Electronic Journal of Applied Statistical Analysis
EJASA (2012), Electron. J. App. Stat. Anal., Vol. 5, Issue 3, 314 – 319
e-ISSN 2070-5948, DOI 10.1285/i20705948v5n3p314

ASSESSING ITEM CONTRIBUTION ON UNOBSERVABLE VARIABLES' MEASURES WITH HIERARCHICAL DATA

Marica Manisera, Marika Vezzoli[*]

Department of Quantitative Methods, University of Brescia, Italy

Received 02 July 2012; Accepted 07 October 2012
Available online 16 November 2012

Abstract: *This paper aims at measuring the contribution of each item used to construct composite indicators of unobservable variables when data come from multi-item scales and have a hierarchical structure. To this end, we combine the MultiLevel NonLinear Principal Components Analysis with the CRAGGING algorithm, and then extracting its MultiLevel Mean Decrease in Accuracy measure of variable importance. The first algorithm is used to realize a composite indicator of the latent variable, while the second is an ensemble method suitable for hierarchical data and able to provide a variable importance measure. The proposed procedure takes account of the data structure, thus offering a new way to assess the items' contribution on the hierarchical-based unobservable variables' measure.*

Keywords: *Latent variables, variable importance measures, multilevel, nonlinear principal components analysis, CRAGGING.*

1. Introduction

One of the main goals in the social and economic research is to measure individuals' perceptions and attitudes (such as customer and job satisfaction). This requires statistical instruments to deal with unobservable (or latent) variables, i.e., complex concepts measured indirectly by means of observable variables.

In order to measure unobservable variables, researchers usually collect data through questionnaires with several items referring to the different aspects of the concept under inspection. Responses often indicate the degree of agreement with each statement, where higher scores reflect greater agreement with the assertion.

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

These data are usually organized within a hierarchical structure. Indeed, individuals (first-level units) are clustered, or nested, within groups (second-level units), which can be gathered in third-level units, and so on. For example, in data coming from surveys on job satisfaction, workers can be nested in organizations, which can be grouped in geographical areas. The number of individuals within each group is often not constant, namely the clusters are unbalanced.

Among several statistical techniques useful to measure unobservable variables, we focus our attention on the MultiLevel NonLinear Principal Components Analysis (ML-NLPCA, [6]), useful to construct a composite indicator taking account of the ordinal nature of the variables, their (possible) nonlinear relationships, and the nesting of individuals in higher-order groups.

In the construction of composite indicators for latent variables, a challenging question refers to the contribution (and thus the importance) of each item in the definition of such indicators. A slightly different issue was considered by some authors (e.g., [7]), which identified the drivers of latent variables using the variable importance measures proposed in the framework of the ensemble learning models (e.g., Random Forests (RF) [1], CRAGGING [8]).

This paper aims at measuring the importance of each item used to construct composite indicators of the latent variables starting from multi-item scales when data have a hierarchical structure. To do this, we combine the ML-NLPCA, used to realize a composite indicator of the latent variable, with the CRAGGING algorithm, which is a recent ensemble learning introduced to deal with hierarchical data. From CRAGGING we then extract a measure of variable importance, that we named MultiLevel Mean Decrease in Accuracy (ML-MDA). The procedure proposed in this study is conceived to maintain the structure in the data, thus offering a new way to assess the items' contribution on the hierarchical-based latent variables' measure.

This procedure was applied to real data referring to workers (first-level units) employed in the social cooperatives (second-level units) sampled in the ICSI^{2007} survey on the Italian social cooperatives [2].

The paper is organized as follows. Section 2 describes the ML-NLPCA and the ML-MDA measure while Section 3 reports the results coming from the application on real data and some concluding remarks.

2. Combining multilevel algorithmic techniques

In this section, we briefly describe the algorithmic procedures used in this paper: the ML-NLPCA and the ML-MDA variable importance measure introduced in the framework of CRAGGING (the latter is described in detail in [7,8]). We consider to observe m (categorical) variables X_j, j=1,2,...,m, on N subjects clustered in K groups, with n_k subjects per group, k=1,2,...,K, with $\sum_{k=1}^{K} n_k = N$.

2.1 MultiLevel NonLinear Principal Components Analysis (ML-NLPCA)

NLPCA [4] is one of the statistical methods useful to provide quantitative measures of the latent variables underlying a multiple-item scale. NLPCA is the nonlinear equivalent of classical PCA conceived to deal with nonlinearly related categorical and numerical variables. It aims at optimally reducing a large number m of categorical (or mixed) variables into a smaller number c of composite variables (the principal components or object scores), useful to represent latent variables. Simultaneously with data reduction, NLPCA transforms the original variables into quantified variables by assigning optimally scaled values to the categories. Such category quantifications are optimal in the sense that the variance accounted for in the transformed variables, given the number c of components, is maximized. The variance accounted for is often expressed in percentage (Percentage of Variance Accounted For, PVAF) and is a global measure of the goodness of the NLPCA solution.

In the literature [6], NLPCA and, more generally, homogeneity analysis was formally extended to a multilevel sampling design framework in order to obtain models that take advantage of the clustering of the subjects and to examine how variables are related across groups and how groups vary.

The approach developed in [6] is very general, allowing to generate several models and incorporate prior knowledge, and other multilevel extensions of homogeneity analysis can be derived from this framework. It is worth noting that under normalization of object scores within every group, ML-NLPCA is equivalent to applying the ordinary NLPCA algorithm to each of the K groups separately[1]. It is straightforward that the basic geometric properties of the NLPCA continue to hold for every group. According to [6], the "overall" PVAF is computed as weighted average of the PVAF$_k$'s, k=1,2,...,K obtained in the groups, with weights given by n_k/N.

Like NLPCA, ML-NLPCA is used as a descriptive data analysis technique. In the literature, stability studies on NLPCA results were obtained by a nonparametric approach, consistent with the weak distributional assumptions. With reference to ML-NLPCA, the internal stability of the composite indicators could be assessed by means of a bootstrap study on the NLPCA solution in each of the K groups separately, thus consistent with the ML-NLPCA philosophy [5].

2.2 MultiLevel Mean Decrease in Accuracy (ML-MDA) measure

In many applied problems, the identification of the most important variables associated to the response Y is a relevant issue. This topic was mainly developed in the context of the ensemble methods (e.g., [1]) that use multiple models (usually trees) in order to obtain accurate predictors. When the data have a hierarchical structure, the well-known ensemble methods (Bagging, Random Forests, Boosting, etc.) do not provide appreciable results. For this reason, Vezzoli and Stone [8] proposed a multiple tree-based model, called CRAGGING, able to deal with structured data. Following the philosophy of the main ensemble methods, CRAGGING combines many binary decision trees built on several samples obtained perturbing the data without destroying the hierarchical structure. The goodness of fit is evaluated by a loss function L, depending on the nature of the response Y. When Y is continuous and the regression trees are then grown in the ensemble, the Mean Square Error (MSE) is often adopted as a loss function. Otherwise, when the Y is categorical, alternative loss functions are used, usually based on the confusion matrix.

In the context of CRAGGING, Vezzoli and Zuccolotto [9] proposed a modified version of the Mean Decrease in Accuracy measure of variable importance[2]. In detail, they conceived it

[1] Since the NLPCA solution is rotationally invariant, different group solutions can have different orientations of the axes. Therefore, in order to fairly compare groups, their axes must be rotated to a target solution by means of a Procrustes orthogonal rotation.

[2] The rationale of this measure is as follows: the association between the j-th variable and the response Y is broken when X_j is randomly permuted. When the permuted variable together with the remaining non-permuted variables are used to predict the response Y, the prediction accuracy decreases substantially if the original variable was associated with Y. As a measure for variable importance, Breiman [1] suggested to use the difference in prediction accuracy (measured by L) before and after permuting X_j, averaged over all the trees of the ensemble.

introducing a randomization of the *j*-th variable without destroying the structure in the data. For this reason, we define this new measure as MultiLevel Mean Decrease in Accuracy (ML-MDA). Formally, for each variable X_j, with $j=1,2,\dots,m$, a permutation $p=\{p_1,p_2,\dots,p_K\}$ of the index set $\{1,2,\dots,K\}$ is randomly selected. The values of X_j are randomized in the dataset according to the following rule

$$\{x_{jki}\}_{k\in\{1,2,\dots,K\},\,i=1,2,\dots,n_k} = s(\mathbf{x}_{jp_k})$$

where $s(\cdot)$ denotes a sampling with replacement from a set of values and $\mathbf{x}_{jp_k} = \{x_{jp_k i}\}_{i=1,2,\dots,n_{p_k}}$.
This way of randomizing the values of X_j is particularly useful when the groups are homogeneous relative to X_j, as frequently happens in the application domain of CRAGGING. The resampling procedure is repeated V times and the ML-MDA measure for the *j*-th variable is given by

$$\text{ML-MDA}_j = \frac{1}{V}\sum\nolimits_v \left(L_{j,v} - L\right),$$

where $L_{j,v}$ is the value of the loss function when the *j*-th variable is perturbed in the *v*-th replication, while L is simply the loss function computed on the original data.
To make the interpretation easier, the measure is often expressed in relative terms based upon its observed maximum (multiplied by 100).

3. Application and concluding remarks

ML-NLPCA was applied to construct a Job Satisfaction (JS) indicator that summarizes 11 categorical ordinal variables measuring different JS facets for 1,804 workers employed in 115 social cooperatives. Missing values were imputed according to [3]. From the whole sample, the cooperatives with less than 10 workers were removed to both improve the ML-NLPCA stability and avoid resampling problems in the computation of the ML-MDA measures.
The data used in this study result from a preliminary Rasch analysis [2], which identified the 11 selected JS items as related to a “global” JS and suggested to merge response categories to obtain a 5-point response scale for each item, ranging from 1=*very dissatisfied* to 5=*very satisfied*, with mid-point 3=*neither dissatisfied nor satisfied*. The variables refer to the satisfaction of workers with extrinsic aspects, i.e., the work characteristics (“`variety` and creativity”, “vocational training and professional `growth`”, “decisional and operative `independence`”, and “`career` promotions”) as well as intrinsic and relational aspects (“personal `fulfilment`”, “`transparency` in the relation with the cooperative”, “recognition by coworkers - `coworkers.recognition`”, “recognition by cooperative - `coop.recognition`”, “`involvement` in the decision”, “relations with `team`”, “relations with `superior`”).
In order to obtain a one-dimensional JS composite indicator, ML-NLPCA was applied in each of the K groups with all of the variables scaled ordinally, to keep, in the quantified variables, the

grouping and the ordering information in the original categorical variables (PVAF=56). The stability of such composite indicator was verified by means of a bootstrap study.

The JS indicator was then used as the response variable in the CRAGGING (MSE=0.03) and the ML-MDA measure was obtained for each of the 11 items (in this study, *V*=5). Results are shown in Figure 1. The variables that mostly contribute to the definition of the JS indicator are, in order, `growth`, `fulfilment`, and `transparency`, while some other items (for example, `career`, `team`, and `co-workers.recognition`) do not contribute at all. These results are in line with previous studies, confirming the role of intrinsic and relational aspects of work in determining JS, although the most important variable can be considered mainly related to extrinsic JS.

Using the same response variable, we grew 5000 trees by means of the RF algorithm [1] extracting the corresponding MDA measure which, differently from the ML-MDA measure, comes from a procedure that permutes the variables without taking account of the hierarchical structure in the data. The comparison with ML-MDA shows that the results are quite different. The correlation coefficient between the RF MDA and the ML-MDA measures equals 0.66. In particular, 8 variables (`fulfilment`, `transparency`, `coop.recognition`, `career`, `superiors`, `growth`, `variety`, `involvement`) out of 11 show a relative MDA measure higher than 60. Instead, only 2 variables (`growth` and `fulfilment`) have a ML-MDA measure higher than 60. These results suggest a better ability of the CRAGGING in detecting few most important variables among the covariates used in the algorithm. Moreover, comparing the loss functions of the two algorithms, we observed that the RF MSE (0.37) is much greater than the CRAGGING MSE (0.03), highlighting the best performance of CRAGGING over RF in this empirical study.

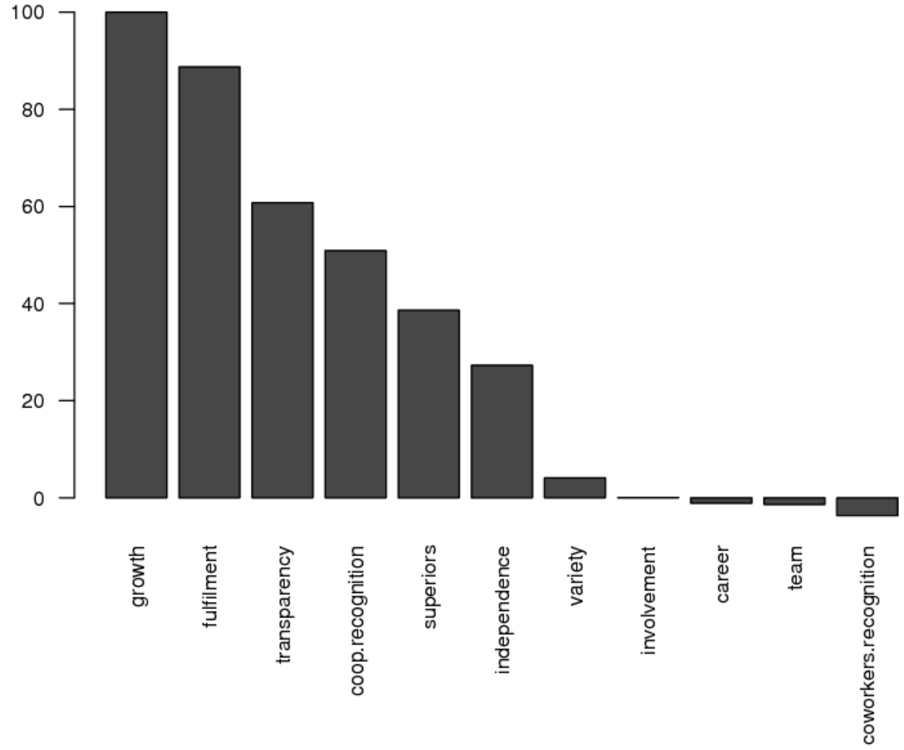

Figure 1. ML-MDA measures of the 11 job satisfaction items

In this study, we preferred to use a JS indicator coming from a data analysis technique rather than other possible more sophisticated stochastic models (for example, related to the Item Response Theory), in order to deal with simpler procedures suited for hierarchical data which do not require strong assumptions.

This preliminary study on algorithmic models to measure variable importance in the definition of composite indicators when data are hierarchical gave rise to some methodological issues that will be soon investigated. For example, it will be interesting to compare the ML-MDA measure with a simple variable importance measure from NLPCA. In fact, in NLPCA the contribution of the j-th variable to the composite indicator is the Variance Accounted For *per variable j* (VAF_j with j=1,2,...,m), which is obtained by the sum of the squared loadings over components. However, we prefer not measure the variables' importance by the loadings in order to avoid multiple measures for each variable when the number c of components retained in the solution is higher than 1. The use of the VAF_j's overcomes this problem, but is still unsatisfactory due to their descriptive nature. More importantly, in the multilevel framework, a set of loadings and, then, of VAF_j's is computed within each group and they may substantially vary across groups. Moreover, a bootstrap study may reveal that, when dealing with ML-NLPCA, the loadings (and the VAF_j's) can be quite unstable, especially within very small groups.

Acknowledgement

A preliminary version of this paper has been presented at the 46-th Scientific Meeting of the Italian Statistical Society, Rome, 20-22 June 2012.

References

[1]. Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32.

[2]. Carpita, M., Golia, S. (2011). Measuring the Quality of Work: The Case of the Italian Social Cooperatives. *Quality and Quantity*. In press.

[3]. Carpita, M., Manisera, M. (2011). On the Imputation of Missing Data in Surveys with Likert-Type Scales. *Journal of Classification*, 28, 93–112.

[4]. Gifi, A. (1990). *Nonlinear Multivariate Analysis*. Chichester: Wiley.

[5]. Manisera, M. (2012). Assessing Stability in NonLinear PCA with Hierarchical Data, in *New Perspectives in Statistical Modeling and Data Analysis*, eds. P. Giudici, S. Ingrassia, M. Vichi, Heidelberg: Springer, *Forthcoming*.

[6]. Michailidis, G., de Leeuw, J. (2000). Multilevel Homogeneity Analysis with Differential Weighting. *Computational Statistics and Data Analysis*, 32, 411–442.

[7]. Vezzoli, M. (2011). Exploring the Facets of Overall Job Satisfaction through a Novel Ensemble Learning. *Electronic Journal of Applied Statistical Analysis*, 4, 23–38.

[8]. Vezzoli, M., Stone, C.J. (2007). CRAGGING, in *Book of Short Papers CLADAG 2007*. EUM, University of Macerata, 12 – 14 September 2007, 363–366.

[9]. Vezzoli, M., Zuccolotto, P. (2011). CRAGGING Measures of Variable Importance for Data with Hierarchical Structure, in *New Perspectives in Statistical Modeling and Data Analysis*, eds. S. Ingrassia, R. Rocci, M. Vichi, Heidelberg: Springer, 393–400.