ON THE CONSTRUCTION OF COMPOSITE INDICES BY PRINCIPAL COMPONENTS ANALYSIS[1]

Matteo Mazziotta, Adriano Pareto

1. Introduction

Social and economic phenomena such as development, poverty, quality of life, innovation and competitiveness, are very difficult to measure and evaluate since they are characterized by a multiplicity of aspects or dimensions. The complex and multidimensional nature of these concepts requires the definition of intermediate objectives whose achievement can be observed and measured by individual indicators. A mathematical combination (or aggregation as it is termed) of a set of indicators that represent the different dimensions of a phenomenon to be measured is called *composite index* (Saisana *et al*., 2002).

The literature on the methods for constructing composite indices is vast and the number of composite indices in the world is growing year after year (Bandura, 2008). Examples of well-known composite indices are the United Nations' *Human Development Index* – HDI – (UNDP, 2010) and the European Commission's *Regional Competitiveness Index* – RCI – (Annoni and Kozovska, 2010).

Producing a single composite index has advantages, such as simplicity, although a set of individual indicators might be preferable for other reasons, such as completeness of the information. However, the procedure for constructing a composite index is very far from being aseptic and requires a number of subjective decisions to be taken (OECD, 2008; Mazziotta and Pareto, 2013).

A fundamental issue often overlooked in composite index construction is the definition of the model measurement, in order to specify the relationship between the concept to be measured and its measures (individual indicators). In this respect, the direction of the relationship is either from the concept to the measures - *reflective model* - or from the measures to the concept - *formative model* (Maggino, 2014).

[1] The paper is the result of combined work of the authors: Matteo Mazziotta has written Sects. 1 and 5; Adriano Pareto has written Sects. 2, 3 and 4.

In this paper, we compare the two approaches, and we show that factorial methods, such as Principal Components Analysis (PCA), may fail if improperly used.

2. Formative versus reflective measurement models

As is known, a model of measurement can be conceived through two different conceptual approaches: reflective or formative (Diamantopoulos *et al*., 2008).

The most popular approach is the reflective model, according to which individual indicators denote effects (or manifestations) of an underlying latent variable. Therefore, causality is from the concept to the indicators and a change in the phenomenon causes variation in all its measures. In this model, the concept exists independently of awareness or interpretation by the researcher, even if it is not directly measurable.

Specifically, the latent variable R represents the common cause shared by all indicators X_i reflecting the concept, with each indicator corresponding to a linear function of the underlying variable plus a measurement error:

$$X_i = \lambda_i R + \varepsilon_i \tag{1}$$

where X_i is the indicator *i*, λ_i is a coefficient (loading) capturing the effect of R on X_i. and ε_i is the measurement error for the indicator *i*. Measurement errors are assumed to be independent and unrelated to the latent variable.

A fundamental characteristic of reflective models is that indicators are interchangeable (the removal of an indicator does not change the essential nature of the underlying concept) and correlations between indicators are explained by the measurement model. A typical example is the intelligence of a person: it is the 'intelligence level' that determines the answers to a questionnaire for measuring attitude, not vice versa.

The second approach is the formative model, according to which individual indicators are causes of an underlying latent variable, rather than its effects. Therefore, causality is from the indicators to the concept and a change in the phenomenon does not necessarily imply variations in all its measures. In this model, the concept is defined by, or is a function of, the observed variables.

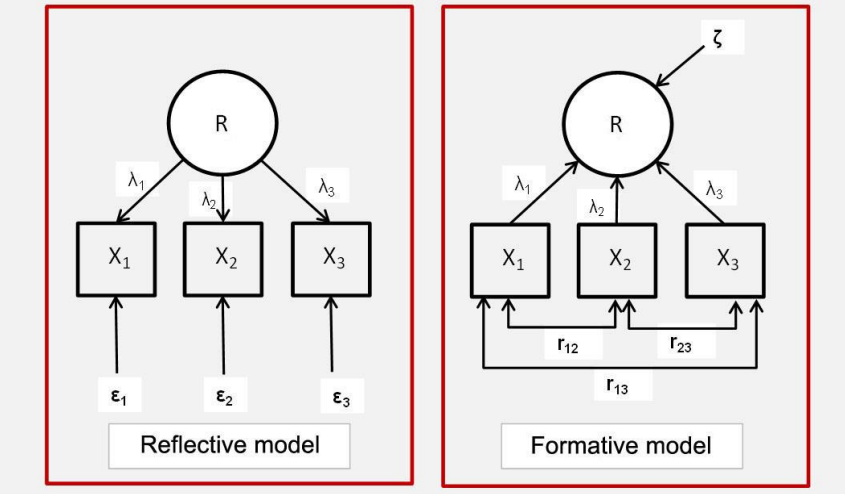
The specification of the formative model is:

$$R = \sum_i \lambda_i X_i + \zeta \tag{2}$$

where λ_i is a coefficient capturing the effect of X_i on R, and ζ is an error term.

In this case, indicators are not interchangeable (omitting an indicator is omitting a part of the underlying concept) and correlations between indicators (r_{ij}, $i \neq j$) are not explained by the measurement model. A typical example is the socio-economic status of a person: it is the 'status level' that depends on education, income, occupation, and residence, not vice versa.

Figure 1 – *Alternative measurement models*



Note that (1) is a simple regression equation where the individual indicator is the dependent variable and the latent variable is the explanatory variable; whereas (2) represents a multiple regression equation where the latent variable is the dependent variable and the indicators are the explanatory variables.

In fig. 1, the two different approaches are graphically represented. Although the reflective model dominates the psychological and management sciences, the formative model is common in economics and sociology (Coltman *et al*., 2008).

3. A numerical example

Imagine that we want to construct a composite index of development in the work dimension, for several countries or regions, based on:

X_1 = Employment rate;

X_2 = Incidence rate of occupational injuries.

Indicator X_1 has positive polarity[2] (it is positively correlated with the development), whereas indicator X_2 has negative polarity (it is negatively correlated with the development).

Suppose also that X_1 and X_2 are positively correlated, so that high employment rates tend to be associated with higher rates of occupational injuries.

In a formative view, we can aggregate the data by arithmetic mean, whereas in a reflective view, the first principal component is the best solution.

In table 1 is reported an example where five countries are considered. The table also provides the normalized indicators[3] Z_1 and Z_2, the arithmetic mean of the normalized values M_1, and the first component[4] score PC_1. Note that $r(X_1, X_2)=0.45$, whereas $r(Z_1, Z_2)=-0.45$, because the polarity of X_2 has been inverted in order to construct the composite index[5].

Table 1 – *Comparing arithmetic mean and first component score as composite indices*

Country	Original values		Normalized values		Ranks		M_1		PC_1	
	X_1	X_2	Z_1	Z_2	R_1	R_2	Value	Rank	Value	Rank
1	80.0	1.9	1.58	-0.71	1	4	0.44	2	1.20	1
2	50.0	2.0	0.00	-1.41	3	5	-0.71	5	0.74	2
3	50.0	1.8	0.00	0.00	3	3	0.00	3	0.00	3
4	50.0	1.6	0.00	1.41	3	1	0.71	1	-0.74	4
5	20.0	1.7	-1.58	0.71	5	2	-0.44	4	-1.20	5
Mean	**50.0**	**1.8**	**0.00**	**0.00**						
Std Dev	**19.0**	**0.1**	**1.00**	**1.00**						

As we can see, units 2, 3, and 4 have the same employment rate ($X_1=50.0$) and decreasing values of the rate of occupational injuries. Nevertheless, unit 2 ranks 5th according to M_1 and ranks 2nd according to PC_1, whereas unit 4 ranks 1st according to M_1 and ranks 4th according to PC_1.

[2] The 'polarity' of a individual indicator is the sign of the relation between the indicator and the phenomenon to be measured. For example, in the case of development, "Life expectancy" has positive polarity, whereas "Infant mortality rate" has negative polarity.

[3] Normalization is required to make the indicators comparable as they often have different measurement units or polarities. In this case, we transformed individual indicators into *z*-scores and we changed the sign if the polarity is negative.

[4] The first principal component accounts for 72.4% of the variance in the data.

[5] When a composite index must be constructed, all the individual indicators must have positive polarity, so that an increase in each indicator corresponds to an increase in the composite index (Mazziotta and Pareto, 2013).

So, the average Spearman rank correlation coefficient between the composite index and the individual indicators is 0.52 for M_1 and 0.05 for PC_1. This is due to the fact that PCA ignores the polarity of the indicators and all normalized indicators, in a reflective measurement model, must be positively intercorrelated.

Therefore, the use of PC_1 for aggregating X_1 and X_2 is incorrect either from a theoretical point of view (PC_1 is a reflective composite index, whereas M_1 is a formative composite index) or from a purely numerical point of view (PC_1 is concordant with both X_1 and X_2, whereas M_1 is concordant with X_1 and discordant with X_2).

4. Some other issues

The PCA has a number of excellent mathematical properties (Kendall and Stuart, 1968). The most important property is that the index obtained from the first principal component explains the largest portion of variance of the individual indicators. This is obtained by maximizing the sum of the squares of the coefficients of correlation between the composite index and the individual indicators. However, the first principal component accounts for a limited part of the variance in the data (in the previous example, 72.4%), so we can lose a consistent amount of information.

Moreover, the PCA based index is often *elitist* (Mishra, 2008), with a strong tendency to represent highly intercorrelated indicators and to neglect the others, irrespective of their possible contextual importance. So many highly important but poorly intercorrelated indicators may be unrepresented by the composite index. On many occasions, it is found that some (evidently) very important indicators are roughly dealt with by PCA, simply because those variables exhibited widely distributed scatter or they did not fall within a narrow band around a straight line (Mishra, 2007).

On the other hand, PCA is a blindly empiricist method based on the observed correlations and it ignores the polarity of the individual indicators. Therefore, if the normalized indicators are not all positively intercorrelated, the results are not correct, as shown above.

Finally, it should be noted that the amount of variance accounted for, and the weights computed by PCA change over time, so the results of different PCAs are not easily comparable.

5. Conclusions

The construction of composite indices for assessing multidimensional phenomena is a central issue in data analysis, particularly in economics and sociology. Researcher cannot solve this question simply by using PCA or related methods, such as Factor Analysis, since they are typically used for a reflective approach and they ignore the polarity of the individual indicators.

Often, a formative approach is required, where the index to be constructed does not exist as an independent entity, but is a composite measure directly determined by a set of non-interchangeable individual indicators. It is the case of the HDI and the RCI. The first index does not use PCA, whereas the second one uses PCA 'only' for selecting indicators. So, in order to obtain valid and reliable results, it is absolutely essential to define the theoretical framework with an appropriate measurement model.

This paradigm should always be considered when the objective of the research is to measure a multidimensional phenomenon through composite indices. And this is even more valid if the phenomenon to be measured is well-being, progress or quality of life. Indeed, these latent factors depend on the individual indicators that represent them and not the contrary. Therefore, the use of PCA for the measurement of these phenomena is at all improper.

References

ANNONI P., KOZOVSKA, K., 2010. *EU Regional Competitiveness Index 2010*. JRC Scientific and Technical Reports. Luxembourg: Publications Office of the European Union.

BANDURA R., 2008. *A Survey of Composite Indices Measuring Country Performance: 2008 Update*. New York: UNDP/ODS Working Papers.

COLTMAN T., DEVINNEY T.M., MIDGLEY D.F., VENAIK S., 2008. Formative versus reflective measurement models: Two applications of formative measurement, *Journal of Business Research*, Vol. 61, pp. 1250-1262.

DIAMANTOPOULOS A., RIEFLER P., ROTH, K.P., 2008. Advancing formative measurement models, *Journal of Business Research*, Vol. 61, pp. 1203-1218.

KENDALL, M.G., STUART A., 1968. *The Advanced Theory of Statistics, Vol. 3*. London: Charles Griffin & Co.

MAGGINO F., 2014. Indicator Development and Construction. In MICHALOS, A. C. (Ed) *Encyclopedia of Quality of Life and Well-Being Research*, Dordrecht: Springer, pp. 3190-3197.

MAZZIOTTA M., PARETO A., 2013. Methods for Constructing Composite Indices: One for all or all for one, *Rivista Italiana di Economia Demografia e Statistica*, Vol. LXVII, n. 2, pp. 67-80.

MISHRA, S.K., 2007. A Comparative Study of Various Inclusive Indices and the Index Constructed by the Principal Components Analysis. *MPRA Paper*, No. 3377. Available at MPRA: http://mpra.ub.uni-muenchen.de/3377.

MISHRA, S.K., 2008. On Construction of Robust Composite Indices by Linear Aggregation. Available at SSRN: http://ssrn.com/abstract=1147964.

SAISANA M., TARANTOLA, S., 2002. *State-of-the-art report on current methodologies and practices for composite indicator development*. European Commission-JRC, EUR 20408 EN, Ispra.

OECD, 2008. *Handbook on Constructing Composite Indicators. Methodology and user guide*. Paris: OECD Publications.

UNDP 2010. *Human Development Report 2010. The Real Wealth of Nations: Pathways to Human Development*. New York: Palgrave Macmillan.

SUMMARY

Principal Components Analysis (PCA) is one of the most commonly used multivariate statistical technique in construction of composite indicators. However, PCA and related methods, such as Factor Analysis, are based on a reflective model where the individual indicators (manifest variables) are seen as functions of a latent variable (principal component or factor). When individual indicators are causes of the latent variable, rather than its effects, a formative model should be adopted. In this paper, we compare the two approaches, and we show by a numerical example that factorial methods, such as PCA, may fail if improperly used.

Matteo MAZZIOTTA, Italian National Institute of Statistics, mazziott@istat.it
Adriano PARETO, Italian National Institute of Statistics, pareto@istat.it