Hierarchical Disjoint Non-Negative Factor Analysis: environment and waste management in the EU

Analisi Fattoriale Gerarchica Disgiunta e Non-Negativa: dati ambientali e gestione dei rifiuti in UE

Carlo Cavicchia, Pasquale Sarnacchiaro and Maurizio Vichi

1 Introduction

We are more than 7.5 billion people on our planet, and we are producing waste every day. Although the management of the waste keeps improving in the EU, many estimations tell us that half of that waste is not collected, treated or safely disposed of. That is why policymakers need consistent and useful tools to measure and monitor quality and efficiency of waste collection and management.

A multidimensional phenomenon like waste management is described by a huge quantity of information useful for making strategical decisions and the demand for statistics on waste generation and treatment has grown considerably in recent years. This amount of information needs to be synthesised by studying relationships among manifest indicators. It is important to find the relationships among dimensions and indicators in order to synthesise the information and have a response on the conduct of each country to achieve the priority goals set by Europe, reducing waste generation and maximising recycling and re-using. Identifying these relationships could be fundamental to understand where each country can focus its actions and what impacts each action could have. A "good" waste management is vital for global suistainable development, it is connected with SDGs.

A usual way to synthesise a big amount of information is using Composite Indicators (CIs), that is, non-observable latent variables, linear combinations of observed variables [1].

Carlo Cavicchia
University of Rome Unitelma Sapienza, Viale Regina Elena 295, Rome,
e-mail: carlo.cavicchia@unitelmasapienza.it

Pasquale Sarnacchiaro
University of Rome Unitelma Sapienza, Viale Regina Elena 295, Rome,
e-mail: pasquale.sarnacchiaro@unitelmasapienza.it

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

Bini M., Amenta P., D'Ambra A., Camminatiello I. (Editors). Statistical Methods for Service quality evaluation. Cuzzolin Publisher.

In this paper we propose a CI for quality and efficiency of waste collection and management in Europe by using a model-based approach. The model has a hierarchical structure formed by factors associated to subsets of manifest variables with positive loadings. This approach guarantees to comply with all the good properties on which an indicator - summarising a multidimensional phenomenon - should be based. Such properties might be: model-based, statistically estimated (i.e., non-normative), with a hierarchical structure, scale-invariant, and non-compensable. Moreover the hierarchical structure allows us to find a set of specific composite indicators which are unidimensional and reliable.

2 Hierarchical Disjoint Non-Negative Factorial Analysis

2.1 Model

Hierarchical Disjoint Non-Negative Factorial Analysis (HDNFA) [3] is a factorial model that considers two typologies of latent unknown constructs: H specific factors and a single (nested) general factor. HDNFA is identified by the two simultaneous equations:

$$\mathbf{x} - \mu_{\mathbf{x}} = \mathbf{A}\mathbf{y} + \mathbf{e}_{\mathbf{x}} \tag{1}$$

$$\mathbf{y} = \mathbf{c}\mathbf{g} + \mathbf{e}_{\mathbf{y}} \tag{2}$$

where $\mathbf{A}$ is the $J \times H$ matrix of unknown specific factors loadings, $\mathbf{c}$ is the $H \times 1$ vector of unknown general factor loadings, $\mathbf{e}_{\mathbf{x}}$ and $\mathbf{e}_{\mathbf{y}}$ are a $J \times 1$ and a $H \times 1$ random vector of errors, respectively.

Let include model 2 into model 1 and considering the loading matrix $\mathbf{A}$ is restricted to the product $\mathbf{A} = \mathbf{B}\mathbf{V}$ [2], the HDFA model is defined

$$\mathbf{x} - \mu_{\mathbf{x}} = \mathbf{B}\mathbf{V}(\mathbf{c}\mathbf{g} + \mathbf{e}_{\mathbf{y}}) + \mathbf{e}_{\mathbf{x}} \tag{3}$$

Let rewrite the model 3 in matrix form

$$\mathbf{X} = \mathbf{g}\mathbf{c}'\mathbf{V}'\mathbf{B} + \mathbf{E}_{\mathbf{x}} \tag{4}$$

The variance-covariance structure related to the model 3 is

$$\Sigma_{\mathbf{x}} = \mathbf{B}\mathbf{V}(\mathbf{c}\mathbf{c}' + \Psi_{\mathbf{y}})\mathbf{V}'\mathbf{B} + \Psi_{\mathbf{x}} \tag{5}$$

where

$$\Sigma_{\mathbf{y}} = \mathbf{c}\mathbf{c}' + \Psi_{\mathbf{y}} \tag{6}$$

such that

$$\mathbf{V} = [\mathbf{v}_{jh} : \forall \mathbf{v}_{jh} \in \{0,1\}] \tag{7}$$

$$\mathbf{V1}_H = \mathbf{1}_J \tag{8}$$

$$\mathbf{B} = diag(b_1, \ldots, b_J) \; with \; b_j^2 > 0 \tag{9}$$

$$\mathbf{V'BBV} = diag(b_{\cdot 1}^2, \ldots, b_{\cdot H}^2) \; with \; b_{\cdot h}^2 = \sum_{j=1}^{J} b_{jh}^2 > 0 \tag{10}$$

It is assumed that $\mathbf{y}$ and $\mathbf{g}$ are standard normal distributed. $\mathbf{e_x}$ and $\mathbf{e_x}$ are normal distributed with a J-dimensional and H-dimensional diagonal positive definite variance-covariance matrix, respectively. In addition it is assumed that errors in the two models are uncorrelated $Cov(\mathbf{e_x}, \mathbf{e_y}) = 0$; and errors and factors are uncorrelated, i.e.,$Cov(\mathbf{e_x}, \mathbf{g}) = 0$ and $Cov(\mathbf{e_y}, \mathbf{g}) = 0$.

2.2 Estimation

Suppose that a random sample of $n > J$ multivariate observations of $\mathbf{x}$ is observed, the maximisation of the log-likelihood with respect to μ gives the sample mean, thus the reduced log-likelihood is as follows

$$L(\mathbf{x}_i, \mathbf{A}, \Psi_{\mathbf{x}}, \Psi_{\mathbf{y}}) = \tag{11}$$
$$= -\frac{nJ}{2} ln 2\pi - \frac{n}{2}\{ln|\mathbf{A}(\mathbf{cc}' + \Psi_{\mathbf{y}})\mathbf{A}' + \Psi_{\mathbf{x}}| + tr\{[\mathbf{A}(\mathbf{cc}' + \Psi_{\mathbf{y}})\mathbf{A}' + \Psi_{\mathbf{x}}]^{-1}\mathbf{S}\}\}$$

where $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \mu_{\mathbf{x}})'\Sigma_{\mathbf{x}}^{-1}(\mathbf{x}_i - \mu_{\mathbf{x}})$
This is equivalent to the minimization of the discrepancy function

$$D(\mathbf{x}_i, \mathbf{A}, \Psi_{\mathbf{x}}, \Psi_{\mathbf{y}}) = ln|\mathbf{A}(\mathbf{cc}' + \Psi_{\mathbf{y}})\mathbf{A}' + \Psi_{\mathbf{x}}| + tr\{[\mathbf{A}(\mathbf{cc}' + \Psi_{\mathbf{y}})\mathbf{A}' + \Psi_{\mathbf{x}}]^{-1}\mathbf{S}\} \tag{12}$$

This is a discrete and continuous problem that cannot be solved by a quasi-Newton type algorithm, it is solved by a descendent coordinate algorithm. A general composite indicator should be composed by consistent and reliable specific composite indicators; thus we require that loadings must be positive during the estimation of $\mathbf{Y}$ and $\mathbf{g}$. So the discrepancy function 12 is minimised with respect to $\mathbf{B}_h = diag(\mathbf{b}_h)$ by

$$\widehat{\mathbf{b}}_h = \widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} \mathbf{u}_{1h}(\lambda_{1h} - 1)^{\frac{1}{2}} \tag{13}$$

where λ_{1h} and $\mathbf{u}_{1h}$ are respectively the largest eigenvalue and the corresponding eigenvector of the variance-covariance matrix $\widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} \mathbf{S}_h \widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}}$ corresponding to variables identified by $\mathbf{v}_{\cdot h}$, that corresponds to h-th column of $\mathbf{V}$. It is important to notice that λ_{1h} and $\mathbf{u}_{1h}$ minimise the function

$$||\mathbf{X}_h \widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} - \sqrt{\lambda_{1h}} \mathbf{y}_h \mathbf{u}_{1h}'||^2 \tag{14}$$

where $\mathbf{X}_h$ is the centred data matrix. That can be solved by an Alternate Non-Negative LS algorithm, such that $\widehat{\mathbf{y}}_h$ is estimated by a step of a normal ALS while the estimations of $\widehat{\mathbf{u}}_{1h}$ consists . thus given $\widehat{\mathbf{u}}_{1h}$, $\widehat{\mathbf{y}}_h$ is computed by

$$\widehat{\mathbf{y}}_h = \mathbf{X}_h \widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} \widehat{\mathbf{u}}_{1h} (\widehat{\mathbf{u}}_{1h}' \widehat{\mathbf{u}}_{1h})^{-1} \tag{15}$$

and given $\mathbf{y}_h$, $\mathbf{u}_{1h}$ is computed by

$$\widehat{\mathbf{u}}_{1h} = \begin{cases} \mathbf{X}_{h+} \widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} \widehat{\mathbf{y}}_h (\widehat{\mathbf{y}}_h' \widehat{\mathbf{y}}_h)^{-1} \\ 0, \qquad \text{otherwise} \end{cases} \tag{16}$$

where X_{h+} is the set of variable that if considered into the estimation of $\mathbf{u}_{1h}'$ in 14, they return a positive value.

3 Application

The Hierarchical Disjoint Non-Negative Factor Analysis has been applied on a dataset composed by 39 indicators about waste generation, circular economy and recycling, for the 28 EU countries. Many variables about the characteristic of countries have been considered in order to help the interpretation of the results. The indicators into the dataset come from different sources: Eurostat, Joint Research Centre (JRC), Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs (DG GROW) and the European Patent Office. Eurostat regularly updates the indicators on its website and they are free.

In our application, we propose hierarchically aggregated index that best represents the quality of the waste collection and treatment in EU, via the statistical identification of reliable and unidimensional specific composite indicators.

Some variables are included into the analysis in order to enrich the information about countries and their performance in waste management (e.g., density population, number of days of rain per year, average annual temperature, etc, ...).

References

1. OECD: The OECD-JRC Handbook on Practices for Developing Composite Indicators, paper presented at the OECD Committee on Statistics, 7-8 June 2004, OECD, Paris.
2. Vichi, M.: Disjoint factor analysis with cross-loadings. Advances in Data Analysis and Classification (2017) doi: 10.1007/s11634-016-0263-9
3. Vichi, M., Cavicchia, C.: Hierarchical Disjoint Non-Negative Factor Analysis for Modelling Composite Indicators. Unpublished manuscript. Last date modified: January 2019 . Microsoft Word file.