DIMENSIONALITY REDUCTION VIA HIERARCHICAL FACTORIAL STRUCTURE

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

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

ABSTRACT: Manifold multidimensional concepts are explained via a tree-shape structure by taking into account the nested hierarchical partition of variables. The root of the tree is a general concept which includes more specific ones. In order to detect the different specific concepts at each level of the hierarchy, we can identify two different features regarding groups of variables: the internal consistency of a concept and the correlation between concepts. Thus, given a data positive correlation matrix, we reconstruct the latter via an ultrametric correlation matrix which detects hierarchical concepts by looking for their internal consistency and the correlation between them measured by relative indices.

KEYWORDS: ultrametric matrix, hierarchical latent concepts, correlation matrix, partition of variables.

1 Introduction

Many relevant multidimensional phenomena are represented via a tree-structure (for example well-being, sustainable development, poverty, climate change). We can hypothesize a Dimensionality Reduction model with a hierarchical structure that goes from disjoint sets of Manifest Variables (MVs) to the General Concept (GC). In other words we build a parsimonious hierarchy of classes of variables starting from a reduced number, (i.e., latent dimensions) which measure specific concepts describing the main components of the phenomenon under study up to the definition of the GC. Each cluster of MVs may be related with a factor which best represents its dimension. This is not new in many fields of research, for instance Revelle (1979) introduced a hierarchical cluster analysis method very useful to detect clusters of variables in a hierarchical approach. Our proposal can be considered into the Dimensionality Reduction framework for its ability of summarizing a big quantity of information by way of many steps of aggregation. In order to detect the hierarchy of variables, i.e., the different specific concepts at each level of the hierarchy, we identify two different features regarding clusters of variables: the internal consistency (i.e., reliability of the concept) and the correlation between concepts. Thus, given a data correlation matrix, we reconstruct the latter via an ultrametric correlation matrix which detects hierarchical concepts with the highest internal consistency and with the highest correlation between them in order to justify their fusion. The *internal consistency of concept* (i.e., variable cluster), is the global consistency of MVs based on their correlations within cluster. This is also called internal reliability and it is commonly measured by Cronbach's alpha (Cronbach, 1951). On the one hand, the reliability is connected to the concept of unidimensionality, which, on the other hand, evaluates to what extent a single latent indicator has been measured with a set of MVs. Reliability and unidimensionality are more realistic for specific dimensions, whereas, when considering a general factor, we have to hypothesize the presence of a GC (Cavicchia & Vichi, 2019). A common error is to interpret a measure of reliability as a measure of unidimensionality. Although being connected, they cannot consider as the same thing. Unidimensionality involves the homogeneity of a set of items, and internal consistency is certainly necessary for homogeneity, but it is not sufficient. We can see that, therefore, the improving of the internal consistency leads to an improvement of unidimensionality as well, but we cannot use the same index to measure both. By supposing that no variable can belong to two clusters at the same time, such that, all the clusters are disjoint at each level, we can consider another important feature which is *the correlation between clusters of variables*. This latter represents a function of the pairwise relationships between the items of the two groups and determines the bottom-up agglomerations of variable clusters. Hence, we are supposing a nested hierarchy where, starting from Q clusters of variables, all the possible combinations are taken into consideration in order to identify the aggregations which best detect reliable concepts at all levels.

2 Internal Consistency and Correlation Between

2.1 A Measure of Internal Consistency

The internal consistency of a cluster of MVs is the ability of all variables to measure the same latent concept. It is usually measured by indices based on the correlations between the MVs within the cluster. Many measures of internal consistency are reviewed by Revelle & Zinbarg (2009). In our framework, by starting from Q variable clusters at the bottom level, we have $\frac{Q(Q-1)}{2}$ clusters along the hierarchy, and as many internal consistency indexes. For each level

$q = Q, \dots, 1$, the $(J \times q)$ membership matrix $\mathbf{V}_q$, where J is equal to the total number of MVs, tells us for each cluster which variable belongs to. Given $\mathbf{V}_q$, Cavicchia *et al.* (2019) proposed a measure of internal consistency for non-negative data correlation matrices, arranged in a $(q \times q)$ diagonal matrix as follows:

$$\widehat{\mathbf{R}}_q^W = diag\big(dg(\mathbf{V}_q'(\mathbf{R} - \mathbf{I}_J)\mathbf{V}_q)\big)[(\mathbf{V}_q'\mathbf{V}_q)^2 - \mathbf{V}_q'\mathbf{V}_q]^{-1}. \tag{1}$$

In Eq. 1 $\mathbf{R}$ represents the $(J \times J)$ observed correlation matrix and $\mathbf{I}_J$ is the identity matrix of order J; furthermore $dg(\cdot)$ produces a vector whereas $diag(\cdot)$ builds a diagonal matrix. It is important to notice that $\widehat{\mathbf{R}}_q^W$ has q non-zero elements which are the internal consistency measures, one for each cluster. $\widehat{\mathbf{R}}_q^W$ corresponds to the Least Squares solution for reconstructing $\mathbf{R}$ via an ultrametric correlation matrix composed by a matrix which explains the internal consistency of concepts and a matrix which explains the correlation between concepts. Each value ${}_W\widehat{r}_{ll}$ $(l = 1, \dots, q)$ of $\widehat{\mathbf{R}}_q^W$ belongs to the interval $[0, 1]$, recalling that $\mathbf{R}$ has all non-negative values, thus it may be considered as a relative index. An important characteristic of the values of $\widehat{\mathbf{R}}_q^W$ is that they are not function of the number of MVs of each cluster, thus they are not affected by the size of clusters.

2.2 A Measure of Correlation Between Clusters of Variables

In order to detect all the levels of the hierarchy, it is crucial to define the correlation between clusters of MVs, each one representing a latent concept.

For each level $q = Q, \dots, 1$ it is possible to compute the correlation between clusters of variables, and the internal consistency within clusters as well, but it is important to stress the fact that the Q-level (i.e., the level with Q variable clusters at the bottom level) is the optimal one in order to reconstruct the data correlation matrix $\mathbf{R}$. Given $\mathbf{V}_q$ and the diagonal matrix of internal consistency measures $\widehat{\mathbf{R}}_q^W$, Cavicchia *et al.* (2019) proposed a measure of correlation between clusters of MVs for non-negative data correlation matrices, arranged in a $(q \times q)$ correlation matrix as follows:

$$\widehat{\mathbf{R}}_q^B = (\mathbf{V}_q'\mathbf{V}_q)^{-1}\mathbf{V}_q'\bar{\mathbf{R}}\mathbf{V}_q(\mathbf{V}_q'\mathbf{V}_q)^{-1}. \tag{2}$$

In Eq. 2, $\bar{\mathbf{R}} = \mathbf{R} - \mathbf{V}_q'\widehat{\mathbf{R}}_q^W\mathbf{V}_q + diag\big(dg(\mathbf{V}_q'\widehat{\mathbf{R}}_q^W\mathbf{V}_q)\big) - \mathbf{I}_J + \mathbf{V}_q'\mathbf{I}_Q\mathbf{V}_q$. The off-diagonal values within $\widehat{\mathbf{R}}_q^B$ are the between-concepts correlation whereas the diagonal elements are equal to one. $\widehat{\mathbf{R}}_q^B$ is the LS solution with respect to the matrix which explains the correlation between concepts. As for $\widehat{\mathbf{R}}_q^W$, each value ${}_B\widehat{r}_{kf}$ $(k = 1,\ldots,q;\ f = 1,\ldots,q;\ k \neq f)$ of $\widehat{\mathbf{R}}_q^B$ belongs to the interval $[0,1]$ and it turns out to be a relative index.

3 Conclusions

A correlation matrix **R** may be reconstructed via a ultrametric hierarchical structure which highlights two crucial characteristic regarding clusters of variables: the internal consistency and the correlation between clusters. In order to detect the ultrametric structure of the latent concepts, it is important to investigate in depth the reliability of each cluster of MVs and all the relations among them. For correlation matrices **R** which are composed only by non-negative elements, as common in psychometric applications, Cavicchia *et al.* (2019) presented a model that considers two main matrices, the first one which contains non-zero element only on the diagonal, that is the internal consistency measure for the related cluster, and the second one which is a correlation matrix with the off-diagonal elements that represent the correlation between clusters. The Dimensionality Reduction model with a hierarchical structure that goes from disjoint sets of Manifest Variables (MVs) to the General Concept (GC) is given by detecting consistent clusters and by following correlation between them.

References

CAVICCHIA, C., & VICHI, M. 2019. Statistical Model-based Composite Indicators for Complex Socio-Demographic and Economic Phenomena: Models, Properties and Features for tracking coherent policy conclusions. *Submitted, Social Indicators Research.*

CAVICCHIA, C., VICHI, M., & ZACCARIA, G. 2019. The Ultrametric Correlation Matrix for Modelling Hierarchical Latent Concepts. *Submitted, Advances in Data Analysis and Classification.*

CRONBACH, L. J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika.*, **16**, 297–334.

REVELLE, L. J. 1979. Hierarchical Cluster Analysis and the Internal Structure of Tests. *Multivariate Behavioral Research.*, **14**, 57–74.

REVELLE, L. J., & ZINBARG, R. 2009. Coefficients Alpha, Beta, Omega, and the glb: Comments on Sijtsma. *Psychometrika.*, **74**(1), 145–154.