Electronic Journal of Applied Statistical Analysis
EJASA (2012), Electron. J. App. Stat. Anal., Vol. 5, Issue 3, 320 – 326
e-ISSN 2070-5948, DOI 10.1285/i20705948v5n3p320

CONSTRUCTING INDICATORS OF UNOBSERVABLE VARIABLES FROM PARALLEL MEASUREMENTS

Maurizio Carpita , Marica Manisera[*]

Department of Quantitative Methods, University of Brescia, Italy

Received 10 July 2012; Accepted 07 October 2012
Available online 16 November 2012

Abstract: *The social and economic research often focuses on the construction of composite indicators for unobservable (or latent) variables using data from a questionnaire with Likert-type scales. Within the variety of procedures, we focus on the data analysis technique of Principal Components Analysis, in its Linear and NonLinear versions. This paper shows that when the variables are parallel measurements of the same latent unobservable variable, Linear and NonLinear Principal Components Analyses practically lead to the same composite indicators.*

Keywords: *Principal components analysis, ordinal variables, nonlinearity, latent variables, probabilistic gauge, Monte Carlo gauge*

1. Introduction

The social and economic research is often focused on the construction of a one-dimensional composite indicator for an unobservable variable (or latent variable) starting from data from Likert-type scales. Many different statistical models and methods allow this goal to be reached. Stochastic models like those developed in the framework of the Rasch Analysis or the Item Response Theory [3] show good statistical properties, but usually require assumptions that are often violated or difficult to satisfy. Another very simple and widespread procedure to construct composite indicators is the *summated rating scale*, which suggests to add up the quantifications (usually the first integers) assigned to the ordered response categories of *m* variables (items) and use this (weighted or unweighted) sum as a composite measure of a latent construct. This procedure is based on the Classical Test Theory idea that if the variables are *parallel measurements* (i.e., are *homogeneous* variables), their sum will tend to cancel out measurement

[*] E-mail: manisera@eco.unibs.it

errors [7]. Among the several proposals existing in the literature to obtain a one-dimensional indicator from parallel measurements, in this paper we follow a data analysis approach, which requires weak statistical and distributional assumptions, and consider the Principal Components Analysis (PCA), in its Linear (L-PCA [8]) and NonLinear (NL-PCA [6, 7, 4]) versions. L-PCA and NL-PCA provide indicators easy to compute, requiring weak assumptions, and often resulting in measures highly correlated with those obtained from a more sophisticated Rasch model [5, 1]. NL-PCA aims at the same goals of traditional L-PCA, but it is suited for variables of nominal, ordinal, numerical measurement levels that may not be linearly related to each other. The NL-PCA model is the same linear model as in traditional L-PCA, but it is applied to nonlinearly transformed data, obtained by assigning optimal scale values (the *quantifications*) to the categories. While L-PCA assigns equally spaced numbers (usually the first positive integers) to the categories, NL-PCA finds category quantifications that are optimal in the sense that the overall variance accounted for in the transformed variables, given the number of components, is maximized.

When using L-PCA and NL-PCA in order to obtain a composite indicator of unobservable variables, these two data analysis techniques are within the same framework as the *summated rating scale* procedure. Indeed, the final composite indicator is obtained by a weighted sum of the quantified variables, with weights identified by the algorithm. The difference between the two algorithmic procedures is that L-PCA assigns *linear* (equally spaced) *quantifications*, while NL-PCA assigns *nonlinear* (i.e., not necessarily ordered nor equally spaced) *quantifications* to the ordered categories; NL-PCA quantifications take also into account possible *nonlinear* relationships among variables [7]. In this paper, we consider to apply NL-PCA with the *ordinal scaling level*, meaning that variables are transformed according to monotonic nonlinear transformations. Therefore, the nonlinear quantifications are ordered and the only difference with the linear quantifications is that the former are not necessarily equally spaced.

However, in some applications, NL-PCA leads to the same results of L-PCA, suggesting that the assumptions of L-PCA are not a practical problem. Starting from these considerations, this paper aims at showing that, when data come from parallel measurements of the same unobservable variable, the composite indicators obtained from NL-PCA and L-PCA do not practically differ. We pursued this goal by using the *gauging* approach ([6], p. 34): (1) a *Probabilistic gauge* to construct a population of homogenous data and compare the L-PCA and NL-PCA composite indicators; (2) a *Monte Carlo gauge*, to compare the sampling performance of L-PCA and NL-PCA in recovering the population parameters of interest. A wider investigation comparing the L-PCA and NL-PCA solutions in the population data as well as in the sampling performance can be found in [2], a very extensive simulation study oriented to evaluate the level of nonlinearity existing in homogeneous data. In the present paper, we focus the comparison on the resulting composite indicators, which can be interpreted as a global result from the two techniques. Other parameters were considered in [2].

2. Simulation design

In order to have a realistic data structure with known properties allowing the comparison of the L-PCA and NL-PCA results, we used the gauging approach [6].

Population and sampling performances of the two techniques are compared focusing on the composite indicator that can be used as a measure of the unobservable variable or latent variable.
The method we used to construct population homogeneous data with a one-dimensional latent variable underlying m ordinal variables was proposed by [10] and is based on the discretization of m continuous variables following a multivariate standard normal distribution with equal correlations ρ. In this case, the m continuous variables are linearly related each other and the correlation matrix has a dominant eigenvalue given by $\lambda_+= [1 + (m - 1) \rho]/ m$ ([10], p. 7).
In this study, we considered $m = 4$ and $\rho = 0.4, 0.6, 0.8$, corresponding to three situations with underlying one-dimensional latent variables having different levels of strength $\lambda_+=0.55$, 0.70, 0.85. The small value of m was chosen to simplify computations but also to introduce some instability in the results, with the aim to stress the differences between L-PCA and NL-PCA.
The continuous variables were then discretized, by means of discretization cuts to map continuous intervals into ordinal categories. We considered three discretization forms resulting from nonlinear monotonic transformations: (1) an optimal discrete distribution O, which resembles the original normal distribution rather closely, (2) a right-skewed discrete distribution (R, with positive skewness), and (3) a left-skewed discrete distribution (L, with negative skewness).
According to [10], we chose k=5 categories for each variable with corresponding frequencies (0.11; 0.24; 0.30; 0.24; 0.11) for the O distribution. Two different versions of the skewed variables were considered: (*i*) (0.45; 0.25; 0.15; 0.10; 0.05) and (0.05; 0.10; 0.15; 0.25; 0.45) for the R and the L distributions, respectively, and (*ii*) (0.65; 0.15; 0.10; 0.05; 0.05) and (0.05; 0.05; 0.10; 0.15; 0.65) for the R and the L distributions, respectively.
The corresponding frequency plots are displayed in Figures 1-2.

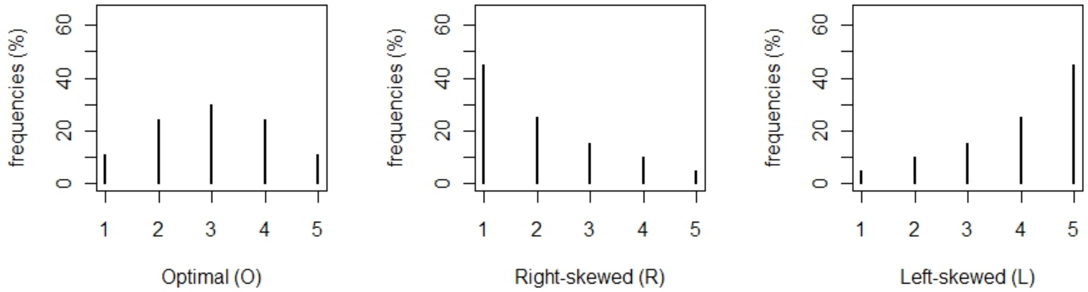

Figure 1. Frequency distributions for the Optimal, Right- and Left-skewed - *version* (*i*) - discretized variables.

With (*ii*), we chose to stress the presence of high frequencies on the mode category, because quantifications typically show nonlinearity and instability in the presence of low frequencies on some categories, like in the R and L distributions [9].

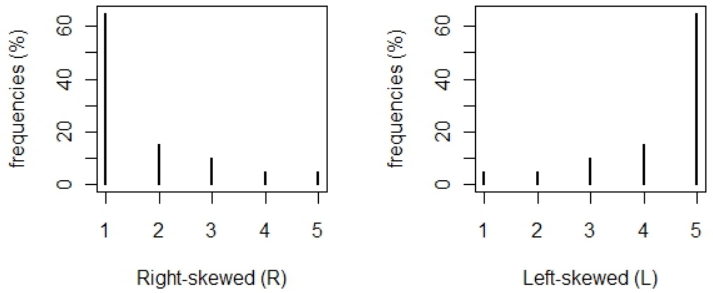

Figure 2. Frequency distributions for the Right- and Left-skewed - *version* (*ii*) - discretized variables.

Moreover, when the distributions of the analysed ordinal variables are very different, these can be thought to be not linearly related. To evaluate the interaction of ordinal variables with different distributions, we combined optimal and right- and left-skewed variables, obtaining 9 distinct Cases: OOOO, OOOL, OOLL, OLLL, LLLL, LLLR, LLRR, LROO, LLRO. Since the R variable follows the reversed frequency pattern of the L variable, OOOR, OORR and ORRR Cases can be skipped from the analysis, because they give the analogous results of the OOOL, OOLL and OLLL Cases already considered; the same holds for the LRRR and RRRR combinations, equivalent to the LLLR and LLLL Cases, respectively.

Using the discretization cuts, we computed the probability distribution for the multinomial distribution by multidimensional integration of the multivariate normal distribution. Then, a population composed of 100,000 units was obtained for each of the 9x3=27 considered combinations - 9 different Cases for O-L-R and 3 different values of ρ - for the 2 different *versions* (*i*) and (*ii*) of the skewed R and L distributions, and both L-PCA and NL-PCA were applied to population data.

For each of the considered combinations, the *population scores*, providing the “measure” of the underlying latent variable, were obtained by computing the weighted sum of the quantifications assigned by L-PCA or NL-PCA with loadings as weights; one score is determined for each of the unique $k^m=5^4=625$ population profiles or response patterns.

The resulting population scores were compared by computing the linear correlation coefficient $\rho_{L;NL}$ between the L-PCA and NL-PCA 625 population scores (weighted by their population frequencies).

In order to check the sample stability of L-PCA and NL-PCA, we derived the *Monte Carlo gauge* from the *Probabilistic gauge* described above: we run 1000 replications of simple random samples with three different sample sizes n=250,500,1000. To compare the *Monte Carlo scores*, we computed the linear correlation coefficient $r_{L;NL}$ between linear and nonlinear scores of all the 625 response patterns (weighted by the corresponding population frequencies to consider their sampling probabilities) in each replication for every considered combination. Then we computed the mean correlation $\bar{r}_{L;NL}$, with the associated standard error, over the replications.

As mentioned in Section 1, in [2] the comparison between the two solutions concerned the quantifications of each variable, by means of the so-called *NL* index [2], the dominant eigenvalue λ_+, the mean correlation and the Cronbach’s alpha, both related to λ_+ ([7], p. 187).

3. Results

Although L-PCA and NL-PCA solutions differ with reference to several parameters (for example, quantifications, dominant eigenvalue, etc., see [2]), the analysis of population scores showed that the linear correlation coefficient $\rho_{L;NL}$ was higher than 0.96 in all the considered configurations. This means that L-PCA and NL-PCA practically provide the same measure of the latent variable underlying the data. The minimum value 0.96 came out in the LLRR Case with *version* (*ii*) and $\rho = 0.8$: when the one-dimensionality is stronger, the flexibility of NL-PCA to assign nonlinear quantifications to the categories of very skewed ordinal variables results in a lower loss of information, and it is easier for NL-PCA to "overperform" and then to obtain a (slightly) different composite indicator, with respect to L-PCA.

Results of the *Monte Carlo gauge* suggested that the NL-PCA and L-PCA performances were comparable in terms of accuracy and efficiency in estimating the population scores, although the *Monte Carlo scores* obtained by NL-PCA showed slightly higher instability. The correlation coefficient $r_{L;NL}$ between linear and nonlinear scores is higher than 0.94 for every sample size *n* and in all the 9 different Cases for the 3x2=6 considered combinations - 3 values of ρ, 2 *versions* (*i*) and (*ii*). For every *n*, in each of these 6 combinations, RRLL is always the Case associated with the lowest $r_{L;NL}$. Figure 3 displays the box-plots of $r_{L;NL}$ obtained for *n*=250, RRLL Case, and the 3x2 combinations of ρ and *versions* (*i*) and (*ii*). As expected, for fixed ρ, *version* (*ii*) is associated with lower values of $r_{L;NL}$ and larger variability than *version* (*i*), while, for fixed *version* (*i*) or (*ii*), as ρ increases, $r_{L;NL}$ decreases as well as its variability.

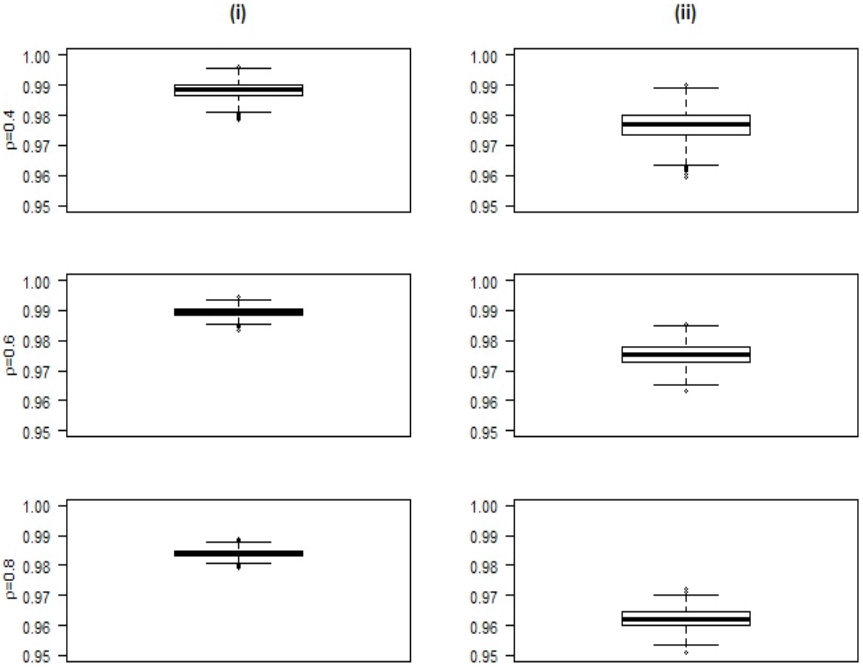

Figure 3. Box-plots of the correlation coefficient $r_{L;NL}$ between L-PCA and NL-PCA scores, for the 3x2=6 considered combinations - 3 values of ρ (row), 2 *versions* (*i*) and (*ii*) (column) - for *n*=250 and RRLL Case.

The mean correlation $\bar{r}_{L;NL}$ between linear and nonlinear scores, obtained averaging over the 1,000 replications, was always higher than 0.96 considering the three sample sizes. As expected, it increased as the sample size increased and, for each sample size, showed the lowest values in correspondence of the LLRR Case with *version* (*ii*) and ρ=0.8, like in the *Probabilistic gauge*. Table 1 displays the values of bias and standard error of the correlation coefficient estimator between L-PCA and NL-PCA scores.

Results show that $\rho_{L;NL}$ was well estimated also with the smallest sample size, and this suggests that L-PCA and NL-PCA were comparable in their ability to estimate the measure of the latent variable underlying the data.

Table 1. Bias and Standard Error of the correlation coefficient estimator between L-PCA and NL-PCA scores for the Monte Carlo gauge (1,000 replications).

n	Bias			Standard Error		
	min	**mean**	**max**	**min**	**mean**	**max**
250	0.001	0.006	0.015	0.002	0.006	0.010
500	0.000	0.003	0.008	0.002	0.003	0.007
1000	0.000	0.002	0.005	0.001	0.002	0.005

4. Concluding Remarks

This study showed that when population data come from parallel measurements of the same unobservable variable and when the focus is on the construction of a composite indicator for it, L-PCA and NL-PCA results do not practically differ, even when variables have very different distributions. For this reason, according to the Occam's razor, the use of the simplest method can be preferred, although ordinal data would require appropriate statistical modelling and variables having different (optimal and skewed) distributions would suggest to use the more complex nonlinear technique. This result relies on the characteristics of the chosen Probabilistic gauge: in a next study, we want to identify different situations recommending the use of NL-PCA.

References

[1]. Brentari, E., Golia, S., Manisera, M. (2007). Models for categorical data: a comparison between the Rasch model and Nonlinear Principal Component Analysis. *Statistica & Applicazioni*, 5, 53-77.

[2]. Carpita, M., Manisera, M. (2011). On the nonlinearity of homogeneous ordinal variables. In *New Perspectives in Statistical Modeling and Data Analysis*, eds. S. Ingrassia, R. Rocci, M. Vichi, Heidelberg: Springer, 489-496.

[3]. DeMars C. (2010). Item response theory. Oxford University Press Inc., New York.

[4]. Ferrari, P.A., Barbiero, A. (2012). Nonlinear principal component analysis, in *Modern Analysis of Customer Surveys: with applications using R*, eds. R.S. Kennett and S. Salini, Chichester: John Wiley, 333-356.

[5]. Ferrari, P.A., Salini, S. (2011). Complementary use of Rasch models and Nonlinear Principal Components Analysis in the assessment of the opinion of Europeans about utilities. *Journal of classification*, 28, 53-69.

[6]. Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester: John Wiley.

[7]. Heiser, W.J., Meulman, J.J. (1994). Homogeneity analysis: exploring the distribution of variables and their nonlinear relationships, in *Correspondence analysis in the social sciences*, eds. M. Greenacre and M. Blasius, New York: Academic Press, 179-209.

[8]. Jolliffe, I.T. (2002). *Principal component analysis*, 2nd Ed. New York: Springer.

[9]. Linting, M., Meulman, J.J., Groenen, P.J.F., Van der Kooij, A. (2007). Stability of Nonlinear Principal Components Analysis: An empirical study using the balanced bootstrap. *Psychological Methods*, 12, 359-379.

[10]. van Rijckevorsel, J., Bettonvil, B., de Leeuw, J. (1985). Recovery and stability in nonlinear PCA, Dept. Data Theory, Leiden Univ, RR-85-21.