A composite indicator via hierarchical disjoint factor analysis for measuring the Italian football teams' performances

Carlo Cavicchia [a,b], Pasquale Sarnacchiaro [a], Maurizio Vichi [b]
[a] Department of of Law and Economics, University of Rome Unitelma Sapienza, Rome, Italy;
[b] Department of Statistical Sciences, University of Rome La Sapienza, Rome, Italy.

1. Introduction

In the last years, with the data revolution and the use of new technologies, phenomena are frequently described by a huge quantity of information useful for making strategical decisions. In the current "big data" era, the interest of statistics into sports is increasing over the years. Football is assuredly the most popular sport in Italy, with millions of supporters and amateurs in every cities. Football has also become one of the most profitable industries, with a significant economic impact in infrastructure development, sponsorships, TV rights and transfers of players.

Sportive and economic data are collected for all teams which use statistical analysis in order to measure and improve their performances. The main goal of any Football Championship club is to achieve sport results, by trying to increase their turnover as well.

For dealing with all this amount of information, an appropriate statistical analysis is needed. A priority is having statistical tools useful to synthesise the information arised from the data. Such tools are represented by composite indicators (CIs), that is, non-observable latent variables and linear combinations of observed variables.The strategy of construction of a CI used in this paper is based on a non-negative disjoint and hierarchical model for a set of quantitative variables. This is a factor model with a hierarchical structure formed by factors associated to subsets of manifest variables with non-negative loadings.

In according to the Handbook on Constructing Composite Indicators of the OECD OECD (2004), where the Factor Analysis (FA) methodology is presented as a weighting method used to combine observed indicators, we propose a hierarchical model with the non-negative loadings which best reconstructs the observed indicators according to the common factor model estimated by Maximum Likelihood Estimation (MLE) method. Therefore, loadings are not subjective, but statistically estimated summarizing the observed common relation among data. By hypothesising a two levels hierarchy, the complete system of loadings that best reconstruct the data according to the model is simultaneously estimated.

In this paper, we propose a CI for measuring the Italian football teams' performances, in terms of both sportive and economic variables.

The paper is organised as follows. In Section 2 a description of the methodology is provided. The real application on Italian football teams' performances is presented in Section 3.

2. Hierarchical Disjoint Non-Negative Factorial Analysis

Hierarchical Disjoint Non-Negative Factorial Analysis (*HDNFA*) Cavicchia et al. (2019) 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 **A** is the $(J \times H)$ matrix of unknown specific factors loadings, **c** is the $(H \times 1)$ vector of unknown general factor loadings, $\mathbf{e_x}$ and $\mathbf{e_y}$ 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 **A** is restricted to the product $\mathbf{A} = \mathbf{BV}$ Vichi (2017), the HDFA model is defined

$$\mathbf{x} - \mu_{\mathbf{x}} = \mathbf{BV}(\mathbf{cg} + \mathbf{e_y}) + \mathbf{e_x} \tag{3}$$

Let rewrite the model 3 in matrix form

$$\mathbf{X} = \mathbf{gc'V'B} + \mathbf{E_x} \tag{4}$$

The variance-covariance structure related to the model 3 is

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

where

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

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 $\mathbf{y} \sim N_H(0, \Sigma_{\mathbf{y}})$ where $\Sigma_{\mathbf{y}}$ is the correlation matrix of the specific factors since they are standardised, and $\mathbf{e_x} \sim N_J(0, \Psi_{\mathbf{x}})$, where $Cov(\mathbf{e_x}) = \Psi_{\mathbf{x}}$ is the J-dimensional diagonal positive definite variance-covariance matrix of the error of model 1 and $Cov(\mathbf{e_x}, \mathbf{y}) = 0$. Furthermore, **g** is the random general factor with mean 0 and variance $\sigma_{\mathbf{g}}^2 = 1$ denoting the composite indicator related to a reduced set of specific factors. In addition, $\mathbf{e_y}$ is a non-observable $(H \times 1)$ random vector of errors. It is assumed that $\mathbf{g} \sim N(0, 1)$ and $\mathbf{e_y} \sim N_H(0, \Psi_{\mathbf{y}})$, where where $Cov(\mathbf{e_y}) = \Psi_{\mathbf{y}}$ is the H-dimensional diagonal positive definite variance-covariance matrix of the error of model 2. 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 $Cov(\mathbf{e_y}, \mathbf{g}) = 0$.

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

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

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 passive variables. Thus, this is an active set algorithm, where the H inequality constraints are active if the regression coefficient $\mathbf{u}'_{1h}$ in 14 will be negative (or zero) when estimated unconstrained, otherwise constraints are passive. The non-negative solution of 10 with respect to $\mathbf{u}_{1h}$ will simply be the unconstrained least squares solution using only the variables corresponding to the passive set, setting the regression coefficients of the active set to zero.

3. Application

Football teams' performances are complex phenomena, described by a huge quantity of information regarding sportive results and economic indices. The main goal of any Football club is to aim sport results, by taking under control the economic aspect. It is more and more important to find the way to measure football teams' performances in order to provide support for decision making. The number of statistics and measures related to sports is expanding every year and the need of build aggregated index to monitor the teams' behavior is even more important.

The Hierarchical Disjoint Non-Negative Factor Analysis has been applied on a dataset obtained from the financial statements filed by the Serie A football teams.

The indicators into the dataset come from different sources: Engsoccerdata, Opta and Transfermarkt. They are regularly updated and they are free.

In our application, we propose a hierarchically aggregated index that best represents the performances of the football teams in terms of sportive and economic conduct, via the statistical identification of reliable and unidimensional specific composite indicators, which are dimensions that measure specific concepts describing the main components of the football italian teams' performances.

In particular, we analyse the impact that all variables have on points made by football teams participating in the series A championship. Some variables are included into the analysis in order to enrich the information about teams. This approach guarantees good properties for the GCI such as (scale-invariance, non-compensability, non-negativity, reliability, unidimensionality, . . .).

References

OECD (2004). *The OECD-JRC Handbook on Practices for Developing Composite Indicators, paper presented at the OECD Committee on Statistics*.

Cavicchia, C., Vichi, M. (2019). Hierarchical Disjoint Non-Negative Factor Analysis. *Submitted manuscript*.

Vichi, M. (2017). Disjoint factor analysis with cross-loadings. *Advances in Data Analysis and Classification*, **11**(3), pp. 563–591.