Dimensions of well-being and their statistical measurements

Carla Ferrara, Francesca Martella, Maurizio Vichi

Abstract Nowadays, a relevant challenge regards the assessment of a global measure of well-being by using composite indicators of different features such as level of wealth, comfort, material goods, quality and availability of education, living standard, etc. The employ of a unique measure, as a consensus of several indicators, in general allows to better understand and synthesize the underlying processes with a more accurate picture of the social progress and it is useful for giving a better information to citizens and policy makers. In this paper, we focus on statistical methodologies designed to build composite indicators of well-being by detecting latent components and assessing the statistical relationships among indicators. We will consider Principal Component Analysis (PCA) and a constrained PCA version, which allows to specify disjoint classes of variables with an associated component of maximal variance. Furthermore, we will take into account the Structural Equation Model (SEM). These methodologies will be compared by using a data set from 34 member countries of the OECD [4].

1 Introduction

In recent years, the assessment of the subjective perception of overall well-being of citizens has received increasing attention. Governments and economists are aware that Gross Domestic Product (GDP) and other traditional measures of economic progress fail to measure the kind of progress that increase the quality of the life. Well-being depends on a number of factors not related exclusively to the economic and materials elements but also to lifestyle, food choices, health condition and environment. In this empirical context, a potential goal of statistical approaches is the identification of new methods which can synthesize various aspects of well-being: both social and economic and its sustainability. Since different layers of well-being are complex and hide underlying features, several indicators have been proposed to resume them [7].

In the literature on building indicators, mainly two aspects have been investigated: i) the identification of key indicators to be used; ii) the ways in which these indicators can be brought together to make a coherent system of information. While the first aspect is studied by economists, sociologists and psychologists, the second one is a field of research for statisticians, who have to provide rigorous tools to aggregate and synthesize indicators in order to build composite indicators. Whereas the use of a unique measure obtained by combining single indicators is an appealing challenge to capture well-being reality, its building may face specific questions. Problems occur in how to choose, aggregate and weight the single indicators among a suite of indicators available, or in how to identify the components driving the composite indicators. Therefore, the selection of the weights and the way the indicators are combined are very sensitive and important questions to not oversimplify a complex system and to not give potentially misleading signals.

In this paper, different methodologies on constructing composite indicators of well-being are discussed. Starting from Principal Component Analysis (PCA) we consider a constrained version which allows to specify disjoint classes of variables and a component of maximal variance for each class [10]. Structural Equation Model (SEM, [1]) is also considered to assess relationships among variables.

2 Constrained PCA

The first approach, as a dimensional reduction method, seems to be the most natural tool to compute composite indicators. In fact, by using PCA, the computation of the weights becomes not subjective, but based on the common statistical relations among single indicators. Moreover, this technique allows us to assess the impact of each single indicator on the composite indicator and, to be used as a underlying well-being latent dimension. However, often the different indicators used in the construction of a composite indicator express different aspects of a complex phenomenon, and, therefore, they could be conceptually split in several blocks of indicators, where each block can be resumed by a composite indicator.

Given the $(n \times J)$ two-way two-mode (objects and variables) data matrix $\mathbf{X}=[x_{ij}]$, describing the J-variate profiles of n objects. Variables are supposed commensurate, and therefore if they are expressed by different units of measurements they are standardized to have mean zero and unit variance;

The model associated to the *disjoint principal component analysis* (DPCA) can be formally written as follows

$$\mathbf{X}= \mathbf{XAA}' + \mathbf{E}, \quad (1)$$

where **A** is the component loading matrix with generally reduce rank, i.e., rank(**A**)=$Q \leq J$, satisfying constraints

$$\sum_{j=1}^{J} a_{jq}^2 = 1 \qquad q=1, \ldots, Q \tag{2}$$

$$\sum_{j=1}^{J} (a_{jq} a_{jr})^2 = 0 \qquad q=1,\ldots,Q\text{-}1;\ r=q+1,\ldots,Q$$

(3)

Matrix **A** is orthonormal; while matrix **Y**=**XA** specifies a reduced set of non observable latent variables corresponding to composite indicators of subsets of variables. **E** is a error component matrix. Note that constraint (3) is more restrict than the usual orthogonal constraint. Model (1) is the factorial model specifying the dimensionality reduction via the component loading matrix **A**, which allows to partition variables into classes summarized by an orthonormal linear combination with maximal variance.

Matrix **A** may be re-parameterized into the product of two matrices: **A**=**BV**, where **V** = $[v_{jq}]$, is a $(J \times Q)$ binary and row stochastic matrix defining a partition of variables into Q clusters, with v_{jq}=1, if the j^{th} variable belongs to q^{th} cluster, v_{jq}=0, otherwise; while, **B** is a $(J \times J)$ diagonal matrix weighting variables and such that

$$\sum_{j=1}^{J} v_{jq} b_j^2 = 1 \qquad \sum_{q=1}^{Q} \sum_{j=1}^{J} v_{jq} b_j^2 = Q\,.$$

Matrix **A** has the pattern of a membership binary classification matrix. Therefore, constraints (2) and (3) can be equivalently rewritten as

V binary and row stochastic (4)

$\mathbf{v}_q' \mathbf{B}' \mathbf{B} \mathbf{v}_q = 1$, for $q=1,\ldots,Q$. (5)

Model (1) subject to constraints (2) and (3) or (4) and (5) can be considered the non clustering version of the CDPCA [10].

3 SEM and PLS Path-Modeling

Of course PCA and DPCA do not describe the potential causal relationships existent among composite indicators. On the other hand, to make more flexible the system of composite indicators and in order to model causal relationships among them, SEM can be used. In its more general form, SEM is mainly used to assess relations among observed and latent variables. It arises from the combination of Path Analysis model [6] and Factor Analysis model [9]. The former allows to describe the causal relationships among variables, while the latter allows to describe complex phenomenon through observed variables by using latent variables. Therefore, SEM enables to simultaneously analyze the latent aspects underlying specific indicators and their potential causal relationships. In fact, it takes into account not only the multiplicity of causes that act on dependent variables, but also the connections between different causes. In the literature on SEM framework, there are two different approaches for estimating SEM parameters: the covariance-based and the component-based techniques.

The first approach, which includes Maximum Likelihood estimation method (ML-SEM or LISREL, [3]), has been for many years the only estimation method aiming at reproducing the sample covariance matrix of the observed variables through the model parameters. On the other hand, the second approach, also known as PLS Path Modeling (PLS-PM, [8]), was developed as an alternative approach to LISREL, as a more flexible technique for the treatment of a huge amount of data characterized by missing values, highly correlated variables and small sample sizes with respect to the number of variables. It provides estimates of the latent variables in such a way that they are the most correlated with each other, according to a path diagram structure, and the most representative of each corresponding block of manifest variables. In this prospective, it allows us to build each composite indicator as the most representative of each corresponding indicator and the most correlated with the others linked composite indicators.

In this paper, we discuss and compare and combine PCA, DPCA and SEM (in particular by using the PLS-PM approach) performances by using the OECD data set described in [4]. In particular, this work analyzes 34 member countries considering well-being under three aspects: material living conditions, quality of life, sustainability, similarly to the paper proposed by [2]. Based on previous works of leading the reflection on better ways to measure progress [5], OECD [4] identified key topics which are essential to well-being in terms of material living conditions (housing, income, jobs) and quality of life (community, education, environment, governance, health, life satisfaction, safety and work-life balance).

The obtained results highlight the differences underlying the proposed methodologies and affirm the appropriateness of PLS-PM approach together with PCA and DPCA for future well-being research lines.

References

1. Bollen, K. A.: Structural equations with latent variables, Wiley, New York (1989)
2. Hall, J., Giovannini, E., Morrone, A. and Ranuzzi, G.: "A Framework to Measure the Progress of Societies", OECD Statistics Working Papers, No. 2010/05, OECD, Paris (2010)
3. Jöreskog, K.: A general method for analysis of covariance structure, Biometrika, 57, 239-251 (1970)
4. OECD, How's Life?: Measuring well-being, OECD Publishing (2011)
5. OECD, Compendium of OECD well-being indicators, OECD Publishing, Paris (2011a)
6. Retherford, R., Choe, M.K.: Statistical models for causal analysis, Wiley-Interscience Pubblication (1993)
7. Saisana, M., Tarantola, S.: State-of-the-art Report on Current Methodologies and Practices for Composite Indicator Development, EUR 20408 EN, European Commission-JRC (2002)
8. Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.M., Lauro, C.: PLS path modeling, Computational Statistics and Data Analysis, 48, 159-205 (2005)
9. Thurstone, L.L.: Multiple-Factor Analysis. Chicago: University of Chicago Press (1947)
10. Vichi, M., Saporta, G.: Clustering and Disjoint Principal Component. Computational Statistics & Data Analysis, vol. 53, 8, pp. 3194-3208 (2009)