Exploring drug consumption via an ultrametric correlation matrix

Una analisi del consumo di droghe mediante una matrice di correlazione ultrametrica

Giorgia Zaccaria and Maurizio Vichi

Abstract In many real applications, the existence of a general concept (a multidimensional phenomenon) composed of nested specific ones is often theorised. In the specialised literature, different sequential methodologies have been proposed to identify a hierarchy of latent dimensions. In this paper, we investigate drug consumption via an ultrametric correlation matrix which allows to detect different, nonoverlapping groups of drugs and their hierarchical relationships, starting from the correlation matrix of the observed data. Since its social and economic relevance, a model-based approach to drug consumption can provide an in-depth understanding of this challenging phenomenon, which turns out to be fundamental to address policies aimed at reducing it.

Abstract *In molte applicazioni l'ipotesi dell'esistenza di un concetto generale (un fenomeno multidimensionale), definito mediante concetti più specifici, è spesso avvalorata. In letteratura, molteplici metodologie di tipo sequenziale sono state proposte con lo scopo di identificare una gerarchica di dimensioni latenti. In questo articolo indaghiamo il fenomeno del consumo di droghe mediante una matrice di correlazione ultrametrica, che permette di individuare diversi, disgiunti gruppi di droghe e le loro relazioni gerarchiche, a partire dalla matrice di correlazione dei dati osservati. Data la sua rilevanza sociale ed economica, un approccio basato su modello per lo studio del consumo di droghe può fornire una conoscenza più approfondita di tale fenomeno, che a sua volta può risultare fondamentale nella definizione di politiche volte alla sua riduzione.*

Key words: Hierarchical structures, drug consumption, ultrametric correlation matrix, dimensionality reduction

---

Giorgia Zaccaria
University of Rome La Sapienza, P.le Aldo Moro 5 00185, Rome
e-mail: giorgia.zaccaria@uniroma1.it

Maurizio Vichi
University of Rome La Sapienza, P.le Aldo Moro 5 00185, Rome
e-mail: maurizio.vichi@uniroma1.it

1 Introduction

The identification of a hierarchy of nested latent concepts is a considerable aspect in the study of phenomena composed of different facets. Manifold methodologies as higher-order factor models [1, 10] and hierarchical factor models [9, 11] deal with the problem of the construction of a general latent concept via a hierarchy of more specific ones. In order to detect consistent groups of variables and their hierarchical factorial structure, [2] propose a novel exploratory, parsimonious and simultaneous model which is based upon the estimation of an ultrametric correlation matrix to reconstruct the relationships between the observed variables, i.e., within groups of variables and between them.

In many fields as the psychometric and marketing ones, the detection of latent dimensions with different relationship intensities is a crucial need for a correct and all-around understanding of the phenomenon under study, along with the dimensionality reduction of the variable space. In this paper, we demonstrate the large-scale applicability of the model proposed by [2] with its application to the phenomenon of drug consumption. The latter is one of the most challenging problems in the modern societies. Indeed, drug consumption contributes to rise the risk of poor health, crimes, social harm, environmental damage and it has become a social problem over years - especially among young people - governments have to face with. Many studies have been developed to analyse drug consumption, its individual and community effects, e.g., [8]. Therefore, an in-depth analysis of this phenomenon through the aforementioned model - which identifies groups of drugs highly correlated and their (hierarchical) relationships - can contribute to its better understanding and to consequently implement policies aimed at reducing it.

The paper is organised as follows. In Section 2 the methodology used to investigate the phenomenon under study is presented and in Section 3 it is applied on the Drug Consumption data set to stress its usefulness. Finally, in Section 4 some conclusions end the paper.

2 Methodology

The exploratory, parsimonious and simultaneous model, called *Ultrametric Correlation Model* (UCM) and proposed by [2], introduces a novel approach to the identification of hierarchical structures of latent variables (concepts). Indeed, starting from a nonnegative correlation matrix, the UCM estimates highly correlated, nonoverlapping groups of variables and different levels of relationships among them in a least-squares framework. The model is mathematically represented by an ultrametric correlation matrix, whose definition gives rise to a hierarchical structure of variable groups as detailed below. Let us consider the following correlation matrix of order J which is composed of 3 variable groups such that $J_1 + J_2 + J_3 = J$, where J_q, $q = 1,2,3$, is the number of variables in the q^{th} group (g_q).

$$\mathbf{R} = \begin{bmatrix} \begin{matrix} 1 & & & \\ & 1 & & r_{g_1}^{\text{Within}} \\ & & \cdots & \\ & & & 1 \end{matrix} & r_{g_1,g_2}^{\text{Between}} & r_{g_1,g_3}^{\text{Between}} \\ & \begin{matrix} 1 & & & \\ & 1 & & r_{g_2}^{\text{Within}} \\ & & \cdots & \\ & & & 1 \end{matrix} & r_{g_2,g_3}^{\text{Between}} \\ & & \begin{matrix} 1 & & & \\ & 1 & & r_{g_3}^{\text{Within}} \\ & & \cdots & \\ & & & 1 \end{matrix} \end{bmatrix}$$

Within each block of the matrix $\mathbf{R}$, the correlation between variables assumes the same value: $r_{g_q}^{\text{Within}}$, $q = 1,2,3$ represents the correlation within the q^{th} group, whereas $r_{g_q,g_p}^{\text{Between}}$, $q,p = 1,2,3\,(p \neq q)$, represents the correlation between the q^{th} and the p^{th} group. $\mathbf{R}$ is an ultrametric correlation matrix if it satisfies the following conditions[1]:

(i) $r_{g_q,g_p}^{\text{rel}} = r_{g_p,g_q}^{\text{rel}}$ for rel $\in \{\text{Within, Between}\}$, $p,q = 1,2,3$ (symmetry);

(ii) $r_{g_q}^{\text{Within}} \geq \max\{r_{g_q,g_p}^{\text{rel}} : q = 1,2,3\}$ for all $p = 1,2,3$, where rel = Within if $q = p$, rel = Between, otherwise (column pointwise diagonal dominance);

(iii) $r_{g_q,g_p}^{\text{rel}} \geq \min\{r_{g_q,g_h}^{\text{rel}}, r_{g_p,g_h}^{\text{rel}}\}$, for all $q,p,h = 1,2,3$ and rel $\in \{\text{Within, Between}\}$ (ultrametric inequality).

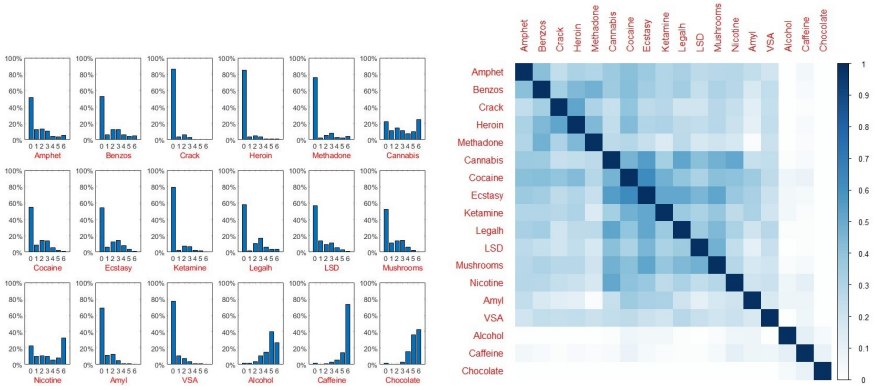
Thus, the ultrametric correlation matrix $\mathbf{R}$ of the UCM model identifies higher correlations within groups of variables and lower (and hierarchically ordered) correlations between those groups by giving rise to a hierarchy of nested latent concepts, each one associated with a variable group. The aforementioned conditions can be easily generalised to Q groups of variables.

3 Drug consumption: evaluation of the hierarchical relationships between groups of drugs

The data set analysed in the paper herein[2] [6] contains information on 1885 respondents, mainly coming from UK (55.58%), USA (29.55%), Canada (4.62%) and Australia (2.86%) and aged from 18 years old, on their drug consumption. Specifically, the use of 18 legal (alcohol, caffeine, chocolate, nicotine) and illegal (am-

[1] The definition of the ultrametric correlation matrix provided herein is based upon [5, pp.58-59].

[2] Drug consumption (quantified) data set available at: `https://archive.ics.uci.edu/ml/datasets/Drug+consumption+\%28quantified\%29`.



(a) Bar chart of the 18 ordinal variables

(b) Correlation matrix of the quantified data via the CatPCA

Fig. 1: Drug consumption data set.

phetamines, amyl nitrite, benzodiazepine, cannabis, cocaine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, Volatile Substance Abuse) drugs is investigated in terms of ordinal variables. The response classes are the following: *Never Used, Used over a Decade Ago, Used in Last Decade, Used in Last Year, Used in Last Month, Used in Last Week* and *Used in Last Day*.

In order to apply the methodology described in Section 2 to investigate the correlation structure among drugs, the ordinal variables - each one representing consumption of a specific drug - have to be quantified. This quantification is implemented via the Categorical Principal Component Analysis (CatPCA) [7] and the correlation matrix of the corresponding quantitative variables is computed. Six correlation coefficients assume negative values (not lower than[3] -0.05) which turn out to be statistically nonsignificant; whereas, the variable *Chocolate* has negative correlations with all the other drugs (Figure 1a) - except for *Alcohol* and *Caffeine* - which are not lower than -0.09 and considered nonsignificant in literature [6]. For this reason, in both cases the negative correlations are set to zero such that the non-negativity condition necessary for the UCM holds (Figure 1b). Furthermore, the number of the variable groups necessary to implement the exploratory, parsimonious model described in Section 2 is set according to the scree plot and it is equal to five. It is worthy of remark that hierarchical clustering methods could be implemented to study the correlation between usage of different drugs, but they would not guarantee the correct identification of the underlying hierarchical structure [3].

The application of the model described in Section 2 to the aforementioned data set provides a representation of drug consumption through the identification of different groups of drugs mostly correlated (Figure 1b), and broader ones defined by merging the initial five groups (Figure 2). In this framework, a model-based ap-

[3] In this case, the term *not lower than* refers to small negative correlation coefficients close to zero.

Table 1: Initial five groups identified by the Ultrametric Correlation Model.

Group	Group Name	Variables
Group 1	Depressant and Artificial Drugs	Ampeth, Benzodiazepine, Crack, Heroine, Methadone
Group 2	Stimulant Drugs and Hallucinogens	Cannabis Cocaine, Ecstasy, Ketamine, Legal highs, LSD, Mushrooms, Nicotine
Group 3	Inhalant Drugs	Amyl nitrite, Volatile Substance Abuse
Group 4	Legal Drugs of Daily Use	Alcohol, Caffeine
Group 5	Chocolate	Chocolate

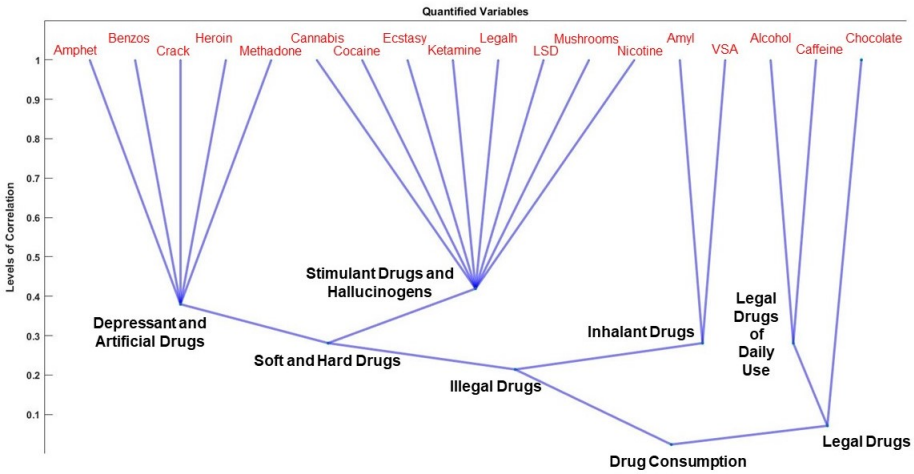


Fig. 2: Path diagram representation of the drug consumption.

proach to analyse correlations between variables can back up the experts' theories on this phenomenon. The initial five groups identified by the model are reported in Table 1. All of them are reliable according to the Cronbach's alpha (α) [4], except for *Inhalant Drugh* and *Legal Drugs of Daily Use*. It is worthy of remark that the Cronbach's alpha of a group is affected by its number of variables.

The hierarchy over the five groups gives rise to broader concepts: *Soft and Hard Drugs* obtained by lumping together Group 1 and Group 2 ($\alpha = 0.87$); *Illegal Drugs* obtained by merging the latter with Group 3 ($\alpha = 0.87$); *Legal Drugs* obtained by lumping together Group 4 and Group 5 ($\alpha < 0.7$). The existence of a general construct representing *Drug Consumption* is assessed through the Cronbach's alpha of the whole data set, which is equal to 0.84. These results turn out to be coherent with the specialised literature on drug consumption (e.g., [6]).

4 Conclusions

In many real applications, the existence of a general concept composed of nested specific ones is often theorised. Manifold sequential methodologies aim at building

a hierarchy of dimensions starting from the observed variables. [2] propose a novel parsimonious and simultaneous model in order to pinpoint latent concepts, each one associated with a variable group, and explore their hierarchical relationships by investigating the correlation matrix of the observed data. In this paper, the aforementioned model is applied to a Drug Consumption data set to study the relationships between groups of drugs. Since its social and economic relevance, a model-based approach to drug consumption analysis can provide a better understanding of the phenomenon which can be fundamental to address policies aimed at reducing it. The results of the application of the model proposed by [2] on the aforementioned data set pinpoint a hierarchy of drug groups which is coherent with the studies on drug consumption (e.g., [6]); this confirms the importance of the UCM in applications where a hierarchical factorial structure can be estimated.

References

1. Cattell, R.B.: The scientific use of factor analysis in behavioral and life sciences. Plenum, New (1978)
2. Cavicchia, C., Vichi, M., Zaccaria, G.: The ultrametric correlation matrix for modelling hierarchical latent concepts. Advances in Data Analysis and Classification, *Accepted*
3. Cavicchia, C., Vichi, M., Zaccaria, G.: Exploring Hierarchical Concepts: Theoretical and Application Comparisons. In: T. Imaizumi, A. Nakayama, S. Yokoyama (eds.) Advanced Studies in Behaviormetrics and Data Science, Springer, Singapore, ISBN: 978-981-15-2699-2, *in press* (2020)
4. Cronbach, L.J.: Coefficient alpha and the internal structure of tests. Psychometrika **16**(3), 297–334 (1951)
5. Dellacherie, C., Martinez, S., San Martin, J.: Inverse M-matrices and ultrametric matrices. Springer International Publishing, Lecture Notes in Mathematics (2014)
6. Fehrman, E., Muhammad, A.K., Mirkes, E.M., Egan, V., Gorban, A.N.: The Five Factor Model of personality and evaluation of drug consumption risk, arXiv (2015) Available at `https://arxiv.org/abs/1506.06297`
7. Gifi, A.: Nonlinear Multivariate Analysis. Wiley, New York (1990)
8. McGinnis, J.M., Foege, W.H.: Actual causes of death in the United States. JAMA **270**(18), 2207—2212 (1993)
9. Schmid, J., Leiman, J.M.: The development of hierarchical factorial solutions. Psychometrika **22**(1), 53–61 (1957)
10. Thompson, G.H.: The factorial analysis of human ability. Houghton Mifflin, New York (1948)
11. Wherry, R.J.: Hierarchical factorial solutions without rotation. Psychometrika **24**(1), 45–51 (1959)