Complex Dimensionality Reduction: ultrametric models for mixed-type data

Marco Mingione, Maurizio Vichi and Giorgia Zaccaria

Abstract The factorial latent structure of variables, if present, can be complex and generally identified by nested latent concepts ordered in a hierarchy, from the most specific to the most general one. This corresponds to a tree structure, where the leaves represent the observed variables and the internal nodes coincide with latent concepts defining the general one (i.e., the root of the tree). Although several methodologies have been proposed in the literature to study hierarchical relationships among quantitative variables, very little has been done for more general mixed-type data sets. Hence, it is of the utmost importance to extend these methods and make them suitable to the even more frequent availability of mixed-type data matrices, as complex real phenomena are often described by both qualitative and quantitative variables. In this work, we propose a new exploratory model to study the hierarchical statistical relationships among variables of mixed-type nature by fitting an ultrametric matrix to the general dependence matrix, where the former is one-to-one associated with a hierarchical structure.

1 Introduction

In many statistical applications, the presence of mixed-type data brings an additional level of complexity to the analysis as they usually result in intricate dependence structures (De Leon and Chough, 2013). When this is the case, the straightforward implementation of standard techniques is generally inappropriate, and specific solutions are required to deal with the mixed nature of the data depending on the final objective of the statistical analysis. For instance, the majority of conventional tools that are used both in the exploratory and the modeling phase relies on the assumption that the data,

Marco Mingione (✉)
University of Rome "La Sapienza", P.le A. Moro 5, 00185, Rome (Italy)
e-mail: marco.mingione@uniroma1.it

Maurizio Vichi
University of Rome "La Sapienza", P.le A. Moro 5, 00185, Rome (Italy)
e-mail: maurizio.vichi@uniroma1.it

Giorgia Zaccaria
University of Rome Unitelma Sapienza, P.zza Sassari 4, 00161, Rome (Italy)
e-mail: giorgia.zaccaria@unitelmasapienza.it

or at least a suitable transformation of them, follow a Normal distribution: an assumption that no longer applies in such contexts. For these reasons, several methods have been developed to deal with mixed-type data in many applications, especially concerning dimensionality reduction (McParland and Gormley, 2016; Van de Velden et al., 2019; Vichi et al., 2019). These methodologies mostly rely on the computation of the general dependence matrix, say $\tilde{\mathbf{\Sigma}}$, which altogether includes the correct pairwise statistical relationship measures, namely: χ^2 in the case of qualitative pairs; ρ^2 between pairs of quantitative variables; η^2 in the mixed-type case. This general dependence matrix can be proven to be positive semi-definite and it can be paired with the matrix of the p-values assessing the statistical significance of the estimated relationship. Hence, while being useful as an exploratory tool, $\tilde{\mathbf{\Sigma}}$ has been already used to extend Factor Analysis (FA) and Principal Component Analysis (PCA) to the case of mixed-type data (Pagès, 2004; Chavent et al., 2011) in order to detect latent structures underlying the data. However, if complex phenomena are considered, FA and PCA for mixed-type data are not able to pinpoint the hierarchical relationships among variables defining nested dimensions (concepts). In this paper, we introduce a new simultaneous, exploratory model for identifying a hierarchy of latent concepts. The proposal aims at reconstructing the general dependence matrix $\tilde{\mathbf{\Sigma}}$ via an ultrametric dependence matrix by extending the model proposed by Cavicchia et al. (2020b) to mixed-type data.

2 Background

Let $\mathbf{x}_i$ be a $(J \times 1)$ random vector corresponding to a generic multivariate observation i, where $i = 1, \ldots, N$ and $N > J$. Without loss of generality, $\mathbf{x}_i$ can be ordered such that the first P values are realizations of qualitative (categorical) variables, while the remaining $J - P$ values come from quantitative (numerical) variables. Hence, we can rewrite $\mathbf{x}_i = [{}_q\mathbf{x}_i', {}_n\mathbf{x}_i']'$, where ${}_q\mathbf{x}_i$ and ${}_n\mathbf{x}_i$ are the $(P \times 1)$ and the $((J-P) \times 1)$ vectors of the qualitative and quantitative variables, respectively. Specifically, ${}_q\mathbf{x}_i$ has elements ${}_qx_{ij} \in \{1, \ldots, c_j\}$, where $c_j \geq 2$ represents the number of distinct categories of variable j, for $j = 1, \ldots, P$. For the moment, we assume that the categories are not ordered, therefore defining the qualitative variables as *nominal* variables.

The sampled observations can be stacked together in the matrix $\mathbf{X} = [{}_q\mathbf{X}, {}_n\mathbf{X}]$, formed by two sub-matrices of dimensions $(N \times P)$ and $(N \times (J-P))$, having a qualitative and quantitative part.

The computation of $\tilde{\mathbf{\Sigma}}$ relies on the suitable standardization of matrix $\mathbf{X}$, which can be defined as in Vichi et al. (2019):

$$ {}_s\mathbf{X} = [{}_s\mathbf{G}, \frac{1}{\sqrt{N}}\mathbf{Z}] = [{}_s\mathbf{G}_1, \ldots, {}_s\mathbf{G}_P, \frac{1}{\sqrt{N}}\mathbf{z}_1, \ldots, \frac{1}{\sqrt{N}}\mathbf{z}_{J-P}], \tag{1} $$

where ${}_s\mathbf{G}$ is the $(N \times C)$ standardized matrix referring to the first P nominal variables, $C = \sum_{j=1}^{P} c_j$ and $\mathbf{Z} = \mathbf{J}_n\mathbf{X}\text{diag}(\text{diag}(\Sigma_{n\mathbf{X}}))^{-\frac{1}{2}}$ is the $(N \times (J-P))$ standardized matrix referring to the $J-P$ quantitative variables. ${}_s\mathbf{G}$ is obtained by appending together the single ${}_s\mathbf{G}_j = \mathbf{J}\mathbf{G}_j(\mathbf{G}_j'\mathbf{G}_j)^{-\frac{1}{2}}, j = 1, \ldots, P$, as in multiple correspondence analysis (Greenacre, 2017), while $\mathbf{Z}$ has 0 mean and unit sum of square columnwise; $\mathbf{J} = \mathbf{I} - (\frac{1}{N})\mathbf{1}_N\mathbf{1}_N'$ is the centering idempotent matrix. Note that ${}_s\mathbf{X}$ has dimension $(N \times (C + J - P))$.

In the following section, we compute the general dependence matrix that includes the relationships among variables of different nature, qualitative and quantitative.

3 Statistical relationships between mixed-type variables

Considering the matrix ${}_s\mathbf{X}$ in Eq. (1), we compute the matrix of the *correct* pairwise statistical relationships among mixed-type variables as follows

$$ {}_e\mathbf{\Sigma}_{s\mathbf{X}} = {}_s\mathbf{X}'{}_s\mathbf{X}. $$

${}_e\mathbf{\Sigma}_{s\mathbf{X}}$ is a square, positive semi-definite matrix of dimension $C + J - P$, and it can be divided into four blocks

$$ {}_e\mathbf{\Sigma}_{s\mathbf{X}} = \begin{pmatrix} \mathbf{\Phi}_{s\mathbf{G}} & \mathbf{H}_{s\mathbf{GZ}} \\ \mathbf{H}'_{s\mathbf{GZ}} & \mathbf{R}_{n\mathbf{X}} \end{pmatrix}, $$

where

- $\mathbf{\Phi}_{s\mathbf{G}} = [{}_s\mathbf{G}_j'\,{}_s\mathbf{G}_m : j, m = 1, \ldots, P]$ is a square matrix of order C, containing positive values if two modalities of ${}_q\mathbf{X}_j$ and ${}_q\mathbf{X}_m$, say l and h, are jointly assumed, negative otherwise;
- $\mathbf{R}_{n\mathbf{X}} = [\frac{1}{N}\mathbf{z}_k'\,\mathbf{z}_r : k, r = P+1, \ldots, J]$ is the correlation matrix of order $J-P$ with elements $|\rho_{kr}| \leq 1$;
- $\mathbf{H}_{s\mathbf{GZ}} = [{}_s\mathbf{G}_j{}'\,\mathbf{z}_k : j = 1, \ldots, P, k = P+1, \ldots, J]$ is the matrix of order $(C \times (J-P))$ accounting for the association between the qualitative variable ${}_q\mathbf{X}_j$ and the quantitative variable ${}_n\mathbf{X}_k$.

The dependence between qualitative and quantitative variables can be defined by considering the matrix with elements equal to the square of elements of ${}_e\mathbf{\Sigma}_{s\mathbf{X}} = {}_s\mathbf{X}'{}_s\mathbf{X}$, i.e.

$$ {}_e\mathbf{\Sigma}^2_{s\mathbf{X}} = \begin{pmatrix} \mathbf{\Phi}^2_{s\mathbf{G}} & \mathbf{H}^2_{s\mathbf{GZ}} \\ \mathbf{H}^{2\prime}{}_{s\mathbf{GZ}} & \mathbf{R}^2_{n\mathbf{X}} \end{pmatrix}. $$

The latter is evidently in connection with well-known dependence/correlation measures.

- $\mathbf{\Phi}^2_{s\mathbf{G}}$ can be exploited to compute

$$ {}_r\chi^2_{jm} = tr({}_s\mathbf{G}_j\, {}_s\mathbf{G}'_j\, {}_s\mathbf{G}_m\, {}_s\mathbf{G}'_m)/min(c_j - 1, c_m - 1); $$

- $\mathbf{R}^2_{{}_n\mathbf{X}} \rightarrow$ is the matrix of the squared Pearson's correlation coefficients measuring the relationship between ${}_n\mathbf{X}_k$ and ${}_n\mathbf{X}_r$;
- $\mathbf{H}^2_{{}_s\mathbf{GZ}}$ is the matrix of the correlation ratios between the qualitative variable ${}_q\mathbf{X}_j$ and the quantitative variable ${}_n\mathbf{X}_k$.

It has to be noticed that all the three matrices have elements ranging between 0 and 1, where the former represents the case of independence and the latter that one of perfect dependence.

The matrix ${}_e\mathbf{\Sigma}^2_{{}_s\mathbf{X}}$ can be therefore synthesized into a $J \times J$ matrix of general dependence as follows

$$ \tilde{\mathbf{\Sigma}} = \begin{pmatrix} {}_r\chi^2_{{}_s\mathbf{G}} & \boldsymbol{\eta}^2_{{}_s\mathbf{GZ}} \\ \boldsymbol{\eta}^{2\prime}_{{}_s\mathbf{GZ}} & \mathbf{R}^2_{{}_n\mathbf{X}} \end{pmatrix}. \tag{2} $$

In the following section, we introduce a model for studying the relationships occurring among quantitative and quantitative variables into $\tilde{\mathbf{\Sigma}}$.

4 Methods

The ultrametricity notion is well-known in statistics for hierarchical clustering. Nevertheless, this notion has been introduced in mathematics with regard to the p-adic number theory (Dellacherie et al., 2014) associated with a generic matrix, which is not necessarily a distance matrix – specifically, a reverse relationship occurs with an ultrametric distance matrix. Hereinafter, we recall the definition of an *ultrametric* matrix.

Definition 1 A nonnegative[1] matrix $\mathbf{U}$ of order J is said to be ultrametric if

1. $u_{ij} = u_{ji}, i, j = 1, \ldots, J$ (symmetry);
2. $u_{jj} \geq \max\{u_{ij} : i = 1, \ldots, J\}, j = 1, \ldots, J$ (column pointwise diagonal dominance);
3. $u_{ij} \geq \min\{u_{il}, u_{jl}\}, i, j, l = 1, \ldots, J$ (ultrametricity).

Condition 3. can be rewritten as follows

- for each triplet i, j, l, there exists a reordering $\{i, j, l\}$ of the elements s.t. $u_{ij} \geq u_{il} = u_{jl}$,

by unraveling that an ultrametric matrix $\mathbf{U}$ is composed of a reduced number of distinct values subject to a specific order. This engenders an interesting feature of an ultrametric matrix, that is of being associated with a hierarchy

[1] A nonnegative matrix $\mathbf{M} = [m_{ij}]$ is a matrix with nonnegative values, i.e., $m_{ij} \geq 0$.

over variables. Moreover, every ultrametric matrix is positive semi-definite (Dellacherie et al., 2014, pp. 60-61).

Dimensionality reduction for mixed-type data can be performed by Factor Analysis (Pagès, 2004). However, FA is not able to reconstruct hierarchical relationships among variables (both qualitative and quantitative) and thus to unravel concepts of higher-order. Cavicchia et al. (2020b) introduced a model to study the hierarchical relationships among variables by reconstructing a nonnegative correlation matrix via an ultrametric correlation one. However, their proposal pertains *quantitative* variables only. In this paper, we extend the ultrametric model proposed by Cavicchia et al. (2020b) in order to detect hierarchical structures on variables of different nature (i.e., quantitative and qualitative). Specifically, we fit an ultrametric matrix to the general dependence matrix $\tilde{\mathbf{\Sigma}}$ defined in Eq. (2) – that is nonnegative by definition – as follows

$$\tilde{\mathbf{\Sigma}} = \tilde{\mathbf{\Sigma}}_{\mathrm{u}} + \mathbf{E}, \tag{3}$$

where $\tilde{\mathbf{\Sigma}}$ is an ultrametric dependence matrix of order P and $\mathbf{E}$ is a residual (error) matrix of the same order. $\tilde{\mathbf{\Sigma}}_{\mathrm{u}}$ pinpoints a hierarchical structure by defining a reduced number of variable groups associated with latent concepts and identifying broader concepts as bottom-up aggregations of the lower-order one. Model (3) can be estimated in the Least-Squares context or in the Maximum-Likelihood framework under suitable specific assumptions on the distribution of the error.

It is worthy to highlight that the proposal differs from hierarchical clustering methods applied on variables (see Cavicchia et al., 2020a, for the quantitative case). Indeed, even if the latter can be implemented on variables after a proper transformation of similarity measures into dissimilarity ones, they are sequential and greedy procedures affected by misclassification at the bottom of the hierarchy that can have an effect on higher levels. Other than detecting specific features related to variables, the proposed methodology overcomes the aforementioned drawbacks of hierarchical clustering procedures since this is a simultaneous and parsimonious model.

5 Application

We apply the ultrametric model described in Section 4 on the well-known benchmark data set `mtcars` (Henderson and Velleman, 1981), available in the package `MASS` of `R` statistical software (Team et al., 2013). The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of design and performance for 32 cars. Specifically, the data set includes 6 quantitative variables:

- miles per gallon (`mpg`): a proxy for the fuel consumption efficiency;

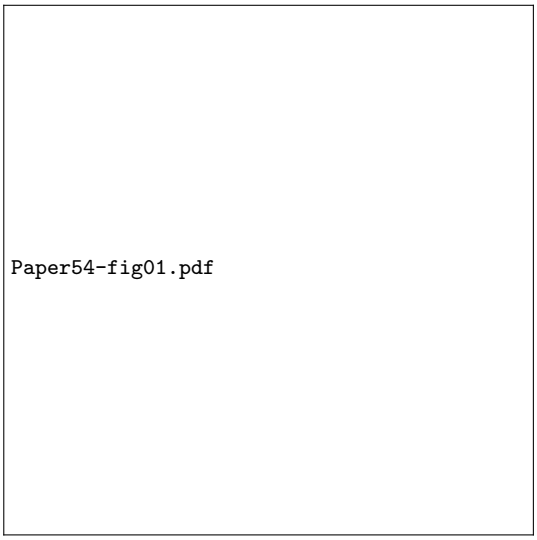


Fig. 1: Observed $\tilde{\mathbf{\Sigma}}$ for the `mtcars` data set.

- displacement (`disp`): the total volume (in cubic inches) of all the cylinders in an engine;
- gross horsepower (`hp`): a measurement of engine output, taken at the flywheel;
- rear axle ratio (`drat`): the number of revolutions the driveshaft must make to spin the axle one full turn;
- weight (`wt`): the weight of the vehicle (in 1000 lbs.);
- quarter per mile (`qsec`): the shortest time from a standing start to the end of a straight 1/4 mile track;

and 5 qualitative variables:

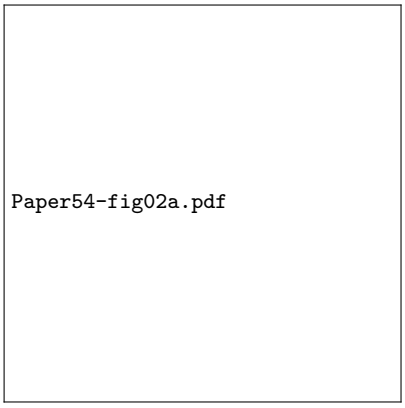
- number of cylinders (`cyl`): low, medium, high;
- engine (`vs`): V-shaped or straight;
- transmission (`am`): automatic or manual;
- number of gears (`gear`): low, medium, high;
- number of carburetors (`carb`): single, low, medium, high.

We computed $\tilde{\mathbf{\Sigma}}$ as described in Section 3. From Fig. 1, we can clearly see an overall high dependence among the variables in the data set: the largest χ^2 is observed for the *number of cylinders* and the *engine shape*; the largest η^2 between the *number of cylinders* and the *displacement*; the largest ρ^2 between the *displacement* and the *weight*.

The results of the ultrametric model on the `mtcars` data set are provided in Fig. 2. The model applied herein considers a complete hierarchy over the 11 variables. Looking at Fig. 2b, we can highlight four main groups of variables,

each one associated with a latent concept: the first composed of *displacement*, *number of cylinders*, *weight*, *miles per gallon* and *gross horsepower*, thus representing the *efficiency* of a car; the second group included *rear axle ration*, *number of gears* and *transmission*, thus identifying the *gearshift features*; the third group formed by *quarter per mile* and *engine*, hence depicting the *performances on the standing start*; the fourth group solely defined by *number of carburetors*, representing a singleton. Remark that we model the dependence between mixed-type data measured into $\tilde{\mathbf{\Sigma}}$, whose elements are non-negative and range between 0 and 1. In defining the hierarchical structure, the ultrametric model therefore considers the strength of the dependence among variables, but not its sign. Nonetheless, the direction of dependence can be taken into account for interpreting the results. For instance, in defining the first latent concept, i.e., *efficiency*, negative relationships manifest between *miles per gallon* and *weight*, and *miles per gallon* and *gross horsepower* since the lower the weight and the horsepower of a car are, the higher the performances of a car are and, consequently, its efficiency.

The existence of these four latent concepts, with highly dependent variables defining them, is clearly visible in Fig. 2a, as well as the order of their aggregations. Indeed, the first aggregation occurs between *efficiency* and *gearshift features*, the second one between the *performances on the standing start* and *number of carburetors* and the last one defines the aggregation of the aforementioned two broader groups.



(a) *Heatmap of the fitted* $\tilde{\mathbf{\Sigma}}_{\mathrm{u}}$ *on the* `mtcars` *data set*

(b) *Path diagram representation*

Fig. 2

6 Discussion

In this work, we firstly revised the importance of the general dependence matrix $\tilde{\mathbf{\Sigma}}$, which depicts pairwise relationships among variables of different nature. The latter is a fundamental tool for the application of methodologies for dimensionality reduction, such as Factor Analysis and Principal Component Analysis, to mixed-type data. Even if these methodologies are suitable to detect latent structures underlying multivariate data, they are not able to unravel hierarchical relationships among variables. In this paper, we extended the ultrametric model proposed by Cavicchia et al. (2020b) to mixed-type data with the aim of inspecting the relationships among dimensions with different levels of abstraction defining a complex phenomenon.

Further developments of this work can be explored. Indeed, it would be interesting to compare the performances of our proposal with the ones of traditional hierarchical clustering algorithms, i.e., to extended the proof in Cavicchia et al. (2020a) to the more general case of mixed-type data.

References

Cavicchia C, Vichi M, Zaccaria G (2020a) Exploring hierarchical concepts: theoretical and application comparison. In: Imaizumi T, Nakayama A, Yokoyama S (eds) Advanced Studies in Behaviormetrics and Data Science. Behaviormetrics: Quantitative Approaches to Human Behavior, vol 5, Springer, Singapore, pp 315–328

Cavicchia C, Vichi M, Zaccaria G (2020b) The ultrametric correlation matrix for modelling hierarchical latent concepts. Advances in Data Analysis and Classification 14(4):837–853

Chavent M, Kuentz V, Liquet B, Saracco J (2011) Clustering of variables via the pcamix method. In: International Classication Conference, p 1

De Leon AR, Chough KC (2013) Analysis of mixed data: methods & applications. CRC Press

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

Greenacre M (2017) Correspondence analysis in practice, 2nd edn. Chapman & Hall/CRC

Henderson HV, Velleman PF (1981) Building multiple regression models interactively. Biometrics pp 391–411

McParland D, Gormley IC (2016) Model based clustering for mixed data: clustmd. Advances in Data Analysis and Classification 10(2):155–169

Pagès J (2004) Analyse factorielle de données mixtes. Revue de statistique appliquée 52(4):93–111

Team RC, et al. (2013) R: A language and environment for statistical computing

Van de Velden M, Iodice D'Enza A, Markos A (2019) Distance-based clustering of mixed data. Wiley Interdisciplinary Reviews: Computational Statistics 11(3):e1456

Vichi M, Vicari D, Kiers HA (2019) Clustering and dimension reduction for mixed variables. Behaviormetrika 46(2):243–269