DSSV 2023
Data Science, Statistics & Visualisation
ECDA 2023
European Conference on Data Analysis

Hierarchical variable clustering using singular value decomposition

Jan O. Bauer[1]

[1]Departmentof Econometrics and Data Science, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

April 30, 2023

In multivariate analysis, finding latent variables serves as an initial step to interpret data. However, simplifying the underlying population by a reduced number of latent variables skims only the surface. Detecting nested structures by using hierarchical variable clustering are further steps to facilitate relations among variables and therefore to deepen the understanding of the underlying random vector.

The relation among variables is represented by their shared variance. Among others, [1] recently proposed a factor model to identify the hierarchical variable structure that fits the sample best. The covariance matrix of the respective models is in the shape of a block diagonal matrix because each block represents the nested structure of a latent variable.

In this work, we provide a new concept that detects the underlying hierarchical variable structure using singular value decomposition. Following the Davis-Kahan theorem, [2] showed that singular vectors can be exploited to detect the underlying block diagonal structure of a covariance matrix. We extent this approach to find the nested structure of the latent variables by iteratively cracking each block into smaller blocks and therefore develop a divisive clustering approach. Also, no assumptions about the distribution of the underlying random vector are made, which makes it feasible for all kinds of cross-sectional data.

The hierarchical clustering structure that is easiest to interpret is not necessarily be the one that fits the underlying sample. There are measures to evaluate if the given structure represents the sample reasonably. However, these measures take into account the whole clustering structure and therefore do not provide information about the fit of single clusters. We provide a measure that evaluates each cluster by considering the ratio of the variance of each block to its conditional variance. This measure guides to the structure that reasonably reflects the sample by adjusting the nesting of variables.

We note that in the high dimensional case when the number of variables is larger than the number of observations, the conditional covariance matrix provides no meaningful contribution. We address this case using the connection between linear regression and the detection of a block diagonal covariance matrix.

We further illustrate the performance of the new concept for hierarchical variable clustering as well as our contributed evaluation measure with simulations and on real datasets.

Keywords: Covariance Structure, Dimensionality Reduction, Hierarchical Models, Latent Variables, Principal Loading Analysis.

[1] C. Cavicchia and M. Vichi, "Second-order disjoint factor analysis," *Psychometrika*, vol. 87, p. 289–309, 2022.
[2] J. O. Bauer and B. Drabant, "Principal loading analysis," *J. Multivariate Anal.*, vol. 184, 2021.