ULTRAMETRIC GAUSSIAN MIXTURE MODELS WITH PARSIMONIOUS STRUCTURES

Giorgia Zaccaria [1]

[1] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, (e-mail: giorgia.zaccaria@unimib.it)

ABSTRACT: Multidimensional phenomena are usually characterized by nested latent dimensions associated, in turn, with observed variables. These phenomena, for instance, poverty, well-being, and sustainable development, can often differ across countries, or cities within countries, in terms of dimensions, other than in their relationships to each other, on the one hand, and their importance in the definition of the general concept, on the other hand. This paper discusses several parsimonious structures of the covariance matrix reconstructing relationships among variables which can be implemented in Gaussian mixture models to study complex phenomena in heterogeneous populations.

KEYWORDS: ultrametricity, Gaussian mixture models, parsimony, hierarchical structures

1 Introduction

Nested latent dimensions associated with observed variables usually characterize multidimensional phenomena. The hierarchical structure underlying them is composed of *specific* and *higher-order* dimensions; therefore, they give rise to a hierarchy of latent concepts, whose root is represented by the general one. These phenomena concern several fields such as economy, sustainability, health, but also differ in their definition across countries. To reconstruct hierarchical relationships among variables in heterogeneous populations, Cavicchia *et al.*, 2022, introduced a Gaussian mixture model with a specific hierarchical structure of the component covariance matrix. The latter corresponds to an extended ultrametric covariance matrix, whose main property is to be one-to-one-associated with a hierarchy of latent concepts. Differently from the mixture of factor analyzers model (McLachlan *et al.*, 2003), where a factorial structure in uncorrelated factors is identified, the methodology proposed by Cavicchia *et al.*, 2022, is able to detect correlated latent concepts, each one associated with a group of observed variables, and to delve deeper into their relationships.

Notwithstanding the general formulation of an extended ultrametric covariance structure is useful to study hierarchies composed of their maximum number of internal nodes, i.e., the number of the specific dimensions and their aggregations in pairs, more parsimonious structures can be considered. In this paper, different configurations of the extended ultrametric covariance structure are discussed, as well as their properties and main features (Section 2). Final considerations conclude the paper in Section 3.

2 Ultrametric Gaussian mixture models: parsimonious structures

Let $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ be a random sample of size n, where $\boldsymbol{x}_i$ $(i = 1, \ldots, n)$ takes value in $\mathcal{R}^p$. Suppose that $\boldsymbol{x}_i$ follows a finite mixture of G Gaussian distributions, whose pdf is given by

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

where $\pi_1, \ldots, \pi_G$ are positive weights (mixing proportions of the mixture) such that $\sum_{g=1}^{G} \pi_g = 1$, $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$, $g = 1, \ldots, G$, are the mean vectors and the component covariance matrices of the multivariate Gaussian distributions $\phi(\cdot|\cdot)$. In the Ultrametric Gaussian Mixture model, the covariance matrix of the gth component of the mixture is parameterized as

$$\begin{aligned}
\boldsymbol{\Sigma}_g = {} & \left(\boldsymbol{V}_g \begin{bmatrix} {}_{V_g}\sigma_{11} & 0 & \ldots & 0 \\ 0 & {}_{V_g}\sigma_{22} & \cdots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & {}_{V_g}\sigma_{QQ} \end{bmatrix} \boldsymbol{V}'_g \right) \odot \boldsymbol{I}_p \\
& + \left(\boldsymbol{V}_g \begin{bmatrix} {}_{W_g}\sigma_{11} & 0 & \ldots & 0 \\ 0 & {}_{W_g}\sigma_{22} & \cdots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & {}_{W_g}\sigma_{QQ} \end{bmatrix} \boldsymbol{V}'_g \right) \odot (\mathbf{1}_p \mathbf{1}'_p - \boldsymbol{I}_p) \\
& + \left(\boldsymbol{V}_g \begin{bmatrix} 0 & {}_{B_g}\sigma_{12} & \cdots & {}_{B_g}\sigma_{1Q} \\ {}_{B_g}\sigma_{12} & 0 & \cdots & {}_{B_g}\sigma_{2Q} \\ \ldots & \ldots & \ldots & \ldots \\ {}_{B_g}\sigma_{1Q} & {}_{B_g}\sigma_{2Q} & \cdots & {}_{B_g}\sigma_{QQ} \end{bmatrix} \boldsymbol{V}'_g \right).
\end{aligned} \tag{2}$$

Each addend of Eq. (2) depends on the matrix $\boldsymbol{V}_g$, which represents the membership matrix determining the partition of the variable space into $Q < p$ groups, and on one of the three parameters characterizing the variable groups.

The first addend corresponds to the diagonal elements of $\boldsymbol{\Sigma}_g$, where ${}_{V_g}\sigma_{11}, \ldots, {}_{V_g}\sigma_{QQ}$ are the variances of the Q groups in $\boldsymbol{V}_g$; whereas, the off-diagonal elements of $\boldsymbol{\Sigma}_g$ are defined by ${}_{W_g}\sigma_{qq}$ and ${}_{B_g}\sigma_{qh}$, $q,h = 1,\ldots,Q, h \neq q$, in the second and third addend of Eq. (2), respectively. The latter represent the covariances within and between the Q groups. Specific constraints on these parameters let the extend ultrametric covariance matrix in Eq. (2) be one-to-one associated with a hierarchy. Specifically, an ordering exists among ${}_{V_g}\sigma_{qq}, {}_{W_g}\sigma_{qq}$ and ${}_{B_g}\sigma_{qh}$ so that the group variance is greater than the covariance within the group, which in turn is not lower than the maximum covariance between the groups.

Even if suitable in different situations to represent the hierarchical relationships among variables, the parameterization in Eq. (2) can be further constrained to obtain more parsimonious structures. By setting the membership matrix $\boldsymbol{V}_g$ to be the same across mixture components, the other three sets of parameters can be fixed or left free to vary across them. Therefore, the latter structures pinpoint specific dimensions that are equal across the subpopulations of the mixture while their aggregations, thus higher-order dimensions, can differ across them. We can delve into an example of these hierarchical configurations by considering well-being. OECD identifies eleven key dimensions for measuring it throughout the countries[*]. Nonetheless, despite sharing the same specific dimensions, the characterization of this complex phenomenon can vary across countries. For instance, the education level is more associated with the possibility of having a better job in less developed economies and more related to a higher civic engagement in more developed economies.

In both cases in which the specific dimensions are equal or not across components, they can be aggregated altogether at the same level, i.e., a unique value occurs in the matrix of the covariances between groups. This structure gives rise to a second-order hierarchy, studied by Cavicchia & Vichi, 2022, in the factor analysis framework. An interesting case that arises from this configuration corresponds to a formative model (Bollen, 2001), where the unique value ${}_{B}\sigma$ – depending or not on g – equals zero. Indeed, in this hierarchical structure, the specific dimensions result to be uncorrelated and, thus, formed the general concept as unique and not interchangeable part of it. Several examples of formative concepts exist in the literature, such as human development, which is measured by three specific dimensions, i.e., long and healthy life, education, and decent standards of living, usually uncorrelated to each other.

[*] https://www.oecd.org/wise/measuring-well-being-and-progress.htm

3 Conclusions

When studying multidimensional phenomena, the hierarchical structures of latent dimensions underlying them have to be analyzed to build an index for their measurement. To this aim, Cavicchia *et al.*, 2022 proposed an ultrametric Gaussian mixture model which is able to delve into hierarchical relationships among latent dimensions, on the one hand, and to study different characterization of concepts in heterogeneous populations, on the other hand. In this paper, several parsimonious structures of the component covariance matrices are discussed together with the analysis of their corresponding hierarchies.

References

BOLLEN, K. A. 2001. Indicator: Methodology. *Pages 7282–7287 of:* SMELSER, N. J., & BALTES, P. B. (eds), *International Encyclopedia of the Social and Behavioral Sciences*. Elsevier Science, Oxford.

CAVICCHIA, C., & VICHI, M. 2022. Second-order disjoint factor analysis. *Psychometrika*, **87**, 289–309.

CAVICCHIA, C., VICHI, M., & ZACCARIA, G. 2022. Gaussian Mixture Model with an extended ultrametric covariance structure. *Advances in Data Analysis and Classification*, **16**, 399–427.

MCLACHLAN, G. J., PEEL, D., & BEAN, R. 2003. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis*, **41**(3), 379–388.