Grade of Membership and Principal Components Analysis: a Comparative Empirical Study

Suleman, Abdul
Instituto Universitário de Lisboa (ISCTE - IUL), DMQ, UNIDE, Lisboa, Portugal
Av. Forças Armadas
Lisbon (1649 - 026), Portugal
E-mail: abdul.suleman@iscte.pt

We aim to contribute to the discussion about the parallelism between principal components (PC) and a typological grade of membership (GoM) analysis. Wachter (1999) tested empirically the close relationship between both analysis. In his empirical work, the author considered two typologies for up to nine dichotomous variables and realized that individual GoM scores and the first PC scores were highly correlated. In addition, both proved effective in recovering the underlying gradient among individuals present in the dataset. Our contribution is empirical as well. We used a dataset that comes from a survey designed to study the reward of skills of retail bankers in Portugal. It comprises thirty polythomous variables structured over three typologies. Even though we had no prior information on data distribution, we show that both analysis lead retail bankers to be individually ranked by skill. However, the hierarchical skill structure of bankers becomes apparent posterior to the application of GoM analysis to data. The PC analysis, by itself, can hardly provide information about that hierarchy.

Introduction

The purpose of this paper is to contribute to the discussion around the relationship between principal component analysis (PCA) and the typological GoM analysis introduced by Woodbury and Clive (1974). This issue was addressed for the first time by Wachter (1999). In his empirical work, the author used sets of three to nine dichotomous variables, drawn from U.S. National Survey of Families and Households (NSFH), carrying implicitly age gradient among individuals. He found that the first PC and the GoM individual scores were highly correlated and also that both analysis proved effective to recover age gradient known a priori. In sequel, he attempted to recast the underlaying GoM model with a geometrical formulation, suggesting it as a version of PC under certain metrics. In any case, he fixed the number of typologies to two for any number of dichotomous variables.

Since it was first presented in 1974 by Woodbury and Clive, GoM model has been mentioned very often as a statistical tool to represent fuzzy partitions. This seems to be the most used formulation of the model both in theoretical and applied frameworks (e.g. Tolley and Manton, 1992; Berkman, Singer and Manton, 1989). Sometimes it is referred to as an alternative to PCA for discrete data, but still under fuzzy framework (Manton and Gu, 2005). In literature, it is also referred to as a member of discrete PCA models class (Buntine and Jakulin, 2004). In any case, we did not find any other attempt to connect GoM to PCA as in Wachter (1999).

The work presented in this paper is an extension of Suleman and Suleman (2011), and it is an attemp to link empirically GoM to PCA in a higher dimension than in Wachter study. We kept the original formulation of the GoM model, that is, the one of fuzzy point of view. We did not consider herein the geometrical formulation referred above. For the purpose of our study, we used a dataset that came from an original survey (Suleman, 2007) specially designed to study the reward of skills of retail bankers in Portugal. In this survey, supervisors were asked to assess skills of each retail banker from a list of thirty skill items in a 5-point Likert scale. We fixed the number of typologies to three prior to applying the GoM model to the dataset. The GoM analysis unveiled an hierarchical skill fuzzy partition. The particular distribution of retail bankers on this structure led, under certain conditions, to the definition of an utility function that assigns each banker a number so as to estimate his / her position in a skill hierarchy. The utility function is indeed a linear combination of GoM scores. However, the coefficients of such combination should meet some specific conditions.

In the second stage of our study we submitted our dataset to a PCA. We realized that the first PC is remarkably correlated with the utility function used to rank retail bankers by skill. Next, we estimated a linear regression model for the first PC to check the extent in which this quantity can be predicted by GoM scores under ordinary least squares (OLS) method. The results achieved show the same model fitness as in the previous case. In addition, the estimated GoM scores coefficients meet the conditions that lead the first PC behaving as an utility function. Of course, the number assigned to each individual depends on the utility function used.

However, we must emphasize that the hierarchical skill structure becomes apparent only after the aplication of GoM analysis to the dataset. The ranking provided by first PC scores, by itself, gives no indication about the underneath hierarchical structure. Thus, we do not see PCA as replacing GoM analysis in our study.

This paper is organized as follows. In the next section we present the data used in our empirical work. The third section describes the GoM model in a fuzzy sets theory perspective. In forth section we show the results achieved as the output of GoM model application. The fifth section is devoted to compare results from GoM and PCA. Finally, in the sixth section we present some concluding remarks.

The Data

The data used in our empirical analysis were compiled from a survey conducted by supervisors of the banking sector in Portugal. The survey was intended to analyse skills rewards of retail bankers in this country (Suleman, 2007). The supervisors were asked to assess each retail banker in thirty different skill items. They used a Likert scale from 1 to 5, with the following meaning: 1: Very Low; 2: Low; 3: Medium; 4: High; and 5: Very High. The thirty variables comprise different skill dimensions namely Knowledge, Behaviour and Attitude toward Others, Behaviour and Attitude toward the Organization and Cognitive and Technical Skills. The final sample size is $N = 593$.

In addition to those variables, the survey provides detailed information on human capital, demographic characteristics, job position and earnings of retail bankers. Besides the number of categories of educational variable of human capital, we did not use that information in the present study. Indeed, to proceed with GoM analysis, we set to three the number of typologies in accordance with the number of categories of educational variable, that is, Lower than Secondary, Secondary and Higher. This procedure follows the "tradition in economics of education to understand the sort of skills related to education" as pointed out by Suleman and Suleman (2011).

Grade of Membership Analysis

A pioneer statistical model based on fuzzy $K-$partitions, and known by the acronym GoM, was introduced by Woodbury and Clive (1974). It assumes that the population under study can be decomposed into $K \geq 2$ fuzzy sets or typologies where each individual, say individual i, is represented by his / her vector of GoM scores

$$(1) \qquad \mathbf{g}_i = (g_{i1}, g_{i2}, ..., g_{iK})$$

This coordinate vector belongs to the unit simplex

$$(2) \qquad S_K = \left\{ \mathbf{a} = (a_1, a_2, ..., a_K) : a_k \geq 0 \wedge \sum_{k=1}^{K} a_k = 1 \right\}$$

The generic GoM score g_{ik} of the vector $\mathbf{g}_i$, in (1), stands for the grade of membership of individual i in k^{th} typology. Typologies' crisp elements are referred to as pure types. In the GoM model, the number of typologies, K, is fixed a priori. The model can be formulated as follows. Let

$$\mathbf{X}_i = (X_{i1}, X_{i2}, ..., X_{iJ}), 1 \leq i \leq N$$

be the vector of outcomes of individual i in J variables, where $X_{ij} \in \{1, 2, .., L_j\}, 1 \leq j \leq J$, is a categorical variable with $L_j \geq 2$ number of categories. In our case, $N = 593$, $J = 30$, X_{ij} is the individual i assessment on j^{th} skill item and $L_j = 5$ for all skill items. The number of typologies is set to three, i.e. $K = 3$, for the reasons explained earlier.

In GoM model the outcomes X_{ij} are, by assumption, ruled in latent form by the vector $\mathbf{g}_i$, as in (1). Given $\mathbf{g}_i$, they are considered independent from each other. This means, $X_{ij}|\mathbf{g}_i$ and $X_{ij'}|\mathbf{g}_i$ $(j \neq j')$ are independent random variables. Denote by λ_{kjl} the probability of a k-typology pure type has the outcome l in j^{th} variable, i.e.,

$$\lambda_{kjl} = \Pr[X_{ij} = l \mid g_{ik} = 1]$$

where $1 \leq i \leq N; 1 \leq j \leq J; 1 \leq k \leq K; 1 \leq l \leq L_j$. Being probabilities, the λ_{kjl} verify the two conditions $\lambda_{kjl} \geq 0$ and $\sum_{l=1}^{L_j} \lambda_{kjl} = 1$, for each k and each j. The estimates of λ_{kjl} are used in practice to set-up the k^{th} typology. The basic assumption of GoM is that, given $\mathbf{g}_i$, the probability p_{ijl} of individual i having the outcome l in j^{th} variable is

$$(3) \qquad p_{ijl} = \Pr[X_{ij} = l \mid \mathbf{g}_i] = \sum_{k=1}^{K} g_{ik}\lambda_{kjl}$$

Parameters in (3), namely g_{ik} and λ_{kjl}, are estimated by the method of maximum likelihood subjected to the above referred constraints (for details about GoM model, see Manton, Woodbury and Tolley, 1994).

Empirical Findings

We split the analysis of GoM model output into two parts. The first part gives a brief account on the typological structure that emerged from the model application, i.e. the estimates of λ_{kjl}. The second part deals with the distribution of retail bankers on that structure using the estimates of g_{ik}.

The results achieved unveils an hierarchical fuzzy 3-partition. Indeed, we found two extreme typologies, one gathering the lower skill categories and the other the higher skill categories. The third typology positions between those two extremes. For example, in one variable of Knowledge dimension, the skill categories Very Low and Low were found to be predominat in say Low skills typology, the category Medium in Medium skills typology and the categories High and Very High in High skills typology. We index the emerged typologies by the numbers 1,2, and 3, respectively.

Now we concern with the estimated distribution of retail bankers on fuzzy 3-partition. We recall that a fuzzy 3-partition can be represented geometrically by an equilateral triangle, corresponding to the unit simplex S_3. The pure types lie on the vertices, the individuals that share exactly two typologies lie on the edges and the ones that share three typologies lie in the interior of the triangle. Thus, skill typologies are mapped onto triangle vertices. The individuals that lie on the edge connecting Low to Medium skill typologies are increansingly more skilled as we move from the former to the later typology along the edge. The same holds for the individuals lying on the edges Medium-High or Low-High.

We depicted the estimates of $\mathbf{g}_i = (g_{i1}, g_{i2}, g_3), 1 \leq i \leq N$, and realize that 76% of retail bankers lie on the path of edges Low-Medium and Medium-High, and only approximately 1% lie on the edge

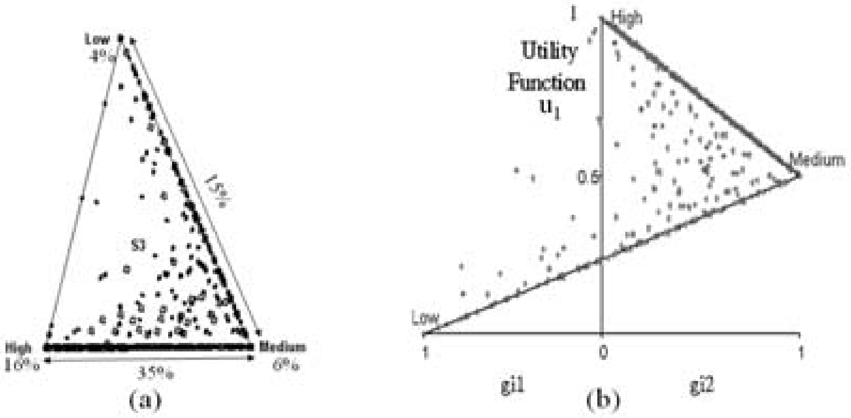

Figure 1: (a) Distribution of retail bankers on a fuzzy 3-partition; (b) Mapping of fuzzy 3-partition onto interval [0,1] through utility function u_1 (adapted from Suleman and Suleman, 2011)

Low-High (Figure 1 (a)). If we consider a vicinity of 0.1 of that path, i.e. individuals whose GoM scores verify any of the conditions $g_{i1} + g_{i2} \geq 0.9$ or $g_{i2} + g_{i3} \geq 0.9$ the rate increases to 93%. With this particular distribution of retail bankers on a fuzzy 3-partition we can use the utility function

$$u_1\left(g_{i1}, g_{i2}, g_{i3}\right) = 1 - \sum_{k=1}^{2} \eta\left(k\right) g_{ik} \tag{4}$$

where $\eta : \mathbb{R}^+ \longrightarrow [0,1]$ is a strictlty decreasing function, to get estimates of individual skill ranks (see Suleman and Suleman, 2011, for a demonstration). In this context, we set subjectively $\eta\left(k\right) = \frac{1}{k}$ in (4). We can use alternatively the utility function

$$u_2\left(g_{i1}, g_{i2}, g_{i3}\right) = \alpha_0 + \sum_{k=1}^{3} \alpha\left(k\right) g_{ik} \tag{5}$$

where α_0 is a real constant and $\alpha\left(k\right)$ a strictly increasing function, for the same purpose. A mapping of retail bankers skill fuzzy partition onto the unit interval $[0,1]$, through the utility function (4), with $\eta\left(k\right) = \frac{1}{k}$, is shown in Figure 1 (b).

Principal Components Approach

This section is devoted to the innovative aspect of our paper. We aim to compare GoM and PCA outputs in our particular setting. Before presenting the numerical results from PCA, we recall that our data are categorical and PCA in not designed for such data type. So, we assume now that the skill items are vectors of $\mathbb{R}^{30}$, where the Euclidean distance makes sense, thus allowing the calculation of an empirical covariance matrix. Although this is convenient for practical purpose, we have no formal statistical justification for such assumption. This follows closely Wachter (1999).

After submitting our dataset to a PCA, we calculated the first PC scores for each individual. The PC scores are linear combinations of the original variables. The coefficients used in those combinations are the PC loadings. In our case, the loadings for the first PC are all positive. The Pearson correlation shows that the first PC is highly correlated with utilily function u_1 (4), being the correlation coefficient $R = 0.96$ ($R^2_{adj} = 0.92$). However, we found many individuals concentrated on typologies ($u_1 = 0$ or $u_1 = 0.5$ or $u_1 = 1$) though with different PC scores (Fig. 2).

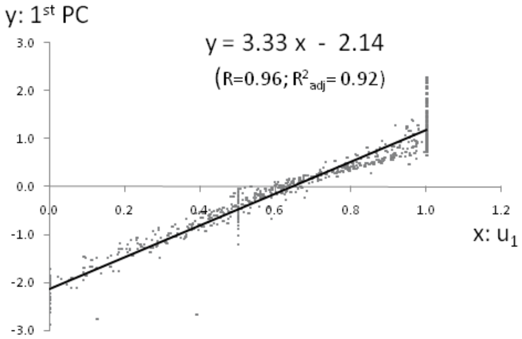

Figure 2: Association between first PC and utility function u_1

There is some possible reason for this to happen. The PC scores are unrestricted and therefore can assume any real value. The same does not hold for GoM scores as they are constrained to lie within the unit simplex S_K (2), S_3 in our case. In the estimation process, the search for an optimal unit simplex may prevent GoM analysis from discriminating "close" cases, provided the cases are effectively different. This might happen in other situations as well. Without going into detail, we give an example that helps to illustrate the idea. We found two retail bankers, call them $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$, assessed equally in all but four skill items. The observed differences for the pair $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ are as follows: $(4,5)$, $(4,5)$, $(3,4)$, and $(4,5)$. They were both estimated as pure types of High skills typology, and thus $u_1\left(\mathbf{g}_{\xi_1}\right) = u_1\left(\mathbf{g}_{\xi_2}\right) = 1$. However, their projection onto the first PC is, respectively, 0.77 and 0.99. Details about practical estimation aspects of GoM model are found in Manton, Woodbury and Tolley (1994).

In the final stage of PCA of our dataset, we estimated a regression model for the first PC so as to better understand how this quantity can be predicted by GoM scores. The model is represented by the equation

$$(6) \qquad 1^{st}PC\,(i) = \beta_0 + \beta_1 g_{i1} + \beta_2 g_{i2} + \varepsilon_i$$

where β_0 is the intercept, β_1 and β_2 are regression coefficients and ε_i is the error term. We omit the term for g_{i3} because it is redundant. Table 1 displays the model (6) parameters' estimates under OLS, as well as the associated relevant statistics. The estimated model fits adequately as measured by the adjusted coefficient of determination, $R^2_{adj} = 0.92$. Furthermore, we fail to reject the hyphotesis $\beta_1 \leq \beta_2$ (p-value $\simeq 1.0$). From this statement we can infer that the first PC behaves as an utility function similar to u_2, in (5), where $\alpha(3)$ is set to zero and noting that β_1 and β_2 are both estimated with negative sign and are statistically significant (at 5% level, see Table 1).

The results achieved reinforce the role of first PC as a ranking device. However, such ranking would probably sound different if we had no information on the latent data structure as provided in GoM analysis (see Fig. 1 (a)). Additionally, we estimated a regression model similar to (6) for all the remaining twenty nine PC. In any case the adjusted coefficient of determination was found to be nearly zero. The highest value was 0.015. In our particular study, we do not see PCA as replacing GoM but we otherwise see it as a complementary analysis.

			95% Confident Interval	
Coefficient		Std. Error	Lower Bound	Upper Bound
β_0	1.17	0.021	1.13	1.21
β_1	−3.37	0.044	−3.46	−3.28
β_2	−1.60	0.036	−1.66	−1.52

Table 1: First PC regression model estimates

Concluding Remarks

In this study, we tried to contribute to the discussion about the relationship between PCA and GoM in a similar way as Wachter (1999). Differently from this author, we used a more complex dataset comprising thirty polythomous variables structured over three typologies. We subjected it to both analysis methods and realized that, despite having some common features, GoM provided a different and perhaps deeper insigth into data structure. In his empirical work, Wachter realized that GoM scores and first PC were "remarkably close" in the framework of the lowest possible dimension for an GoM analysis, i.e. $K = 2$. In a slightly higher dimension framework ($K = 3$), we showed that the first PC behaves as a linear function of two non redundant GoM scores. Consequently, we do not see how to replace GoM with PCA in our particular study, without loss of detail.

However, both studies give room and incentive for a research to find out the real reasons that make GoM so closely related to PCA. We are particularly focused on developing a research on this area.

Acknowledgments

The author would like to thank Fátima Suleman for providing the data on skill assessment in banking sector in Portugal, and Marina Coutinho and Elizabeth Reis for their suggestions and comments on the earlier version of this paper.

REFERENCES

Berkman, L., Singer, B., & Manton, K. (1989). Black / White Differences in Health Status and Mortality Among Eldery. *Demography* 26, 661-678.

Buntine, W., & Jakulin, A. (2004). Applying Discrete PCA in Data Analysis. *Proceeding UAI '04 Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 59-66 (available on-line http://portal.acm.org/citation.cfm?id=1036851).

Manton, K.G., Woodbury, M.A., & Tolley, D. (1994). Statistical Applications Using Fuzzy Sets. John Wiley & Sons, Inc.

Manton, K.G., & Gu, X. (2005). Disability Declines and Trends in Medicare Expendidure. *Ageing Horizons*, Issue No. 2, 25-34.

Suleman, F. (2007). O valor das Competências: Um Estudo Aplicado ao Sector Bancário. Livros Horizonte. Lisboa.

Suleman, A., & Suleman, F., (2011). Ranking by Competence using a Fuzzy Approach. *Quality and Quantity* (Forthcoming). DOI: 10.1007/s11135-010-9357-1.

Tolley, D., & Manton, K.G. (1992). Large Sample properties of Estimates of a Discrete Grade of Membership Model. *Annals of Institute of Statistical Mathematics* 44, 85-95.

Wachter, K.W. (1999). Grade of Membership Models in Low Dimensions. *Statistical Papers* 40, 439-457.

Woodbury, M.A., & Clive, J., (1974). Clinical pure types as a fuzzy partition. *Journal of Cybernetics* 4, 111-121.