Statistics and Computing (2024) 34:108
https://doi.org/10.1007/s11222-024-10405-9

ORIGINAL PAPER

Parsimonious ultrametric Gaussian mixture models

Carlo Cavicchia[1] · **Maurizio Vichi**[2] · **Giorgia Zaccaria**[3]

Received: 4 August 2023 / Accepted: 9 February 2024

Abstract
Gaussian mixture models represent a conceptually and mathematically elegant class of models for casting the density of a heterogeneous population where the observed data is collected from a population composed of a finite set of G homogeneous subpopulations with a Gaussian distribution. A limitation of these models is that they suffer from the curse of dimensionality, and the number of parameters becomes easily extremely large in the presence of high-dimensional data. In this paper, we propose a class of parsimonious Gaussian mixture models with constrained extended ultrametric covariance structures that are capable of exploring hierarchical relations among variables. The proposal shows to require a reduced number of parameters to be fit and includes constrained covariance structures across and within components that further reduce the number of parameters of the model.

Keywords Model-based clustering · Hierarchical models · Ultrametricity · Parsimony · Dimensionality reduction

1 Introduction

Finite mixture models lay their foundations in the assumption that data is collected from a finite set of G populations and that data within each population is shaped by a statistical model. Finite Gaussian Mixture Models (GMMs) provide a widely used probabilistic approach to group continuous multivariate data, and assume a Gaussian structure for each population. When considering a p-dimensional random vector $\boldsymbol{x}$ with a GMM distribution, the pdf model is, therefore, of the form

$$f(\boldsymbol{x} \mid \boldsymbol{\Theta}) = \sum_{g=1}^{G} \pi_g \, \phi(\boldsymbol{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{1}$$

where $\phi(\cdot \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ denotes the density of a multivariate Gaussian distribution with a p-dimensional mean vector $\boldsymbol{\mu}_g$ and a covariance matrix $\boldsymbol{\Sigma}_g$ of order p. The mixing proportions (prior probabilities) $\pi_1, \ldots, \pi_G$ are such that $\pi_g > 0$ and $\sum_{g=1}^{G} \pi_g = 1$, and $\boldsymbol{\Theta} = \{\pi_1, \ldots, \pi_G, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G\}$ is the overall parameter set. Relevant specialized literature and complete reviews of GMMs are found in, among others, Titterington et al. (1985), McLachlan and Basford (1988), McLachlan and Peel (2000a), Fraley and Raftery (2002), Bouveyron et al. (2019).

Although GMMs offer the benefit of quantifying uncertainty through probabilities, their practical usability is jeopardized in high-dimensional spaces because of the estimation of a large number of parameters. As a consequence, several solutions have been proposed, often relying on matrix decomposition or variable selection strategies. In detail, Banfield and Raftery (1993), Bensmail and Celeux (1996) extensively worked on the eigen-decomposition of the covariance matrix of the form $\boldsymbol{\Sigma}_g = \lambda_g \boldsymbol{D}_g \boldsymbol{A}_g \boldsymbol{D}_g'$, where $\lambda_g = |\boldsymbol{\Sigma}_g|^{1/p}$, $\boldsymbol{A}_g$ is a diagonal matrix such that $|\boldsymbol{A}_g| = 1$, and $\boldsymbol{D}_g$ is an orthogonal matrix of eigenvectors. These elements specify the scale (or volume), shape, and orientation of the Gaussian components, respectively. The decomposition provides different GMMs by using from one to $Gp(p+1)/2$ parameters, while imposing different geometric characteristics to the component covariance structure and/or by constraining the covariance matrices to be equal or unequal across components (Gaus-

✉ Giorgia Zaccaria
giorgia.zaccaria@unimib.it

Carlo Cavicchia
cavicchia@ese.eur.nl

Maurizio Vichi
maurizio.vichi@uniroma1.it

[1] Econometric Institute, Erasmus University Rotterdam, Burgemeester Oudlaan 50, 3062 PA Rotterdam, Netherlands

[2] Department of Statistical Sciences, Sapienza University of Rome, Piazzale Aldo Moro 5, 20100 Rome, Italy

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

Published online: 01 April 2024

sian Parsimonious Clustering Models, GPCMs, Celeux and Govaert 1995; Fraley and Raftery 1998, 2002). In the high-dimensional context, the High-Dimensional Data Clustering model (Bouveyron et al. 2007, HDDC) provides an alternative eigen-decomposition of the covariance matrix of the form $\boldsymbol{\Sigma}_g = \boldsymbol{Q}_g \boldsymbol{\Lambda}_g \boldsymbol{Q}_g'$, where $\boldsymbol{Q}_g$ is an orthogonal matrix of eigenvectors, $\boldsymbol{\Lambda}_g = \text{diag}([a_{g1}, \ldots, a_{gd_g}, b_g, \ldots, b_g]')$ is a diagonal matrix of eigenvalues and $d_g \in \{1, \ldots, p-1\}$ is the intrinsic dimension of the gth component. HDDC also defines a family of models by constraining covariance structures between and/or within components whose number of parameters varies from $d(p-(d+1)/2)+3$, where d is the intrinsic dimension shared by the G components, to $\sum_{g=1}^{G} d_g(p-(d_g+1)/2)+2G+\sum_{g=1}^{G} d_g$. McNicholas and Murphy (2008, 2010) developed a class of GMMs, called Parsimonious Gaussian Mixture Models (PGMMs), that assume a latent structure per component. This approach extends both the mixtures of factor analyzers (Ghahramani and Hinton 1997; McLachlan and Peel 2000b; McLachlan et al. 2003) and the mixtures of probabilistic principal component analyzers (Tipping and Bishop 1999a, b), and assumes a component covariance structure of the form $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$, where $\boldsymbol{\Lambda}_g$ is a $(p \times m)$, with $m \ll p$, factor loading matrix and $\boldsymbol{\Psi}_g$ is a p-dimensional diagonal covariance matrix of error. These models can have from $pm - m(m-1)/2 + 1$ to $G(pm - m(m-1)/2 + p)$ parameters with $m \in \{1, 2, \ldots, p\}$, by considering equal or unequal covariance structures among the components. Finally, Cavicchia et al. (2022) proposed a parameterization of the covariance matrix that aims to model multidimensional phenomena—usually defined by hierarchically nested latent concepts—by assuming an *extended ultrametric covariance matrix* for each component. This peculiar structure is associated with a hierarchy of concepts and can explore hierarchical relationships among variables. The model introduced by Cavicchia et al. (2022) is based upon the general parameterization of an extended ultrametric covariance matrix that requires a reduced number of parameters to fit; however, this number can be further reduced.

In this paper, we therefore propose a class of thirteen parsimonious GMMs with *constrained* extended ultrametric covariance structures between and within components that further reduce the number of the ultrametric model parameters. The models belonging to this family, named Parsimonious Ultrametric Gaussian Mixture Models (PUGMMs), can thus have from $p+3$ to $G(p+3m-1)$ parameters for the covariance structure, with $m \in \{1, 2, \ldots, p\}$. Furthermore, we introduce computational improvements on the estimate of the extended ultrametric covariance matrix. First, we facilitate its implementation in GMMs by adapting the results of Archakov and Hansen (2020) that make the computation of its determinant and inverse remarkably faster and easier. Second, we consider the polar decomposition on the extended ultrametric covariance matrix to ensure its positive definiteness, whenever necessary. The clustering performance of the thirteen PUGMMs is compared to that of the existing methodologies mentioned above on thirteen benchmark data sets. Additionally, two real-world applications, respectively on FIFA football player skills and the use of web open collaborative environments for teaching within universities, are presented by inspecting the hierarchical relationships of variables featuring them.

The paper is organized as follows. Section 2 presents the full development of the extended ultrametric covariance structure and its application as parameterization for GMMs. In Sect. 3, the collection of parsimonious ultrametric GMMs is given. Pivotal computational aspects are presented in Sect. 4. Section 5 features real data examples where the proposed models are numerically illustrated. A final discussion completes the paper in Sect. 6.

2 GMM with an extended ultrametric covariance structure

In this section, we briefly recall the parameterization introduced by Cavicchia et al. (2022) to model the component covariance matrices of a GMM. Let us consider a random vector $\boldsymbol{x}$ drawn from a finite set of G populations as in (1), where the covariance matrix $\boldsymbol{\Sigma}_g$ is parameterized through an Extended Ultrametric Covariance Structure (EUCovS) defined as follows

$$\boldsymbol{\Sigma}_g = \boldsymbol{V}_g(\boldsymbol{\Sigma}_{W_g} + \boldsymbol{\Sigma}_{B_g})\boldsymbol{V}_g' + \text{diag}(\boldsymbol{V}_g(\boldsymbol{\Sigma}_{V_g} - \boldsymbol{\Sigma}_{W_g})\boldsymbol{V}_g'). \quad (2)$$

Equation (2) depends on four parameters: the binary and row-stochastic $(p \times m)$ variable-group membership matrix defining the partition of variables into $m \leq p$ groups $(\boldsymbol{V}_g = [v_{jq} : j = 1, \ldots, p, q = 1, \ldots, m])$, the diagonal $(m \times m)$ group variance matrix $(\boldsymbol{\Sigma}_{V_g} = [{}_V\sigma_{qq(g)} : q = 1, \ldots, m])$, the diagonal $(m \times m)$ within-group covariance matrix $(\boldsymbol{\Sigma}_{W_g} = [{}_W\sigma_{qq(g)} : q = 1, \ldots, m])$ and the symmetric $(m \times m)$ between-group covariance matrix $(\boldsymbol{\Sigma}_{B_g} = [{}_B\sigma_{qh(g)} : q, h = 1, \ldots, m])$ for each component $g = 1, \ldots, G$. To guarantee the ultrametricity of $\boldsymbol{\Sigma}_g$ in (2), its parameters must comply with some constraints: (i) $\boldsymbol{\Sigma}_{B_g}$ is a hollow matrix[1] whose off-diagonal triplets respect the ultrametric inequality (i.e., ${}_B\sigma_{qh(g)} \geq \min\{{}_B\sigma_{qs(g)}, {}_B\sigma_{hs(g)}\}, q, h, s = 1, \ldots, m, s \neq h \neq q$), (ii) the lowest diagonal value of $\boldsymbol{\Sigma}_{W_g}$ is always greater than (or equal to) the highest off-diagonal value of $\boldsymbol{\Sigma}_{B_g}$ (that is, $\min\{{}_W\sigma_{qq(g)}, q = 1, \ldots, m\} \geq \max\{{}_B\sigma_{qh(g)}, q, h = 1, \ldots, m, h \neq q\}$), (iii) each group variance is greater than the absolute value of the corresponding within-group

[1] The hollow matrix is a matrix with diagonal entries equal to zero.

covariance (i.e., ${}_V\sigma_{qq(g)} > |{}_W\sigma_{qq(g)}|, q = 1, \ldots, m$). By complying with these constraints, EUCovS results to be a Weak Extended Ultrametric Matrix (Cavicchia et al. 2022, Definition 1) since it is symmetric, nonnegative on the diagonal, ultrametric and column pointwise diagonally dominant. Furthermore being a covariance matrix $\boldsymbol{\Sigma}_g$ must be positive semidefinite. As demonstrated by Dellacherie et al. (2014), an ultrametric matrix is guaranteed to be positive semidefinite if all its entries are nonnegative. The column pointwise diagonal dominance condition required by the definition of an ultrametric matrix results pivotal for obtaining the positive semi-definiteness. As discussed by Cavicchia et al. (2022), when the entries are not nonnegative, a stronger condition (i.e., diagonal dominance) is needed to guarantee this property. Since the diagonal dominance condition is very strong and may lead to an overestimation of the parameter $\boldsymbol{\Sigma}_{V_g}$, in this paper, the positive semi-definiteness of $\boldsymbol{\Sigma}_g$ is ensured by a procedure that considers its polar decomposition (Higham 1986) as discussed in detail in Sect. 4.3.

The goal of the EUCovS parameterization is twofold. This enables, on the one hand, to reduce the dimensionality of the data by merging the p variables into a reduced number m of groups, and on the other hand, to identify the hierarchical structure over them. Each variable group is then characterized by three different features: the variance of the group, the covariance within the group and the covariance between itself and the remaining groups. As shown by Cavicchia et al. (2022, Corollary 1), EUCovS is one-to-one associated with a hierarchy of m latent concepts that arise from the collection of variables in m groups. In detail, ${}_V\sigma_{qq(g)}, q = 1, \ldots, m$, define the initial levels of the hierarchy, ${}_W\sigma_{qq(g)}, q = 1, \ldots, m$, are associated with the levels at which the variables are grouped and represent the covariance within the m groups. Finally, values ${}_B\sigma_{qh(g)}, q, h = 1, \ldots, m$, identify the remaining $m-1$ levels and represent the covariance between groups of variables. Thus, the ultrametric property that holds for the relationship between $\boldsymbol{\Sigma}_{W_g}$ and $\boldsymbol{\Sigma}_{B_g}$ (constraint ii, meaning that the variables belonging to the same group are more concordant than the variables belonging to two different groups), and within $\boldsymbol{\Sigma}_{B_g}$ (constraint i, meaning that the groups aggregation in pairs occurs from the most concordant to the least concordant) guarantees the formation of a hierarchy that depicts the relationships within and between groups of variables, from the most concordant to the most discordant.

The parameterization in Eq. (2) requires a reduced number of parameters. Specifically, EUCovS needs at most $p+3m-1$ parameters for each component covariance matrix to be estimated, where p parameters derive from $\boldsymbol{V}_g$, $2m$ from $\boldsymbol{\Sigma}_{V_g}$ and $\boldsymbol{\Sigma}_{W_g}$, and $m-1$ from $\boldsymbol{\Sigma}_{B_g}$. It should be noted that, even when $m = p$, $\boldsymbol{\Sigma}_g$ has a parsimonious structure owing to the ultrametricity of $\boldsymbol{\Sigma}_{B_g}$. In that case, $\boldsymbol{\Sigma}_{V_g} = \boldsymbol{\Sigma}_{W_g}$ and $\boldsymbol{\Sigma}_{B_g}$ provides a hierarchy over the p singletons.

The ultrametric Gaussian mixture model with the covariance structure in (2) is estimated using a grouped coordinate ascent algorithm (Zangwill 1969), which Hathaway (1986) demonstrated to be equivalent to an Expectation–Maximization (EM) algorithm (Dempster et al. 1977) to estimate the GMM parameters. Therefore, by considering a random sample $\boldsymbol{x} = (\mathbf{x}_1, \ldots, \boldsymbol{x}_n)'$ composed of n p-dimensional vectors, the Hathaway log-likelihood function to maximize is

$$\ell(\boldsymbol{W}, \boldsymbol{\Theta}) = \sum_{i=1}^{n}\sum_{g=1}^{G} w_{ig}\Big(\log \pi_g + \log \phi(\boldsymbol{x}_i|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\Big) - \sum_{i=1}^{n}\sum_{g=1}^{G} w_{ig}\log w_{ig}, \tag{3}$$

where $w_{ig}, i = 1, \ldots, n, g = 1, \ldots, G$, are the posterior probabilities considered as parameters and such that $w_{ig} \in [0, 1]$, $\sum_{g=1}^{G} w_{ig} = 1$, and $0 < \sum_{i=1}^{n} w_{ig} < n$, and $\boldsymbol{\Theta} = \{\pi_1, \ldots, \pi_G, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G\}$. $\boldsymbol{\Sigma}_g$ is parameterized as in Eq. (2) and therefore consists of $\boldsymbol{V}_g$, $\boldsymbol{\Sigma}_{V_g}$, $\boldsymbol{\Sigma}_{W_g}$, and $\boldsymbol{\Sigma}_{B_g}$. Henceforth, for the sake of simplicity, we use ℓ instead of $\ell(\boldsymbol{W}, \boldsymbol{\Theta})$.

By maximizing ℓ in (3) with respect to $\boldsymbol{W}$ and $\boldsymbol{\Theta}$ one at a time, keeping the other parameter fixed each time, we obtain the estimates of the posterior probabilities and the ultrametric GMM parameters. PUGMMs introduced in the following section provide more parsimonious EuCovS parameterizations and encompass the one in (2) as the least constrained case. Thus, its estimation details are provided in the next section.

3 Parsimonious Ultrametric Gaussian Mixture Models

Constraints on the parameterization in Eq. (2) bring PUGMMs into being and allow the hierarchical structures of the variables to vary or be equal across and within the mixture components. The class of models proposed herein includes *thirteen* cases obtained by constraining the EUCovS parameters $\boldsymbol{V}_g$, $\boldsymbol{\Sigma}_{V_g}$, $\boldsymbol{\Sigma}_{W_g}$ and $\boldsymbol{\Sigma}_{B_g}$ to be equal within and/or across components or free to vary between them. PUGMMs are coded with a combination of four letters, that is, one for each EUCovS parameter: the first one refers to the variable-group membership matrix, which can be equal (E) or free to vary (F) across components; the second, third, and fourth terms indicate whether the matrix of the group variances, the matrix of the covariances within the groups, and the matrix of the covariances between the groups, respectively, are unique (U, equal within and across components), isotropic (I, equal within components), equal (E, equal across components) or

Table 1 Model ID, nomenclature, covariance structure, and number of covariance parameters for each PUGMM with an example (in brackets, the total number of parameters including $\boldsymbol{\pi}_g$ and $\boldsymbol{\mu}_g$)

Model ID	Nomenclature			Covariance structure	# of covariance parameters	$G = 4, m = 3,$ $p = 100$
	$\boldsymbol{\Sigma}_V$	$\boldsymbol{\Sigma}_W$	$\boldsymbol{\Sigma}_B$			
***V** Equal*						
EUUU	Unique	Unique	Unique	$\boldsymbol{V}(\sigma_W \boldsymbol{I}_m + \sigma_B(\mathbf{1}_m\mathbf{1}_m' - \boldsymbol{I}_m))\boldsymbol{V}' + \text{diag}(\boldsymbol{V}(\sigma_V - \sigma_W)\boldsymbol{I}_m\boldsymbol{V}')$	$p+3$	103 (506)
EUUE	Unique	Unique	Equal	$\boldsymbol{V}(\sigma_W \boldsymbol{I}_m + \boldsymbol{\Sigma}_B)\boldsymbol{V}' + \text{diag}(\boldsymbol{V}(\sigma_V - \sigma_W)\boldsymbol{I}_m\boldsymbol{V}')$	$p+m+1$	104 (507)
EUEE	Unique	Equal	Equal	$\boldsymbol{V}(\boldsymbol{\Sigma}_W + \boldsymbol{\Sigma}_B)\boldsymbol{V}' + \text{diag}(\boldsymbol{V}(\sigma_V \boldsymbol{I}_m - \boldsymbol{\Sigma}_W)\boldsymbol{V}')$	$p+2m$	106 (509)
EEEU	Equal	Equal	Unique	$\boldsymbol{V}(\boldsymbol{\Sigma}_W + \sigma_B(\mathbf{1}_m\mathbf{1}_m' - \boldsymbol{I}_m))\boldsymbol{V}' + \text{diag}(\boldsymbol{V}(\boldsymbol{\Sigma}_V - \boldsymbol{\Sigma}_W)\boldsymbol{V}')$	$p+2m+1$	107 (510)
EEEE	Equal	Equal	Equal	$\boldsymbol{V}(\boldsymbol{\Sigma}_W + \boldsymbol{\Sigma}_B)\boldsymbol{V}' + \text{diag}(\boldsymbol{V}(\boldsymbol{\Sigma}_V - \boldsymbol{\Sigma}_W)\boldsymbol{V}')$	$p+3m-1$	108 (511)
EEEF	Equal	Equal	Free	$\boldsymbol{V}(\boldsymbol{\Sigma}_W + \boldsymbol{\Sigma}_{B_g})\boldsymbol{V}' + \text{diag}(\boldsymbol{V}(\boldsymbol{\Sigma}_V - \boldsymbol{\Sigma}_W)\boldsymbol{V}')$	$p+2m+G(m-1)$	114 (517)
EEFF	Equal	Free	Free	$\boldsymbol{V}(\boldsymbol{\Sigma}_{W_g} + \boldsymbol{\Sigma}_{B_g})\boldsymbol{V}' + \text{diag}(\boldsymbol{V}(\boldsymbol{\Sigma}_V - \boldsymbol{\Sigma}_{W_g})\boldsymbol{V}')$	$p+m+G(2m-1)$	123 (526)
EFFF	Free	Free	Free	$\boldsymbol{V}(\boldsymbol{\Sigma}_{W_g} + \boldsymbol{\Sigma}_{B_g})\boldsymbol{V}' + \text{diag}(\boldsymbol{V}(\boldsymbol{\Sigma}_{V_g} - \boldsymbol{\Sigma}_{W_g})\boldsymbol{V}')$	$p+G(3m-1)$	132 (535)
***V** Free*						
FIII	Isotropic	Isotropic	Isotropic	$\boldsymbol{V}_g(\sigma_{W_g} \boldsymbol{I}_m + \sigma_{B_g}(\mathbf{1}_m\mathbf{1}_m' - \boldsymbol{I}_m))\boldsymbol{V}_g' + \text{diag}(\boldsymbol{V}_g(\sigma_{V_g} - \sigma_{W_g})\boldsymbol{I}_m\boldsymbol{V}_g')$	$G(p+3)$	412 (815)
FIIF	Isotropic	Isotropic	Free	$\boldsymbol{V}_g(\sigma_{W_g} \boldsymbol{I}_m + \boldsymbol{\Sigma}_{B_g})\boldsymbol{V}_g' + \text{diag}(\boldsymbol{V}_g(\sigma_{V_g} - \sigma_{W_g})\boldsymbol{I}_m\boldsymbol{V}_g')$	$G(p+m+1)$	416 (819)
FIFF	Isotropic	Free	Free	$\boldsymbol{V}_g(\boldsymbol{\Sigma}_{W_g} + \boldsymbol{\Sigma}_{B_g})\boldsymbol{V}_g' + \text{diag}(\boldsymbol{V}_g(\sigma_{V_g} \boldsymbol{I}_m - \boldsymbol{\Sigma}_{W_g})\boldsymbol{V}_g')$	$G(p+2m)$	424 (827)
FFFI	Free	Free	Isotropic	$\boldsymbol{V}_g(\boldsymbol{\Sigma}_{W_g} + \sigma_{B_g}(\mathbf{1}_m\mathbf{1}_m' - \boldsymbol{I}_m))\boldsymbol{V}_g' + \text{diag}(\boldsymbol{V}_g(\boldsymbol{\Sigma}_{V_g} - \boldsymbol{\Sigma}_{W_g})\boldsymbol{V}_g')$	$G(p+2m+1)$	428 (831)
FFFF	Free	Free	Free	$\boldsymbol{V}_g(\boldsymbol{\Sigma}_{W_g} + \boldsymbol{\Sigma}_{B_g})\boldsymbol{V}_g' + \text{diag}(\boldsymbol{V}_g(\boldsymbol{\Sigma}_{V_g} - \boldsymbol{\Sigma}_{W_g})\boldsymbol{V}_g')$	$G(p+3m-1)$	432 (835)

free to vary across components (F). The thirteen PUGMMs are listed in detail in Table 1 together with the parameterization of the corresponding covariance structure. It has to be noticed that for the F$\cdots$ models (i.e., the ones for which $\boldsymbol{V}_g$ is let free to vary across components) the last three parameters cannot be constrained across components since the variable partition into groups changes for each component.

As in the general case illustrated in Sect. 2, PUGMMs are estimated using a grouped coordinate ascent algorithm. The updating formula of w_{ig} for $i = 1, \ldots, n$ and $g = 1, \ldots, G$ is

$$\widehat{w}_{ig} = \frac{\widehat{\pi}_g \, \phi(\boldsymbol{x}_i \mid \widehat{\boldsymbol{\mu}}_g, \widehat{\boldsymbol{\Sigma}}_g)}{\sum_{h=1}^{G} \widehat{\pi}_h \, \phi(\boldsymbol{x}_i \mid \widehat{\boldsymbol{\mu}}_h, \widehat{\boldsymbol{\Sigma}}_h)},$$

where the covariance structure depends on PUGMMs. The estimates of π_g and $\boldsymbol{\mu}_g$ for $g = 1, \ldots, G$ are equal for all the thirteen PUGMMs and are detailed below. By omitting additive constant terms w.r.t. $\pi_1, \ldots, \pi_G$, the updating formula of the mixing proportions is obtained by maximizing

$$\ell = \sum_{i=1}^{n} \sum_{g=1}^{G} \widehat{w}_{ig} \log(\pi_g),$$

which leads to

$$\widehat{\pi}_g = \frac{\sum_{i=1}^{n} \widehat{w}_{ig}}{n}.$$

Henceforth, we use $n_g = \sum_{i=1}^{n} \widehat{w}_{ig}$.

By neglecting additive constant terms w.r.t. $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G$, the maximization of the following log-likelihood function

$$\ell = -\frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \widehat{w}_{ig} \mathrm{tr}\Big((\boldsymbol{x}_i - \boldsymbol{\mu}_g)(\boldsymbol{x}_i - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}_g^{-1} \Big)$$

gives rise to the update formula of the component mean vectors

$$\widehat{\boldsymbol{\mu}}_g = \frac{\sum_{i=1}^{n} \widehat{w}_{ig} \, \boldsymbol{x}_i}{n_g}.$$

The features and updating formulas of the component covariance matrices depend on the parsimonious model considered. They are provided in the following subsections by grouping the thirteen PUGMMs in two families: *unique and equal models* (EUUU, EUUE, EUEE, EEEU, EEEE) and *isotropic and free models* (EEEF, EEFF, EFFF, FIII, FIIF, FIFF, FFFI, FFFF). However, regarding the estimates of the variable-group membership matrices, there exist two possible configurations, the E$\cdots$ models and the F$\cdots$ models. For the E$\cdots$ models (i.e., the ones for which $\boldsymbol{V}_g$ is equal across components), the mixtures components share the same variable partition into m groups and $\boldsymbol{V}_g = \boldsymbol{V}$ is estimated row by row across components, i.e., for each $\boldsymbol{v}_j$, $j = 1, \ldots, p$, as follows

$$\begin{cases} \widehat{v}_{jq} = 1 & \text{if } \; q = \underset{q'=1,\ldots,m}{\arg\max} \, \ell \\ \widehat{v}_{jq} = 0 & \text{otherwise} \end{cases},$$

where $\ell = \ell(\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{\Theta}}_{-\boldsymbol{V}}, [\widehat{\boldsymbol{v}}_1, \ldots, \boldsymbol{v}_j = \boldsymbol{i}_{q'}, \ldots, \widehat{\boldsymbol{v}}_p]')$. In the latter, $\widehat{\boldsymbol{\Theta}}_{-\boldsymbol{V}}$ contains all PUGMM parameters, that is, the mixing proportions, the component mean vectors and covariance structures except for $\boldsymbol{V}$, and $\boldsymbol{i}_{q'}$ is the q'th row of the identity matrix of order m. For the F$\cdots$ models, the mixture components can differ for the variable partition into m groups and $\boldsymbol{V}_g$ is estimated row by row per component, i.e., for each $\boldsymbol{v}_{j(g)}$, $j = 1, \ldots, p$, $g = 1, \ldots, G$, as follows

$$\begin{cases} \widehat{v}_{jq(g)} = 1 & \text{if } \; q = \underset{q'=1,\ldots,m}{\arg\max} \, \ell \\ \widehat{v}_{jq(g)} = 0 & \text{otherwise} \end{cases},$$

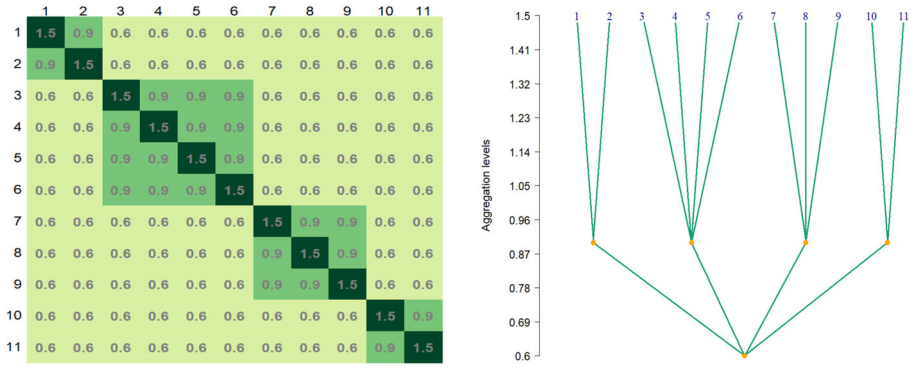
where $\ell = \ell(\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{\Theta}}_{-\boldsymbol{V}_g}, [\widehat{\boldsymbol{v}}_{1(g)}, \ldots, \boldsymbol{v}_{j(g)} = \boldsymbol{i}_{q'}, \ldots, \widehat{\boldsymbol{v}}_{p(g)}]')$ with $\widehat{\boldsymbol{\Theta}}_{-\boldsymbol{V}_g}$ containing all the PUGMM parameters except $\boldsymbol{V}_g$.

3.1 Unique and equal models

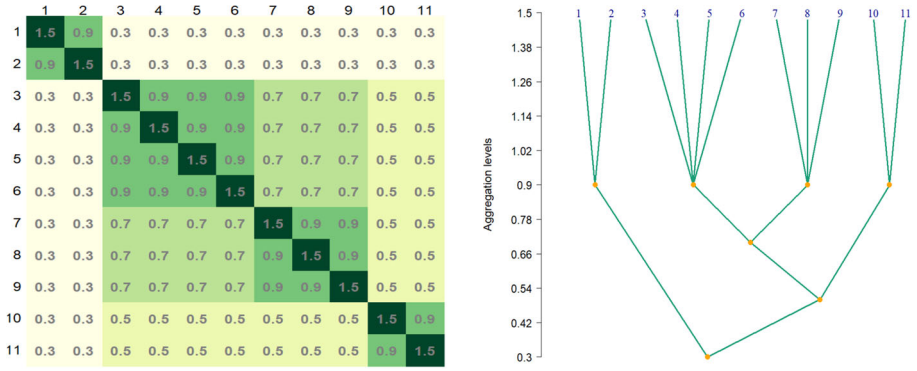
The unique and equal models include the PUGMM cases where the variable-group membership matrices are the same across components, and the group variance matrices, the within-group covariance matrices, and the between-group covariance matrices are characterized by a single unique value each within and across components (EUUU, EUUE, EUEE, EEEU) or are equal across components (EEEE). These models lead to the greatest reduction in the number of parameters involved in the covariance structure estimates compared the freest model, i.e., FFFF (see Table 1). For this family, $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$, and by omitting constant terms as regards the EUCovS parameters, the log-likelihood to maximize is

$$\ell = -\frac{n}{2} \Big[\log(|\boldsymbol{\Sigma}|) + \mathrm{tr}(\boldsymbol{\Sigma}^{-1} \bar{\boldsymbol{S}}) \Big], \tag{4}$$

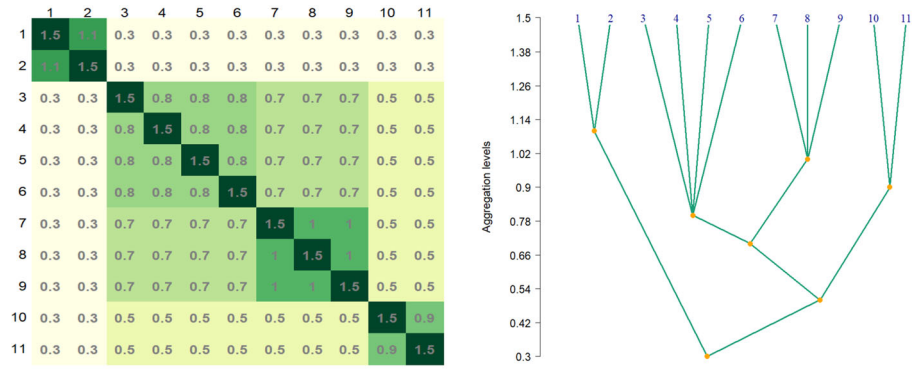
where $\bar{\boldsymbol{S}} = \sum_{g=1}^{G} \hat{\pi}_g \boldsymbol{S}_g$ and $\boldsymbol{S}_g = (1/n_g) \sum_{i=1}^{n} \widehat{w}_{ig} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_g)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_g)'$. The features of the unique and equal models are examined below, while the details of the estimates of their covariance structure parameters are provided in Appendix A.1. In detail, Fig. 1 displays the one-to-one correspondence between the extended ultrametric covariance matrix, through its graphical representation as a heatmap, and the path diagram for the five unique and equal models (EUUU, EUUE, EUEE, EEEU and EEEE). The diagonal



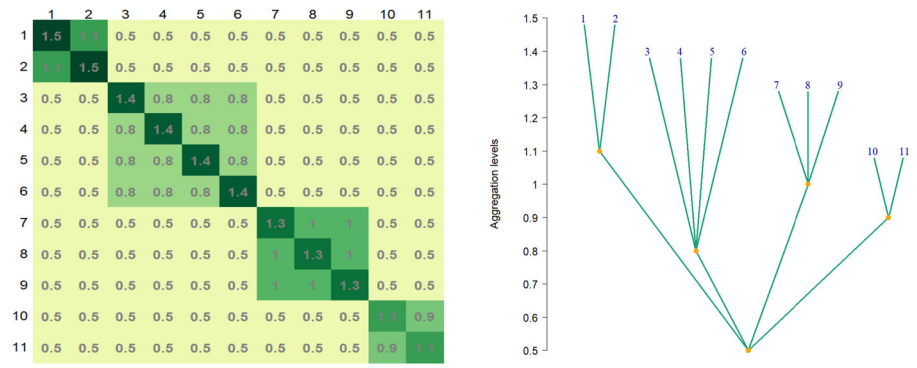
(a) EUUU model



(b) EUUE model



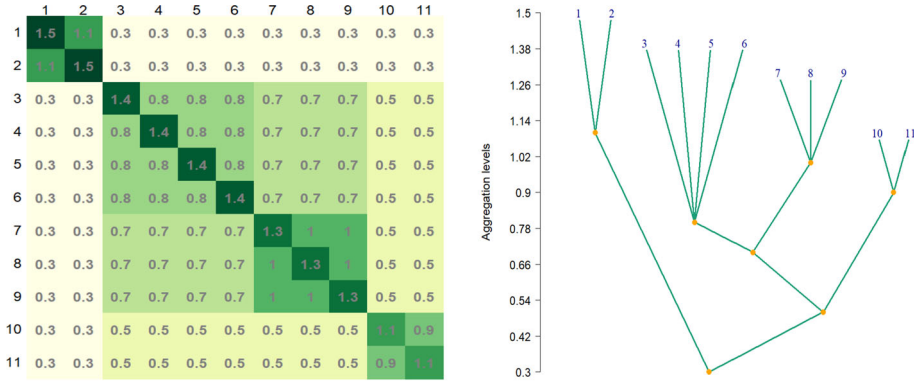
(c) EUEE model



(d) EEEU model

Fig. 1 Example of the heatmap and the corresponding path diagram of the unique and equal models when $m = 4$ and $p = 11$

Fig. 1 continued



(e) EEEE model

elements of the heatmaps consist of the diagonal elements of $\mathbf{\Sigma}_V$, while the elements of diagonal blocks are the diagonal ones of $\mathbf{\Sigma}_W$ and the off-block-diagonal elements consist of the off-diagonal elements of $\mathbf{\Sigma}_B$. Therefore, the aggregation levels of the path diagram correspond exactly to these covariance values.

EUUU model: This is the most constrained case among the thirteen presented here. Indeed, the covariance structure of this model is characterized by a unique value for the main diagonal of $\mathbf{\Sigma}_V$ and $\mathbf{\Sigma}_W$ and for the off-diagonal elements of $\mathbf{\Sigma}_B$. The resulting $\mathbf{\Sigma}$ has at most 3 different values and is the same across components. The main feature of this model is represented by the reduced number of aggregations among variable groups. In fact, in addition to having the same variance and covariance with them, the m variable groups are aggregated at the same level, i.e., σ_B. An example of the EUUU model covariance structure and the corresponding hierarchy is depicted in Fig. 1a. This model represents the situation where all variables enter in the hierarchy at the same level (constant value on the diagonal of $\mathbf{\Sigma}_V$) and are aggregated in m groups with the same intensity (constant value on the diagonal of $\mathbf{\Sigma}_W$), and the aggregation level among the groups is also constant (constant value on the off-diagonal entries of $\mathbf{\Sigma}_B$). Constraints (iii) and (ii) described in Sect. 2 on the relationships between the EUCovS parameters reduce to $\sigma_V > |\sigma_W|$ and $\sigma_W \geq \sigma_B$, respectively. Therefore, the relationships among the variables are modeled such that the aggregation level between the m variable groups is weaker than the aggregation level within the groups, and the latter is in turn weaker than the variance level.

EUUE model: In this case, the group variance matrix $\mathbf{\Sigma}_V$ and the within-group covariance matrix $\mathbf{\Sigma}_W$ have the same value each on their main diagonal, i.e., σ_V and σ_W respectively, whereas $\mathbf{\Sigma}_B$ can have at most $m-1$ different off-diagonal values. This means that the covariances among the m variable groups can differ, leading to different levels of aggregation in the corresponding hierarchical structure. Therefore, the higher number of parameters compared to the EUUU case depends on $\mathbf{\Sigma}_B$, and specifically on m. An example of the EUUE model is shown in Fig. 1b. For this model, the constraint on the relationship between $\mathbf{\Sigma}_V$ and $\mathbf{\Sigma}_W$ remains the same as in the EUUU model, while the constraint on $\mathbf{\Sigma}_W$ and $\mathbf{\Sigma}_B$ becomes $\sigma_W \geq \max_{q,h=1,\ldots,m,h\neq q} {}_B\sigma_{qh}$.

EUEE model: For this model, only the group variance matrix $\mathbf{\Sigma}_V$ is limited to having a unique value on its main diagonal. The other two covariance parameters are free to vary within components (not across them), that is, the m variable groups have different levels of covariance within them other than different levels of aggregation between them. The increase in the number of parameters compared to the EUUE model depends on the m diagonal values of $\mathbf{\Sigma}_W$. An example of the EUEE model is provided in Fig. 1c. In this case, the constraint between $\mathbf{\Sigma}_W$ and $\mathbf{\Sigma}_B$ is the one introduced in Sect. 2—except for the reference to the component—while that between $\mathbf{\Sigma}_V$ and $\mathbf{\Sigma}_W$ becomes $\sigma_V > \max_{q=1,\ldots,m} |{}_W\sigma_{qq}|$.

EEEU model: In this case, only the between-group covariance matrix $\mathbf{\Sigma}_B$ is restricted to have a unique off-diagonal value, whereas $\mathbf{\Sigma}_V$ and $\mathbf{\Sigma}_W$ are free to vary within components. Therefore, the m groups of variables are characterized by m values of the variance, different (at most m) covariances within groups, but they are aggregated at the same level σ_B. An example of the EEEU model is given in Fig. 1d; herein, it can be seen that the "entry level" of the variables corresponding to the group variances can differ across groups, as well as the internal nodes representing the within-group covariances, whereas the aggregation among groups is unique. This is an interesting case that can occur in reality when the hierarchy of latent concepts is composed of only two "levels", one representing specific concepts and the other identifying the general concept. For instance, two important indexes such as the global Multidimensional Poverty Index (MPI, Alkire and Foster 2011; Alkire et al. 2015) and the Human Development Index (HDI, Alkire 2010) follow a two-level model of this type. The constraints of the EEEU model are equal to those displayed in Sect. 2, where σ_B is unique.

EEEE model: This model is the least constrained within the family of unique and equal models. In fact, even if restricted to being the same across components, the EUCovS parameters are free to vary within them. Therefore, the increase in the number of parameters with respect to the EEEU model depends on the $m-1$ different values out of the main diagonal of $\boldsymbol{\Sigma}_B$. An example of the EEEE model is provided in Fig. 1e. Its constraints are equal to those reported in Sect. 2, where $\boldsymbol{\Sigma}$ does not depend on g as well as its parameters.

3.2 Isotropic and free models

The family of the isotropic and free models encompasses two subgroups: the first referring to the models where the variable-group membership matrices are equal across components (EEEF, EEFF, EFFF) and the second where they are free to vary (FIII, FIIF, FIFF, FFFI, FFFF). Although in the second group, the other EUCovS parameters can be characterized by a single value each, these values must differ across components. It should be noted that the number of parameters of the isotropic model FIII (resp. FIIF, FIFF and FFFI) corresponds to that of the equal model EUUU (resp. EUUE, EUEE and EEEU) multiplied by the number of mixture components, whereas that of the EEEF, EEFF, and EFFF cases is in-between. The least constrained model is the FFFF case described in Sect. 2.

By omitting constant terms with respect to the EUCovS parameters, the log-likelihood to maximize for this family is

$$\ell = -\frac{1}{2}\left[\sum_{g=1}^{G} n_g \log(|\boldsymbol{\Sigma}_g|) + \sum_{g=1}^{G} n_g \mathrm{tr}\left(\boldsymbol{\Sigma}_g^{-1}\boldsymbol{S}_g\right)\right]. \tag{5}$$

The features of the isotropic and free models are examined below, while the details of their estimates are provided in Appendix A.2.

EEEF model: Unlike the EEEE model, in this case the between-group covariance matrix varies across components, i.e., $\boldsymbol{\Sigma}_{B_g}$. This means that the aggregation levels between variable groups can change across components, even if the variable partition is the same, affecting the number of parameters to estimate. The latter has to take into account the $G(m-1)$ values of $\boldsymbol{\Sigma}_{B_1}, \ldots, \boldsymbol{\Sigma}_{B_G}$. For simplicity reasons, we do not display an example of this model since the reader might infer it by replicating Fig. 1e for G components with different values of covariance between the variable groups across components. With respect to the EEEE model, the constraint on the relationship between the covariance within and between groups changes: this turns out to be $\sigma_W \geq \max_{g=1,\ldots,G}\left\{\max_{q,h=1,\ldots,m,h\neq q} {}_B\sigma_{qh(g)}\right\}$, since $\boldsymbol{\Sigma}_{B_g}$ varies across components, whereas $\boldsymbol{\Sigma}_W$ does not.

EEFF model: By relaxing the constraint on the within-group covariance matrix that holds in the EEEF model, this case is obtained. Therefore, since $\boldsymbol{\Sigma}_{W_g}$ varies across components, the EEFF model encompasses also the Gm values of $\boldsymbol{\Sigma}_{W_1}, \ldots, \boldsymbol{\Sigma}_{W_G}$ as parameters to estimate. Likewise the EEEF case, it is easy to derive an example for this model by replicating Fig. 1e while changing the covariance values within and between the variable groups in the G components. Constraint (ii) in Sect. 2 holds for the relationship between $\boldsymbol{\Sigma}_{W_g}$ and $\boldsymbol{\Sigma}_{B_g}$, whereas that one between $\boldsymbol{\Sigma}_V$ and $\boldsymbol{\Sigma}_{W_g}$ turns out to be ${}_V\sigma_{qq} > \max_{g=1,\ldots,G} |{}_W\sigma_{qq(g)}|$ for $q = 1, \ldots, m$. It must be recalled that the variable-group membership matrix does not vary between components, therefore, the row-by-row comparison between the elements of $\boldsymbol{\Sigma}_V$ and $\boldsymbol{\Sigma}_{W_g}$, $g = 1, \ldots, G$, is reasonable.

EFFF model: This model differs from the EEFF case, as also the group variance matrix is free to vary across components, i.e., $\boldsymbol{\Sigma}_{V_g}$, thus increasing the number of parameters to take into account its Gm values. Equivalently, the reader can easily derive an example for this model through the replication of Fig. 1e, where the values of the covariance matrix change between the G components although the variable partition remains the same. The constraints on the parameters of the EFFF case correspond to those presented in Sect. 2.

FIII model: Unlike the EUUU model, in this case, the variable-group membership matrix differs across components, that is $\boldsymbol{V}_g$, and the unique values in the other EUCovS parameters have to be component-specific. Consequently, the number of parameters of the EUUU model must be multiplied by the number of mixture components G to obtain the one of the FIII model. The reader can easily infer an example of this case by replicating the covariance structure shown in Fig. 1a with different variable partitions into groups along with distinct parameters values per component. The constraints on $\boldsymbol{\Sigma}_{V_g}$, $\boldsymbol{\Sigma}_{W_g}$ and $\boldsymbol{\Sigma}_{B_g}$ turn out to be $\sigma_{V_g} > |\sigma_{W_g}|$ and $\sigma_{W_g} \geq \sigma_{B_g}$ for $g = 1, \ldots, G$.

FIIF model: The difference between this isotropic model and the unique model EUUE figures in $\boldsymbol{V}_g$, which varies across components by making free the other EUCovS parameters across them. As well as the FIII case, an example of this model can be easily derived from its unique counterpart depicted in Fig. 1b. The number of parameters in this model corresponds to that of the EUUE case multiplied by the number of components G. The constraints on the relationship between $\boldsymbol{\Sigma}_{V_g}$ and $\boldsymbol{\Sigma}_{W_g}$ remain the same from the FIII model, whereas those between $\boldsymbol{\Sigma}_{W_g}$ and $\boldsymbol{\Sigma}_{B_g}$ are taken from EUUE and transformed to be component-specific, that is, $\sigma_{W_g} \geq \max_{q,h=1,\ldots,m,h\neq q} {}_B\sigma_{qh(g)}$ for $g = 1, \ldots, G$.

FIFF model: This model is the isotropic counterpart of the EUEE model. As for the other two previous models, in the

FIFF case, the partition of variables into groups changes throughout the components. This entails differences in the other EUCovS parameters across components, whose number is equal to G times that of the EUEE model. The reader can easily think of an example of this case by generalizing it to that depicted in Fig. 1c. For the FIFF model, the constraint between $\mathbf{\Sigma}_{W_g}$ and $\mathbf{\Sigma}_{B_g}$ is the one introduced in Sect. 2, while that between $\mathbf{\Sigma}_{V_g}$ and $\mathbf{\Sigma}_{W_g}$ is $\sigma_{V_g} > \max_{q=1,\ldots,m} |_W\sigma_{qq(g)}|$ for $g = 1, \ldots, G$.

FFFI model: As well as for the previous three cases, this model has its counterpart among the unique models, that is, EEEU. The FFFI model constrains only the between-group covariance matrix to have one single value *within* components, even if this value can vary across them, i.e., σ_{B_g}, as well as $\mathbf{\Sigma}_{V_g}$, $\mathbf{\Sigma}_{W_g}$ and $\boldsymbol{V}_g$. An example of this model can be easily obtained by considering the structure depicted in Fig. 1d with a different variable partition and variance and covariance values per component. The constraints of the FFFI case are equal to those displayed in Sect. 2, where the parameters of EUCovS depends on g and σ_{B_g} is a unique value within components.

FFFF model: This is the least constrained case among the thirteen presented herein. Indeed, the FFFF model corresponds to that illustrated in Sect. 2 and introduced by Cavicchia et al. (2022), where examples of the FFFF covariance structure are provided. This case represents the most general ultrametric model, where all the EUCovS parameters differ throughout the mixture components.

As formerly described, constraints (iii) and (ii) on $\mathbf{\Sigma}_{V_g}$, $\mathbf{\Sigma}_{W_g}$, and $\mathbf{\Sigma}_{B_g}$ hold for any case. For instance, for the FFFF model, they are implemented as follows: ${}_V\sigma_{qq(g)} = |_W\sigma_{qq(g)}| + 1.5 \times 10^{-8}$ for some (or all) q and g, and $\min\{_W\sigma_{qq(g)}, q = 1, \ldots, m\} = \max\{_B\sigma_{qh(g)}, q, h = 1, \ldots, m, h \neq q\}$ for some (or all) g, respectively, when necessary. In the more constrained cases, i.e., unique and isotropic models, where the parameters have a unique value, they are straightforwardly applied without considering the reference to the variable groups.

4 Computational aspects of the PUGMMs algorithm and model section

4.1 Initialization

The proposed algorithm for the estimation of PUGMMs should be run multiple times with different initial values of $\boldsymbol{W}$ and $\boldsymbol{V}$ (or $\boldsymbol{V}_g, g = 1, \ldots, G$, for the F$\cdots$ models) to increase the chance of reaching the global optimum of the log-likelihood function. The initial values of $\boldsymbol{W} = [w_{ig}]$ can be randomly selected such that $w_{ig} \in [0, 1]$, $\sum_{g=1}^{G} w_{ig} = 1$ and $0 < \sum_{i=1}^{n} w_{ig} < n$, for all i. As an alternative, the solution of k-means (MacQueen 1967) with $k = G$ (default in our experiments) or fuzzy c-means (Bezdek 1974, 1981) as the initial values of w_{ig} can be used.

The initial values of $\boldsymbol{V}$, or $\boldsymbol{V}_g$ for $g = 1, \ldots, G$, can be randomly chosen so that the variable-group membership matrix turns out to be binary and row-stochastic, or obtained from the solution of an adapted UCM algorithm (Cavicchia et al. 2020) applied to covariance matrices, as reported in Cavicchia et al. (2022). The initial values of π_g and $\boldsymbol{\mu}_g$ can then be calculated, as well as those of $\mathbf{\Sigma}_{V_g}$, $\mathbf{\Sigma}_{W_g}$ and $\mathbf{\Sigma}_{B_g}$ according to the chosen PUGMM case.

4.2 Canonical representation

The calculation of the log-likelihood for GMMs can be computationally demanding when dealing with high-dimensional data, as it requires computing the determinant and inverse of the covariance matrix. The ultrametric correlation and covariance matrices introduced by Cavicchia et al. (2020, 2022) have a reduced number of parameters to be fit due to their block structure, allowing us to use its peculiar form to save computational power when fitting PUGMMs.

In detail, Archakov and Hansen (2020) proposed a canonical representation for block matrices that facilitates the computation of operators, such as, determinant and inverse, for this type of matrices. This representation is a semispectral decomposition of a block matrix that is diagonalized with the exception of a single diagonal block, whose dimension is given by the number of blocks. The covariance structure in (2) can always be written as a block structure, after reordering the variables into groups, where the off-diagonal blocks have identical entries (i.e., the off-diagonal values of $\mathbf{\Sigma}_{B_g}$), while the diagonal blocks consist of variance—on the diagonal—and covariance—off-diagonal—within each group of variables, i.e., the values stored in $\mathbf{\Sigma}_{V_g}$ and $\mathbf{\Sigma}_{W_g}$, respectively. This means that each off-diagonal block has dimensions $n_q \times n_h$ for $q, h = 1, \ldots, m, h \neq q$, and each diagonal one $n_q \times n_q$ for $q = 1, \ldots, m$, where n_q represents the number of variables in the qth group. For simplicity, we introduce the canonical representation of $\mathbf{\Sigma}_g$ by omitting the reference to the component, i.e., g.

Let the orthonormal rotation matrix $\boldsymbol{Q}$ be defined as

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{v}_{n_1} & \mathbf{0} & \ldots & \mathbf{0} & \boldsymbol{v}_{n_1\perp} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{v}_{n_2} & \ddots & \mathbf{0} & \mathbf{0} & \boldsymbol{v}_{n_2\perp} & \ddots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{v}_{n_m} & \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{v}_{n_m\perp} \end{bmatrix},$$

where $\boldsymbol{v}_{n_q} = n_q^{1/2}\mathbf{1}_{n_q}$ is a vector of dimension n_q that spans the eigenspace corresponding to the qth eigenvalue and $\boldsymbol{v}_{n_q\perp}$ is a $n_q \times (n_q - 1)$ matrix that is the orthogonal complement

to $\boldsymbol{v}_{n_q}$. Therefore, the block diagonal matrix $\boldsymbol{D} = \boldsymbol{Q}'\boldsymbol{\Sigma}\boldsymbol{Q}$ has the following form

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \lambda_1 \boldsymbol{I}_{n_1-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \dots & \boldsymbol{0} & \lambda_m \boldsymbol{I}_{n_m-1} \end{bmatrix},$$

where $\boldsymbol{A}$ has order m with $a_{qh} = {}_B\sigma_{qh}\sqrt{n_q n_h}$ for $q \neq h$, and $a_{qq} = {}_V\sigma_{qq} + {}_W\sigma_{qq}(n_q - 1)$. Moreover, $\lambda_q = {}_V\sigma_{qq} - {}_W\sigma_{qq}$.

Given the canonical representation of $\boldsymbol{\Sigma}$, such that $\boldsymbol{\Sigma} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}'$, we can rewrite Eq. (3) as follows

$$\ell = \sum_{i=1}^{n}\sum_{g=1}^{G} w_{ig}\left[\log \pi_g + \log\left(2\pi^{-p/2}|\boldsymbol{D}_g|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_g)'\boldsymbol{Q}_g\boldsymbol{D}_g^{-1}\boldsymbol{Q}_g'(\boldsymbol{x}_i - \boldsymbol{\mu}_g)\right)\right)\right] - \sum_{i=1}^{n}\sum_{g=1}^{G} w_{ig}\log(w_{ig}). \quad (6)$$

In this way, instead of inverting the $p \times p$ matrix $\boldsymbol{\Sigma}_g$ and computing $|\boldsymbol{\Sigma}_g|$, it suffices to invert the smaller $m \times m$ matrix, $\boldsymbol{A}_g$, and evaluate $|\boldsymbol{A}_g|$. It should be noted that the eigenvalues of $\boldsymbol{\Sigma}_g$ correspond to those of $\boldsymbol{D}_g$ (see Archakov and Hansen 2020).

4.3 Polar decomposition

When EUCovS turns out not to be positive definite after the estimation of its parameters, this property is obtained via the polar decomposition. Specifically, consider a Hermitian matrix $\boldsymbol{\Sigma}$ of order p—the reference to the component g is removed in this section for simplicity reasons—its polar decomposition is $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{H}$, where $\boldsymbol{U}$ is a unitary matrix of order p and $\boldsymbol{H}$ is a Hermitian positive semidefinite matrix of the same order. The latter equals the spectral decomposition of $\boldsymbol{\Sigma}$ with its negative eigenvalues taken in absolute terms. Higham (1986) demonstrated that the matrix $(\boldsymbol{\Sigma} + \boldsymbol{H})/2$ is the nearest Hermitian positive semidefinite matrix to $\boldsymbol{\Sigma}$ in the 2-norm.

It has to be noticed that this procedure does not guarantee the ultrametricity of the resulting matrix. For this reason, the nearest EUCovS matrix is fit to $(\boldsymbol{\Sigma} + \boldsymbol{H})/2$, depending on the PUGMM case. If this new ultrametric matrix results not to be positive semidefinite, its smallest eigenvalue taken in absolute value is added to its diagonal elements, plus a small constant ($\approx 1.5 \times 10^{-8}$ in our experiments) to comply with the positive definiteness of the estimated matrix, as proposed by Cailliez (1983) and implemented by Cavicchia et al. (2022). Therefore, the resulting matrix, obtained via the latter procedure, whenever needed, is the nearest extended ultrametric positive semidefinite (adding the aforementioned constant, positive definite) matrix to $\boldsymbol{\Sigma}$ in the 2-norm, while applying the proposal by Cailliez (1983) directly to $\boldsymbol{\Sigma}$ only does guarantee to reap an ultrametric and positive definite matrix by inflating the diagonal values of $\boldsymbol{\Sigma}_V$ and potentially ending up to an overestimation of this parameter.

4.4 Model selection

A major question in GMMs is model selection. In detail, for PUGMMs, we have to determine both the number of components to include in the mixture (G) and the number of variable groups (m), as well as the covariance structure to assume for the components (EUUU, EUUE, …, FFFF). To address all these issues we consider information criteria, as the Bayesian Information Criterion (BIC, Schwarz 1978) and Integrated Completed Likelihood (ICL, Biernacki et al. 2000).

The main feature of the information criteria is that they are based on penalized forms of the log-likelihood. Therefore, a penalty term for the number of free parameters to estimate is subtracted from the log-likelihood, which increases with the addition of more components. BIC is a popular choice in the context of GMMs. For PUGMMs, the BIC formula takes the form

$$\text{BIC} = 2\ell_{\max} - \nu \log n, \quad (7)$$

where $\ell_{\max}$ is the maximized log-likelihood value and ν is the total number of free parameters in the model. The latter accounts for the $G - 1$ and Gp parameters for estimating the mixing proportions and the component mean vectors, respectively, common to all PUGMM cases, and varies depending on the covariance structure, which is therefore chosen, together with G and m, to maximize BIC. Specifically, in the FFFF model, $\nu = 2G(p + m) - 1 - (c_{V,W} + c_{W,B})$. Other than the aforementioned free parameters for estimating $\pi_1, \dots, \pi_G$, and $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_G$, ν is also given by p parameters minus m constraints (non-empty groups) for the estimation of each $\boldsymbol{V}_g$, m parameters for the diagonal values of each $\boldsymbol{\Sigma}_{V_g}$ and $\boldsymbol{\Sigma}_{W_g}$ (i.e., $2m$), $m - 1$ parameters for the off-diagonal elements of each $\boldsymbol{\Sigma}_{B_g}$ (i.e., number of levels in the hierarchy), and subtracting from this number $c_{V,W} + c_{W,B}$ that represent the constraints activated in the algorithm to satisfy constraints (iii) and (ii), respectively (see Sect. 2). Evidently, ν decreases as more constrained PUGMM models are considered. For all cases, the maximum number of $c_{V,W}$ and $c_{W,B}$ per model that can be activated is reported in Appendix B.

ICL is based on BIC and penalizes it by deducting an entropy term that assesses the overlap between the components. ICL favors solutions where the overlap among the components is not too large (Celeux et al. 2018). However,

because an asymptotic approximation of the log-posterior probability of the models exists (Kass and Raftery 1995), we suggest using BIC to select the number of mixture components, the number of variable groups, and the covariance case. Although mixture models typically do not satisfy the regularity requirements for the asymptotic approximation utilized in the formulation of BIC, Kerebin (1998, 2000) demonstrated that it yields consistent estimates of the number of components in a mixture model. Furthermore, in the same field, Fraley and Raftery (1998, 2002) gave examples to demonstrate how BIC works effectively for model selection.

4.5 Aitken's acceleration

As shown by Lindstrom and Bates (1988), the relative change in the log-likelihood function usually adopted in the EM algorithm does not represent a proper stopping criterion, but rather a lack-of-progress criterion. Following McLachlan and Krishnan (2008), we implement the Aitken acceleration-based stopping rule in our algorithm. In detail, the Aitken's acceleration at iteration $t-1$ is given by

$$a^{(t-1)} = \frac{\ell^{(t)} - \ell^{(t-1)}}{\ell^{(t-1)} - \ell^{(t-2)}},$$

where $\ell^{(t-2)}$, $\ell^{(t-1)}$ and $\ell^{(t)}$ are the log-likelihood values at iterations $t-2$, $t-1$ and t, respectively. The log-likelihood asymptotic estimate at iteration t needed to compute the stopping criterion is given by

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

The PUGMMs algorithm can be considered to have converged when $\ell_\infty^{(t)} - \ell^{(t-1)} < \epsilon$ (McNicholas et al. 2010), where $\epsilon > 0$ is the desired tolerance (e.g., 1.5×10^{-8} in our experiments). Alternative stopping criteria are, for example, $\ell_\infty^{(t)} - \ell_\infty^{(t-1)} < \epsilon$ (Böhning et al. 1994) and $\ell_\infty^{(t)} - \ell^{(t)} < \epsilon$ (Lindsay 1995).

5 Applications

We evaluate the performance of PUGMMs on thirteen benchmark data sets (Sect. 5.1), where the theoretical clustering structure is known a priori, and on two real-world applications: the first provides insights on the features of the FIFA player roles (Sect. 5.2), whereas the latter inspects the use of Wikipedia as a teaching approach within the university (Sect. 5.3). The analysis of these benchmark data sets shows the potential of PUGMMs in recovering clustering structure, whereas the further two applications illustrate the deeper ability of PUGMMs to estimate the features of the covariance structures and their interpretation. It is worth noting that, in this section, we refer to components as "clusters" since we obtain a partition of the unit space according to the Maximum A Posteriori (MAP) classification.

The benchmark data sets and the two real-world applications display our proposal's performance when no assumption on the data generating process is made; however, we provide in the Supplementary Materials a simulation study where we generate data from PUGMMs.

5.1 Benchmark data sets

In this section, we compare PUGMMs with GPCMs (R package `mclust`, Fraley and Raftery 1999; Scrucca et al. 2016), PGMMs (R package `pgmm`, McNicholas et al. 2019) and HDDC (R package `HDclassif`, Bergé et al. 2012) in correctly detecting the expected clustering structure. For each model, we select the triplet $(G, m, case)$ according to BIC. The value of G ranges from 1 to $G^* + 2$, where G^* represents the theoretical value. Furthermore, whenever the "true" number of clusters is not correctly identified, we also run the models by fixing $G = G^*$. For PUGMMs and PGMMs, we choose m in $\{1, \ldots, 10\}$; however, if $p < 10$, m is at most equal to p. The model *case* is chosen among those illustrated in Sect. 3 for PUGMMs; for competitors, the reader can refer to the references reported in Sect. 1. We assess the clustering performance of the models using the Adjusted Rand Index (ARI, Hubert and Arabie 1985), which quantifies the similarity between the theorized and estimated classifications by reaching 1 in the case of perfect agreement.

We analyze thirteen benchmark data sets retrieved from different sources, which are detailed in Appendix C. The variables of all these data sets are standardized to z-score, moreover it should be noted that in these data sets, a hierarchical structure is not assumed or inspected beforehand. The results synthesized in Table 2 show comparable performance of PUGMMs to the competitors in classification recovery. Particularly, PUGMMs achieve this goal while often estimating a smaller number of parameters. In the first four data sets, PUGMMs correctly detect the theoretical number of clusters and achieves similar results w.r.t. the other models in terms of ARI. We also achieve compelling results on the remaining nine data sets. Specifically, we can split the results into two distinct sets: first, cases where PUGMMs fail to identify G^* while at least one of the competitors does succeed to correctly detect it; and second, cases where none of the models correctly select the number of clusters. On the former, the proposal can accurately identify the theoretical clustering structure when the value of G is set to G^*, achieving the maximum ARI for *Ceramic*, while displaying comparable and better results for *Tetragonula* and *Sobar*, respectively. For the other set composed of *Kidney*, *Economics*, *Ais*, *Banknotes*, and *Coffee*, G differs from the theoretical one, despite

Table 2 Clustering results on the benchmark data sets per model, where G, m (for PUGMMs and PGMMs) and Case are chosen according to BIC, and ν (number of free parameters) and ARI are computed accordingly

Data set	Features			PUGMMs					GPCMs			
	n	G^*	p	G	m	Case	ν	ARI	G	Case	ν	ARI
Penguin	342	3	4	3	1	EEFF	21	0.97	4 (3)	VEE (VEE)	32 (26)	0.81 (0.96)
Wine 13	178	3	13	3	8	FFFF	103	0.93	3	VVE	158	0.93
Wine 27	178	3	27	3	9	FFFI	189	0.88	3	VVE	515	0.93
Thyroid	215	3	5	3	3	FFFI	44	0.85	3	VVI	32	0.89
Kidney	203	2	11	4 (2)	7 (6)	FFFI (FFFI)	112 (58)	0.81 (0.01)	3 (2)	VVI (VVI)	68 (45)	0.78 (0.88)
Economics	55	2	12	3 (2)	5 (8)	FFFI (FFFI)	89 (57)	0.81 (1.00)	3 (2)	VVI (VVI)	74 (49)	0.32 (1.00)
Tetragonula	236	9	4	7 (9)	3 (2)	FFFF (EFFF)	87 (80)	0.80 (0.81)	9	VVE	86	0.71
Diabetes	145	3	3	5 (3)	2 (2)	FFFI (FFFI)	44 (29)	0.68 (0.51)	3	VVV	29	0.66
Ais	202	2	11	4 (2)	7 (8)	FFFI (EEEE)	107 (49)	0.59 (0.96)	4 (2)	EVE (EEV)	143 (144)	0.63 (0.76)
Ceramic	88	2	17	3 (2)	6 (9)	FFFI (FFFI)	125 (89)	0.56 (1.00)	2	VVE	205	0.59
Banknotes	200	2	6	4 (2)	5 (5)	EEEE (EEEE)	38 (23)	0.56 (0.96)	4 (2)	VEE (EEV)	51 (49)	0.68 (0.98)
Coffee	43	2	12	4 (2)	2 (5)	EFFF (EEEE)	77 (41)	0.38 (1.00)	3 (2)	VEI (VEI)	52 (38)	0.38 (1.00)
Sobar	72	2	19	3 (2)	6 (5)	FFFI (FFFI)	136 (89)	0.21 (0.40)	4 (2)	VEI (VVE)	101 (248)	0.30 (0.16)

Data set	Features			PGMMs					HDDC			
	n	G^*	p	G	m	Case	ν	ARI	G	Case	ν	ARI
Penguin	342	3	4	3	1	CUC	21	0.96	3	AJBQD	23	0.97
Wine 13	178	3	13	3	2	CUU	105	0.92	3	AKBKQKD	147	0.93
Wine 27	178	3	27	3	6	CUU	311	0.93	4 (3)	ABQD (AJBQD)	212 (187)	0.83 (0.93)
Thyroid	215	3	5	– (3)	– (1)	– (UUCU)	– (45)	– (0.86)	3	AKBKQKDK	52	0.88
Kidney	203	2	11	4 (2)	1 (2)	CUU (UUCU)	102 (86)	0.78 (0.90)	3 (2)	AKJBKQKDK (AKJBKQKDK)	174 (128)	0.88 (0.88)
Economics	55	2	12	– (2)	– (5)	– (UUU)	– (149)	– (1.00)	4 (2)	ABKQKD (AKBQKDK)	101 (89)	0.32 (0.89)
Tetragonula	236	9	4	9	2	CUU	87	0.81	9	AKJBKQKD	117	0.82
Diabetes	145	3	3	3	2	UCU	29	0.64	3	AKJBKQKD	30	0.67
Ais	202	2	11	3 (2)	4 (4)	UCU (UUCU)	160 (120)	0.51 (0.81)	3 (2)	AKJBQKDK (AKJBKQKDK)	146 (102)	0.36 (0.35)
Ceramic	88	2	17	2	9	UUCU	302	1.00	2	AKJBQKDK	133	1.00
Banknotes	200	2	6	4 (2)	2 (2)	CCUU (UUCU)	47 (46)	0.65 (0.98)	4 (2)	AJBQD (AJBQD)	44 (30)	0.70 (0.98)
Coffee	43	2	12	– (2)	– (1)	– (CCUU)	– (50)	– (1.00)	3 (2)	ABQD (AJBQD)	52 (39)	0.38 (1.00)
Sobar	72	2	19	2	4	CUCU	146	0.36	4 (2)	AJBQD (ABQD)	118 (77)	0.22 (0.33)

Results in brackets are obtained by fixing G to the theoretical value G^*

"–" represents the occurrence of an error while running the experiment in R, making impossible to report the results

Table 3 Variables studied for FIFA

ID	Variable name	ID	Variable name
1	Crossing	16	Shot power
2	Finishing	17	Jumping
3	Heading accuracy	18	Stamina
4	Short passing	19	Strength
5	Volleys	20	Long Shots
6	Dribbling	21	Aggression
7	Curve	22	Interceptions
8	FK Accuracy	23	Positioning
9	Long passing	24	Vision
10	Ball control	25	Penalties
11	Acceleration	26	Composure
12	Sprint speed	27	Marking
13	Agility	28	Standing tackle
14	Reactions	29	Sliding tackle
15	Balance		

there may be cases where some or all models select the same value, e.g., on *Banknotes* all the competitors choose $G = 4$ and on *Ais* PUGMMs and GPCMs single out $G = 4$, whereas PGMMs and HDDC choose $G = 3$. In these benchmarks, it can be reasonable to assume that the "true" clusters do not align with distinguishable patterns in the data, as none of the models, regardless of their component covariance structure, can accurately select the "true" value of G (Hennig 2022).

5.2 Grouping soccer players with similar skill-sets in FIFA

Table 4 First four best players per cluster

1	Robert Lewandowski	5	Neymar Jr
1	Harry Kane	5	David Silva
1	Edison Cavani	5	Mohamed Salah
1	Gonzalo Higuan	5	Lorenzo Insigne
2	Kevin De Bruyne	6	Marcelo
2	Lukas Modric	6	Jordi Alba
2	Tony Kroos	6	Alex Sandro
2	Paul Pogba	6	Joshua Kimmich
3	Giorgio Chiellini	7	N'Golo Kante
3	Kalidou Koulibaly	7	Sergio Busquets
3	Medhi Benatia	7	Casemerio
3	Milan Škriniar	7	Marco Verratti
4	Sergio Ramos	8	Lionel Messi
4	Diego Godin	8	Cristiano Ronaldo
4	Mats Hummels	8	Eden Hazard
4	Thiago Silva	8	Luis Surez

The Fédération Internationale de Football Association (FIFA) is a governing body of football (sometimes, especially in the USA, called soccer). FIFA is also a series of a football simulation games developed by EA Sports which faithfully reproduces the characteristics of real players. The main characters of the video game are the football players, and players in the video game are meant to be as close to the real ones, both physically and in skills. This set of skills also determines the position they play on the field. FIFA ratings of football players from the video game is contained in Giordani et al. (2020) and can be downloaded from the R package `datasetsICR`. In detail, the data set contains the attributes for every player registered in the latest edition of FIFA 19 database, and consists of 18,207 observations on 80 variables measured on a 0–100 scale. However, for this application, we select the 1398 best outfield players—those with the variable *Overall* larger than 75—and 29 variables representing their skills. Goalkeepers, being characterized only by specific variables, are discarded because they form a separated cluster that can be easily detected by any clustering method. Table 3 reports the complete list of variables considered in this application.

The thirteen PUGMMs are fitted to the data for $G = 1, 2, \ldots, 10$, and $m = 1, 2, \ldots, 10$. The model with the highest BIC, equal to -271716.7, is the FFFI model with $G = 8$ and $m = 10$. Specifically, the first cluster is characterized by the strikers, while the second cluster consists of attacking midfielders. The third and fourth clusters are composed by stoppers and central backs, respectively. The fifth cluster is characterized by the *number ten* players who are very talented and have an evident attacking predisposition. The sixth and seventh clusters consist of back wings and midfielders, respectively. Finally, the forward players compose the eighth cluster. Table 4 reports the 4 best players per cluster, while Fig. 2 displays a MultiDimensional Scaling (MDS) scatterplot of the same players. Furthermore, Fig. 3 shows the complete MDS scatter plot of all players by highlighting the 8 clusters. The players presented in Table 4 clearly illustrate that the clusters are formed based on players' positions, which, in turn, are determined by their skills. It is worth noting that a player's football position is not always definitive, as they may be capable of playing in various positions, and their understanding of a specific position can be subjective. For instance, we can have both highly offensive and defensive wing backs occupying the same assigned position, despite their distinct skill sets. The same holds for midfielders, and this explains why they appear in close proximity and overlap in Fig. 3. Cluster 5 (*number tens*) and Cluster 8 (forwards) are also extremely similar, the main difference we can observe is that players in Cluster 8 have a greater propensity to score many goals.

The FFFI model suggests a second-order model for every cluster, with cluster-specific variable groups, group vari-

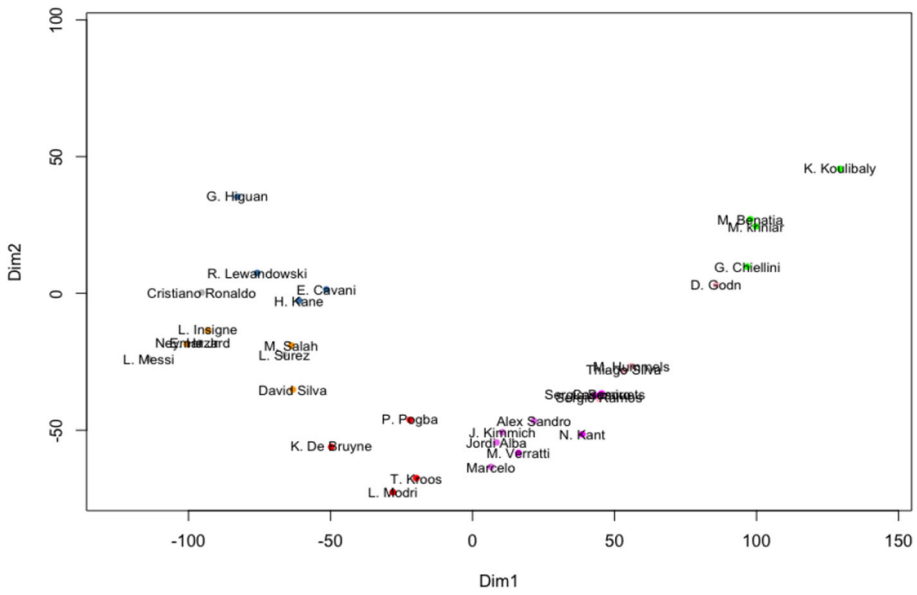


Fig. 2 MDS scatterplot (first two MDS dimensions) of the 4 best players per cluster as listed in Table 4

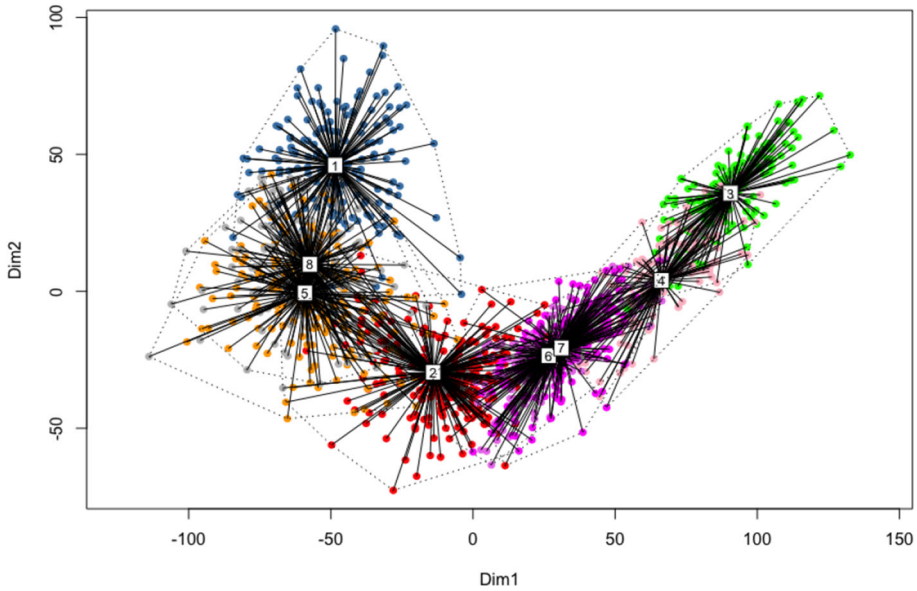


Fig. 3 MDS scatterplot (first two MDS dimensions) of all players

ances, within-group covariances, and between-group covariances (Fig. 4). It is evident that distinct clusters or players' positions require different variable structures. Despite the significant differences among the eight hierarchies in terms of variable group identification and aggregation, it is observable that certain variables are frequently grouped together due to their high correlation and association with the same macro-skills. For example, variables related to speed and running are grouped together, as well as variables associated with technical skills are grouped into one group. The same principle applies to defensive skills.

5.3 The use of Wikipedia for teaching within universities

The further real-world application of PUGMMs delves into the use of Wikipedia for teaching within universities (UCI repository, Aibar et al. 2015). We focus on variables directly related to the use of Wikipedia for higher education teaching activities that refer to six latent concepts (Table 5). The variables are measured on a 5-point Likert scale that reflects the level of disagreement or agreement with a statement or to the frequency of certain actions, depending on the nature of the question, ranging from 1 (strongly disagree or never) to 5 (strongly agree or always). The data set consists of 800 members (full, associate, and assistant professors, lecturers, instructors, adjuncts, all called "professors" hereinafter) of

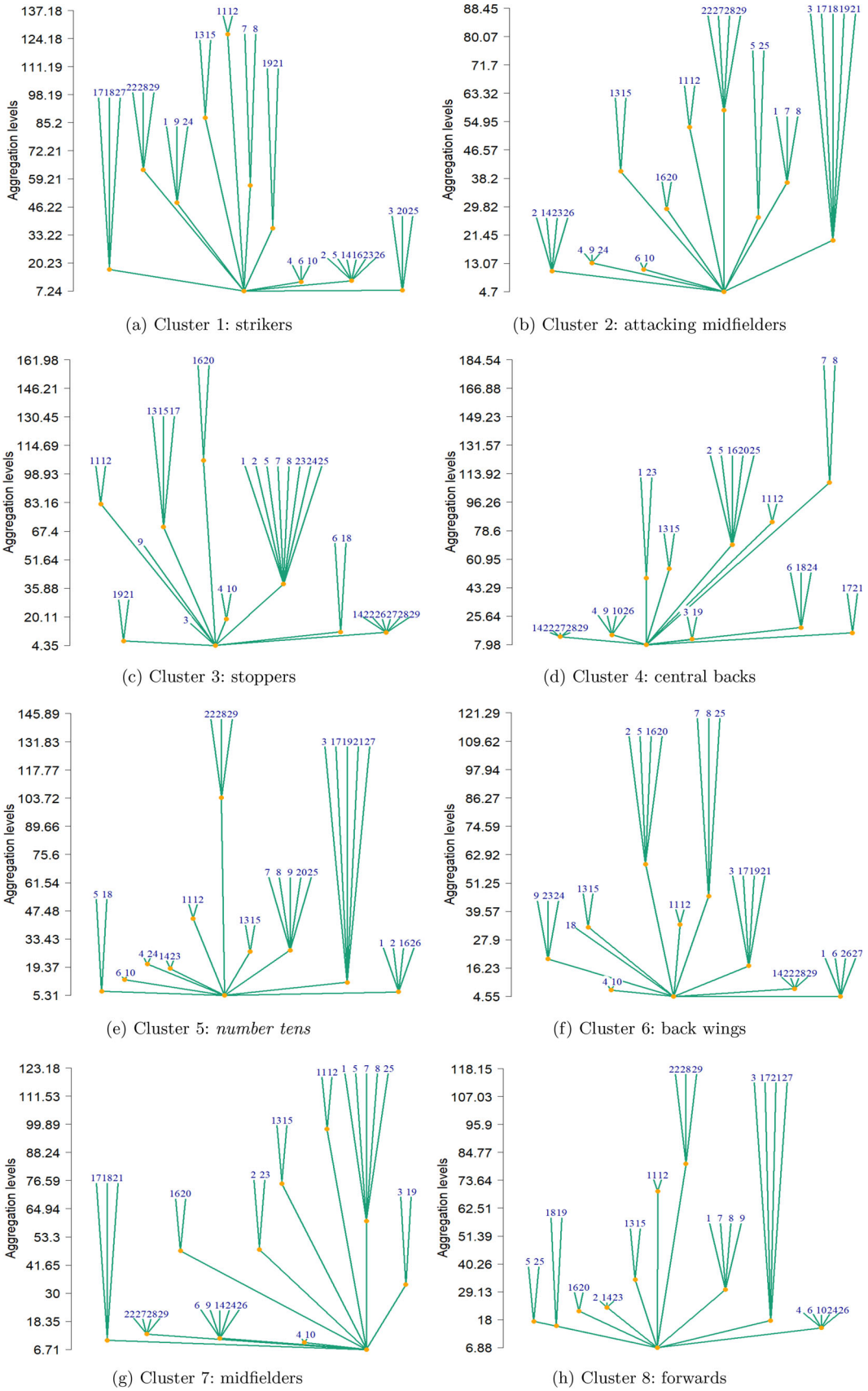


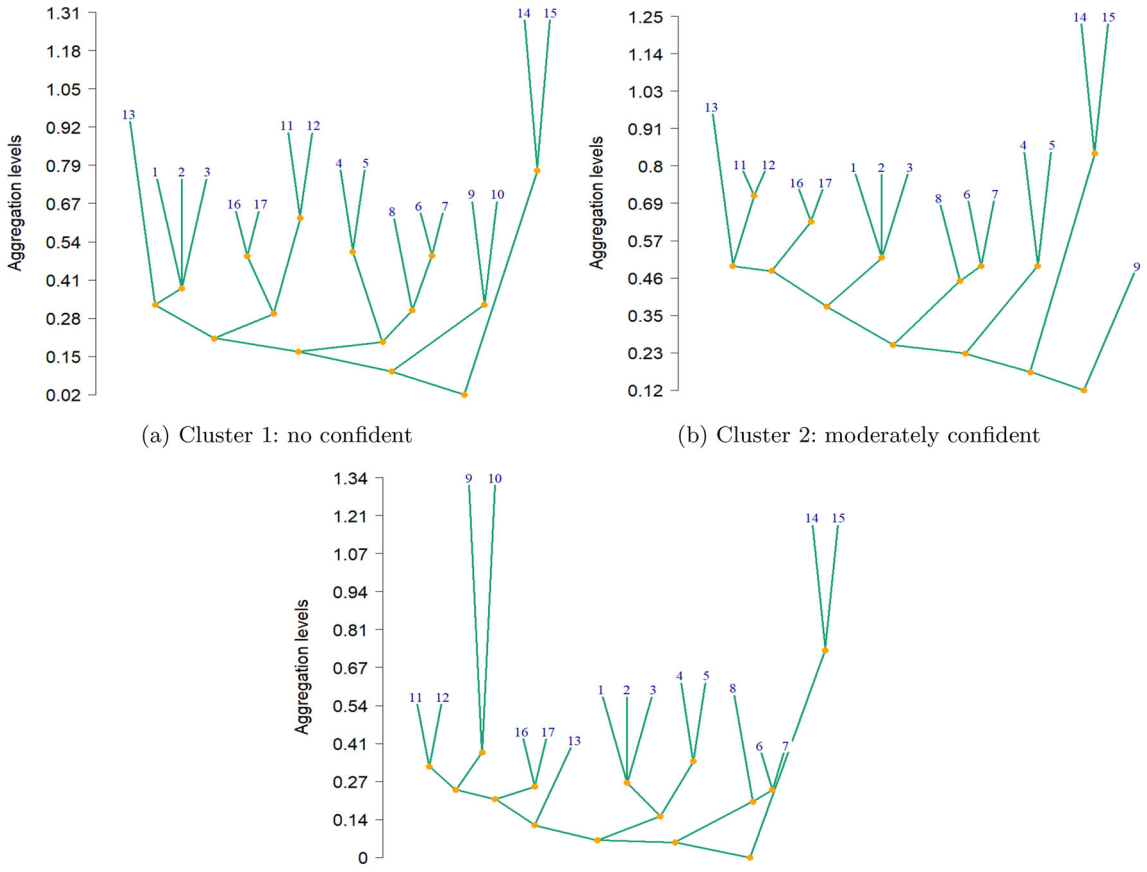
Fig. 4 Hierarchical structures of variables per cluster of FIFA players

Table 5 Variables studied for Wikipedia usage in teaching grouped with reference to the corresponding latent concept, estimated variable-group membership and their mean per cluster of Wikipedia users

Group	ID	Variable name	No conf	Moderate conf	Strong conf
Perceived usefulness					
1	1	The use of Wikipedia makes it easier for students to develop new skills	2.88	2.96	3.82
1	2	The use of Wikipedia improves students' learning	2.92	2.97	3.81
1	3	Wikipedia is useful for teaching	3.19	3.18	4.33
Perceived enjoyment					
2	4	The use of Wikipedia stimulates curiosity	3.72	3.65	4.19
2	5	The use of Wikipedia is entertaining	3.73	3.73	4.08
Quality					
3	6	Articles in Wikipedia are reliable	2.99	3.04	3.71
3	7	Articles in Wikipedia are updated	3.26	3.28	3.84
4	8	Articles in Wikipedia are comprehensive	2.81	2.86	3.40
Use behaviour					
5	9	I use Wikipedia to develop my teaching materials	2.11	1.73	3.05
5	10	I use Wikipedia as a platform to develop educational activities with students	1.50	1.73	2.55
6	11	I recommend my students to use Wikipedia	2.22	2.38	3.81
6	12	I recommend my colleagues to use Wikipedia	2.16	2.28	3.62
7	13	I agree my students use Wikipedia in my courses	3.02	3.09	4.05
Job relevance					
8	14	My university promotes the use of open collaborative environments in the Internet	3.69	3.71	3.89
8	15	My university considers the use of open collaborative environments in the Internet as a teaching merit	3.17	3.15	3.23
Behavioral intention					
9	16	In the future I will recommend the use of Wikipedia to my colleagues and students	2.71	2.71	3.87
9	17	In the future I will use Wikipedia in my teaching activity	2.73	2.71	4.04

the Universitat Oberta de Catalunya, a Spanish online university that offers bachelor degrees, master's degrees, and postgraduate courses. Since few missing values occur in the data, we impute them using the K-nearest neighbors method by setting $K = 5$, using the Euclidean distance as metric, and assuming the missing completely at random mechanism (Rubin 1976).

We run PUGMMs with both G and $m \in \{1, \ldots, 10\}$. The EFFF model returns the highest BIC value of -25226.06 with $G = 3$ and $m = 9$. The three clusters reveal varying degrees of confidence in the use of Wikipedia for educational purposes, from no or little confidence (Cluster 1) to strong confidence (Cluster 3) with moderate confidence in between (Cluster 2). Analyzing the mean vectors in Table 5, the first and second clusters appear closer, indicating a smaller dissimilarity between professors who are not confident and those who are moderately confident in the use of Wikipedia for teaching. As shown in Fig. 5, the variable configuration in groups and corresponding latent concepts remain constant across clusters, while remarkable differences occur in the aggregation of these concepts as a result of the EFFF model. It is worth noting that the variable-group membership matrix essentially identifies the six theoretical concepts, such as *perceived usefulness*, by splitting *quality* in two groups and *use behavior* into three groups. Indeed, for *quality*, the variable *Articles in Wikipedia are comprehensive* (8) is a singleton, while for *use behavior*, the variables *I use Wikipedia to develop my teaching materials* (9) and *I use Wikipedia as a platform to develop educational activities with students* (10) are lumped together in a group representing the core variables of this latent concept. Additionally, the variables *I recommend my students to use Wikipedia* (11) and *I recommend my colleagues to use Wikipedia* (12) form a separate group representing *recommendation for the use of Wikipedia*, with the remaining variable (13) as a singleton.



(a) Cluster 1: no confident

(b) Cluster 2: moderately confident

(c) Cluster 3: strongly confident

Fig. 5 Hierarchical structures of variables per cluster of Wikipedia users. In Fig. 5b, variable 9 represents both itself and variable 10 since $v\sigma_{55(2)} = w\sigma_{55(2)}$

Looking at Fig. 5, we can notice several differences in the aggregations of the variable groups. For example, in Fig. 5a, the first aggregation concerns *perceived usefulness* with the variable *I agree my students use Wikipedia in my courses*. This indicates that professors who have less confidence in Wikipedia pay greater attention to its employment in higher education courses by disagreeing with it. In Cluster 2 (Fig. 5b), the model first combines the groups representing *recommendation for the use of Wikipedia* and the singleton composed of *I agree my students use Wikipedia in my courses*, which all refer to *use behavior*, then this broader group with the one indicating *behavioral intention*. Therefore, professors who have moderate confidence in Wikipedia do not necessarily recommend its use to students and colleagues, although they themselves employ this tool for preparing teaching material. This could be attributed to a caution in the use of Wikipedia when recommended to others, which brings Cluster 1 closer to Cluster 2. In Cluster 3, the first aggregation involves the variables related to *recommendation for the use of Wikipedia* with the core variables of *use behavior* and then *behavioral intention*, as shown in Fig. 5c, by highlighting an opposite approach to the no confident professors.

Overall, we can observe that, after some aggregations, two broader groups with a low level of covariance are identified for the clusters of no and strongly confident professors. This evidence demonstrates the presence of "higher-order" low-correlated dimensions at a certain level of the hierarchy that distinctively and uniquely contribute to defining confidence in the use of Wikipedia. On the contrary, for Cluster 2 the aggregation seems to be smoother by pinpointing a unique, internally consistent dimension representing confidence in the use of Wikipedia.

6 Conclusions

We have introduced a new class of parsimonious ultrametric GMMs with the aim of inspecting hierarchical relationships among variables while further reducing the number of parameters compared to the existing GMMs in the literature. The

proposed thirteen models are obtained by constraining the covariance structure, which is extended ultrametric, to be equal within and/or across components. We have also proposed computational improvements for the estimation of the extended ultrametric covariance structure. Specifically, we have used its canonical representation based upon the result of Archakov and Hansen (2020) to obtain faster computation of its determinant and inverse. Moreover, we have enhanced the strategy chosen by Cavicchia et al. (2022) to guarantee the positive definiteness of the extended ultrametric covariance matrix by considering a procedure based on its polar decomposition (Higham 1986), that turns into the nearest solution in the 2-norm.

We have evaluated the performance of PUGMMs on both benchmark and real-world data sets. In the case of benchmark data sets, we have compared our proposal to GPCMs, PGMMs and HDDC. Our results showed comparable cluster recovery performance with a significantly reduced number of parameters, often substantially lower. Moreover, we have inspected the features of the FIFA football players to identify clusters related to their roles and variable groups associated with different kind of skills. We have obtained interesting outcomes also on the second application where we have studied the use of Wikipedia for teaching purposes within universities. We have discovered three clusters of professors with varying levels of confidence in web-based open collaborative environments. Additionally, we have identified a hierarchical structure of latent concepts, where the first nine remained constant across all clusters. Therefore, these applications illustrate the potential of PUGMMs in effectively recovering sub-populations within data sets. Furthermore, they highlight the proposal's ability to detect hierarchical latent concepts. These results are achieved by employing constrained covariance structures, wherein the number of parameters scales linearly with both the data dimension and the number of variable groups.

All the functions to estimate PUGMMs are implemented in the R package `PUGMM`, which is available on https://github.com/giorgiazaccaria/PUGMM and will be soon released on CRAN. A further extension of the PUGMM models to cases where the variable groups are assumed to be uncorrelated, i.e., $\boldsymbol{\Sigma}_{B_g}$ is sparse or fully equal to zero, will be studied in a separate future work to investigate their specific properties.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11222-024-10405-9.

Author contributions C.C., M.V. and G.Z. all have made substantial contributions to the conception and design of the work. C.C. and G.Z. have drafted the paper and developed the R codes for PUGMMs implementation. C.C., M.V. and G.Z. have revised the paper.

Funding Open access funding provided by Università degli Studi di Milano - Bicocca within the CRUI-CARE Agreement. The authors received no financial support for the research, the authorship and/or the publication of this article.

Data availability The data that support the findings of this study are available on R packages and UCI repository as detailed in the article.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Code availability The source code of PUGMMs for data analyses is openly available at https://github.com/giorgiazaccaria/PUGMM. The R scripts for the experiments reported in Sect. 5 and in the Supplementary Materials are available at https://github.com/giorgiazaccaria/PUGMM-Paper-Experiments.

Appendix A

The ML estimates of the PUGMMs covariance parameters are presented in the following sections, considering the two identified families of models: the *unique and equal models* and the *isotropic and free models*.

A.1 Unique and equal models

For unique and equal models, the ML estimates of the covariance parameters are obtained by differentiating ℓ in Eq. (4) with respect to each parameter separately, i.e., $\boldsymbol{\Sigma}_D$ where $D = \{V, W, B\}$. Applying the results of Lütkepohl (1996), we obtain

$$\begin{aligned}\frac{\partial \ell}{\partial \boldsymbol{\Sigma}_D} &= -\frac{n}{2}\left[\frac{\partial \log(|\boldsymbol{\Sigma}|)}{\partial \boldsymbol{\Sigma}_D} + \frac{\partial \operatorname{tr}(\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{S}})}{\partial \boldsymbol{\Sigma}_D}\right] \\ &= -\frac{n}{2}\left[\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \boldsymbol{\Sigma}_D} - \boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \boldsymbol{\Sigma}_D}\bar{\boldsymbol{S}}\boldsymbol{\Sigma}^{-1}\right]. \end{aligned} \tag{8}$$

Setting to zero the partial derivative of ℓ in Eq. (8) with respect to $\boldsymbol{\Sigma}_D$, $D = \{V, W, B\}$ one-at-a-time given the oth-

ers, leads to

$$\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \boldsymbol{\Sigma}_D} - \boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \boldsymbol{\Sigma}_D}\bar{\boldsymbol{S}}\boldsymbol{\Sigma}^{-1} = 0,$$

that holds if and only if $\bar{\boldsymbol{S}}\,\boldsymbol{\Sigma}^{-1} = \boldsymbol{I}_p$. This means that $\bar{\boldsymbol{S}} = \boldsymbol{\Sigma}$ where $\boldsymbol{\Sigma}^{-1}$ is nonsingular,[2] which is the starting point to compute the ML estimates of the unique and equal model parameters. The same result is easily obtained when $\boldsymbol{\Sigma}_D$ depends on a scalar (unique models), since the trace operator occurs in Eq. (8).

A.1.1. EUUU model

For the EUUU model, $\boldsymbol{\Sigma}_V = \sigma_V \boldsymbol{I}_m$, $\boldsymbol{\Sigma}_W = \sigma_W \boldsymbol{I}_m$ and $\boldsymbol{\Sigma}_B = \sigma_B(\mathbf{1}_m\mathbf{1}_m' - \boldsymbol{I}_m)$. Then the covariance structure is

$$\boldsymbol{\Sigma} = \boldsymbol{V}\big(\sigma_W \boldsymbol{I}_m + \sigma_B(\mathbf{1}_m\mathbf{1}_m' - \boldsymbol{I}_m)\big)\boldsymbol{V}' + \operatorname{diag}\big(\boldsymbol{V}(\sigma_V - \sigma_W)\boldsymbol{I}_m\boldsymbol{V}'\big).$$

Given $\widehat{\boldsymbol{V}}$ and pre- and post-multiplying $\bar{\boldsymbol{S}} = \boldsymbol{\Sigma}$ for matrices that depend on $\widehat{\boldsymbol{V}}$, we obtain the ML estimate of σ_V by solving

$$\operatorname{diag}(\widehat{\boldsymbol{V}}\sigma_V \boldsymbol{I}_m \widehat{\boldsymbol{V}}') = \operatorname{diag}(\bar{\boldsymbol{S}}),$$

that returns

$$\widehat{\sigma}_V = \frac{\operatorname{tr}\big(\widehat{\boldsymbol{V}}^{+}\operatorname{diag}(\bar{\boldsymbol{S}})\widehat{\boldsymbol{V}}\big)}{m}. \tag{9}$$

Note that $\widehat{\boldsymbol{V}}^{+}$ represents the Moore–Penrose inverse of the matrix $\widehat{\boldsymbol{V}}$.

With the same reasoning, given $\widehat{\boldsymbol{V}}$ and $\widehat{\sigma}_V$, the ML estimate of σ_W is derived by solving the following problem

$$\widehat{\boldsymbol{V}}\sigma_W \boldsymbol{I}_m \widehat{\boldsymbol{V}}' - \operatorname{diag}(\widehat{\boldsymbol{V}}\sigma_W \boldsymbol{I}_m \widehat{\boldsymbol{V}}') = \bar{\boldsymbol{S}} - \widehat{\sigma}_V \boldsymbol{I}_p - \widehat{\boldsymbol{V}}\widehat{\sigma}_B\big(\mathbf{1}_m\mathbf{1}_m' - \boldsymbol{I}_m\big)\widehat{\boldsymbol{V}}',$$

that results in

$$\widehat{\sigma}_W = \frac{\operatorname{tr}\Big(\widehat{\boldsymbol{V}}'\big(\bar{\boldsymbol{S}} - \widehat{\sigma}_V \boldsymbol{I}_p\big)\widehat{\boldsymbol{V}}\big((\widehat{\boldsymbol{V}}'\widehat{\boldsymbol{V}})^2 - \widehat{\boldsymbol{V}}'\widehat{\boldsymbol{V}}\big)^{-1}\Big)}{m}. \tag{10}$$

Finally, the ML estimate of σ_B given $\widehat{\boldsymbol{V}}$ is

$$\widehat{\sigma}_B = \frac{\operatorname{tr}\Big(\big(\widehat{\boldsymbol{V}}^{+}\bar{\boldsymbol{S}}\widehat{\boldsymbol{V}}'^{+} \odot (\mathbf{1}_m\mathbf{1}_m' - \boldsymbol{I}_m)\big)(\mathbf{1}_m\mathbf{1}_m' - \boldsymbol{I}_m)^{-1}\Big)}{m} \tag{11}$$

which is derived from

$$\widehat{\boldsymbol{V}}\sigma_B(\mathbf{1}_m\mathbf{1}_m' - \boldsymbol{I}_m)\widehat{\boldsymbol{V}}' = (\bar{\boldsymbol{S}} - \widehat{\sigma}_V \boldsymbol{I}_p) \odot (\mathbf{1}_p\mathbf{1}_p' - \boldsymbol{I}_p).$$

[2] Since $\boldsymbol{\Sigma}$ is positive definite by definition, $\boldsymbol{\Sigma}^{-1}$ is positive definite in turn (Lütkepohl 1996, p. 152, property 7c) and thus nonsingular.

A.1.2 EUUE model

In the EUUE model, $\boldsymbol{\Sigma}_V$ and $\boldsymbol{\Sigma}_W$ are constrained to have a unique value each on their diagonal, whereas $\boldsymbol{\Sigma}_B$ has at most $m-1$ different values. Hence the covariance structure is

$$\boldsymbol{\Sigma} = \boldsymbol{V}(\sigma_W \boldsymbol{I}_m + \boldsymbol{\Sigma}_B)\boldsymbol{V}' + \operatorname{diag}(\boldsymbol{V}(\sigma_V - \sigma_W)\boldsymbol{I}_m\boldsymbol{V}').$$

It is easy to prove that the ML estimates of σ_V and σ_W equal those in Eqs. (9) and (10), respectively. The ML estimate of $\boldsymbol{\Sigma}_B$ is

$$\widetilde{\boldsymbol{\Sigma}}_B = \widehat{\boldsymbol{V}}^{+}\bar{\boldsymbol{S}}(\widehat{\boldsymbol{V}}')^{+} \odot (\mathbf{1}_m\mathbf{1}_m' - \boldsymbol{I}_m), \tag{12}$$

where the Hadamard product derives from the constraint that $\operatorname{diag}(\boldsymbol{\Sigma}_B) = \mathbf{0}$. If Eq. (12) does not satisfy the ultrametricity condition, $\widehat{\boldsymbol{\Sigma}}_B$ is obtained from $\widetilde{\boldsymbol{\Sigma}}_B$ by applying an adapted UPGMA algorithm for covariances to it (see Cavicchia et al. 2022, for further details).

A.1.3 EUEE model

In the EUEE model, $\boldsymbol{\Sigma}_V$ is the only parameter constrained within the component, i.e., $\boldsymbol{\Sigma}_V = \sigma_V \boldsymbol{I}_m$, and hence the covariance structure is

$$\boldsymbol{\Sigma} = \boldsymbol{V}(\boldsymbol{\Sigma}_W + \boldsymbol{\Sigma}_B)\boldsymbol{V}' + \operatorname{diag}(\boldsymbol{V}(\sigma_V \boldsymbol{I}_m - \boldsymbol{\Sigma}_W)\boldsymbol{V}').$$

It is easy to prove that the ML estimates of σ_V and $\boldsymbol{\Sigma}_B$ equal those in Eqs. (9) and (12), respectively. The ML estimate of $\boldsymbol{\Sigma}_W$ is obtained as follows

$$\widehat{\boldsymbol{\Sigma}}_W = \operatorname{diag}\big(\widehat{\boldsymbol{V}}'(\bar{\boldsymbol{S}} - \widehat{\sigma}_V \boldsymbol{I}_p)\widehat{\boldsymbol{V}}\big)\big((\widehat{\boldsymbol{V}}'\widehat{\boldsymbol{V}})^2 - \widehat{\boldsymbol{V}}'\widehat{\boldsymbol{V}}\big)^{-1}. \tag{13}$$

A.1.4 EEEU model

The EEEU model constrains $\boldsymbol{\Sigma}_B = \sigma_B(\mathbf{1}_m\mathbf{1}_m' - \boldsymbol{I}_m)$, whereas $\boldsymbol{\Sigma}_V$ and $\boldsymbol{\Sigma}_W$ can have m different diagonal values. Therefore, the covariance structure is

$$\boldsymbol{\Sigma} = \boldsymbol{V}(\boldsymbol{\Sigma}_W + \sigma_B(\mathbf{1}_m\mathbf{1}_m' - \boldsymbol{I}_m))\boldsymbol{V}' + \operatorname{diag}(\boldsymbol{V}(\boldsymbol{\Sigma}_V - \boldsymbol{\Sigma}_W)\boldsymbol{V}').$$

The ML estimate of $\boldsymbol{\Sigma}_V$ is obtained as follows

$$\widehat{\boldsymbol{\Sigma}}_V = \widehat{\boldsymbol{V}}^{+}\operatorname{diag}(\bar{\boldsymbol{S}})\widehat{\boldsymbol{V}}, \tag{14}$$

whereas that of $\boldsymbol{\Sigma}_W$ can be easily derived from Eq. (13) by substituting $\widehat{\sigma}_V \boldsymbol{I}_p$ for $\operatorname{diag}(\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Sigma}}_V\widehat{\boldsymbol{V}}')$, as in Eq. (17). The unique off-diagonal value of $\boldsymbol{\Sigma}_B$ is estimated as in Eq. (11).

A.1.5 EEEE model

The covariance structure of the EEEE model is

$$\boldsymbol{\Sigma} = \boldsymbol{V}(\boldsymbol{\Sigma}_W + \boldsymbol{\Sigma}_B)\boldsymbol{V}' + \text{diag}(\boldsymbol{V}(\boldsymbol{\Sigma}_V - \boldsymbol{\Sigma}_W)\boldsymbol{V}')$$

and the ML estimates of $\boldsymbol{\Sigma}_V$ and $\boldsymbol{\Sigma}_W$ are equal to those presented in the EEEU model, whereas $\boldsymbol{\Sigma}_B$ is estimated as in Eq. (12).

A.2 Isotropic and free models

For the isotropic and free models, the ML estimates of the covariance parameters are obtained by differentiating ℓ in Eq. (5) with respect to each parameter separately. For the EEEF and EEFF models, the partial derivative of ℓ with respect to $\boldsymbol{\Sigma}_D$ with $D = \{V, W\}$, i.e., the parameters that do not vary between components in these cases, is given by

$$\begin{aligned}\frac{\partial \ell}{\partial \boldsymbol{\Sigma}_D} &= -\frac{1}{2}\sum_{g=1}^{G} n_g \left[\frac{\partial \log(|\boldsymbol{\Sigma}_g|)}{\partial \boldsymbol{\Sigma}_D} + \frac{\partial \,\text{tr}(\boldsymbol{\Sigma}_g^{-1}\boldsymbol{S}_g)}{\partial \boldsymbol{\Sigma}_D}\right] \\ &= -\frac{1}{2}\sum_{g=1}^{G} n_g \left[\boldsymbol{\Sigma}_g^{-1}\frac{\partial \boldsymbol{\Sigma}_g}{\partial \boldsymbol{\Sigma}_D} - \boldsymbol{\Sigma}_g^{-1}\frac{\partial \boldsymbol{\Sigma}_g}{\partial \boldsymbol{\Sigma}_D}\boldsymbol{S}_g\boldsymbol{\Sigma}_g^{-1}\right].\end{aligned} \tag{15}$$

If we set the partial derivative of ℓ to zero in Eq. (15) with respect to $\boldsymbol{\Sigma}_D$, $D = \{V, W\}$ one-at-a-time given the others, we obtain

$$\sum_{g=1}^{G} n_g \left[\boldsymbol{\Sigma}_g^{-1}\frac{\partial \boldsymbol{\Sigma}_g}{\partial \boldsymbol{\Sigma}_D} - \boldsymbol{\Sigma}_g^{-1}\frac{\partial \boldsymbol{\Sigma}_g}{\partial \boldsymbol{\Sigma}_D}\boldsymbol{S}_g\boldsymbol{\Sigma}_g^{-1}\right] = 0.$$

This holds if and only if $\sum_{g=1}^{G} n_g \boldsymbol{S}_g = \sum_{g=1}^{G} n_g \boldsymbol{\Sigma}_g$ since $\boldsymbol{\Sigma}_g$ is nonsingular. From this equivalence, it is possible to derive the ML estimates of $\boldsymbol{\Sigma}_D$, $D = \{V, W\}$, for the EEEF and EEFF cases.

For the other models and parameters of this family, the partial derivative of ℓ in Eq. (5) with respect to $\boldsymbol{\Sigma}_{D_g}$, $D_g = \{V_g, W_g, B_g\}$, is

$$\begin{aligned}\frac{\partial \ell}{\partial \boldsymbol{\Sigma}_{D_g}} &= -\frac{n_g}{2}\left[\frac{\partial \log(|\boldsymbol{\Sigma}_g|)}{\partial \boldsymbol{\Sigma}_{D_g}} + \frac{\partial \,\text{tr}(\boldsymbol{\Sigma}_g^{-1}\boldsymbol{S}_g)}{\partial \boldsymbol{\Sigma}_{D_g}}\right] \\ &= -\frac{n_g}{2}\left[\boldsymbol{\Sigma}_g^{-1}\frac{\partial \boldsymbol{\Sigma}_g}{\partial \boldsymbol{\Sigma}_{D_g}} - \boldsymbol{\Sigma}_g^{-1}\frac{\partial \boldsymbol{\Sigma}_g}{\partial \boldsymbol{\Sigma}_{D_g}}\boldsymbol{S}_g\boldsymbol{\Sigma}_g^{-1}\right],\end{aligned} \tag{16}$$

whose first-order condition is $\boldsymbol{S}_g = \boldsymbol{\Sigma}_g$. The same result is easily obtained when $\boldsymbol{\Sigma}_{D_g}$ is restricted to have a single value, since the trace operator occurs in Eq. (16).

A.2.1 EEEF model

In the EEEF model, $\boldsymbol{\Sigma}_{V_g} = \boldsymbol{\Sigma}_V$ and $\boldsymbol{\Sigma}_{W_g} = \boldsymbol{\Sigma}_W$, whereas the between-group covariance matrix is free to vary across components. Thus, the covariance structure is

$$\boldsymbol{\Sigma}_g = \boldsymbol{V}(\boldsymbol{\Sigma}_W + \boldsymbol{\Sigma}_{B_g})\boldsymbol{V}' + \text{diag}(\boldsymbol{V}(\boldsymbol{\Sigma}_V - \boldsymbol{\Sigma}_W)\boldsymbol{V}').$$

It is easy to prove that the ML estimate of $\boldsymbol{\Sigma}_V$ is consistent with Eq. (14).

The ML estimate of $\boldsymbol{\Sigma}_W$ is obtained by solving the following equation

$$\begin{aligned}&\widehat{\boldsymbol{V}}\boldsymbol{\Sigma}_W\widehat{\boldsymbol{V}}' - \text{diag}(\widehat{\boldsymbol{V}}\boldsymbol{\Sigma}_W\widehat{\boldsymbol{V}}') \\ &\quad = \bar{\boldsymbol{S}} - \text{diag}(\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Sigma}}_V\widehat{\boldsymbol{V}}') - \widehat{\boldsymbol{V}}\left(\sum_{g=1}^{G}\hat{\pi}_g\widehat{\boldsymbol{\Sigma}}_{Bg}\right)\widehat{\boldsymbol{V}}',\end{aligned}$$

that results in

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_W = \text{diag}\Big(\widehat{\boldsymbol{V}}'\big(\bar{\boldsymbol{S}} - \text{diag}(\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Sigma}}_V\widehat{\boldsymbol{V}}')\big)\widehat{\boldsymbol{V}}\Big) \\ \times \big((\widehat{\boldsymbol{V}}'\widehat{\boldsymbol{V}})^2 - \widehat{\boldsymbol{V}}'\widehat{\boldsymbol{V}}\big)^{-1}.\end{aligned} \tag{17}$$

Finally, the ML estimate of $\boldsymbol{\Sigma}_{B_g}$ is

$$\widetilde{\boldsymbol{\Sigma}}_{Bg} = \widehat{\boldsymbol{V}}^{+}\boldsymbol{S}_g(\widehat{\boldsymbol{V}}')^{+} \odot (\boldsymbol{1}_m\boldsymbol{1}_m' - \boldsymbol{I}_m), \tag{18}$$

where the Hadamard product derives from the constraint that $\text{diag}(\boldsymbol{\Sigma}_{B_g}) = \boldsymbol{0}$. If Eq. (18) does not satisfy the ultrametricity condition, $\widehat{\boldsymbol{\Sigma}}_{Bg}$ is obtained from $\widetilde{\boldsymbol{\Sigma}}_{Bg}$ by applying an adapted UPGMA algorithm for covariances to it.

A.2.2 EEFF model

In the EEFF model, $\boldsymbol{\Sigma}_V$ is the only parameter constrained across components and, therefore, the covariance structure is

$$\boldsymbol{\Sigma}_g = \boldsymbol{V}(\boldsymbol{\Sigma}_{W_g} + \boldsymbol{\Sigma}_{B_g})\boldsymbol{V}' + \text{diag}(\boldsymbol{V}(\boldsymbol{\Sigma}_V - \boldsymbol{\Sigma}_{W_g})\boldsymbol{V}').$$

It is easy to prove that the ML estimates of $\boldsymbol{\Sigma}_V$ and $\boldsymbol{\Sigma}_{B_g}$ correspond to Eqs. (14) and (18), respectively, while the ML estimate of $\boldsymbol{\Sigma}_{W_g}$ is

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{Wg} = \text{diag}\Big(\widehat{\boldsymbol{V}}'\big(\boldsymbol{S}_g - \text{diag}(\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Sigma}}_V\widehat{\boldsymbol{V}}')\big)\widehat{\boldsymbol{V}}\Big) \\ \times \big((\widehat{\boldsymbol{V}}'\widehat{\boldsymbol{V}})^2 - \widehat{\boldsymbol{V}}'\widehat{\boldsymbol{V}}\big)^{-1}.\end{aligned} \tag{19}$$

A.2.3 EFFF model

The covariance structure of the EFFF model is

$$\boldsymbol{\Sigma}_g = \boldsymbol{V}(\boldsymbol{\Sigma}_{W_g} + \boldsymbol{\Sigma}_{B_g})\boldsymbol{V}' + \text{diag}(\boldsymbol{V}(\boldsymbol{\Sigma}_{V_g} - \boldsymbol{\Sigma}_{W_g})\boldsymbol{V}'),$$

where the component covariance matrices share only the variable-group membership matrix parameter. In this case, the ML estimates of $\boldsymbol{\Sigma}_{W_g}$ and $\boldsymbol{\Sigma}_{B_g}$ equal those in Eqs. (19) and (18), respectively, while the ML estimate of $\boldsymbol{\Sigma}_{V_g}$ is

$$\widehat{\boldsymbol{\Sigma}}_{Vg} = \widehat{\boldsymbol{V}}^{+}\text{diag}(\boldsymbol{S}_g)\widehat{\boldsymbol{V}}. \tag{20}$$

A.2.4 FIII model

In the FIII model, $\boldsymbol{\Sigma}_{V_g} = \sigma_{V_g}\boldsymbol{I}_m$, $\boldsymbol{\Sigma}_{W_g} = \sigma_{W_g}\boldsymbol{I}_m$, $\boldsymbol{\Sigma}_{B_g} = \sigma_{B_g}(\mathbf{1}_m\mathbf{1}_m' - \boldsymbol{I}_m)$, and the covariance structure is

$$\begin{aligned}\boldsymbol{\Sigma}_g = {} & \boldsymbol{V}_g\big(\sigma_{W_g}\boldsymbol{I}_m + \sigma_{B_g}(\mathbf{1}_m\mathbf{1}_m' - \boldsymbol{I}_m)\big)\boldsymbol{V}_g' \\ & + \text{diag}\big(\boldsymbol{V}_g(\sigma_{V_g} - \sigma_{W_g})\boldsymbol{I}_m\boldsymbol{V}_g'\big).\end{aligned}$$

The ML estimates of σ_{V_g}, σ_{W_g} and σ_{B_g} are easily derivable; their expression corresponds to Eqs. (9), (10) and (11), respectively, where all the involved parameters depend on g and $\bar{\boldsymbol{S}}$ is replaced by $\boldsymbol{S}_g$.

A.2.5 FIIF model

In the FIIF model, only $\boldsymbol{\Sigma}_{V_g}$ and $\boldsymbol{\Sigma}_{W_g}$ are constrained to have a single value each on their diagonal, even if these values differ across components. The covariance structure is

$$\begin{aligned}\boldsymbol{\Sigma}_g = {} & \boldsymbol{V}_g(\sigma_{W_g}\boldsymbol{I}_m + \boldsymbol{\Sigma}_{B_g})\boldsymbol{V}_g' \\ & + \text{diag}(\boldsymbol{V}_g(\sigma_{V_g} - \sigma_{W_g})\boldsymbol{I}_m\boldsymbol{V}_g').\end{aligned}$$

For the ML estimates of σ_{V_g} and σ_{W_g}, what is reported for the FIII model holds, whereas the ML estimate of $\boldsymbol{\Sigma}_{B_g}$ equals Eq. (18), with $\widehat{\boldsymbol{V}}$ replaced by $\widehat{\boldsymbol{V}}_g$.

A.2.6 FIFF model

In the FIFF model, $\boldsymbol{\Sigma}_{V_g}$ is the only parameter restricted to having a single value on its diagonal that varies between components. The covariance structure is

$$\begin{aligned}\boldsymbol{\Sigma}_g = {} & \boldsymbol{V}_g(\boldsymbol{\Sigma}_{W_g} + \boldsymbol{\Sigma}_{B_g})\boldsymbol{V}_g' \\ & + \text{diag}(\boldsymbol{V}_g(\sigma_{V_g}\boldsymbol{I}_m - \boldsymbol{\Sigma}_{W_g})\boldsymbol{V}_g').\end{aligned}$$

The considerations reported for the FIII and FIIF models remain valid for the ML estimates of σ_{V_g} and $\boldsymbol{\Sigma}_{B_g}$. The ML estimate of $\boldsymbol{\Sigma}_{W_g}$ corresponds to Eq. (13), where all the involved parameters depend on g and $\bar{\boldsymbol{S}}$ is replaced by $\boldsymbol{S}_g$.

A.2.7 FFFI model

The FFFI model constrains $\boldsymbol{\Sigma}_{B_g} = \sigma_{B_g}(\mathbf{1}_m\mathbf{1}_m' - \boldsymbol{I}_m)$, whereas $\boldsymbol{\Sigma}_{V_g}$ and $\boldsymbol{\Sigma}_{W_g}$ vary within and across components. The covariance structure is

$$\begin{aligned}\boldsymbol{\Sigma}_g = {} & \boldsymbol{V}_g(\boldsymbol{\Sigma}_{W_g} + \sigma_{B_g}(\mathbf{1}_m\mathbf{1}_m' - \boldsymbol{I}_m))\boldsymbol{V}_g' \\ & + \text{diag}(\boldsymbol{V}_g(\boldsymbol{\Sigma}_{V_g} - \boldsymbol{\Sigma}_{W_g})\boldsymbol{V}_g').\end{aligned}$$

The ML estimates of $\boldsymbol{\Sigma}_{V_g}$ and $\boldsymbol{\Sigma}_{W_g}$ correspond to Eqs. (20) and (19), respectively, where all the parameters depend on g. The unique off-diagonal value of $\boldsymbol{\Sigma}_{B_g}$ is estimated as the component-specific counterpart of Eq. (11), where $\bar{\boldsymbol{S}}$ is replaced by $\boldsymbol{S}_g$ and $\widehat{\boldsymbol{V}}$ by $\widehat{\boldsymbol{V}}_g$.

A.2.8 FFFF model

The FFFF case is the most general model, whose covariance structure is described in Sect. 2. The ML estimates of $\boldsymbol{\Sigma}_{V_g}$, $\boldsymbol{\Sigma}_{W_g}$ and $\boldsymbol{\Sigma}_{B_g}$ correspond to Eqs. (20), (19) and (18), respectively, where all the parameters involved depend on g.

Appendix B

In Table 6, we provide the *maximum* number of constraints per model that can be activated in the PUGMMs algorithm. This is a useful information for the user since it is needed for the computation of the number of free parameters in Sect. 4.4. We recall that $c_{V,W}$ concerns the relationship between $\boldsymbol{\Sigma}_{V_g}$ and $\boldsymbol{\Sigma}_{W_g}$ (constraint iii in Sect. 2), whereas $c_{W,B}$ does that between $\boldsymbol{\Sigma}_{W_g}$ and $\boldsymbol{\Sigma}_{B_g}$ (constraint ii in Sect. 2).

Table 6 Maximum number of constraints activated in the PUGMMs algorithm

Model ID	$c_{V,W}$	$c_{W,B}$
$\boldsymbol{V}$ Equal		
EUUU	1	1
EUUE	1	1
EUEE	1	m
EEEU	m	m
EEEE	m	m
EEEF	m	m
EEFF	m	Gm
EFFF	Gm	Gm
$\boldsymbol{V}$ Free		
FIII	G	G
FIIF	G	G
FIFF	G	Gm
FFFI	Gm	Gm
FFFF	Gm	Gm

Table 7 Benchmark data sets source

Name	Original name	Source
Penguin	Penguin dataset	Kaggle
Wine 13	Wine	R-`HDclassif`
Wine 27	Wine	R-`pgmm`
Thyroid	R-`mclust`	
Kidney	ckd	R-`teigen`
Economics	Economics	R-`datasetsICR`
Tetragonula	Tetragonula	(Hennig 2022, Supplementary File 2)
Diabetes	Diabetes	R-`mclust`
Ais	ais	R-`sn`
Ceramic	Chemical Composition of Ceramic Samples	UCI repository
Banknotes	Banknote	R-`mclust`
Coffee	Coffee	R-`pgmm`
Sobar	Cervical Cancer Behavior Risk	UCI repository

Appendix C

Table 7 provides information on the source of the benchmark data sets used for the analysis in Sect. 5.1.

References

Aibar, E., Llads, J., Meseguer, A., Minguilln, J., Lerga, M.: wiki4HE (2015). https://doi.org/10.24432/C50031

Alkire, S.: Human development: definitions, critiques, and related concepts. In: Human Development Research Papers (2009 to present), Human Development Report Office (HDRO), United Nations Development Programme (UNDP) (2010)

Alkire, S., Foster, J.: Counting and multidimensional poverty measurement. J. Public Econ. **95**(7), 476–487 (2011)

Alkire, S., Foster, J., Seth, S., Santos, M., Roche, J., Ballon, P.: Multidimensional Poverty Measurement and Analysis. Oxford University Press, Oxford (2015)

Archakov, I., Hansen, P.: A canonical representation of block matrices with applications to covariance and correlation matrices (2020). arXiv:2012.02698

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

Bensmail, H., Celeux, G.: Regularized Gaussian discriminant analysis through eigenvalue decomposition. J. Am. Stat. Assoc. **91**(436), 1743–1748 (1996)

Bergé, L., Bouveyron, C., Girard, S.: HDclassif: an R package for model-based clustering and discriminant analysis of high-dimensional data. J. Stat. Softw. **46**(6), 1–29 (2012)

Bezdek, J.: Cluster validity with fuzzy set. J. Cybern. **3**(3), 58–73 (1974)

Bezdek, J.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum, New York (1981)

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., Lindsay, B.: The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. Ann. Inst. Stat. Math. **46**(2), 373–388 (1994)

Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans. Pattern Anal. Mach. Intell. **22**(7), 719–725 (2000)

Bouveyron, C., Girard, S., Schmid, C.: High-dimensional data clustering. Comput. Stat. Data Anal. **52**(1), 502–519 (2007)

Bouveyron, C., Celeux, G., Murphy, T., Raftery, A.: Model-Based Clustering and Classification for Data Science: With Applications in R. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge (2019)

Cailliez, F.: The analytical solution of the additive constant problem. Psychometrika **48**(2), 305–308 (1983)

Cavicchia, C., Vichi, M., Zaccaria, G.: The ultrametric correlation matrix for modelling hierarchical latent concepts. Adv. Data Anal. Classif. **14**(4), 837–853 (2020)

Cavicchia, C., Vichi, M., Zaccaria, G.: Gaussian mixture model with an extended ultrametric covariance structure. Adv. Data Anal. Classif. **16**(2), 399–427 (2022)

Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern Recogn. **28**, 781–793 (1995)

Celeux, G., Frühwirth-Schnatter, S., Robert, C.: Model selection for mixture model—perspectives and strategies. In: Fruhwirth-Schnatter, S., Celeux, G., Robert, C. (eds) Handbook of Mixture Analysis, Chapter 7. Chapman and Hall/CRC (2018)

Dellacherie, C., Martinez, S., Martin, J. S.: Inverse M-matrices and ultrametric matrices. In: Lecture Notes in Mathematics. Springer International Publishing (2014)

Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Stat. Methodol. **39**(1), 1–38 (1977)

Fraley, C., Raftery, A.: How many clusters? Which clustering method? Answers via model-based cluster analysis, and density estimation. Comput. J. **41**(8), 578–588 (1998)

Fraley, C., Raftery, A.: MCLUST: software for model-based cluster analysis. J. Classif. **16**(2), 297–306 (1999)

Fraley, C., Raftery, A.: Model-based clustering, discriminant analysis, and density estimation. J. Am. Stat. Assoc. **97**(458), 611–631 (2002)

Ghahramani, Z., Hinton, G.: The EM algorithm for factor analyzers. Technical Report, University of Toronto, Toronto, 1997. Technical report CRG-TR-96-1

Giordani, P., Ferraro, M., Martella, F.: An Introduction to Clustering with R, 1st edn. Springer, Singapore (2020)

Hathaway, R.: Another interpretation of the EM algorithm for mixture distributions. Stat. Probab. Lett. **4**(2), 53–56 (1986)

Hennig, C.: An empirical comparison and characterisation of nine popular clustering methods. Adv. Data Anal. Classif. **16**(1), 201–229 (2022)

Higham, N.: Computing the polar decomposition–with applications. SIAM J. Sci. Stat. Comput. **7**(4), 1160–1174 (1986)

Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**(1), 193–218 (1985)

Kass, R., Raftery, A.: Bayes factors. J. Am. Stat. Assoc. **90**, 773–795 (1995)

Kerebin, C.: Estimation consistante de l'ordre de modèles de mélange. C. R. Acad. Sci. Paris Ser. I Math. **326**(2), 243–248 (1998)

Kerebin, C.: Consistent estimation of the order of mixture models. Sankhya Ser. A **62**(1), 49–66 (2000)

Lindsay, B.: Mixture Models: Theory. Geometry and Applications. Institute of Mathematical Statistics, Hayward (1995)

Lindstrom, M., Bates, D.: Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. J. Am. Stat. Assoc. **83**(404), 1014–1022 (1988)

Lütkepohl, H.: Handbook of Matrices. Wiley, Chichester (1996)

MacQueen, J.: Classification and analysis of multivariate observations. In: 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297 (1967)
McLachlan, G., Basford, K.: Mixture Models: Inference and Applications to Clustering. Marcel Dekker, New York (1988)
McLachlan, G., Krishnan, T.: The EM Algorithm and Extensions, 2nd edn. Wiley, Hoboken (2008)
McLachlan, G., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
McLachlan, G., Peel, D.: Mixtures of factor analyzers. In: Langley, P. (ed) Proceedings of the Seventeenth International Conference on Machine Learning, pp. 599–606. San Francisco, Morgan Kaufmann (2000b)
McLachlan, G., Peel, D., Bean, R.: Modelling high-dimensional data by mixtures of factor analyzers. Comput. Stat. Data Anal. **41**(3), 379–388 (2003)
McNicholas, P., Murphy, T.: Parsimonious Gaussian mixture models. Stat. Comput. **18**, 285–296 (2008)
McNicholas, P., Murphy, T.: Model-based clustering of microarray expression data via latent Gaussian mixture models. Bioinformatics **26**(21), 2705–2712 (2010)
McNicholas, P., Murphy, T., McDaid, A., Frost, D.: Serial and parallel implementations of model-based clustering via parsimonious gaussian mixture models. Comput. Stat. Data Anal. **54**(3), 711–723 (2010)
McNicholas, P., ElSherbiny, A., Jampani, K., McDaid, A., Murphy, T., Banks, L.: PGMM: Parsimonious Gaussian mixture models 2019. In: R package version 1.2.4. https://cran.r-project.org/web/packages/pgmm/
Rubin, D.: Inference and missing data. Biometrika **63**(3), 581–592 (1976)
Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**(2), 461–464 (1978)
Scrucca, L., Fop, M., Murphy, T., Raftery, A.: mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. R J. **8**(1), 289–317 (2016)
Tipping, M., Bishop, C.: Probabilistic principal component analysis. J. R. Stat. Soc. Ser. B **61**, 611–622 (1999)
Tipping, M., Bishop, C.: Mixtures of probabilistic principal component analysers. Neural Comput. **11**, 443–482 (1999)
Titterington, D., Smith, A., Makov, U.: Statistical Analysis of Finite Mixture Models. Wiley, Chichester (1985)
Zangwill, W.: Nonlinear Programming: A Unified Approach. Prentice-Hall, Englewood Cliffs (1969)