Model-Based Clustering with Parsimonious Covariance Structure

Carlo Cavicchia[1] , Maurizio Vichi[2] and Giorgia Zaccaria[2]

[1] Econometric Institute, Erasmus University Rotterdam, Rotterdam, The Netherlands, (e-mail: cavicchia@ese.eur.nl)

[2] Department of Statistical Sciences, Sapienza University of Rome, Rome, Italy, (e-mail: maurizio.vichi@uniroma1.it, giorgia.zaccaria@uniroma1.it)

Abstract: Complex multidimensional concepts are often explained by a tree-shape structure by considering nested partitions of variables, where each variable group is associated with a specific concept. Recalling that relations among variables can be detected by their covariance matrix, this paper introduces a covariance structure that reconstructs hierarchical relationships among variables highlighting three features of the variable groups. We finally present an application of the latter covariance structure to the model-based clustering.

Keywords: Gaussian mixture model, hierarchical latent concepts, partition of variables

1 Introduction

The main goal of Factor Analysis (FA, Spearman, 1904) is to reconstruct the covariance matrix of variables by computing a reduced number of factors while preserving as much information as possible. However, since FA is unable to reconstruct hierarchical relations, a model with a hierarchical form is therefore required. Among several models based on the sequential application of FA addressing the same problem, Cavicchia *et al.* (2020) proposed a model to reconstruct a nonnegative correlation matrix via an ultrametric one. The model results in a simultaneous procedure which is able both to detect the best variable partition in a reduced number of groups and build the hierarchy upon them. The latter model ensues particularly suitable for complex hierarchical multidimensional concepts due to the one-to-one relation between a hierarchy of concepts and an ultrametric correlation matrix (Dellacherie *et al.*, 2014). Our paper overcomes the limitations of the model presented by Cavicchia *et al.* (2020) extending the same idea to a general covariance matrix and applies this special covariance structure in the Gaussian Mixture Models (GMMs) framework.

Since GMMs can easily fall into the so-called “curse of dimensionality” because of the large number of parameters dedicated to covariance structures, in the specialized literature several different parametrizations are present. One of the most used is the eigen-decomposition (Banfield & Raftery, 1993) of the form $\mathbf{\Sigma} = \lambda \mathbf{DAD}'$, where λ is a scalar determining the cluster volume, $\mathbf{A}$ is a diagonal matrix controlling the cluster shape, and $\mathbf{D}$ is an orthogonal matrix which specifies the cluster orientation. Another parameterization is proper of the mixture of factor analyzers (Ghahramani & Hilton, 1997) and assumes a cluster covariance structure of the form $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$, where p is the number of variables, Q is the number of factors, $\mathbf{\Lambda}$ is the $p \times Q$ factor loading matrix and $\mathbf{\Psi}$ is the p-dimensional diagonal covariance matrix of the error. Our proposal aims to implement a new parameterization of a covariance matrix via a hierarchical covariance one for each cluster that can be extremely parsimonious.

2 Features of the covariance structure

Multidimensional phenomena are often composed of nested dimensions characterized by distinct levels of abstraction. Each dimension is uniquely connected to a group of variables and represents a specific concept. Merging two dimensions together gives rise to a broader dimension up to the general one such that the hierarchical structure underlying a multidimensional phenomenon is detected. In order to model the hierarchical relationships among the dimensions, we introduce three main features of a variable group: the variance of the variable group, the covariance within the variable group, which measures the internal concordance among variables belonging to the same group, and the covariance between concepts associated with the variable groups. These features are constrained to be “ordered” such that the variance of the groups is greater (in the absolute sense) than the covariance within or between groups, whereas the covariance within groups must be in turn larger than the covariance between groups. These constraints allow to define a hierarchical structure of concepts, from the most concordant to the most discordant. The last aggregations in the hierarchy may occur between: (i) concordant concepts defining a general one; (ii) discordant concepts with negative between-group covariance; (iii) uncorrelated concepts.

Given the number of specific dimensions Q which underlie the multidimensional phenomenon, each level $q = Q, \ldots, 1$ of the hierarchy is characterized by: (i) the $p \times q$ membership matrix $\mathbf{V}_q$, which pinpoints the membership of each variable to a group; (ii) the diagonal matrix $\mathbf{S}_q^V$ of order q, whose main diagonal represents the variance of each group; (iii) the diagonal matrix $\mathbf{S}_q^W$

of order q, whose main diagonal represents the covariance within each group; (iv) the ultrametric matrix $\mathbf{S}_q^B$ of order q, whose diagonal entries are set to zero and off-diagonal ones represent the hierarchical relationships between pairs of concepts. Given $\mathbf{V}_q$, the estimates of the matrices $\mathbf{S}_q^V$, $\mathbf{S}_q^W$ and $\mathbf{S}_q^B$ are

$$\widehat{\mathbf{S}}_q^V = (\widehat{\mathbf{V}}_q'\widehat{\mathbf{V}}_q)^{-1}\widehat{\mathbf{V}}_q'\mathrm{diag}(\mathbf{S})\widehat{\mathbf{V}}_q, \tag{1}$$

$$\widehat{\mathbf{S}}_q^W = [(\widehat{\mathbf{V}}_q'\widehat{\mathbf{V}}_q)^2 - \widehat{\mathbf{V}}_q'\widehat{\mathbf{V}}_q]^{-1}\mathrm{diag}\Big[\widehat{\mathbf{V}}_q'\Big(\mathbf{S} - \mathrm{diag}(\widehat{\mathbf{V}}_q\widehat{\mathbf{S}}_q^V\widehat{\mathbf{V}}_q')\Big)\widehat{\mathbf{V}}_q\Big], \tag{2}$$

$$\widehat{\mathbf{S}}_q^B = \widehat{\mathbf{V}}_q^+\mathbf{S}(\widehat{\mathbf{V}}_q')^+, \tag{3}$$

respectively, where $\mathbf{S}$ represents the $p \times p$ observed covariance matrix, $\mathbf{I}_p$ is the identity matrix of order p and $\mathrm{diag}(\cdot)$ denotes the diagonal matrix whose diagonal elements are those of a parenthesized one.

We implement the parameterization of the covariance matrix based on the aforementioned quantities into the GMMs in order to simultaneously detect homogeneous clusters of units and a hierarchical definition of a multidimensional phenomenon.

3 Application

Our proposal is applied on the "Human Development Index" dataset[*] which consists of 167 countries and 9 variables. The optimal model in terms of Bayesian Information Criterion (BIC, Schwarz, 1978) considers 3 clusters of countries (Fig. 1) and 3 groups of variables. It is worth highlighting that the model requires 71 parameters to be estimated, of which only 14 for each covariance structure. The first cluster is characterized by the countries with high income, gdp per capita and very low child mortality. The second cluster is constituted by the poorest countries with low life expectancy and income, whereas the third one is composed by countries with median performances. Each cluster is characterized by a different hierarchy of the latent concepts associated with the three groups of variables. The group made by the economic variables (income, gdp, exports and imports) in Cluster 1 is the one with the highest value of internal variance, whereas the same group in Cluster 3 is merged with the group considering child mortality and fertility and has the highest covariance within the group. Notwithstanding the latent concepts and their hierarchical relationships are specific per cluster, all the hierarchies end

[*] https://www.kaggle.com/rohan0301/unsupervised-learning-on-country-data

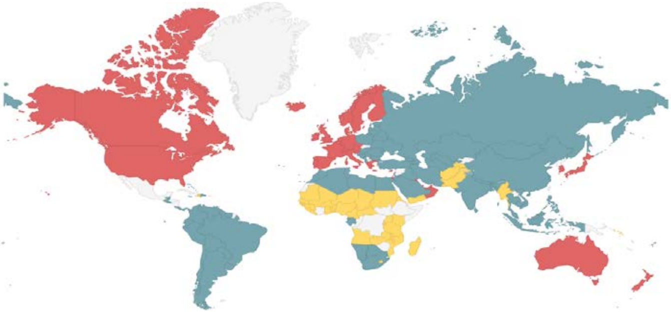

Figure 1: Clusters of countries: Cluster 1 (red), Cluster 2 (yellow) and Cluster 3 (blue)

with a negative between-group covariance highlighting the absence of a unique concordant general concept.

4 Conclusions

This paper proposes a parsimonious GMM which aims at modeling multidimensional phenomena, usually defined by hierarchically nested latent concepts. The application of the method on real data shows its potentialities.

References

BANFIELD, J.D., & RAFTERY, A.E. 1993. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**(3), 803–821.

CAVICCHIA, C., VICHI, M., & ZACCARIA, G. 2020. The ultrametric correlation matrix for modelling hierarchical latent concepts. *Advances in Data Analysis and Classification*, **14**(4), 837–853.

DELLACHERIE, C, MARTINEZ, S, & MARTIN, J SAN. 2014. *Inverse M-matrices and ultrametric matrices*. Lecture Notes in Mathematics. Springer International Publishing.

GHAHRAMANI, Z., & HILTON, G.H. 1997. The EM algorithm for factor analyzers. Technical report CRG-TR-96-1, University of Toronto, Toronto.

SCHWARZ, G. 1978. Estimating the dimension of a model. *Annals of Statistics*, **6**(2), 461–464.

SPEARMAN, C. E. 1904. “General intelligence,’ objectively determined and measured. *The American Journal of Psychology*, **15**(2), 201–293.