Journal of Classification
https://doi.org/10.1007/s00357-024-09492-0

ORIGINAL RESEARCH

Studying Hierarchical Latent Structures in Heterogeneous Populations with Missing Information

Francesca Greselin[1] · **Giorgia Zaccaria**[1]

Accepted: 17 September 2024

Abstract
An ultrametric Gaussian mixture model is a powerful tool for modeling hierarchical relationships among latent concepts, making it ideal for studying complex phenomena in diverse and potentially heterogeneous populations. However, in many cases, only an incomplete set of observations is available on the phenomenon under study. To address this issue, we propose MissUGMM, an ultrametric Gaussian mixture model which takes into account the missing at random mechanism for the unobserved values. Our approach is estimated using the expectation-maximization algorithm and achieves favorable results in comparison to other existing mixture models in simulations conducted with synthetic and benchmark data sets, even without a theorized ultrametric structure underlying the data. Furthermore, MissUGMM is applied to a real-world problem for exploring the sustainable development of cities across countries starting from incomplete information provided by municipalities. Overall, our results demonstrate that MissUGMM is a powerful and versatile model in dealing with missing data and is applicable to a broader range of real-world problems.

Keywords Ultrametricity · Gaussian mixture model · Missing data · Hierarchy of latent concepts · Cities' sustainable development

1 Introduction

In real applications, missing values often occur in the data by requiring specific strategies to treat them. Available-case analysis and imputation beforehand are examples of ad hoc methods used to force an incomplete data set into a rectangular complete data format on which applying statistical methodologies (Little & Rubin, 2019). To introduce proper models able to handle missing data, the mechanism that governs the relationship between a missing variable and its underlying value has to be analyzed. Assuming that the indicator pinpointing the pattern of unobserved values of a variable is random, Rubin (1976) distinguished three cases that differ for the assumptions on the missing indicator distribution. The most restrictive is called missing completely at random (MCAR) and considers the unobserved values not

✉ Giorgia Zaccaria
giorgia.zaccaria@unimib.it

Francesca Greselin
francesca.greselin@unimib.it

[1] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, Milan 20100, Italy

Published online: 04 October 2024

depending on the data, whether they are missing or observed; under the less restrictive missing at random (MAR) assumption, the unobserved values depend only on the observed data, whereas the missing not at random (MNAR) mechanism assumes the distribution of the missing indicator depending on the missing values themselves. In different fields, MCAR and MAR are the prevailing used mechanisms thanks to their easier tractability. Indeed, under these assumptions, the missingness mechanism can be considered to be ignorable, i.e., the parameters of the missing data distribution are distinct from the parameters of the observed data distribution (Schafer, 1997; Little & Rubin, 2019).

In the model-based clustering literature (see, among others, McLachlan & Peel, 2000; Fraley & Raftery, 2002; McNicholas, 2016; Bouveyron et al., 2019, for an overview), several methodologies have been proposed for modeling heterogeneous populations in the presence of missing data by means of finite mixture models. Ghahramani and Jordan (1995) first handled the incomplete data problem in Gaussian mixture models (GMMs), where the mixture components are assumed to be normally distributed with component mean vectors and covariance matrices as parameters. More parsimonious GMMs were proposed with the aim of reducing the number of the model parameters mostly deriving from the component covariance matrices. Specifically, Serafini et al. (2020) extended the work of Ghahramani and Jordan (1995) by including constrained structures for the component covariance matrices that are based upon their eigen-decomposition, as previously proposed by Banfield and Raftery (1993) and Celeux and Govaert (1995) in the presence of complete observations. Furthermore, Wang (2013) and Wang and Lin (2020) dealt with the missing information problem in GMMs with a twofold goal that consists of reaching model parsimony on one hand and dimensionality reduction on the other hand. These authors worked on the extension of the mixtures of (common) factor analyzers (MFA, Ghahramani & Hinton, 1997; McLachlan et al., 2003; Baek et al., 2010) with missing values by assuming a factorial parameterization of the component covariance matrices. It has to be highlighted that these methodologies were developed under the MAR assumption and estimated via the expectation-maximization (EM) algorithm (Dempster et al., 1977; Redner & Walker, 1984) and its extensions (McLachlan & Krishnan, 2008). The EM algorithm is indeed effective for handling the missing data problem, as it can deal with both conceptual and actual missing data.

Notably, the MFA model inspects latent structures underlying the data assuming uncorrelated factors; however, hierarchical relationships among unobserved dimensions can also occur in multidimensional phenomena. To deal with such a case, Cavicchia et al. (2022) introduced an ultrametric Gaussian mixture model, where the hierarchical relationships among *completely observed* variables are modeled via an ultrametric structure with the pivotal feature of being one-to-one associated with a hierarchy of latent concepts. Nonetheless, this model is not able to tackle the occurrence of missing values, that is often recurring in several applications, unless removing or imputing the data beforehand.

In this paper, we introduce MissUGMM, an ultrametric Gaussian mixture model designed to handle missing data. Within this model, we adopt the MAR mechanism for the unobserved values and propose an EM algorithm to effectively take this additional source of missingness into account. The performance of the proposal is assessed through an extensive simulation study both on synthetic data, where a hierarchical structure of variables is assumed, and on benchmark data sets, in which hierarchical relationships among variables have not been inspected beforehand. In terms of classification and missing values imputation, the results provide evidence on the advantages of using this new methodology compared with other existing mixture models in the literature. Moreover, the study on synthetic data assesses the performance of MissUGMM in recovering the hierarchical structure of variables. Finally, having tailored the model proposed by Cavicchia et al. (2022) to account for missing data

allows to analyze a broader manifold of multidimensional phenomena, whose measurement is often affected by lack of information. Specifically, we apply MissUGMM to study the cities' sustainable development in an incomplete data framework. We delve into the characterization of this phenomenon by pinpointing the dimensions that may differently contribute to its definition within and across countries. Overall, the study provides interesting insights into distinct patterns of sustainable development among cities.

The remainder of the paper is organized as follows. In Section 2, we introduce MissUGMM, together with details on the EM algorithm used for its estimation. Section 3 illustrates an extensive simulation study on both synthetic and benchmark data sets, which is further expanded upon in the Supplementary Materials. The application of MissUGMM to a real data set is depicted in Section 4, where we investigate the cities' sustainable development in the world. Section 5 completes the paper and outlines future developments of the proposed methodology.

2 Ultrametric Gaussian Mixture Model with Missing Data

2.1 Background

Let $\mathbf{X} = (X_1, \ldots, X_p)'$ be a p-dimensional random vector and $\mathbf{x} = (x_1, \ldots, x_p)'$ its realization. Suppose that $\mathbf{X}$ follows a finite mixture of G multivariate Gaussian distributions, whose pdf is given by

$$f(\mathbf{x}; \boldsymbol{\Psi}) = \sum_{g=1}^{G} \pi_g \phi_p\big(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\big), \tag{1}$$

where π_g is the mixing proportion of the gth component such that $\pi_g > 0$ for $g = 1, \ldots, G$, and $\sum_{g=1}^{G} \pi_g = 1$, $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ are the mean vector and covariance matrix of the gth mixture component, respectively. The pdf parameters in Eq. 1 are encompassed by $\boldsymbol{\Psi} = \{\boldsymbol{\pi}, \boldsymbol{\theta}\}$, with $\boldsymbol{\pi} = \{\pi_g\}_{g=1}^{G}$ and $\boldsymbol{\theta} = \{\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\}_{g=1}^{G}$. The ultrametric Gaussian mixture model (Cavicchia et al., 2022) parameterizes the component covariance matrix $\boldsymbol{\Sigma}_g$ as follows:

$$\boldsymbol{\Sigma}_g = \mathbf{V}_g(\boldsymbol{\Sigma}_{\mathrm{W}_g} + \boldsymbol{\Sigma}_{\mathrm{B}_g})\mathbf{V}_g' + \mathrm{diag}\big(\mathbf{V}_g(\boldsymbol{\Sigma}_{\mathrm{V}_g} - \boldsymbol{\Sigma}_{\mathrm{W}_g})\mathbf{V}_g'\big). \tag{2}$$

The $(p \times Q)$ variable-group membership matrix $\mathbf{V}_g$ defines a partition of the variable space into a reduced number Q of groups, with $Q \leq p$. The remaining three parameters in Eq. 2 express the features of the variable groups. The diagonal group-wise variance matrix $\boldsymbol{\Sigma}_{\mathrm{V}_g}$ of order Q identifies their variance, denoted by ${}_{V_g}\sigma_{qq}$. Lastly, $\boldsymbol{\Sigma}_{\mathrm{W}_g}$ is the diagonal within-group covariance matrix of order Q representing the covariance within the variable groups, indicated by ${}_{W_g}\sigma_{qq}$, and $\boldsymbol{\Sigma}_{\mathrm{B}_g}$ is the between-group covariance matrix of the same order embodying the relationships among the variable groups, via ${}_{B_g}\sigma_{qh}$. Therefore, the set of parameters $\boldsymbol{\theta}$ is given by $\{\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_{\mathrm{V}_g}, \boldsymbol{\Sigma}_{\mathrm{W}_g}, \boldsymbol{\Sigma}_{\mathrm{B}_g}, \mathbf{V}_g\}_{g=1}^{G}$.

$\boldsymbol{\Sigma}_g$ in Eq. 2 is extended ultrametric (Definition 2, Cavicchia et al., 2022) if the following constraints hold for each component of the mixture:

(i) $\mathbf{V}_g$ is binary and row-stochastic.
(ii) $\boldsymbol{\Sigma}_{\mathrm{B}_g}$ is symmetric, with all diagonal entries equal to zero and triplets of off-diagonal values complying with the ultrametric condition, i.e., ${}_{B_g}\sigma_{qh} \geq \min\{{}_{B_g}\sigma_{qs}, {}_{B_g}\sigma_{hs}\}$, $q, h, s = 1, \ldots, Q, s \neq h \neq q$.
(iii) $\min\{{}_{W_g}\sigma_{qq} : q = 1, \ldots, Q\} \geq \max\{{}_{B_g}\sigma_{qh} : q, h = 1, \ldots, Q, h \neq q\}$.
(iv) ${}_{V_g}\sigma_{qq} > |{}_{W_g}\sigma_{qq}|$ for $q = 1, \ldots, Q$.

(v) $\boldsymbol{\Sigma}_g$ is positive definite (pd).

Constraints (iii) and (iv) impose an ordering among the elements of $\boldsymbol{\Sigma}_{\mathrm{V}_g}$, $\boldsymbol{\Sigma}_{\mathrm{W}_g}$, and $\boldsymbol{\Sigma}_{\mathrm{B}_g}$, which entails a hierarchy over the variable groups, each associated with a latent concept. For satisfying constraint (v), if $\boldsymbol{\Sigma}_g$ is not pd, a coefficient a corresponding to the absolute value of its smallest eigenvalue (Cailliez, 1983) plus an arbitrary small positive constant ξ is added to its main diagonal. It has to be noticed that this corresponds to imposing the constraint $\lambda_p(\boldsymbol{\Sigma}_g) \geq \xi$, where $\lambda_p(\boldsymbol{\Sigma}_g)$ is the pth eigenvalue of $\boldsymbol{\Sigma}_g$ and $\lambda_p(\boldsymbol{\Sigma}_g) \leq \cdots \leq \lambda_1(\boldsymbol{\Sigma}_g)$.

2.2 MissUGMM and Its Parameter Estimation

In many real applications, data can be affected by missing values occurring in the collection process giving $\mathbf{x}$ to be divided into a p^o- and a $(p - p^o)$-dimensional vector containing the observed and missing variables, respectively. The ultrametric Gaussian mixture model with missing data (MissUGMM) assumes that the missing data mechanism is MAR and ignorable. MissUGMM is estimated via the EM algorithm by considering two sources of missing information: the missing values in the data and the unit-component membership. Letting $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ be a random sample of size n, where the ith observation $\mathbf{x}_i$ is divided into $(\mathbf{x}_i^{o_i\prime}, \mathbf{x}_i^{m_i\prime})'$, the missing data correspond to $\{\mathbf{x}_i^{m_i}\}_{i=1}^n$, whereas the unit-component memberships are denoted as $\{\mathbf{z}_i\}_{i=1}^n$ with $\mathbf{z}_i = (z_{i1}, \ldots, z_{iG})'$, where $z_{ig} = 1$ if the ith observation belongs to the gth component, and $z_{ig} = 0$ otherwise. It is worth noticing that the pattern of missing values can differ across the n observations; hence, o_i and m_i depend on i.

The MissUGMM log-likelihood of the complete data $\{\mathbf{x}_i^{o_i}, \mathbf{x}_i^{m_i}, \mathbf{z}_i\}_{i=1}^n$ is

$$\begin{aligned}\ell_c(\boldsymbol{\Psi}) = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\Bigg\{ &\log(\pi_g) - \frac{1}{2}\log\left(|\boldsymbol{\Sigma}_g|\right) \\ &- \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_g^{-1}\left(\begin{bmatrix}\mathbf{x}_i^{o_i}\\ \mathbf{x}_i^{m_i}\end{bmatrix} - \boldsymbol{\mu}_g\right)\left(\begin{bmatrix}\mathbf{x}_i^{o_i}\\ \mathbf{x}_i^{m_i}\end{bmatrix} - \boldsymbol{\mu}_g\right)'\right]\Bigg\},\end{aligned} \tag{3}$$

where $\boldsymbol{\Sigma}_g$ is the extended ultrametric covariance matrix in Eq. 2 subject to constraints (i)–(v) (refer to Section 2.1, omitted henceforth). In Eq. 3 onward, constant terms not depending on the model parameters are omitted and the squared Mahalanobis distance $\delta\left(\mathbf{x}_i, \boldsymbol{\mu}_g; \boldsymbol{\Sigma}_g\right)$ is rewritten by considering the trace properties, i.e., $\delta\left(\mathbf{x}_i, \boldsymbol{\mu}_g; \boldsymbol{\Sigma}_g\right) = (\mathbf{x}_i - \boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_g) = \mathrm{tr}\left((\mathbf{x}_i - \boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_g)\right) = \mathrm{tr}\left(\boldsymbol{\Sigma}_g^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_g)(\mathbf{x}_i - \boldsymbol{\mu}_g)'\right)$, where $\mathbf{x}_i$ is partitioned into observed and missing vectors.

The EM algorithm maximizes the complete data log-likelihood in Eq. 3 by alternating an expectation step (E-step) and a maximization step (M-step) until convergence. The E-step consists of the computation of the expected value of Eq. 3 given the observed data $\{\mathbf{x}_i^{o_i}\}_{i=1}^n$ and the current estimates of the model parameters in $\boldsymbol{\Psi}$; then the M-step involves the maximization of the expected value obtained in the E-step. Before detailing them, we recall some useful results for the E-step, stemming from the properties of the multivariate Gaussian distribution.

Proposition 1 *If a p-dimensional random vector $\mathbf{X}_i$ is partitioned into $(\mathbf{X}_i^{o_i\prime}, \mathbf{X}_i^{m_i\prime})'$, then*

$$\begin{bmatrix}\mathbf{X}_i^{o_i}\\ \mathbf{X}_i^{m_i}\end{bmatrix}\Bigg| z_{ig} = 1 \sim N_p\left(\begin{bmatrix}\boldsymbol{\mu}_g^{o_i}\\ \boldsymbol{\mu}_g^{m_i}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_g^{o_i,o_i} & \boldsymbol{\Sigma}_g^{o_i,m_i}\\ \boldsymbol{\Sigma}_g^{m_i,o_i} & \boldsymbol{\Sigma}_g^{m_i,m_i}\end{bmatrix}\right). \tag{4}$$

Therefore,

$$\mathbf{X}_i^{o_i} \big| z_{ig} = 1 \sim N_{p^{o_i}}\left(\boldsymbol{\mu}_g^{o_i}, \boldsymbol{\Sigma}_g^{o_i,o_i}\right), \tag{5}$$

$$\mathbf{X}_i^{m_i} \big| \mathbf{x}_i^{o_i}, z_{ig} = 1 \sim N_{p-p^{o_i}}\left(\boldsymbol{\mu}_g^{m_i|o_i}, \boldsymbol{\Sigma}_g^{m_i,m_i|o_i}\right), \tag{6}$$

where $\mathbf{x}_i^{o_i}$ *is a realization of* $\mathbf{X}_i^{o_i}$ *and*

$$\boldsymbol{\mu}_g^{m_i|o_i} = \boldsymbol{\mu}_g^{m_i} + \boldsymbol{\Sigma}_g^{m_i,o_i}\left[\boldsymbol{\Sigma}_g^{o_i,o_i}\right]^{-1}(\mathbf{x}_i^{o_i} - \boldsymbol{\mu}_g^{o_i}),$$
$$\boldsymbol{\Sigma}_g^{m_i,m_i|o_i} = \boldsymbol{\Sigma}_g^{m_i,m_i} - \boldsymbol{\Sigma}_g^{m_i,o_i}\left[\boldsymbol{\Sigma}_g^{o_i,o_i}\right]^{-1}\boldsymbol{\Sigma}_g^{o_i,m_i}.$$

Proof See Mardia et al. (1979, Theorem 3.2.4, p. 63), among others. □

Proposition 1 is based upon the theorem we recall below.

Theorem 1 *If* $\boldsymbol{\Sigma}_g$ *is an extended ultrametric covariance matrix partitioned as follows*

$$\begin{bmatrix} \boldsymbol{\Sigma}_g^{o_i,o_i} & \boldsymbol{\Sigma}_g^{o_i,m_i} \\ \boldsymbol{\Sigma}_g^{m_i,o_i} & \boldsymbol{\Sigma}_g^{m_i,m_i} \end{bmatrix}$$

according to the corresponding partition of $\mathbf{X}_i|z_{ig} = 1$, $\boldsymbol{\Sigma}_g^{o_i,o_i}$ *is nonsingular.*

Proof The variables in $\mathbf{X}_i^{o_i}$ can belong to the same variable group or different groups in $\mathbf{V}_g$. In both cases, since $\boldsymbol{\Sigma}_g$ is pd thanks to constraint (v), its principal minor composed of the p^{o_i} variables in $\mathbf{X}_i^{o_i}$ is pd as well (Sylvester's criterion, Gilbert, 1991; Horn & Johnson, 2013, Theorem 7.2.5, p. 439). □

The steps of the EM algorithm for maximizing (3) are described hereinafter. In this section, we omit the reference to the iteration t, for simplicity reasons; then, the estimates represented by the symbols $\hat{\cdot}$ and $\tilde{\cdot}$ refer to the iteration t, e.g., $\widehat{\boldsymbol{\Psi}}$ and $\tilde{z}_{ig}$. At iteration $t+1$, the E-step is formalized as

$$\begin{aligned} Q(\boldsymbol{\Psi}; \widehat{\boldsymbol{\Psi}}) &= \mathbb{E}\left[\ell_c(\boldsymbol{\Psi})|\mathbf{x}_1^{o_1}, \ldots, \mathbf{x}_n^{o_n}; \widehat{\boldsymbol{\Psi}}\right] \\ &= \sum_{i=1}^{n}\sum_{g=1}^{G}\mathbb{E}[Z_{ig}|\mathbf{x}_i^{o_i}; \widehat{\boldsymbol{\Psi}}]\Bigg\{\log(\pi_g) - \frac{1}{2}\log\left(|\boldsymbol{\Sigma}_g|\right) \\ &\quad - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_g^{-1}\mathbb{E}\left[\Pi(\mathbf{x}_i^{o_i}, \mathbf{X}_i^{m_i}, \boldsymbol{\mu}_g)|\mathbf{x}_i^{o_i}, z_{ig} = 1; \widehat{\boldsymbol{\Psi}}\right]\right)\Bigg\}, \end{aligned} \tag{7}$$

where $\Pi(\mathbf{x}_i^{o_i}, \mathbf{X}_i^{m_i}, \boldsymbol{\mu}_g) = \left(\begin{bmatrix}\mathbf{x}_i^{o_i} \\ \mathbf{X}_i^{m_i}\end{bmatrix} - \boldsymbol{\mu}_g\right)\left(\begin{bmatrix}\mathbf{x}_i^{o_i} \\ \mathbf{X}_i^{m_i}\end{bmatrix} - \boldsymbol{\mu}_g\right)'$. Equation 7 requires the computation of the following expected values:

(a) the expected value of Z_{ig} given the observed data $\mathbf{x}_i^{o_i}$ and the current estimate of the overall parameter vector $\widehat{\boldsymbol{\Psi}}$, i.e.,

$$\mathbb{E}[Z_{ig}|\mathbf{x}_i^{o_i}; \widehat{\boldsymbol{\Psi}}] = \frac{\widehat{\pi}_g \phi_{p^{o_i}}\left(\mathbf{x}_i^{o_i}; \widehat{\boldsymbol{\mu}}_g^{o_i}, \widehat{\boldsymbol{\Sigma}}_g^{o_i,o_i}\right)}{\sum_{h=1}^{G}\widehat{\pi}_h \phi_{p^{o_i}}\left(\mathbf{x}_i^{o_i}; \widehat{\boldsymbol{\mu}}_h^{o_i}, \widehat{\boldsymbol{\Sigma}}_h^{o_i,o_i}\right)} := \tilde{z}_{ig}; \tag{8}$$

(b) the expected value of $\Pi(\mathbf{x}_i^{o_i}, \mathbf{X}_i^{m_i}, \boldsymbol{\mu}_g)$ given the observed data $\mathbf{x}_i^{o_i}$, the membership of the ith observation to the gth component, that is $z_{ig} = 1$, and the current estimate of the overall parameter vector $\widehat{\boldsymbol{\Psi}}$, i.e.,

$$\mathbb{E}\big[\Pi(\mathbf{x}_i^{o_i}, \mathbf{X}_i^{m_i}, \boldsymbol{\mu}_g)\big|\mathbf{x}_i^{o_i}, z_{ig} = 1; \widehat{\boldsymbol{\Psi}}\big] = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{C} \end{bmatrix},$$

where

$$\begin{aligned}
\mathbf{A} &= \big(\mathbf{x}_i^{o_i} - \boldsymbol{\mu}_g^{o_i}\big)\big(\mathbf{x}_i^{o_i} - \boldsymbol{\mu}_g^{o_i}\big)' \\
\mathbf{B} &= \big(\mathbf{x}_i^{o_i} - \boldsymbol{\mu}_g^{o_i}\big)\big(\mathbb{E}[\mathbf{X}_i^{m_i}|\mathbf{x}_i^{o_i}, z_{ig} = 1; \widehat{\boldsymbol{\Psi}}] - \boldsymbol{\mu}_g^{m_i}\big)' \\
\mathbf{C} &= \mathbb{E}[(\mathbf{X}_i^{m_i} - \boldsymbol{\mu}_g^{m_i})(\mathbf{X}_i^{m_i} - \boldsymbol{\mu}_g^{m_i})'|\mathbf{x}_i^{o_i}, z_{ig} = 1; \widehat{\boldsymbol{\Psi}}].
\end{aligned}$$

Recalling Proposition 1, the expected values in $\mathbf{B}$ and $\mathbf{C}$ are computed as

$$\begin{aligned}
&\mathbb{E}[\mathbf{X}_i^{m_i}|\mathbf{x}_i^{o_i}, z_{ig} = 1; \widehat{\boldsymbol{\Psi}}] \\
&\quad = \widehat{\boldsymbol{\mu}}_g^{m_i} + \widehat{\boldsymbol{\Sigma}}_g^{m_i,o_i}\big[\widehat{\boldsymbol{\Sigma}}_g^{o_i,o_i}\big]^{-1}(\mathbf{x}_i^{o_i} - \widehat{\boldsymbol{\mu}}_g^{o_i}) := \widehat{\boldsymbol{\mu}}_g^{m_i|o_i}, \qquad (9) \\
&\mathbb{E}[(\mathbf{X}_i^{m_i} - \boldsymbol{\mu}_g^{m_i})(\mathbf{X}_i^{m_i} - \boldsymbol{\mu}_g^{m_i})'|\mathbf{x}_i^{o_i}, z_{ig} = 1; \widehat{\boldsymbol{\Psi}}] \\
&\quad = \widehat{\boldsymbol{\Sigma}}_g^{m_i,m_i|o_i} + (\widehat{\boldsymbol{\mu}}_g^{m_i|o_i} - \boldsymbol{\mu}_g^{m_i})(\widehat{\boldsymbol{\mu}}_g^{m_i|o_i} - \boldsymbol{\mu}_g^{m_i})', \qquad (10)
\end{aligned}$$

where $\widehat{\boldsymbol{\Sigma}}_g^{m_i,m_i|o_i} := \widehat{\boldsymbol{\Sigma}}_g^{m_i,m_i} - \widehat{\boldsymbol{\Sigma}}_g^{m_i,o_i}\big[\widehat{\boldsymbol{\Sigma}}_g^{o_i,o_i}\big]^{-1}\widehat{\boldsymbol{\Sigma}}_g^{o_i,m_i}$. It is worth noticing that the terms $-\widehat{\boldsymbol{\Sigma}}_g^{m_i,o_i}\big[\widehat{\boldsymbol{\Sigma}}_g^{o_i,o_i}\big]^{-1}\widehat{\boldsymbol{\Sigma}}_g^{o_i,m_i}$ in the latter equation and $\widehat{\boldsymbol{\Sigma}}_g^{m_i,o_i}\big[\widehat{\boldsymbol{\Sigma}}_g^{o_i,o_i}\big]^{-1}(\mathbf{x}_i^{o_i} - \widehat{\boldsymbol{\mu}}_g^{o_i})$ in Eq. 9 can be interpreted as the adjustment for imputing the conditions in the expectation computation.

Therefore, we obtain $Q(\boldsymbol{\Psi}; \widehat{\boldsymbol{\Psi}})$ necessary for the M-step, that can be re-written as

$$Q(\boldsymbol{\Psi}; \widehat{\boldsymbol{\Psi}}) = \sum_{i=1}^{n}\sum_{g=1}^{G}\tilde{z}_{ig}\log(\pi_g) - \frac{1}{2}\sum_{g=1}^{G}\tilde{n}_g\Big\{\log\big(|\boldsymbol{\Sigma}_g|\big) + \mathrm{tr}\Big(\boldsymbol{\Sigma}_g^{-1}\mathbf{S}_g\Big)\Big\} \qquad (11)$$

where $\mathbf{S}_g = \dfrac{1}{\tilde{n}_g}\sum_{i=1}^{n}\tilde{z}_{ig}\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{C} \end{bmatrix}$ with the expected values in Eqs. 9 and 10 replaced into $\mathbf{B}$ and $\mathbf{C}$, respectively, and $\tilde{n}_g = \sum_{i=1}^{n}\tilde{z}_{ig}$.

The M-step maximizes $Q(\boldsymbol{\Psi}; \widehat{\boldsymbol{\Psi}})$ with respect to $\boldsymbol{\Psi}$ by updating the estimates of the MissUGMM parameters. Specifically, at iteration $t+1$, denoted by * in the following M-step formulas,

(a) the estimate of the prior probabilities in $\boldsymbol{\pi} = \{\pi_g\}_{g=1}^{G}$ is

$$\widehat{\pi}_g^* = \frac{\tilde{n}_g}{n}; \qquad (12)$$

(b) the estimate of the mean vectors in $\boldsymbol{\theta} = \{\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\}_{g=1}^{G}$ is

$$\widehat{\boldsymbol{\mu}}_g^* = \frac{\sum_{i=1}^{n}\tilde{z}_{ig}\begin{bmatrix} \mathbf{x}_i^{o_i} \\ \widehat{\boldsymbol{\mu}}_g^{m_i|o_i} \end{bmatrix}}{\tilde{n}_g}; \qquad (13)$$

(c) the estimate of the extended ultrametric covariance matrices in $\boldsymbol{\theta} = \{\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\}_{g=1}^{G}$ is

$$\widehat{\boldsymbol{\Sigma}}_g^* = \widehat{\mathbf{V}}_g^*\big(\widehat{\boldsymbol{\Sigma}}_{\mathrm{W}_g}^* + \widehat{\boldsymbol{\Sigma}}_{\mathrm{B}_g}^*\big)\widehat{\mathbf{V}}_g^{*\prime} + \mathrm{diag}\big(\widehat{\mathbf{V}}_g^*(\widehat{\boldsymbol{\Sigma}}_{\mathrm{V}_g}^* - \widehat{\boldsymbol{\Sigma}}_{\mathrm{W}_g}^*)\widehat{\mathbf{V}}_g^{*\prime}\big). \qquad (14)$$

The estimates in Eq. 14 are obtained by plugging $\widehat{\boldsymbol{\mu}}_g^{o_i*}$ and $\widehat{\boldsymbol{\mu}}_g^{m_i*}$ into $\mathbf{S}_g$ to attain $\tilde{\mathbf{S}}_g^*$. The estimation of $\mathbf{V}_g$ is a combinatorial problem (see constraint i). Indeed, $\mathbf{V}_g$ is estimated row-by-row, i.e., for the jth row the one occurs in the qth column if assigning the jth variable to the qth group for the gth component maximizes (11), $j = 1, \ldots, p, q \in \{1, \ldots, Q\}$. Given the actual configuration of the estimate of $\mathbf{V}_g$, maximizing (11) with respect to the other parameters composing $\boldsymbol{\Sigma}_g$, one at a time, gives rise to the following estimates. The group-wise variance matrix is updated by

$$\widehat{\boldsymbol{\Sigma}}_{\mathrm{V}_g}^* = \widehat{\mathbf{V}}_g^{*+}\mathrm{diag}(\tilde{\mathbf{S}}_g^*)\widehat{\mathbf{V}}_g^{*}, \tag{15}$$

subject to constraint (iv) and where $(\cdot)^+$ denotes the Moore-Penrose inverse of a matrix. Given $\widehat{\boldsymbol{\Sigma}}_{\mathrm{V}_g}^*$, the updating formula of the within-group covariance matrix is given by

$$\widehat{\boldsymbol{\Sigma}}_{\mathrm{W}_g}^* = \left((\widehat{\mathbf{V}}_g^{*\prime}\widehat{\mathbf{V}}_g^{*})^2 - \widehat{\mathbf{V}}_g^{*\prime}\widehat{\mathbf{V}}_g^{*}\right)^{+}\mathrm{diag}\left[\widehat{\mathbf{V}}_g^{*\prime}\left(\tilde{\mathbf{S}}_g^* - \mathrm{diag}(\widehat{\mathbf{V}}_g^{*}\widehat{\boldsymbol{\Sigma}}_{\mathrm{V}_g}^{*}\widehat{\mathbf{V}}_g^{*\prime})\right)\widehat{\mathbf{V}}_g^{*}\right], \tag{16}$$

subject to constraint (iii). It is worth noticing that the Moore-Penrose inverse of a diagonal matrix is obtained by taking the reciprocal of its non-zero diagonal elements and letting the others set to zero. Therefore, it avoids the singularity problem of $\left((\widehat{\mathbf{V}}_g^{*\prime}\widehat{\mathbf{V}}_g^{*})^2 - \widehat{\mathbf{V}}_g^{*\prime}\widehat{\mathbf{V}}_g^{*}\right)$ that can occur if variable groups are singleton. In the latter case, the only parameter reflecting the group-specific feature is the variance and thus we set ${}_{W_g}\hat{\sigma}_{qq}^* = {}_{V_g}\hat{\sigma}_{qq}^*$ for the variable groups of size one.
Finally, $\widehat{\boldsymbol{\Sigma}}_{\mathrm{B}_g}^*$ is computed from

$$\tilde{\boldsymbol{\Sigma}}_{\mathrm{B}_g}^* = \widehat{\mathbf{V}}_g^{*+}\tilde{\mathbf{S}}_g^*(\widehat{\mathbf{V}}_g^{*\prime})^{+}, \tag{17}$$

such that constraint (ii) holds.

After convergence, the missing values are imputed using the conditional mean method. The predictor for the missing values of the ith observation is given by

$$\widehat{\mathbf{x}}_i^{m_i} = \mathbb{E}[\mathbf{X}_i^{m_i}|\mathbf{x}_i^{o_i}; \widehat{\boldsymbol{\Psi}}] = \sum_{g=1}^{G}\widehat{z}_{ig}\mathbb{E}[\mathbf{X}_i^{m_i}|\mathbf{x}_i^{o_i}, z_{ig} = 1; \widehat{\boldsymbol{\Psi}}],$$

where $\widehat{z}_{ig}$ are the posterior probabilities computed at convergence and $\mathbb{E}[\mathbf{X}_i^{m_i}|\mathbf{x}_i^{o_i}, z_{ig} = 1; \widehat{\boldsymbol{\Psi}}]$ is obtained as in Eq. 9 by considering the final estimates of the model parameters.

2.3 Algorithm Implementation Details

In the following two sections, we provide some details on the EM algorithm for estimating MissUGMM concerning its initialization, stopping criterion, and model selection.

2.3.1 Initialization

As pointed out by many authors, the log-likelihood function of a finite mixture model can have multiple local maxima by requiring the need for a strategy to increase the chance of obtaining a global optimum (see, for instance, McLachlan & Peel, 2000, Chapter 2.12). To reach this goal, the EM algorithm can be run several times by starting from different initial values for the model parameters and letting it converge each time. Among several solutions, the one with the highest value of the maximized log-likelihood is retained. Under the MAR

assumption and the distinctness of the observed and missing parameters, we can evaluate the observed data log-likelihood for MissUGMM

$$\ell(\widehat{\boldsymbol{\Psi}}^{o}) = \sum_{i=1}^{n} \log \left(\sum_{g=1}^{G} \widehat{\pi}_g \phi_{p^{o_i}} \left(\mathbf{x}_i^{o_i}; \widehat{\boldsymbol{\mu}}_g^{o_i}, \widehat{\boldsymbol{\Sigma}}_g^{o_i,o_i} \right) \right) \tag{18}$$

at each iteration of the EM algorithm (Little & Rubin, 2019) and among solutions obtained with different starting values (e.g., 30 in our experiments).

The starting values for computing the initial parameters of MissUGMM concern the imputation of the missing values $\{\mathbf{x}_i^{m_i}\}_{i=1}^n$, and the resulting initialization of $\{\mathbf{z}_i\}_{i=1}^n$ and $\{\mathbf{V}_g\}_{g=1}^G$ based on the completed data. Missing data can be imputed according to several approaches. One of the most used is the sample mean of the observed data for each variable (see, for instance, Wang, 2013); nonetheless, this approach does not take the heterogeneity of the data into account. For this reason, in the initialization step of our algorithm, we impute missing values according to the k-nearest neighbors method (Fix & Hodges, 1951; Cover & Hart, 1967) by using the Euclidean distance, setting $k = 5$ and considering only complete cases as neighbors. The unit-component membership and the variable-group membership can be initialized randomly or via specific methodologies for detecting clustering structures and variable partitions, respectively. Specifically, we recommend to use k-means (MacQueen, 1967) with $k = G$ to obtain initial values of $\{\mathbf{z}_i\}_{i=1}^n$, and the solution of the ultrametric algorithm proposed by Cavicchia et al. (2020) and adapted for covariance matrices to compute initial values for $\{\mathbf{V}_g\}_{g=1}^G$. Both methods return partitions of objects (observations or variables). It has to be highlighted that the predicted component membership of each unit is evaluated at the EM algorithm convergence (Section 2.3.2) according to the maximum a posteriori, i.e., $i \in g$ if $g = \arg\max\{\widehat{z}_{ig'} : g' = 1, \ldots, G\}$. Since $\widehat{z}_{ig}$ takes values in [0, 1], a feasible alternative to the k-means initialization is fuzzy k-means (Bezdek, 1974).

2.3.2 Stopping Criterion and Model Selection

The convergence of the EM algorithm is often evaluated through the relative change of the log-likelihood in Eq. 18 in two sequential iterations. However, Lindstrom and Bates (1988) stated that this is not a proper measure of convergence, but rather of lack of progress. For this reason, we select Aitken's acceleration procedure as the stopping criterion of the MissUGMM algorithm, which was first inspected by Böhning et al. (1994) (see also McLachlan & Krishnan, 2008, Section 4.9). By considering $\ell(\widehat{\boldsymbol{\Psi}}^{o(t)}) = \ell^{(t)}$ for simplicity reasons, the algorithm converges if $\ell_\infty^{(t+1)} - \ell^{(t)} < \epsilon$ (McNicholas et al., 2010), where ϵ is a small arbitrary positive constant (e.g., 1.5×10^{-8} in our experiments). $\ell_\infty^{(t+1)}$ is the Aitken accelerated estimate of the log-likelihood at iteration $t + 1$, which is given by

$$\ell_\infty^{(t+1)} = \ell^{(t)} + \frac{1}{1 - a^{(t)}} (\ell^{(t+1)} - \ell^{(t)}),$$

where $a^{(t)}$ represents the ratio of successive increments, i.e.,

$$a^{(t)} = \frac{\ell^{(t+1)} - \ell^{(t)}}{\ell^{(t)} - \ell^{(t-1)}}.$$

Alternative stopping criteria based on Aitken's acceleration procedure are illustrated in McNicholas et al. (2010).

The algorithm for estimating MissUGMM can be run when fixing the number of mixture components G and variable groups Q. Their choice is thus a crucial issue to cope with. If no

prior information exists on the clustering structure and the specific dimensions composing the phenomenon under study, we can use the Bayesian information criterion (BIC, Schwarz, 1978) to select the pair (G, Q). Generally speaking, BIC is the most prevailing choice of model selection criterion in GMMs thanks to its good performance under certain regularity conditions (Fraley & Raftery, 1998). BIC has the form

$$\text{BIC}_{G,Q} = 2\,\ell_{G,Q}(\widehat{\boldsymbol{\Psi}}^o) - \nu \log n,$$

where $\ell_{G,Q}(\widehat{\boldsymbol{\Psi}}^o)$ is the maximized observed data log-likelihood and $\nu = 2G(p+Q) - 1 - (c_{V,W} + c_{W,B})$ is the number of free parameters in the model. This number is obtained by considering $G-1$ parameters for the estimation of the mixing proportions, Gp for the mean vectors, and $G(p+2Q-1)-(c_{V,W}+c_{W,B})$ for the extended ultrametric covariance matrices. Specifically, the latter include p parameters in $\mathbf{V}_g$ minus Q constraints (non-empty groups), Q values in $\boldsymbol{\Sigma}_{\text{V}_g}$ and $\boldsymbol{\Sigma}_{\text{W}_g}$, $Q-1$ values in $\boldsymbol{\Sigma}_{\text{B}_g}$, $c_{V,W}$ and $c_{W,B}$ constraints corresponding to the cases in which constraints (iv) and (iii), respectively, are activated in the algorithm, i.e., ${}_{V_g}\sigma_{qq} = |{}_{W_g}\sigma_{qq}| + 1.5 \times 10^{-8}$ and $\min\{{}_{W_g}\sigma_{qq}, q = 1, \ldots, Q\} = \max\{{}_{B_g}\sigma_{qh}, q, h = 1, \ldots, Q, h \neq q\}$. For a review of the model selection criteria in GMMs see Celeux et al. (2018) and McLachlan and Rathnayake (2014).

3 Simulation Study

The performance of MissUGMM is evaluated on synthetic (Section 3.1) and benchmark (Section 3.2) data. In both cases, the proposed methodology is compared to GMM, implemented via the fast EM algorithm proposed by Lin et al. (2006), and MFA (Wang & Lin, 2020, with the number of factors fixed in advance) in the presence of missing data. These algorithms are run under the same conditions of the proposal, i.e., initialization (see Section 2.3.1), tolerance value (see Section 2.3.2), and maximum number of iterations (equal to 500 in our experiments) for convergence, while respecting the remaining default options set by the authors.

Synthetic data sets are used for analyzing the proposal's potential in clustering structure recovery, on the one hand, and detection of hierarchical structures on variables, on the other hand. Further comparison on benchmark data sets, where the underlying component covariance structure is not necessarily ultrametric, provides insight into the behavior of MissUGMM outside its natural framework. Finally, additional results on synthetic data generated by considering non-hierarchical covariance structures are reported in the Supplementary Materials.

3.1 Synthetic Data

We illustrate here a simulation study on synthetic data to evaluate the performance of the proposal when the covariance matrices of the mixture components are supposed to be ultrametric. In this setting, the comparison with GMM and MFA in the presence of missing information emphasizes the need for a specific, new methodology able to detect hierarchical relationships among variables within heterogeneous populations.

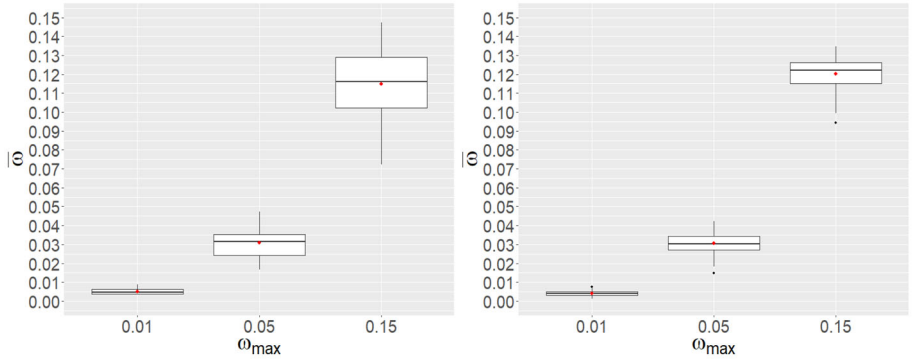
Two scenarios are designed and detailed in Table 1. In both cases, complete data sets are generated first from the pdf in Eq. 1, with the component covariance matrices being extended ultrametric. Specifically, each random sample is obtained by generating $\mathbf{z}_i$, $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g, i = 1, \ldots, n, g = 1, \ldots, G$, as follows. The unit-component membership $\mathbf{z}_i$ results from a multinomial distribution with probabilities corresponding to $\boldsymbol{\pi} = \{\pi_g\}_{g=1}^{G}$ (see Table 1).

Table 1 Design of the simulation study

	Scenario 1	Scenario 2
n	200	400
p	10	30
G	3	5
Q	3	4
$\boldsymbol{\pi}$	(0.25, 0.25, 0.5)	(0.15, 0.15, 0.20, 0.35, 0.15)
${}_{V_g}\sigma_{qq}$	[1.3, 3.3]	[6.7, 10.5]
${}_{W_g}\sigma_{qq}$	[0.7, 2.5]	[2.8, 6.5]
${}_{B_g}\sigma_{qh}$	[−0.9, 1.1]	[−1.5, 4.4]

The component mean vectors $\boldsymbol{\mu}_g$ are generated from a continuous uniform distribution in [0, 10], ensuring that their pairwise Euclidean distance is not lower than 8. Finally, the values of the component covariance matrices $\boldsymbol{\Sigma}_g$ are derived from the four parameters in Eq. 2. Each row of $\mathbf{V}_g$ is engendered from a multinomial distribution with equal probabilities, while the diagonal values of $\boldsymbol{\Sigma}_{\mathrm{V}_g}$ and $\boldsymbol{\Sigma}_{\mathrm{W}_g}$, and the off-diagonal values of $\boldsymbol{\Sigma}_{\mathrm{B}_g}$, take values within the intervals reported in Table 1, plus a uniformly distributed random number in the interval [0, 0.1] and such that conditions (i)–(v) described in Section 2.1 hold.

For each scenario, the complete random samples are generated according to different overlapping levels by considering the measure introduced by Maitra and Melnykov (2010). They defined the overlap between two components, say g and h, i.e. ω_{gh}, as the sum of the two misclassification probabilities $\omega_{g|h}$ and $\omega_{h|g}$, where $\omega_{g|h} = P\big(\pi_h\phi(\mathbf{X}|\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h) < \pi_g\phi(\mathbf{X}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)|\mathbf{X} \sim N_p(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)\big)$. We set three levels of maximum overlapping $\omega_{\max} = 0.01, 0.05, 0.15$ between pairs of components, which correspond to the thresholds defined by Maitra and Melnykov (2010) for well-separated, moderately separated, and poorly separated components, respectively. It can be noticed that, in both scenarios, the average overlapping ($\bar{\omega}$) between all component pairs ranges within the same thresholds by clearly differentiating the three configurations (see Fig. 1). After generating 100 complete random samples for each scenario and overlapping level, as illustrated above, we randomly hide 10%, 20%, and 30% of the values in each complete matrix by altogether obtaining 1800 incomplete data sets (2 scenarios × 3 missing value percentages × 3 maximum overlapping levels × 100 data sets).



(a) Scenario 1 (b) Scenario 2

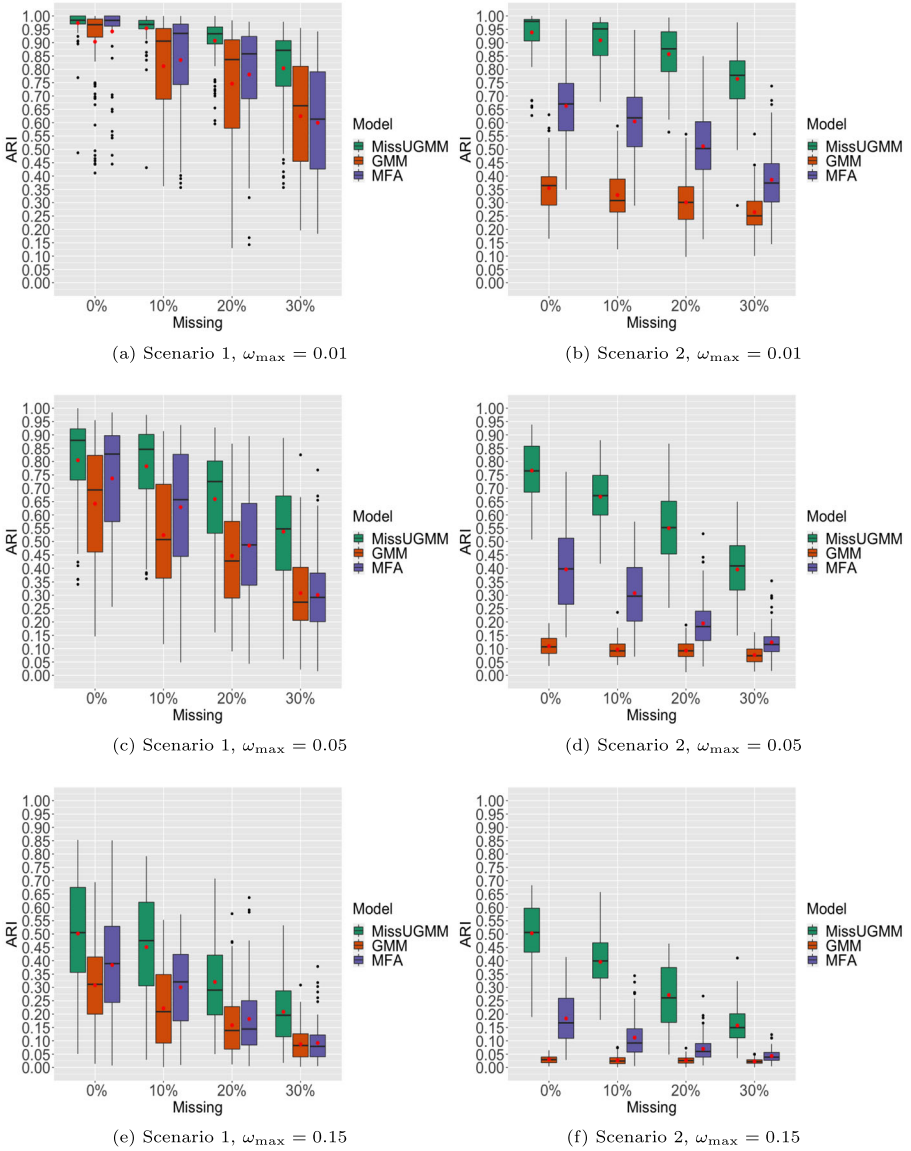
Fig. 1 Boxplot of the distribution of the average overlapping ($\bar{\omega}$) among components for the generated complete data sets per level of maximum overlapping ($\omega_{\max}$) among components

The classification performance of MissUGMM in comparison with GMM and MFA with missing information is evaluated by the adjusted Rand index (ARI, Hubert & Arabie, 1985). The latter measures the similarity between the true, i.e., generated, unit-component membership and the MAP classification estimated by the models, and equals 1 when the perfect agreement is reached. Therefore, the higher the ARI, the better the clustering structure recovery. It has to be noticed that for both MissUGMM and MFA, we set Q to the same value as that used in data generation, as well as G for all the methodologies (see Table 1). We also contrast the proposal with the competitors in terms of missing value imputation by computing the mean absolute error, the mean absolute relative error, and the root mean square error between the incomplete data sets imputed by the models and the corresponding complete data sets, averaged over 100 random samples per percentage of missing values. For assessing the MissUGMM performance in retrieving the model parameters and detecting the hierarchical structure of variables, we evaluate its ability in terms of recovery of the mixing proportions, the component mean vectors, the variable partitions in groups, and the component covariance matrices. To evaluate the reconstruction of π_g and $\boldsymbol{\mu}_g$, we consider the mean square error. To assess the detection of the hierarchy of variables, we compute the mean and standard deviation of ARI between the generated and the estimated variable-group membership matrices at level Q and across the hierarchical levels from 2 to Q for each mixture component. The partition of variables in q groups, $q = Q - 1, \ldots, 2$, is obtained by considering the $Q - 1$ different values in $\boldsymbol{\Sigma}_{\mathrm{B}_g}$ representing the pairwise group aggregations. Regarding the last hierarchical level, i.e., $q = 1$, we do not compute ARI since the resulting partition is always a unitary vector of dimension p, whatever the partition in two groups is. The estimation of the component covariance matrices is assessed via two indices based on those introduced by Di Zio et al. (2007). In detail, given the true component covariance matrix $\boldsymbol{\Sigma}_g$ and the estimated one $\widehat{\boldsymbol{\Sigma}}_g$, we first compute the matrix $\mathbf{D}_g = [{}_g d_{jl} : j, l = 1, \ldots, p]$ with

$$ {}_g d_{jl} = \sqrt{\frac{1}{100} \sum_{s=1}^{100} \frac{({}_g \sigma_{jl}^{(s)} - {}_g \hat{\sigma}_{jl}{}^{(s)})^2}{{}_g \sigma_{jl}{}^{(s)2}}}; $$

then, the variance and covariance preservation indices, say D_{V} and D_{C}, are obtained as the mean of the diagonal and off-diagonal entries of $\mathbf{D}_g$, $g = 1, \ldots, G$, respectively.

Figure 2 displays the results of the simulation study concerning the clustering structure recovery. It is worth noticing that all the outcomes in this section include the performance of the models on complete data sets, i.e., 0% of missing values, as a baseline. Looking at Scenario 1, it is noticeable that the results of the proposal and the competitors are comparable when no missing values occur in the data and the generated components are well-separated. However, the difference among the models becomes increasingly evident, and their performance deteriorates as the level of overlapping and percentage of missing values enlarge. As a whole, MissUGMM outperforms GMM and MFA in terms of ARI, whose variability is often lower than that of the competitors. These results can be foreseen considering the ultrametric nature of the component covariance matrices by stressing the potential of MissUGMM when hierarchical relationships among variables and missing information occur at once in the data. It has to be highlighted that MFA has better performance than GMM, especially with no high percentage of missing values. This can be traced back to the fact that MFA is able to detect a factorial structure—even if not hierarchical (ultrametric)—in a heterogeneous population. We can draw similar conclusions for Scenario 2, where the difference between the proposed methodology and the competitors is more exacerbated because of the high dimensionality of the data that reflects on a higher number of hierarchical relationships among variables to detect. Missing value imputation is evaluated in Table 2, where the three measures of error



(a) Scenario 1, $\omega_{\max} = 0.01$ (b) Scenario 2, $\omega_{\max} = 0.01$

(c) Scenario 1, $\omega_{\max} = 0.05$ (d) Scenario 2, $\omega_{\max} = 0.05$

(e) Scenario 1, $\omega_{\max} = 0.15$ (f) Scenario 2, $\omega_{\max} = 0.15$

Fig. 2 Boxplot of ARI per scenario, model, percentage of observations with missing values, and level of maximum overlapping

previously mentioned are reported. In all cases, GMM has the poorest results for each index, whereas MissUGMM generally reaches the lowest error values.

Besides comparing our proposal with other methodologies in terms of classification performance, Table 3 examines the reconstruction of the model parameters and hierarchical structures on variable groups for Scenario 1. The latter is the distinctive feature of MissUGMM. We highlight that, in this context, the label switching problem has to be solved since the parameters and the hierarchical relationships among variables differ across components. To correctly compute the indices mentioned above, we label the estimated components

Table 2 Evaluation of missing value imputation: mean absolute error (MAE), mean absolute relative error (MARE), and root mean square error (RMSE) for each scenario, model, percentage of missing values, and maximum overlapping level, averaged over 100 random samples

$\omega_{\max}$	% missing	MissUGMM			GMM			MFA		
		MAE	MARE	RMSE	MAE	MARE	RMSE	MAE	MARE	RMSE
Scenario 1										
0.01	10	1.82	1.52	2.31	2.07	1.75	2.65	1.98	1.65	2.54
	20	1.91	1.83	2.45	2.26	2.10	2.93	2.11	1.86	2.73
	30	2.05	3.66	2.65	2.51	4.43	3.27	2.33	5.06	3.03
0.05	10	2.75	3.08	3.51	3.23	4.52	4.15	3.01	3.62	3.85
	20	2.95	2.54	3.78	3.49	3.30	4.50	3.27	2.90	4.21
	30	3.14	4.29	4.03	3.84	6.31	4.94	3.51	5.31	4.52
0.15	10	5.50	2.87	7.05	6.34	3.32	8.12	5.93	3.13	7.60
	20	5.79	3.84	7.38	6.87	5.51	8.82	6.43	3.92	8.24
	30	6.10	3.77	7.81	7.35	4.72	9.43	6.77	4.32	8.65
Scenario 2										
0.01	10	3.51	7.22	4.42	5.61	10.16	7.17	3.87	5.93	4.89
	20	3.60	3.88	4.53	5.90	7.55	7.47	4.08	4.33	5.17
	30	3.74	5.66	4.72	5.92	9.58	7.48	4.35	9.03	5.52
0.05	10	5.75	4.95	7.25	9.01	7.51	11.47	6.43	5.14	8.13
	20	5.92	3.87	7.47	9.44	7.23	11.95	6.77	4.63	8.57
	30	6.21	6.25	7.85	9.37	9.83	11.82	7.12	6.70	9.02
0.15	10	16.34	5.60	20.59	24.92	11.05	31.77	18.31	6.60	23.12
	20	16.99	4.22	21.44	26.22	8.17	33.17	19.15	4.87	24.23
	30	17.58	6.34	22.18	26.03	10.20	32.85	19.98	10.25	25.33

via the complete likelihood-based labelling method introduced by Yao (2015). This approach chooses the permutation of the labels $1, \ldots, G$ maximizing the log-likelihood of the complete data. As the percentage of missing values and the overlapping level increase, the results reported in Table 3 reveal a similar behavior to those for the classification recovery. Specifically, the mean of (hierarchical) ARI is sizable for well-separated components even when a large amount of data entries is missing, and slowly deteriorates when the components overlap more since the underlying ultrametric structures become more challenging to detect. The same trend holds at the indices for measuring the reconstruction of the mixing proportions, component mean vectors, and extended ultrametric covariance matrices. Analogous conclusions can be reached for Scenario 2 by looking at Table 10 in Appendix 1, which is thus omitted herein.

In Appendix 2, we detail the cases where MissUGMM outperforms GMM and/or MFA in terms of BIC on synthetic data sets. Table 11 demonstrates the optimal performance of our proposal, highlighting the importance of employing a specific ultrametric structure to accurately approximate a Gaussian mixture model for data characterized by hierarchical relationships among variables within components. Moreover, by fixing G and Q to the values used in the data generation process, the number of free parameters for estimating the component covariance matrices of MissUGMM (see Section 2.3.2) is consistently much lower than that of both GMM (i.e., $Gp(p+1)/2$), and MFA (i.e., $G(pQ - Q(Q-1)/2 + p)$). For instance, in Scenario 2, where $G = 5$, $Q = 4$ and $p = 30$, the number of free parameters for estimating the covariance matrices is $185 - (c_{V,W} + c_{W,B})$, 2325 and 720 for the three

Table 3 Results of MissUGMM for Scenario 1: mean square error of the mixing proportions and mean vectors (MSE_π and $MSE_{\boldsymbol{\mu}}$), mean and standard deviation of ARI between the generated and the estimated variable partitions at level Q (mARI and sdARI) and across the hierarchical levels from 2 to Q (mhARI and sdhARI), indices of the variance and covariance structure preservation (D_V and D_C) per component, averaged over 100 random samples for each combination of percentage of missing values and maximum overlapping level

$\omega_{\max}$	% Mis.	Comp.	MSE_π	$MSE_{\boldsymbol{\mu}}$	mARI	sdARI	mhARI	sdhARI	D_V	D_C
0.01	0	1	0.00	0.22	0.99	0.05	0.91	0.27	0.15	0.48
		2	0.00	0.30	0.95	0.17	0.97	0.14	0.15	0.42
		3	0.00	0.15	0.99	0.09	0.94	0.23	0.12	0.32
	10	1	0.00	0.23	0.99	0.05	0.90	0.30	0.16	0.48
		2	0.00	0.43	0.96	0.16	0.97	0.14	0.16	0.47
		3	0.00	0.18	0.98	0.11	0.94	0.23	0.12	0.32
	20	1	0.00	0.36	0.98	0.07	0.88	0.31	0.18	0.56
		2	0.00	0.51	0.91	0.21	0.94	0.18	0.17	0.52
		3	0.00	0.17	0.99	0.08	0.95	0.20	0.13	0.34
	30	1	0.00	0.80	0.89	0.28	0.77	0.40	0.19	0.83
		2	0.00	1.51	0.85	0.28	0.85	0.31	0.19	0.74
		3	0.00	0.36	0.97	0.13	0.92	0.24	0.15	0.39
0.05	0	1	0.00	1.20	0.96	0.14	0.83	0.35	0.17	0.68
		2	0.00	1.69	0.88	0.27	0.91	0.26	0.16	0.60
		3	0.01	0.81	0.99	0.08	0.95	0.21	0.14	0.36
	10	1	0.00	1.21	0.93	0.19	0.82	0.36	0.18	0.71
		2	0.00	2.22	0.86	0.28	0.88	0.29	0.18	0.66
		3	0.01	0.71	0.99	0.10	0.95	0.21	0.14	0.35
	20	1	0.01	1.79	0.90	0.23	0.72	0.41	0.20	0.83
		2	0.00	3.34	0.77	0.33	0.77	0.39	0.19	0.84
		3	0.01	1.09	0.95	0.15	0.88	0.31	0.16	0.42
	30	1	0.01	2.21	0.84	0.26	0.71	0.39	0.21	0.92
		2	0.01	3.84	0.63	0.37	0.67	0.41	0.22	0.92
		3	0.01	1.58	0.89	0.21	0.83	0.33	0.18	0.51
0.15	0	1	0.01	6.80	0.92	0.20	0.86	0.31	0.19	0.82
		2	0.00	14.82	0.84	0.26	0.83	0.32	0.22	0.76
		3	0.01	6.11	0.97	0.12	0.90	0.28	0.16	0.46
	10	1	0.01	10.71	0.87	0.26	0.77	0.38	0.20	0.96
		2	0.01	14.61	0.78	0.32	0.81	0.35	0.22	0.79
		3	0.01	6.10	0.92	0.20	0.88	0.28	0.17	0.45
	20	1	0.01	9.90	0.80	0.30	0.70	0.42	0.21	1.02
		2	0.01	21.08	0.67	0.32	0.67	0.38	0.25	0.99
		3	0.01	10.25	0.89	0.22	0.80	0.37	0.20	0.53
	30	1	0.01	13.17	0.64	0.34	0.55	0.42	0.25	1.27
		2	0.01	22.97	0.50	0.32	0.51	0.40	0.28	1.10
		3	0.02	13.30	0.78	0.28	0.68	0.40	0.24	0.65

models, respectively. It is important to note that if $(c_{V,W} + c_{W,B})$ is greater than zero, this number decreases. Conversely, the number of free parameters for the mixing proportions and the component mean vectors remains the same across all models.

3.2 Benchmark Data

In this section, we compare MissUGMM with GMM and MFA on three benchmark data sets to evaluate the proposal's classification performance when no specific assumption on the component covariance matrices exists. The analyses presented herein give an example of the MissUGMM potential even when there may not be necessarily hierarchical relationships among variables. In Section 3.2.1, we generate missing values from two benchmark data sets which are complete. Likewise, in Section 3.1, we first run the models on complete data, where we select Q for our proposal and MFA in the set $\{1, \ldots, 5\}$ according to the *maximum* BIC, while fixing G to the number of clusters known in the literature for these data sets, named G^*. Then, we obtain 100 data sets from the original one by hiding its entries at random under each percentage of missing values, i.e., 10%, 20%, and 30%. On these, we implement our proposal and the competitors with G corresponding to G^* and Q chosen on the complete data set. In Section 3.2.2, we consider a benchmark data set that already contains authentic missing values, not artificially generated.

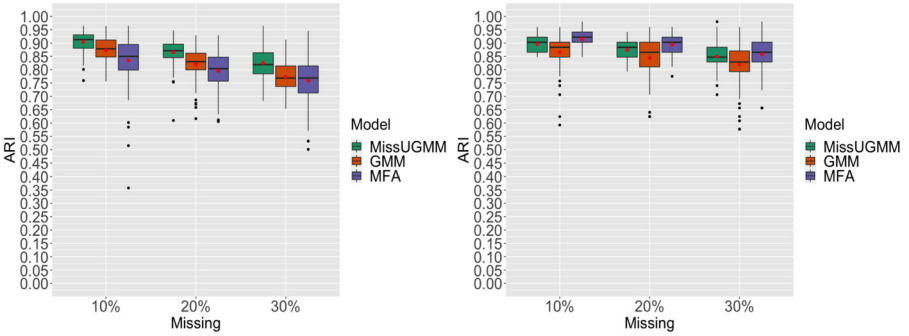
3.2.1 Artificial Missing Values

The first benchmark analyzed herein is wine (available in the R package `gclus` on CRAN), where 13 chemical properties of three types of Italian wines ($G^* = 3$) are measured on 178 observations. The results obtained by implementing the models on the complete data point out that the three methodologies are comparable in terms of ARI, which is equal to 0.95 with 3 misclassifications for MissUGMM and GMM, and 0.93 with 4 misclassifications for MFA. As shown in Table 4, MissUGMM exhibits the highest BIC and the lowest number of free parameters compared to the competitors. It has to be noticed that if we let BIC choose both G and Q—with G in $\{1, \ldots, 5\}$ as well—this correctly identifies G for the two models based on a "factorial" structure, while it selects $G = 2$ for GMM (BIC $= -5643.64$ with 209 free parameters). The distribution of ARI on the incomplete data sets is depicted in Fig. 3a, showing a slightly better performance of the proposal than the competitors as the percentage of missing values rises.

As the second benchmark, we illustrate the results of the three models' implementation on the Kidney data set (available in the R package `teigen` on CRAN, Andrews et al., 2018). This examines the chronic kidney disease (ckd) through 11 characteristics of 203 patients by classifying them into two classes (persons with and without ckd, that is $G^* = 2$). On complete data, ARI of MissUGMM equals 0.90 (5 misclassifications), whereas that of GMM reaches 0.85 (8 misclassifications) and that of MFA attains 0.92 (4 misclassifications), respectively. The comparison of the models in terms of BIC reveals a similar ranking to that of ARI, with MissUGMM consistently having a lower number of parameters. Differently from the wine data set, for this benchmark, BIC fails to correctly choose G for either MissUGMM or MFA. Indeed, the former one splits the original classes into 5 groups, where two count persons with

Table 4 Number of variable groups/factors Q, BIC, and number of free parameters of each model on complete benchmark data sets when $G = G^*$

	Wine data set			Kidney data set		
Model	Q	BIC	# free param.	Q	BIC	# free param.
MissUGMM	5	−5334.89	101	5	−4225.05	63
GMM	–	−5760.13	314	–	−4566.74	155
MFA	1	−5343.65	119	1	−4195.92	67



(a) Wine data set (b) Kidney data set

Fig. 3 Boxplot of ARI per model and percentage of missing values for the benchmark data sets

ckd and three those without ckd (4 misclassifications, BIC $= -3741.89$ and $Q = 3$ with 134 free parameters); instead, MFA only divides the class of persons with ckd into two groups with 1 misclassification (BIC $= -4106.19$ and $Q = 2$ with 131 free parameters). As shown in Fig. 3b, MissUGMM achieves similar results to the competitors, especially to MFA, when we augment the percentage of missing information in the data.

Looking at Table 5, where the evaluation of missing value imputation is reported for both benchmark data sets, we can come to the same conclusions illustrated above. Moreover, in Appendix 2, we report the number of times MissUGMM outperforms GMM and/or MFA in terms of BIC on the incomplete data sets.

3.2.2 Authentic Missing Values

The original Kidney data set (available in the UCI learning repository as "Chronic Kidney Disease") is collected on a bigger sample of 400 patients and contains 14.45% of missing values. The observed features and the number of classes remain the same. In this data set, we let the models choose G, and Q when necessary. MissUGMM selects $G = 4$ and $Q = 4$ with a BIC of -6831.38 and 113 free parameters. In this configuration, individuals with ckd are split into 3 groups, with 48 misclassified (a misclassification rate of 12%). For GMM,

Table 5 Evaluation of missing value imputation: mean absolute error (MAE), mean absolute relative error (MARE), and root mean square error (RMSE) for each benchmark data set, model, and percentage of missing values, averaged over 100 random samples

% missing	MissUGMM			GMM			MFA		
	MAE	MARE	RMSE	MAE	MARE	RMSE	MAE	MARE	RMSE
Wine data set									
10	0.53	2.60	0.72	0.55	3.45	0.75	0.55	3.07	0.75
20	0.54	3.02	0.74	0.62	3.96	0.85	0.56	3.42	0.77
30	0.56	2.88	0.76	0.67	4.18	0.92	0.57	3.44	0.78
Kidney data set									
10	0.51	2.10	0.75	0.55	2.53	0.84	0.52	2.35	0.77
20	0.52	2.13	0.78	0.57	2.71	0.89	0.53	2.27	0.79
30	0.53	2.09	0.79	0.60	2.63	0.95	0.54	2.22	0.80

the optimal G is 3 (BIC $= -7085.91$ with 233 free parameters), resulting in a division of the original class of individuals with ckd into 2 groups, with 66 misclassifications (16.5%). Finally, MFA chooses $G = 4$ and $Q = 2$ (BIC = -6560.82 with 175 free parameters), dividing the ckd group into 3, with 15 misclassifications for the individuals with ckd and 5 for those without ckd, resulting in a misclassification rate of 5%.

When we fix $G = 2$, MissUGMM achieves the best performance in terms of the misclassification rate, misclassifying 88 individuals with ckd (22%), while GMM and MFA have 118 (29.5%) and 101 (25.5%) misclassifications, respectively.

4 Application: Cities' Sustainable Development Analysis

In this section, we apply the proposed methodology to study the cities' sustainable development and the dimensions characterizing it. The data set[1] is composed of 53 cities listed in Table 6 from several countries (Table 7) which distinguish each other for different features. Table 8 illustrates the 12 variables considered for the analysis that concern various present-day aspects (called themes) of the cities' sustainable development, such as economy, education, energy, environment, governance, and urban planning. The latter can be interpreted as domains (latent concepts) defining the multidimensional phenomenon of the cities' sustainable development, making them apt for analysis using an ultrametric model. The data set is not complete, and the missing entries account for 3.52%.

In Table 8, we also report on the variable polarity, which refers to the relationship between each variable and the general concept under study. It is worth highlighting that MissUGMM builds hierarchies by identifying variable groups and their aggregations from the most concordant to the least concordant. For this reason, we reverse the variables with negative polarity ($\{\max x_{ij} : i = 1, \ldots, n\} - x_{ij}$, $j = 1, 2, 6, 8$, in Table 8) so that to apply the proposed methodology on variables that are all positively related to the general concept.

MissUGMM is implemented on the cities' incomplete data set by selecting G and Q—both in $\{1, \ldots, 5\}$—according to BIC. The best model corresponds to that with $G = 3$ and $Q = 4$. Since we use the MAP classification, we refer to groups of cities as clusters. As shown in Table 9, Cluster 1 lumps together all the municipalities of some countries, except for South Africa, and can be considered the cluster of cities with a low level of sustainable development. This cluster encompasses cities that share a common sub-Saharan climate and are located in countries where balanced and equitable growth is not evident. Moreover, it collects almost all the Asian countries in the data set. Cluster 2 and Cluster 3 are primarily characterized by European and North American cities. Notably, they also include the Asian cities of Shanghai (Cluster 2), Tainan city, and Taipei (Cluster 3) which are part of the most developed countries in their geographic area. Cluster 2 contains cities mainly from Eastern and Mediterranean Europe, whereas Cluster 3 those considered the richest countries in the European territory and almost all the Canadian cities. They correspond to the clusters of cities with a medium and high level of sustainable development, respectively. MissUGMM estimates the posterior probabilities for each municipality's cluster membership, revealing some interesting cases. For instance, The Hague in the Netherlands has a 73% probability of belonging to Cluster 2 and a 27% probability of belonging to Cluster 3, while Zagreb in Croatia has a 75% probability of belonging to Cluster 2 and a 25% probability of belonging to Cluster 1.

[1] Data were collected from http://open.dataforcities.org/ on March 15th, 2019, and are available upon request to the corresponding author.

Table 6 List of cities

City	Country	City	Country	City	Country	City	Country
Ahmedabad	IND	Guadalajara	MEX	Oslo	NOR	Tbilisi	GEO
Amman	JOR	Haiphong	VNM	Piedras Negras	MEX	The Hague	NLD
Amsterdam	NLD	Helsinki	FIN	Portland	USA	Toronto	CAN
Barcelona	ESP	Jamshedpur	IND	Porto	PRT	Torreon	MEX
Boston	USA	Johannesburg	ZAF	Pune	IND	Tshwane	ZAF
Brisbane	AUS	Kielce	POL	Quebec City	CAN	Valencia	ESP
Buenos Aires	ARG	León	MEX	Riyadh	SAU	Vaughan	CAN
Cambridge	GBR	London	GBR	San Diego	USA	Vijayawada	IND
Cape Town	ZAF	Los Angeles	USA	Shanghai	CHN	Whitby	CAN
Ciudad Juárez	MEX	Makati	PHL	Sintra	PRT	Zagreb	HVR
Dubai	ARE	Makkah	SAU	Surat	IND	Zwolle	NLD
Eindhoven	NLD	Minna	NGA	Surrey	CAN		
Gdynia	POL	Mississauga	CAN	Tainan city	TWN		
Greater Melbourne	AUS	Oakville	CAN	Taipei	TWN		

The hierarchies of variables identified by MissUGMM vary across clusters, revealing different importance of the dimensions in defining the cities' sustainable development across countries. Figure 4 depicts the hierarchical structure for each cluster, where the aggregation levels are computed as the log-modulus transformation (John & Draper, 1980) of the covariances in $\mathbf{\Sigma}_g$ that preserves their ordering and sign. We use this transformation to appreciate the variable group configuration and their hierarchical relationships, which would be otherwise jammed differently into a small portion of the path diagram due to a singleton with high variance for each cluster. Indeed, the three partitions of variables in Fig. 4 share one group composed of the single variable "Green area (hectares) per 100,000 population" which joints the others last with a negative covariance value. This shows that the urban planning aspect concerning the increase of green areas in the cities is not yet satisfactory both in cities with a low and high level of sustainable development; on the contrary, we can state that the

Table 7 List of countries

Country	ID	N. cities	Country	ID	N. cities
Argentina	ARG	1	Norway	NOR	1
Australia	AUS	2	Philippines	PHL	1
Canada	CAN	7	Poland	POL	2
China	CHN	1	Portugal	PRT	2
Croatia	HVR	1	Saudi Arabia	SAU	2
Finland	FIN	1	South Africa	ZAF	3
Georgia	GEO	1	Spain	ESP	2
Jordan	JOR	1	Taiwan (Republic of China)	TWN	2
India	IND	5	United Arab Emirates	ARE	1
Mexico	MEX	5	United Kingdom	GBR	2
Netherlands	NLD	4	United States of America	USA	4
Nigeria	NGA	1	Vietnam	VNM	1

Table 8 List of variables with information on polarity and % of missing values

ID	Variable name	Polarity	% missing
1	City's unemployment rate	–	9.43
2	Percentage of city population living in poverty	–	9.43
3	Percentage of students completing secondary education: Survival rate	+	9.43
4	Percentage of city population with authorized electrical service	+	11.32
5	Percentage of total energy derived from renewable sources, as a share of the city's total energy consumption	+	18.87
6	Fine Particulate Matter (PM2.5) concentration	–	11.32
7	Women as a percentage of total elected to city-level office	+	7.55
8	Percentage of city population living in slums	–	56.60
9	Average life expectancy	+	1.89
10	Percentage of the city's solid waste that is recycled	+	15.09
11	Green area (hectares) per 100,000 population	+	5.66
12	Percentage of city population with potable water supply service	+	5.66

Table 9 MissUGMM results: clusters of cities and countries' representation per component

	Cities		Cities		Cities
Cluster 1	Ahmedabad	Cluster 2	Amsterdam	Cluster 3	Brisbane
	Amman		Barcelona		Cambridge
	Ciudad Juárez		Boston		Eindhoven
	Dubai		Buenos Aires		Greater Melbourne
	Guadalajara		Cape Town		Helsinki
	Haiphong		Gdynia		London
	Jamshedpur		Kielce		Los Angeles
	Johannesburg		Portland		Mississauga
	León		Porto		Oakville
	Makati		Shanghai		Oslo
	Makkah		Tbilisi		Quebec City
	Minna		The Hague		San Diego
	Piedras Negras		Valencia		Sintra
	Pune		Whitby		Surrey
	Riyadh		Zagreb		Tainan city
	Surat				Taipei
	Torreon				Toronto
	Tshwane				Vaughan
	Vijayawada				Zwolle
Cluster 1	ARE (1/1), JOR (1/1), IND (5/5), MEX (5/5), NGA (1/1), PHL (1/1), SAU (2/2) ZAF (2/3), VNM (1/1)				
Cluster 2	ARG (1/1), CAN (1/7), CHN (1/1), ESP (2/2), GEO (1/1), HVR (1/1), NLD (2/4) POL (2/2), PRT (1/2), USA (2/4), ZAF (1/3)				
Cluster 3	AUS (2/2), CAN (6/7), FIN (1/1), GBR (2/2), NLD (2/4), NOR (1/1), PRT (1/2) TWN (2/2), USA (2/4)				

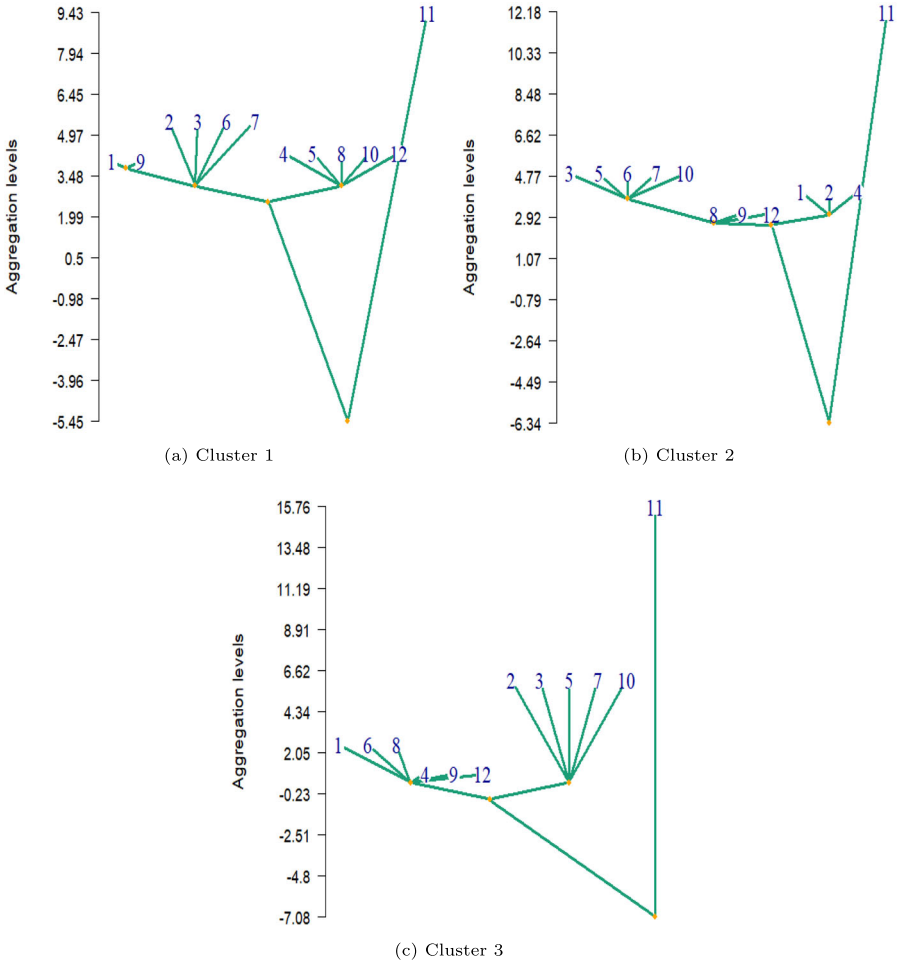


Fig. 4 Hierarchical structures of variables per cluster of cities. The aggregation levels are expressed through the log-modulus transformation of the covariances in $\mathbf{\Sigma}_g$, $g = 1, \ldots, 3$

more the investments for advancement in other aspects toward the city's sustainable development the less those in green areas that remain a target to reach in the future. Concerning the other variable groups, another interesting similarity among the hierarchies is the relation between "Percentage of total energy derived from renewable sources, as a share of the city's total energy consumption" and "Percentage of the city's solid waste that is recycled." These aspects denote the city's governance attention to recycling and renewable energies, which are merged together for all clusters.

For Cluster 1, "City's unemployment rate" (with inverted polarity) and "Average life expectancy" represent the highly internally consistent group since its variance and covariance within it are similar; therefore, in the cities with a low level of sustainable development, the average life expectancy is deeply related to the working conditions. This group is then lumped together with the one measuring poverty (variable 2, with inverted polarity), education (variable 3), environment (variable 6, with inverted polarity), and governance (variable 7),

all aspects representing primary targets for economic development with social responsibility. Their strong relation points out the need for improvement of at least some of these aspects in the cities, and relative countries, with a low level of sustainable development. The second to last aggregation for this cluster merges the latter broader group with the strongly internally consistent one related to energy (variables 4 and 5), living conditions (variable 8), waste management (variable 10), and water supply (variable 12), on which can be done a similar reasoning to the previous group in its improvement. All these variable groups are strongly related to each other; to increase the development of the cities belonging to Cluster 1, it is thus necessary to enhance at least some of these aspects. For Cluster 2, the variable groups, except for the singleton, are highly internally consistent. Specifically, in the cities with a medium level of sustainable development, the variables related to topics of great interest for public debate, such as education, renewable sources, environment, and gender gap in governance, are aggregated in the same group, as well as those associated with living conditions, such as variables 8, 9, and 12. The last element of the partition is composed of "City's unemployment rate," "Percentage of city population living in poverty," and "Percentage of city population with authorized electrical service." The polarity of the former two variables is reversed, and their aggregation points out that, in these cities, working does not necessarily guarantee a decent living. The three groups described for Cluster 2 are lumped together at approximately the same level, before the aggregation with the singleton previously described. Differently, Cluster 3 identifies variable groups, except for the singleton, with levels of covariance within them that approximate zero, as well as their aggregation levels. In the cities with a high level of sustainable development, we can state that all the aspects illustrated in the data set, with the exception of variable 11, represent a not interchangeable part of the general concept and all uniquely contribute to its definition.

5 Conclusions

When studying complex phenomena in heterogeneous populations, it is crucial to consider the hierarchical relationships among the dimensions defining them. An ultrametric Gaussian mixture model is a potential tool for this purpose in the complete data settings. However, incomplete data sets are often encountered in applications. In this paper, we propose MissUGMM, an ultrametric Gaussian mixture model in the incomplete data framework, that enhances its applicability in real-world scenarios by effectively handling the presence of missing at random information. Our proposal is estimated via an expectation-maximization algorithm which is a proper tool to deal with different sources of missingness in the data. We compare MissUGMM with Gaussian mixture models and mixtures of factor analyzers via a simulation study. Our approach performs more effectively than the competitors when analyzing synthetic data with missing entries and increasing levels of component overlapping that affect the identification of variable hierarchies. This highlights the advantage given by an appropriate methodology to detect hierarchical structures of variables in heterogeneous populations. We further demonstrate the potential of MissUGMM on benchmark data sets, where the component covariance structure may not be ultrametric.

Afterward, we apply MissUGMM to a real-world problem to study the sustainable development of cities across multiple countries. The municipalities provide information on aspects related to their energy equipment, population living conditions, citizens' education and health, and environmental sustainability. However, data collection and harmonization worldwide have some intricacies, resulting in a 13.52 missing value percentage in the data set. Using MissUGMM, we identify three clusters representing cities with a low, medium, and high

rate of sustainable growth. These clusters are characterized by diverse dimensions that have different roles and importance in defining the cities' sustainable development. For example, the analysis of this data set makes evident that certain factors need improvement in cities with slow progress towards sustainable growth, such as the unemployment rate, that could correspond to an enhancement of other aspects, such as citizens' average life expectancy and population living conditions. Crosswise, the study reveals that the urban planning of green areas is negatively related to all other determinants and has not been adequately addressed in any cities, regardless of their sustainability level. Overall, the application of MissUGMM provides interesting insights on this phenomenon and its configuration across cities.

Despite the advantages of MissUGMM illustrated throughout the paper, further developments can be made in the field of ultrametric modeling. Firstly, following Boldea and Magnus (2009), Montanari and Viroli (2011), and Wang and Lin (2016), the theoretical and inferential properties of the extended ultrametric covariance matrix estimators can be investigated, allowing the derivation of standard errors for the parameter estimates, both in the complete and subsequently incomplete data framework. Furthermore, real data can be influenced by outliers, e.g., in our application, they could be municipalities with extreme entries in some variables compared to the majority of the cities. While the detection of outliers is out of the scope of this paper, we recognize it as a potential future development for our proposal. In the literature, Tong and Tortora (2022) developed a framework to model mild outliers in the presence of missing information via the mixture of multivariate contaminated normal distributions, whereas Wang (2015) and Wang and Lin (2022b) proposed the mixtures of (common) t factor analyzers, where the factor-analytic representation of the component covariance matrices preserves the inspection of latent structures underlying the data. To deal with gross outliers, trimming is a powerful tool in different frameworks. It consists of removing a small proportion of the observations whose values are the most unlikely to occur when the fitted model is true. In this direction, García-Escudero et al. (2016) introduced the robust mixtures of factor analyzers when complete data are available. Furthermore, when data are likely to be censored, other than contaminated, mixtures of factor and t factor analyzers can be extended to accommodate for them, as proposed by Wang et al. (2019) and Wang and Lin (2022a). In any case, none of these methodologies can inspect hierarchical latent relationships among dimensions defining a multidimensional concept in the presence of outliers and/or missing data. We aim to address this gap in future studies.

Finally, all the methodologies discussed in this paper have been developed under the MAR assumption. Only recently, a pioneering approach to handle MNAR data in GMMs was proposed by Sportisse et al. (2024), paving the way for future research in this direction. Particularly, the authors focused on a specific MNAR model by considering its MAR counterpart. However, verifying the MNAR mechanism from the data remains challenging (Molenberghs et al., 2008), and further investigations would be necessary in the ultrametric context due to the potentially different contribution of the missing variables to the hierarchical structure.

Appendix 1

Further results on the simulation study depicted in Section 3.1 are provided in this appendix. Specifically, Table 10 shows the results of MissUGMM in estimating the model parameters and the hierarchical relationships among variables for Scenario 2.

Table 10 Results of MissUGMM for Scenario 2: mean square error of the mixing proportions and mean vectors (MSE_{π} and $MSE_{\boldsymbol{\mu}}$), mean and standard deviation of ARI between the generated and the estimated variable partitions at level Q (mARI and sdARI) and across the hierarchical levels from 2 to Q (mhARI and sdhARI), indices of the variance and covariance structure preservation (D_V and D_C) per component, averaged over 100 random samples for each combination of percentage of missing values and maximum overlapping level

ω_{max}	% Mis.	Comp.	MSE_{π}	$MSE_{\boldsymbol{\mu}}$	mARI	sdARI	mhARI	sdhARI	D_V	D_C
0.01	0	1	0.00	0.77	0.96	0.11	0.87	0.31	0.11	0.48
		2	0.00	1.65	0.92	0.22	0.87	0.28	0.11	0.78
		3	0.00	0.65	0.95	0.17	0.91	0.23	0.09	0.61
		4	0.00	0.42	1.00	0.04	0.97	0.13	0.08	0.21
		5	0.00	1.45	0.91	0.20	0.83	0.28	0.13	0.47
	10	1	0.00	1.59	0.89	0.19	0.83	0.33	0.12	0.61
		2	0.00	1.74	0.87	0.25	0.82	0.31	0.13	0.82
		3	0.00	0.82	0.92	0.17	0.89	0.24	0.09	0.72
		4	0.00	0.35	1.00	0.01	0.97	0.13	0.08	0.21
		5	0.00	2.31	0.84	0.28	0.78	0.32	0.13	0.62
	20	1	0.00	1.62	0.85	0.22	0.79	0.35	0.12	0.70
		2	0.00	2.94	0.78	0.31	0.72	0.36	0.13	0.98
		3	0.00	0.88	0.88	0.20	0.84	0.27	0.10	0.78
		4	0.00	0.54	0.99	0.06	0.96	0.16	0.09	0.24
		5	0.00	2.65	0.78	0.28	0.74	0.33	0.14	0.63
	30	1	0.00	2.66	0.78	0.25	0.73	0.38	0.14	0.77
		2	0.00	4.89	0.61	0.35	0.58	0.39	0.16	1.26
		3	0.00	1.83	0.77	0.30	0.77	0.34	0.12	1.02
		4	0.00	0.74	0.98	0.08	0.93	0.20	0.10	0.28
		5	0.00	5.42	0.60	0.35	0.58	0.38	0.17	0.88
0.05	0	1	0.00	2.96	0.90	0.19	0.83	0.33	0.12	0.62
		2	0.00	6.76	0.82	0.30	0.77	0.35	0.13	0.93
		3	0.00	2.46	0.88	0.21	0.86	0.26	0.11	0.72
		4	0.00	0.98	0.99	0.05	0.97	0.13	0.09	0.21
		5	0.00	5.17	0.87	0.23	0.82	0.29	0.14	0.54
	10	1	0.00	5.18	0.78	0.26	0.77	0.35	0.13	0.77
		2	0.00	7.77	0.71	0.35	0.65	0.39	0.15	1.08
		3	0.00	4.42	0.79	0.30	0.76	0.35	0.12	1.00
		4	0.00	1.76	0.98	0.09	0.96	0.15	0.09	0.24
		5	0.00	7.10	0.77	0.30	0.75	0.33	0.16	0.71
	20	1	0.00	8.74	0.70	0.29	0.70	0.37	0.15	0.92
		2	0.00	13.51	0.48	0.35	0.44	0.40	0.16	1.32
		3	0.00	6.84	0.61	0.35	0.60	0.39	0.14	1.37
		4	0.00	3.17	0.95	0.13	0.92	0.22	0.11	0.31
		5	0.00	9.89	0.65	0.36	0.65	0.38	0.15	0.77
	30	1	0.00	11.95	0.53	0.32	0.51	0.39	0.19	1.10
		2	0.00	16.04	0.34	0.30	0.32	0.35	0.18	1.42
		3	0.00	11.59	0.40	0.29	0.36	0.35	0.18	1.52
		4	0.01	5.26	0.88	0.19	0.83	0.29	0.14	0.41
		5	0.00	11.54	0.49	0.32	0.52	0.37	0.17	0.86
0.15	0	1	0.00	42.54	0.83	0.24	0.77	0.36	0.15	0.74
		2	0.00	66.19	0.71	0.33	0.68	0.37	0.16	1.11

Table 10 continued

$\omega_{\max}$	% Mis.	Comp.	MSE_{π}	$\text{MSE}_{\boldsymbol{\mu}}$	mARI	sdARI	mhARI	sdhARI	D_{V}	D_{C}
		3	0.00	40.88	0.72	0.32	0.72	0.37	0.13	1.19
		4	0.00	13.77	0.96	0.12	0.93	0.21	0.09	0.26
		5	0.00	55.17	0.79	0.29	0.75	0.34	0.14	0.62
	10	1	0.00	63.05	0.76	0.27	0.73	0.37	0.17	0.83
		2	0.00	83.03	0.53	0.36	0.51	0.41	0.18	1.22
		3	0.00	73.81	0.58	0.35	0.55	0.40	0.15	1.49
		4	0.00	28.83	0.97	0.09	0.92	0.21	0.12	0.34
		5	0.00	67.53	0.68	0.35	0.67	0.37	0.17	0.73
	20	1	0.00	83.08	0.60	0.30	0.58	0.39	0.20	0.98
		2	0.00	118.61	0.36	0.29	0.33	0.34	0.19	1.39
		3	0.00	75.62	0.42	0.33	0.43	0.37	0.17	1.67
		4	0.01	55.73	0.90	0.17	0.85	0.28	0.16	0.42
		5	0.00	82.50	0.51	0.31	0.53	0.38	0.19	0.80
	30	1	0.00	111.94	0.47	0.27	0.46	0.37	0.22	1.06
		2	0.00	114.85	0.27	0.27	0.27	0.32	0.21	1.41
		3	0.00	104.81	0.27	0.23	0.25	0.31	0.20	1.78
		4	0.01	96.13	0.75	0.26	0.68	0.36	0.19	0.50
		5	0.00	102.28	0.38	0.29	0.41	0.35	0.20	0.92

Appendix 2

We report herein the occurrences where BIC of MissUGMM exceeds BIC of both GMM and MFA (Case 1), BIC of GMM only (Case 2), and BIC of MFA only (Case 3) for both the synthetic data sets described in Section 3.1 (Table 11) and benchmark data sets illustrated in Section 3.2.1 (Table 12).

Table 11 Number of times where BIC of MissUGMM exceeds BIC of both GMM and MFA (Case 1), BIC of GMM only (Case 2), and BIC of MFA only (Case 3) across 100 random samples per scenario, percentage of missing values and maximum overlapping level

$\omega_{\max}$	Scenario 1 Case 1				Case 2				Case 3			
	0%	10%	20%	30%	0%	10%	20%	30%	0%	10%	20%	30%
0.01	99	100	100	100	1	0	0	0	0	0	0	0
0.05	93	99	99	99	7	1	1	0	0	0	0	1
0.15	99	100	100	100	1	0	0	0	0	0	0	0
	Scenario 2											
$\omega_{\max}$	Case 1				Case 2				Case 3			
	0%	10%	20%	30%	0%	10%	20%	30%	0%	10%	20%	30%
0.01	100	100	100	100	0	0	0	0	0	0	0	0
0.05	100	100	100	100	0	0	0	0	0	0	0	0
0.15	100	100	100	100	0	0	0	0	0	0	0	0

Table 12 Number of times where BIC of MissUGMM exceeds BIC of both GMM and MFA (Case 1), BIC of GMM only (Case 2), and BIC of MFA only (Case 3) across 100 random samples per percentage of missing values for each benchmark data set

	Case 1			Case 2			Case 3		
% missing	10	20	30	10	20	30	10	20	30
Wine data set	77	74	82	23	26	18	0	0	0
Kidney data set	13	21	32	87	79	68	0	0	0

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00357-024-09492-0.

Acknowledgements The authors would like to express their gratitude to Prof. Tsung-I Lin for providing them with the R codes to implement GMM and MFA in the presence of missing data, that they used for comparison with the proposal in Section 3.

Funding Open access funding provided by Università degli Studi di Milano - Bicocca within the CRUI-CARE Agreement. The authors' research has been supported by Milano-Bicocca University Fund for Scientific Research, 2021-ATE-0707.

Data Availability The benchmark data sets used in Section 3.2 are all available in R packages and the UCI repository, as detailed in the paper. The data that support the findings of the study reported in Section 4 are available from the corresponding author upon request.

Code Availability The source code of MissUGMM for data analyses is openly available at https://github.com/giorgiazaccaria/MissUGMM.

Declarations

Conflict of Interest The authors declare no competing interests.

References

Andrews, J., Wickins, J., Boers, N., & McNicholas, P. (2018). teigen: An R package for model-based clustering and classification via the multivariate *t* distribution. *Journal of Statistical Software, 83*(7), 1–32.

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

Banfield, J., & Raftery, A. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics, 49*(3), 803–821.

Bezdek, J. (1974). Cluster validity with fuzzy set. *Journal of Cybernetics, 3*(3), 58–73.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., & Lindsay, B. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics, 46*(2), 373–388.

Boldea, O., & Magnus, J. (2009). Maximum likelihood estimation of the multivariate normal mixture model. *Journal of the American Statistical Association, 104*(488), 1539–1549.
Bouveyron, C., Celeux, G., Murphy, T., & Raftery, A. (2019). *Model-based clustering and classification for data science: With applications in R*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge.
Cailliez, F. (1983). The analytical solution of the additive constant problem. *Psychometrika, 48*(2), 305–308.
Cavicchia, C., Vichi, M., & Zaccaria, G. (2020). The ultrametric correlation matrix for modelling hierarchical latent concepts. *Advances in Data Analysis and Classification, 14*(4), 837–853.
Cavicchia, C., Vichi, M., & Zaccaria, G. (2022). Gaussian mixture model with an extended ultrametric covariance structure. *Advances in Data Analysis and Classification, 16*(2), 399–427.
Celeux, G., Frühwirth-Schnatter, S., & Robert, C. (2018). Model selection for mixture models - Perspectives and strategies. In: S. Fruhwirth-Schnatter, & C. R. G Celeux (Eds.), *Handbook of mixture analysis* (chap 7, pp. 117–154). Chapman and Hall/CRC.
Celeux, G., & Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition, 28*(5), 781–793.
Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27.
Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 39*(1), 1–38.
Di Zio, M., Guarnera, U., & Luzi, O. (2007). Imputation through finite Gaussian mixture models. *Computational Statistics & Data Analysis, 51*(11), 5305–5316.
Fix, E., & Hodges, J. (1951). Discriminatory analysis. nonparametric discrimination: Consistency properties. Tech. rep., USAF School of Aviation Medicine, Randolph Field, Texas.
Fraley, C., & Raftery, A. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis, and density estimation. *Computer Journal, 41*(8), 578–588.
Fraley, C., & Raftery, A. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association, 97*(458), 611–631.
García-Escudero, L., Gordaliza, A., Greselin, F., Ingrassia, S., & Mayo-Iscar, A. (2016). The joint role of trimming and constraints in robust estimation for mixtures of Gaussian factor analyzers. *Computational Statistics & Data Analysis, 99*, 131–147.
Ghahramani, Z., & Hinton, G. (1997). The EM algorithm for factor analyzers. Tech. Rep. CRG-TR-96-1, University of Toronto, Toronto.
Ghahramani, Z., & Jordan, M. (1995). Learning from incomplete data. Tech. Rep. AI Lab Memo No. 1509, CBCL Paper No. 108, MIT AI Lab.
Gilbert, G. (1991). Positive definite matrices and Sylvester's criterion. *American Mathematical Monthly, 98*(1), 44–46.
Horn, R., & Johnson, C. (2013). *Matrix analysis* (2nd ed.). Cambridge: Cambridge University Press.
Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193–218.
John, J., & Draper, N. (1980). An alternative family of transformations. *Journal of the Royal Statistical Society Series C: Applied Statistics, 29*(2), 190–197.
Lindstrom, M., & Bates, D. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association, 83*(404), 1014–1022.
Lin, T., Lee, J., & Ho, H. (2006). On fast supervised learning for normal mixture models with missing information. *Pattern Recognition, 39*(6), 1177–1187.
Little, R., & Rubin, D. (2019). *Statistical analysis with missing data* (3rd ed.). Hoboken: John Wiley & Sons.
MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281–297). University of California Press, Berkeley.
Maitra, R., & Melnykov, V. (2010). Simulating data to study performance of finite mixture modeling and clustering algorithms. *Journal of Computational and Graphical Statistics, 19*(2), 354–376.
Mardia, K., Kent, J., & Bibby, J. (1979). *Multivariate analysis* (1st ed.). San Diego: Academic Press.
McLachlan, G., & Krishnan, T. (2008). *The EM algorithm and extensions* (2nd ed.). Hoboken: Wiley.
McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.
McLachlan, G., Peel, D., & Bean, R. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis, 41*(3), 379–388.
McLachlan, G., & Rathnayake, S. (2014). On the number of components in a Gaussian mixture model. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 4*(5), 341–355.
McNicholas, P. (2016). Model-based clustering. *Journal of Classification, 33*(3), 331–373.

McNicholas, P., Murphy, T., McDaid, A., & Frost, D. (2010). Serial and parallel implementations of model-based clustering via parsimonious gaussian mixture models. *Computational Statistics & Data Analysis, 54*(3), 711–723.
Molenberghs, G., Beunckens, C., Sotto, C., & Kenward, M. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 70*(2), 371–388.
Montanari, A., & Viroli, C. (2011). Maximum likelihood estimation of mixtures of factor analyzers. *Computational Statistics & Data Analysis, 55*(9), 2712–2723.
Redner, R., & Walker, H. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review, 26*(2), 195–239.
Rubin, D. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592.
Schafer, J. (1997). *Analysis of incomplete multivariate data* (3rd ed.). New York: Chapman and Hall/CRC.
Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464.
Serafini, A., Murphy, T., & Scrucca, L. (2020). Handling missing data in model-based clustering. https://arxiv.org/abs/2006.02954, arXiv:2006.02954
Sportisse, A., Marbac, M., Laporte, F., Celeux, G., Boyer, C., Josse, J., & Biernacki, C. (2024). Model-based clustering with missing not at random data. *Statistical Computing, 34*(135).
Tong, H., & Tortora, C. (2022). Model-based clustering and outlier detection with missing data. *Advances in Data Analysis and Classification, 16*(1), 5–30.
Wang, W. (2013). Mixtures of common factor analyzers for high-dimensional data with missing information. *Journal of Multivariate Analysis, 117*, 120–133.
Wang, W. (2015). Mixtures of common t-factor analyzers for modeling high-dimensional data with missing values. *Computational Statistics & Data Analysis, 83*, 223–235.
Wang, W., Castro, L., Lachos, V., & Lin, T. (2019). Model-based clustering of censored data via mixtures of factor analyzers. *Computational Statistics & Data Analysis, 140*, 104–121.
Wang, W., & Lin, T. (2016). Maximum likelihood inference for the multivariate t mixture model. *Journal of Multivariate Analysis, 149*, 54–64.
Wang, W., & Lin, T. (2020). Automated learning of mixtures of factor analysis models with missing information. *TEST, 29*(4), 1098–1124.
Wang, W., & Lin, T. (2022a). Robust clustering of multiply censored data via mixtures of t factor analyzers. *TEST, 31*, 22–53.
Wang, W., & Lin, T. (2022b). Robust clustering via mixtures of t factor analyzers with incomplete data. *Advances in Data Analysis and Classification, 16*(3), 659–690.
Yao, W. (2015). Label switching and its solutions for frequentist mixture models. *Journal of Statistical Computation and Simulation, 85*(5), 1000–1012.