It is often used as an example of the application of methods, such as MGSEM and clustered SEM (Fordellone & Vichi, 2020; Hair Jr. et al., 2016; Matthews et al., 2016). The survey questions were structured on a Likert scale, and the subjects' responses were obtained on a seven-point scale from "1: not at all true (not satisfied at all)" to "7: most true (very satisfied)". The subjects were service users of the target providers, and the sample size of the data was $N = 344$. These data had 31 variables, which are shown in Table 5, and they were obtained from seven components, evaluations of company's competence (COMP), company's likeability (LIKE), customer loyalty (CUSL), quality (QUAL), performance PERF, corporate social responsibility (CSOR), attractiveness (ATTR), and customer satisfaction constructs (CUSA), similar to Hair Jr. et al. (2016). We used the corporate reputation model as the path structure (Fig. 4) to apply each method, which was proposed by Eberl (2010), and used by Fordellone and Vichi (2020); Hair Jr. et al. (2016), and Matthews et al. (2016).

According to the data and path structure, PS-SEM, MS-SEM, MSEM, and KM+SEM were compared. As the meaning of the latent variable and the relationships among variables are determined in the corporate reputation model (Eberl, 2010), we do not regularize the parameters of the coefficients representing the relationship between the latent variables and the observed variables, but only the parameters of the coefficients representing the relationship between the latent variables, to avoid the interpretation of the latent variables that differ from the corporate reputation model.

As all methods require the number of clusters to be determined in advance, the optimal number of clusters was determined by the information criterion by applying each method with $G = 2, 3$, and 4. Bold letters in Table 6 indicate the part where the log-likelihood is

Table 5 Data description

Latent variable	Variable	Variable description
Competence	comp_1	[The company] is a top competitor in its market
(COMP)	comp_2	As far as I know, [the company] is recognized worldwide
	comp_3	I believe that [the company] performs at premium level
Likeability	like_1	[The company] is a company that I can better identify with than any other company
(LIKE)	like_2	[The company] is a company that I would regret more not having if it no longer existed than I would other companies
	like_3	I regret [the company] as a likeable company
Customer	cusl_1	I would recommend [company] to friends and relatives
Loyalty	cusl_2	If I had to choose again, I would choose [company] as my mobile phone services provider
(CUSL)	cusl_3	I will remain a customer of [company] in the future
Quality	qual_1	The products/services offered by [the company] are of high quality
(QUAL)	qual_2	[The company] is an innovator, rather than an imitator with respect to industry
	qual_3	[The company]'s products/services offer good value for money
	qual_4	The services [the company] offered are good
	qual_5	Customer concerns are held in high regard at [the company]
	qual_6	[The company] is a reliable partner for customers
	qual_7	[The company] is a trustworthy company
	qual_8	I have a lot of respect for [the company]
Performance	perf_1	[The company] is a very well-managed company
(PERF)	perf_2	[The company] is an economically stable company
	perf_3	The business risk for [the company] is modest compared to its competitors
	perf_4	[The company] has growth potential
	perf_5	[The company] has a clear vision about the future of the company
Corporate	csor_1	[The company] behaves in a socially conscious way
Social	csor_2	[The company] is forthright in giving information to the public
Responsibility	csor_3	[The company] has a fair attitude toward competitors
(CSOR)	csor_4	[The company] is concerned about the preservation of the environment
	csor_5	[The company] is not only concerned about profits
Attractiveness	attr_1	[The company] is successful in attracting high-quality employees
(ATTR)	attr_2	I could see myself working at [the company]
	attr_3	I like the physical appearance of [the company] (company, buildings, shops, etc.)
Customer Satisfaction	cusa	If you consider your experiences with [company], how satisfied are you with [company]?
Construct		
(CUSA)		

The actual name of the company was inserted in the bracketed space

Table 6 Information criterion for each method for each number of clusters

Method	G	λ	log L	BIC
PS-SEM	2	0.03	−10.40	**24.83**
	3	0.07	−9.73	25.31
	4	0.07	**−9.23**	26.29
MS-SEM	2	2.00	−13292.37	**27935.46**
	3	1.70	−13152.62	28354.04
	4	1.50	−13069.92	28856.32
MSEM	2	0.00	−13273.97	**27949.88**
	3	0.00	−13130.48	28397.29
	4	0.00	**−13028.7**	28904.24
KM+SSEM	2	0.03	−10.49	**25.01**
	3	0.09	−9.81	25.44
	4	0.08	**−9.31**	26.42

the maximum and the information criterion is the minimum for each method. Table 6 shows the values of the log-likelihood and information criterion for each number of clusters for the four methods and the values of the regularization parameters at that time. According to Table 6, the log-likelihood is maximum at $G = 4$, and both AIC and BIC are minimum for all methods. This result suggests that when the number of clusters is increased, the value of the log-likelihood becomes larger because the data are more segmented; however, the number of parameters also increases at the same time, resulting in a larger value of the information criterion. Due to this result, we report the results with $G = 2$ which is regarded as the best number of clusters. The result of the best number of clusters is the same as Fordellone and Vichi (2020) and Matthews et al. (2016), which use the corporate reputation model. For selecting the tuning parameter for sparseness, we used BIC in the same way as choosing the number of clusters.

Table 7 shows the results of parameter estimation. We state the coefficients representing the relationships among the latent variables because they are sparsely estimated. ⟶ in the table indicates the direction of the arrow in a path diagram, for example, CSOR⟶LIKE represents the relationship from CSOR to LIKE. $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are the coefficient matrices of Cluster 1 and Cluster 2, respectively, and the bold letters in the table indicate coefficients that are 0 by sparse estimation.

Before describing the interpretation of the results, we examine the validity of the estimated model. The GFI, which is one of the fit indices of SEM is 0.436 for PS-SEM, 0.434 for MS-SEM, 0.435 for MSEM, and 0.407 for KM+MGSSEM. From this result, it can be said that the proposed method estimates a model that fits the data while simplifying the path structure by sparse estimation to improve interpretability. Next, we provide the interpretation of the results for each method in Table 7.

First, we discuss the results for PS-SEM. The coefficient representing CSOR⟶LIKE in $\boldsymbol{B}_1$ was estimated to be 0. Due to this, the evaluation of the company's social responsibility by the respondents belonging to Cluster 1 does not affect its likeability. However, the coefficient representing CSOR⟶COMP in $\boldsymbol{B}_2$ was estimated as 0. Hence evidently, the social responsibility of the respondents belonging to Cluster 2 affects its likeability but does not affect its

Table 7 Application results for each method

Structural model	PS-SEM		MS-SEM		MSEM		KM+MGSSEM	
	$\boldsymbol{B}_1$	$\boldsymbol{B}_2$	$\boldsymbol{B}_1$	$\boldsymbol{B}_2$	$\boldsymbol{B}_1$	$\boldsymbol{B}_2$	$\boldsymbol{B}_1$	$\boldsymbol{B}_2$
CSOR⟶LIKE	**0.000**	0.325	**0.000**	0.118	0.017	0.431	1.007	0.194
ATTR⟶LIKE	0.161	0.252	0.098	0.206	0.531	1.932	**0.000**	0.247
PERF⟶LIKE	0.309	0.074	0.520	**0.000**	1.330	0.089	0.292	0.157
QUAL⟶LIKE	0.274	−0.324	0.577	0.284	1.729	0.671	0.257	0.120
CSOR⟶COMP	1.325	**0.000**	0.723	0.327	1.680	1.345	1.694	0.217
ATTR⟶COMP	0.704	**0.000**	0.417	**0.000**	1.290	0.171	**0.000**	**0.000**
PERF⟶COMP	1.221	1.228	0.204	0.343	0.603	0.955	0.817	1.464
QUAL⟶COMP	1.699	0.683	0.241	0.430	0.783	1.431	1.411	0.778
LIKE⟶CUSA	0.040	0.195	0.059	0.287	0.026	0.152	0.019	0.341
COMP⟶CUSA	0.019	0.023	0.020	−0.055	0.003	−0.030	0.023	−0.068
LIKE⟶CUSL	0.288	0.490	0.399	0.790	0.469	0.812	0.487	0.410
COMP⟶CUSL	**0.000**	**0.000**	**0.000**	**0.000**	−0.281	−0.114	**0.000**	**0.000**
CUSA⟶CUSL	3.081	1.689	4.091	1.836	4.938	2.482	2.978	1.496

competence. As the coefficient representing ATTR⟶COMP was 0, evidently, attractiveness also does not affect competence. Moreover, the coefficient representing COMP⟶CUSL was 0 in both clusters, which was the same result for all methods with sparse estimation. Therefore, its competence does not affect customer loyalty overall, and it is suggested that the COMP⟶CUSL relationship is unnecessary in the design of path structure.

Second, in MS-SEM, the part of $\boldsymbol{B}_1$, which is sparsely estimated, is the same as that in PS-SEM, and we can interpret it similarly. In $\boldsymbol{B}_2$, the result that the coefficient representing ATTR⟶COMP is estimated to be 0 is the same as that of PS-SEM, but the difference is that the coefficient of PERF⟶LIKE is estimated to be 0. Consequently, the evaluation of the company's performance does not affect its likeability. In fact, this coefficient is relatively small in the other three methods, and we can conclude that MS-SEM was able to perform sparse estimation validly.

Third, in MSEM all coefficients remain, because they are estimated without penalization. Noteworthy, the coefficients that are sparsely estimated by the proposed methods have small values in MSEM, which indicates that the proposed method estimates the relations with small effects as 0 and improves the interpretability. For example, coefficients corresponding to CSOR⟶LIKE, CSOR⟶LIKE, ATTR⟶COMP and COMP⟶CUSL are estimated as zero. The estimated results have the same tendency as the results of PLS-SEM-KM in the data (Fordellone and Vichi, 2020). In fact, these coefficients of CSOR⟶LIKE, CSOR⟶LIKE, ATTR⟶COMP and COMP ⟶CUSL were estimated as close to zero in PLS-SEM-KM. Therefore, the results of PS-SEM, MS-SEM, and MSEM are consistent with those of the previous study.

Finally, in KM+MGSSEM, the coefficients estimated as 0 by the proposed methods are estimated to be 0. However, the coefficients representing ATTR⟶LIKE and ATTR⟶COMP are estimated to be 0, which is a relatively large value in MSEM. This suggests that the tandem method may perform sparse estimation even for relations with a large influence.

5 Concluding Remarks

In this paper, we propose two novel clustered sparse SEM to estimate the shape of the path diagram of each cluster differently for heterogeneous data and make the interpretation for each cluster more easily. To achieve this purpose, clustered SEM, which is a simultaneous analysis of SEM and clustering, was incorporated in the sparse estimation.

The first method is PS-SEM, which enables sparse estimation by combining existing SSEM with clustering based on Kmeans. The clustering in PS-SEM is performed by assigning individuals to clusters with the maximum log-likelihood, transforming the likelihood of PS-SEM, and then introducing the indicator variable that indicates individual assignment. Therefore, it can provide results for which the characteristics of each cluster are interpretable. The second method is MS-SEM, which is an extension of MSEM, which has been proposed in the framework of Gaussian mixture distribution, with sparse estimation. In MS-SEM, the method of parameter estimation is similar to that of PS-SEM, but the clustering method is different. The responsibility is used for clustering because the clustering in MS-SEM is based on MSEM. The objective functions of PS-SEM and M-SEM are both the penalized log-likelihood. The EM algorithm is used for the maximization of the objective functions. However, in both methods, it is difficult to estimate the proper covariance matrix of the error variables, because the determinant is included in the objective function. Due to this problem, the update formula is derived by using ICF, which is used in the framework of Gaussian graphical modeling.

In Sect. 3, we compared the results of PS-SEM, MS-SEM, MSEM, and KM+MGSSEM by numerical simulation, and it was found that PS-SEM has the best clustering accuracy. In TPR, PS-SEM performed the most accurate sparse estimation, whereas MS-SEM did not perform sparse estimation. However, MS-SEM exhibited the highest estimation accuracy and showed the smallest estimation error. From the TPR results, the proposed methods are considered as providing user-friendly results because the proposed methods can provide sparse path-structure and enhance interpreting features of each cluster.

In the real data application described in Sect. 4, we mainly focused on the estimations of the coefficients as 0 and interpreted each cluster. In the proposed methods, the parts where the coefficients are small in the existing methods are sparsely estimated, and each cluster is interpreted more easily in a way that emphasizes the characteristics of each cluster. In addition, the proposed methods obtained a simpler path structure by sparse estimation while maintaining the goodness-of-fit of the data.

As discussed above, although the form of the objective function is different between PS-SEM and MS-SEM, the flow of estimating each parameter is the same, and the parameter updated formulas of each method are very similar, as described in Subsections 2.2 and 2.3. Furthermore, we focus on the weights $W_{\underline{\theta}_q}$ and W_{θ_q} of PS-SEM and MS-SEM, respectively. In the PS-SEM update formulas, the denominator part of $W_{\underline{\theta}_q}$ contains w_g, which is the proportion of individuals belonging to cluster g. At the same time, in the MS-SEM update formulas, responsibility r_{ng} occurred, which represents the probability that an individual arises from the normal distribution corresponding to the cluster g. Evidently, in both methods, the parameters are updated by weighting the linear combination of the parameters without the terms related to clustering. Therefore, the information about clustering has a significant influence on the parameter estimation, which can be explained by the fact that each evaluation index of KM+MGSSEM is poor in the simulation study.

Based on the above, whether the PS-SEM or MS-SEM method is more appropriate depends on the cluster structure behind the data. This is a similar problem to the comparison between model-based clustering and Kmeans clustering. However, the simulation results suggest that PS-SEM has a relatively high clustering accuracy and can obtain a sparse coefficient matrix, whereas MS-SEM can estimate the coefficients closer to the true causal structure. As mentioned in Subsection 2.2, PS-SEM has the advantage that the computational cost is less than that of MS-SEM because it is not necessary to compute the expected value for all individuals in the E-step. MS-SEM has the advantage of being a more stable estimation because Kmeans-based clustering is more dependent on the initial values and is hard clustering. MS-SEM is also more suitable for cases where the number of individuals in a cluster is not equal.

Finally, we discuss the scope for further research. There are four things for future work. First, there is room for improvement in the regularization term and objective function because MS-SEM did not have sufficient sparse estimation results in the simulation study. The number of clusters and regularization parameters are determined by grid search in this study; therefore, the construction of a rational selection criterion remains an issue. SEM with sparse estimation is also called the semi-confirmatory method (Huang, 2020). This means that it explores the optimal path structure from a set of path structures assumed to some extent in advance by pruning. In short, it is not supposed to discover the optimal path diagram from a path diagram (full path) that assumes relationships among all variables and latent variables or to determine the number of latent variables. This is due to the identifiability of SEM and the fact that the order of pruning by sparse estimation is not uniquely determined. These problems make it difficult to compare the path diagrams estimated from the full path for each cluster. Therefore, the extent to which we can weaken the restriction of the assumed path diagram in advance is also an issue when searching for comparable path diagrams for each cluster in the proposed method. Second, the numerical simulation was designed to evaluate the results of the proposed methods; however, the numerical simulations were performed in a specific situation. For example, the RMSE of MS-SEM is the best among these methods; however, the differences of among them were relatively small. The differences depend on the true path structures and coefficients. To deal with these problems, it needs to improve the computational time and conduct further numerical simulations. Third, in a real example, we adopted GFI as an evaluation index according to previous studies such as (Huang, 2018; Steiger, 1998). However, it needs further consideration from various perspectives. Finally, although the MSEM originally combines the EM algorithm with the conjugate gradient method, MS-SEM does not use the conjugate gradient method. Therefore, it needs to use conjugate gradient method in MS-SEM to improve the computational time.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00357-023-09449-9.

Acknowledgements We appreciate the constructive comments and valuable suggestions from the reviewers.

Data availability The data (Sarstedt, Hair, Cheah, Becker & Ringle, 2019) is included in the R package of “seminr.” The url is as follows: https://cran.r-project.org/web/packages/seminr/

Declarations

Ethical Approval This manuscript has not been published or presented elsewhere in part or in entirety.

Conflict of interest The authors declare no competing interests.

References

Bagozzi, R. P. (1982). A field investigation of causal relations among cognitions, affect, intentions, and behavior. *Journal of Marketing Research, 19*(4), 562–584.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Chaudhuri, S., Drton, M., & Richardson, T. S. (2007). Estimation of a covariance matrix with zeros. *Biometrika, 94*(1), 199–216.

Chin, W. W. (1998). The partial least squares approach to structural equation modeling. *Modern Methods for Business Research, 295*(2), 295–336.

Eberl, M. (2010). An application of pls in multi-group analysis: The need for differentiated corporate-level marketing in the mobile communications industry. in '*Handbook of partial least squares*', pp. 487–514. Springer

Fop, M., Murphy, T. B., & Scrucca, L. (2019). Model-based clustering with sparse covariance matrices. *Statistics and Computing, 29*(4), 791–819.

Fordellone, M., & Vichi, M. (2020). 'Finding groups in structural equation modeling through the partial least squares algorithm'. *Computational Statistics and Data Analysis 147*, online

Galimberti, G., Montanari, A., & Viroli, C. (2009). Penalized factor mixture analysis for variable selection in clustered data. *Computational Statistics and Data Analysis, 53*(12), 4301–4310.

Hair, J. F., Jr., Hult, G. T. M., Ringle, C., & Sarstedt, M. (2016). *A primer on partial least squares structural equation modeling (PLS-SEM)*. Los Angeles: SAGE Publications.

Hair, J. F., Jr., Matthews, L. M., Matthews, R. L., & Sarstedt, M. (2017). Pls-sem or cb-sem: Updated guidelines on which method to use. *International Journal of Multivariate Data Analysis, 1*(2), 107–123.

Heck, R. H., & Thomas, S. L. (2015). *An introduction to multilevel modeling techniques: MLM and SEM approaches using Mplus*. New York: Routledge.

Huang, P.-H. (2018). A penalized likelihood method for multi-group structural equation modelling. *British Journal of Mathematical and Statistical Psychology, 71*(3), 499–522.

Huang, P.-H. (2020). lslx: Semi-confirmatory structural equation modeling via penalized likelihood. *Journal of Statistical Software, 93*(7), 1–37.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193–218.

Hwang, H., DeSarbo, W. S., & Takane, Y. (2007). Fuzzy clusterwise generalized structured component analysis. *Psychometrika, 72*(2), 181–198.

Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science, 16*(1), 39–59.

Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997). Stemm: A general finite mixture structural equation model. *Journal of Classification, 14*(1), 23–50.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 409–426.

Kamakura, W. A., Kim, B.-D., & Lee, J. (1996). Modeling preference and structural heterogeneity in consumer choice. *Marketing Science, 15*(2), 152–172.

Liang, X., & Jacobucci, R. (2020). Regularized structural equation modeling to detect measurement bias: Evaluation of lasso, adaptive lasso, and elastic net. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(5), 722–734.

Lindstrøm, J.C.,& Dahl, F. A. (2020). Model selection with lasso in multi-group structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(1), 33–42.

Matthews, L. M., Sarstedt, M., Hair, J. F., & Ringle, C. M. (2016). 'Identifying and treating unobserved heterogeneity with fimix-pls'. *European Business Review*

Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the em algorithm. *Biometrics, 55*(2), 463–469.

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

Powell, M. J. D. (1977). Restart procedures for the conjugate gradient method. *Mathematical programming, 12*(1), 241–254.

Reinartz, W., Haenlein, M., & Henseler, J. (2009). An empirical comparison of the efficacy of covariance-based and variance-based sem. *International Journal of research in Marketing, 26*(4), 332–344.

Rubin, D. B., & Thayer, D. T. (1982). Em algorithms for ml factor analysis. *Psychometrika, 47*(1), 69–76.

Sarstedt, M., Hair, J. F., Jr., Cheah, J. H., Becker, J. M., & Ringle, C. M. (2019). How to specify, estimate, and validate higher-order constructs in pls-sem. *Australasian Marketing Journal (AMJ), 27*(3), 197–211.

Sarstedt, M., & Mooi, E. (2014). ‘A concise guide to market research: The process, data, and methods using ibm spss statistics’

Smid, S. C., McNeish, D., Miočević, M., & van de Schoot, R. (2020). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(1), 131–161.

Steiger, J. H. (1998). A note on multiple sample extensions of the rmsea fit index. *Structural Equation Modeling: A Multidisciplinary Journal, 5*(4), 411–419.

Xie, B., Pan, W., & Shen, X. (2010). Penalized mixtures of factor analyzers with application to clustering high-dimensional microarray data. *Bioinformatics, 26*(4), 501–508.

Zhou, H., Pan, W., & Shen, X. (2009). Penalized model-based clustering with unconstrained covariance matrices. *Electronic journal of statistics, 3*, 1473.