Multilevel algorithmic models to measure item importance on latent variables' indicators

Marica Manisera and Marika Vezzoli

Abstract This paper aims at measuring the importance of each item used to construct composite indicators of the latent 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 MultiLevel Mean Decrease in Accuracy variable importance measure conceived within the CRAGGING framework. The first algorithm is used to realize a composite indicator of the latent variable, while the second permits to identify the most important items relative to such indicator. The two techniques offer a new way to assess the items' contribution on the hierarchical-based latent variables' measure.

Key words: latent variables, variable importance measures, multilevel, nonlinear Principal Components Analysis, CRAGGING

1 Introduction

One of the main goals in social and economic research is to measure individuals' perceptions and attitudes, which are latent variables. In order to measure latent variables, researchers usually collect data through multiple-item scales. These data are often organized within a hierarchical structure: individuals (first-level units) are clustered, or nested, within groups (second-level units) which in turn can be gathered in higher-level units. The number of individuals within each group is often not constant (unbalanced clustering).

Among the several statistical techniques useful to measure latent variables, we focus on the MultiLevel NonLinear Principal Components Analysis (ML-NLPCA; [5]), an algorithmic model aimed at constructing a composite indicator taking ac-

Department of Quantitative Methods, University of Brescia, C.da S. Chiara, 50 - 25122 Brescia, Italy e-mail: {manisera,vezzoli}@eco.unibs.it

count of the ordinal nature of the variables, their (possible) nonlinear relationships, and the nesting of individuals in higher-order groups.

A challenging question refers to the importance of each item in the construction of the composite indicators for latent variables. In the literature, variable importance measures were mainly proposed in the framework of the ensemble learning methods ([1], [6]).

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 a measure of variable importance (MultiLevel Mean Decrease in Accuracy; ML-MDA [7]), proposed in the context of the CRAGGING [6], a recent ensemble learning dealing with hierarchical data. Both the techniques are conceived to take account of the structure in the data; their combination offers a new way to assess the items' contribution on the hierarchical-based latent variables' measure. The procedure proposed in this study 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. The ML-MDA was used to study the importance of different items relative to a job satisfaction (JS) indicator constructed by the ML-NLPCA.

2 Methods

In this section, we briefly describe the two algorithmic models used in this paper: the ML-NLPCA and the ML-MDA variable importance measure. We consider m (categorical) variables on N subjects (or objects) clustered in K groups, with n_k subjects per group and $\sum_{k=1}^{K} n_k = N$.

1. **ML-NLPCA**
   NonLinear Principal Components Analysis (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 and aims at optimally reducing a large number m of categorical variables into a smaller number c of composite variables (object scores), useful to represent latent variables. Simultaneously with data reduction, NLPCA transforms the original variables into quantified ones by assigning optimally scaled values to the categories. Such category quantifications are optimal in the sense that the overall Variance Accounted For (VAF) in the transformed variables, given the number c of components, is maximized. The VAF is often expressed in Percentage (PVAF) and is a global measure of the goodness of the NLPCA solution. In the literature [5], NLPCA was formally extended to a multilevel sampling design framework. The approach developed in [5] is very general, allowing to generate other multilevel extensions of homogeneity analysis and incorporate prior knowledge. 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 [5], the "overall" PVAF is computed as weighted average of the PVAF$_k$'s, $k = 1,2,\ldots,K$ 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. Therefore, the internal stability [5] of the ML-NLPCA 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.

2. **ML-MDA**
   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 [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 Forest, Boosting) do not provide appreciable results. For this reason, a multiple tree-based model, called CRAGGING [6], was proposed to deal with structured data. Following the philosophy of main ensemble methods, CRAGGING combines many binary decision trees built on several samples obtained perturbing the data without destroying the hierarchical structure. In the context of CRAGGING, a modified version of the Mean Decrease in Accuracy measure of variable importance[2] was proposed [7], starting from the randomization of the j-th variable without destroying the structure of the data. Formally, for each variable X_j a permutation $p = \{p_1,\cdots,p_k\}$ of the set $\mathscr{L} = \{1,\cdots,K\}$ is randomly selected. The values of X_j are randomized in the data set according to $\{x_{jki}\}_{k\in\mathscr{L},\ i=1,2,\cdots,n_k} = s(\mathbf{x}_{jp_k})$, where $s(\cdot)$ denotes a sampling with replacement from a set of values and $\mathbf{x}_{jp_k} = \{x_{jp_ki}\}_{i=1,\cdots,n_{p_k}}$. The resampling procedure is repeated v times and the ML-MDA measure for the j-th variable is given by ML-MDA$_j = \frac{1}{v}\sum_v (L_{j,v} - L)$ where $L_{j,v}$ is 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.

3 Application

ML-NLPCA was applied to construct a JS indicator that summarizes 11 ordinal variables measuring different JS facets for $1,804$ workers employed in 115 social

[1] Since the NLPCA solution is rotationally invariant, different group solutions can fairly be compared only if their axes are rotated to a target solution by means of a Procrustes orthogonal rotation.

[2] 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 other covariates are used to predict the response Y, the prediction accuracy decreases substantially if the X_j is associated with Y.

cooperatives[3]. The 11 items (fully described in [2]) refer to the satisfaction of workers with extrinsic aspects as well as intrinsic and relational aspects.

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 (total PVAF $= 56$). The stability of such indicator was assessed by a bootstrap study. The JS indicator was then used as the response variable in the CRAGGING and the ML-MDA measure was computed for each of the 11 items. The variables that mostly contribute to the definition of the JS indicator are the vocational training and professional growth, personal fulfilment, and transparency in the relation with the cooperative while others (achieved and prospective career promotions, the relations within the team, and the recognition by co-workers) do not contribute at all. These results are in line with previous studies confirming the role of intrinsic and relational aspects in determining JS.

This preliminary work 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, in the NLPCA framework, there is a measure of the contribution of each separate variable to the composite indicator that is the VAF *per variable*. However, due to its descriptive nature and different meaning, we preferred to measure importance by ML-MDA measures. It could be interesting to investigate the relationships between the two measures and examine the possibility to use the VAF *per variable* to pre-select the items to include as covariates in CRAGGING, in order to decrease its running time.

References

1. Breiman, L.: Random Forests. Machine Learning, **45**, 5–32 (2001)
2. Carpita, M., Golia, S.: Measuring the Quality of Work: The Case of the Italian Social Cooperatives. Quality and Quantity. On Line First, DOI 10.1007/s11135-011-9515-0 (2011)
3. Carpita, M., Manisera, M.: On the Imputation of Missing Data in Surveys with Likert-Type Scales. Journal of Classification, **28**, 93–112 (2011)
4. Gifi, A.: Nonlinear multivariate analysis. Wiley, Chichester (1990)
5. Michailidis, G., de Leeuw, J.: Multilevel Homogeneity Analysis with differential weighting. Computational Statistics and Data Analysis, **32**, 411–442 (2000)
6. Vezzoli, M., Stone, C.J.: CRAGGING. In Book of Short Papers CLADAG 2007. EUM, 363–366 (2007)
7. Vezzoli, M., Zuccolotto, P.: CRAGGING measures of variable importance for data with hierarchical structure. In: S. Ingrassia, R. Rocci, M. Vichi (eds.) New Perspectives in Statistical Modeling and Data Analysis. Springer, Heidelberg (2011)

[3] Missing values were imputed according to [3]. Cooperatives with less than 10 workers were removed to improve the ML-NLPCA stability and avoid resampling problems in the ML-MDA. The data used in this study result from a preliminary Rasch analysis [2], which identified the 11 JS items as related to a "global" JS latent trait 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".