Quaderni di Statistica
Vol. 11, 2009

Missing values treatment with interval imputation in satisfaction measurement

Paola Zuccolotto
Dipartimento Metodi Quantitativi, Università degli Studi di Brescia
E-mail: zuk@eco.unibs.it

Summary: In the framework of satisfaction measurement many statistical techniques have been proposed in order to face two main questions: firstly the fact that the analyses usually involve ordered categorical variables, secondly the presence of missing values, due to some subjects' incapacity or unwillingness to give answers to one or more points. In this paper Interval Imputation, a recent technique for missing values treatment, based on the idea of symbolic data analysis, is employed with nonlinear principal component analysis in the context of satisfaction measurement. The performance of the proposed strategy is compared with the traditional procedures by means of a simulation study.

Keywords: Missing values, symbolic data analysis, Interval Imputation, satisfaction measurement.

1. Introduction

In the recent years a wide interest has been devoted to statistical techniques aimed at evaluating and measuring satisfaction. The analyses are usually carried out on the basis of the judgments acquired from the subjects, by means of a structured questionnaire. The main method for acquiring satisfaction judgments is to arrange a set of items, asking for satisfaction judgments about several different aspects of the product/service (facet satisfaction, FS). With this approach, specific statistical techniques are usually employed, in order to obtain one or more composite satisfaction indexes, taking account of all the single FS judgments.

In the construction of a composite satisfaction index, two main prob-

lems have to be faced: (1) the ordinal scale of FS judgments, typically obtained through Likert-type variables and (2) the possible presence of missing values. For both these problems, several techniques have been proposed in the literature.

In this paper we deal with the dimensionality reduction tool called nonlinear principal component analysis (NL-PCA - Gifi, 1990). NL-PCA is widely used in the context of the construction of composite satisfaction indexes because it assigns numerical values to the categories of qualitative variables by means of an optimal scaling procedure. Thus it is able to properly treat ordinal-scaled FS judgements.

From the point of view of missing values treatment, we focus attention on Interval Imputation (InI -Zuccolotto, 2008, 2009), a technique based on the idea of filling blanks using intervals instead of specific imputed values.

The main aim of this paper is to investigate the performance of InI when used with NL-PCA in the framework of satisfaction measurement and to compare it with the approaches to missing values treatment traditionally used with NL-PCA. This is made by means of a simulation study.

The paper is organized as follows: in section 2 the InI technique is recalled and explained using both an illustrative example and a real data case study about job satisfaction measurement. Section 3 illustrates the simulation study and summarizes its main results, some final remarks follow in section 4.

2. Missing values treatment with InI

Interval Imputation (Zuccolotto, 2008, 2009) is a procedure for missing values treatment, consisting in the replacement of every missing value with an interval ranging from the minimum to the maximum value admissible for the concerned variable. The resulting dataset is thus composed by single-valued and interval-valued measurements mixed and has to be processed using specific techniques.

The theoretical apparatus we can refer to is that of symbolic data analysis (SDA - Diday, 1987). SDA deals with data matrices whose generic element x_{ij} is not necessarily a single quantitative or categorical value,

but can be, for example, a distribution (histogram-type variables) as well as an interval (interval-type variables), or a set of values linked by some logical rule. In this framework, some methods have been proposed in order to extend classical dimensionality reduction methods to symbolic data. The problem has been explicitly formalized by Godwa *et al.* (1995), and then further investigated from different points of view. Cazes *et al.* (1997) proposed two popular methods for PCA on interval data, called *vertices method* and *centers method*, later refined by Lauro and Palumbo (2000). About the PCA for histogram data, Rodríguez *et al.* (2000) proposed an algorithm also working if the data table has variables of interval-type and histogram-type mixed. Giordani and Kiers (2004) introduced a method for three-way component analysis of interval data.

In this paper we follow the popular vertices method of Cazes *et al.* (1997), whose basic observation is that a subject measured by means of p interval-type variables, can be considered a p-dimensional hyperrectangle. Symbolic principal component analysis (SPCA) is then carried out by applying traditional PCA to the data matrix $\Xi_{N2^p \times p}$ obtained stacking below each other all the 2^p vertices of the hyperrectangles corresponding to each subject. The dimensionality reduction is obtained by projecting all the hyperrectangles' vertices in the first q $(q < p)$ factorial axes. For example, with $q = 2$, the two-dimensional representation visualizes the projections of all the hyperrectangles' vertices in the plane spanned by the 1st and the 2nd Principal Component (PC). Thus, according to vertices method, each subject is represented by means of an irregularly shaped two-dimensional scattering, which can be visualized by means of its maximum covering area rectangle or its convex hull.

The vertices method can be applied to the particular data matrix generated by InI because the single-valued measurements it contains can be considered as limiting cases of degenerate intervals. Let $\mathbf{X}$ be the original data matrix, containing some missing values ($miss$)

$$\mathbf{X}_{N\times p} = \begin{bmatrix} x_{11} & miss & x_{13} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & x_{N3} & \dots & miss \end{bmatrix},$$

the "imputed" matrix is given by

$$\mathbf{X}(InI)_{N\times p} = \begin{bmatrix} x_{11} & [x_{2,\min}; x_{2,\max}] & x_{13} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & x_{N3} & \dots & [x_{p,\min}; x_{p,\max}] \end{bmatrix},$$

where $[x_{j,\min}; x_{j,\max}]$ denotes the closed interval ranging from the minimum to the maximum value of j-th variable.

Since in the matrix $\mathbf{X}(InI)$ there is a relatively small number of non-degenerate intervals, the $(N2^p) \times p$ matrix $\mathbf{\Xi}$, required by the vertices method, results largely oversized. The presence of many single-valued measurements allows to construct a much smaller matrix, because a subject i with m_i missing values is a m_i-dimensional hyperrectangle in the p-dimensional space and its hyperrectangle has only 2^{m_i} vertices. On the other hand, we need to give the same weight to all the subjects. So, let $M = \max(m_i)$, each subject is expanded into a $2^M \times p$ matrix $\mathbf{X}i$, $(i = 1, 2, \dots, N)$, containing the vertices of its hyperrectangle, replicated 2^{M-m_i} times.

The matrix $\mathbf{\Xi}(InI)$ is then obtained stacking below each other all the $2^M \times p$ matrices corresponding to each subject and traditional PCA can be applied. Dimensionality is reduced by projecting all the vertices in the first q $(q < p)$ factorial axes. We denote with $\mathbf{y}_{hi}$ the $2^M \times 1$ vector containing the projections of the vertices of subject i on the h-th factorial axis. The vector $\mathbf{y}_{hi}$ contains 2^{m_i} different values.

When $q = 2$, subjects can be represented in the plane spanned by the first two PCs. Observations with $m_i > 0$ project into a scatter which can be visualized by means of an irregularly shaped two-dimensional polygon. This highlights the uncertainty deriving from the presence of some missing values. On the other hand, when $m_i = 0$, that is with a complete observation, the degenerate hyperrectangle is a point in the p-dimensional space, as in the traditional data analysis, and it projects into a single point.

The InI method has at least four major advantages:

1. subjects with missing values are maintained in the dataset, and the information they contain, even if partial, is in any case recovered;

2. the procedure does not need to replace the missing with subjectively imputed values;

3. the particular final representation of subjects with missing values allows to appreciate their peculiar fuzzy condition, in addition the larger the number of missing of a given subject, the larger the area of its polygon;

4. for a given subject, the missing information differently affects its graphical representation in the factorial plane, in accordance with the scores of the variables where the missing are located.

In the context of FS, a missing value occurs when a subject is not able, or does not want, to formulate a judgment about a given item. If FS judgments are acquired by means of a Likert scale with k categories, according to InI each missing is replaced by the interval $[1; k]$. In the context of satisfaction measurement NL-PCA can be applied to the vertices matrix in place of PCA, in order to take account of the ordinal scale of Likert-type variables.

The composite indexes measuring satisfaction levels derive from one-dimensional representations of the subjects, for example projections onto the first factorial axis. In many empirical analyses we are interested in the average of these indexes within groups defined according some categorical variable G, whose influence on satisfaction we want to investigate. In order to take account of the fuzzy condition of incomplete observations, the average indexes should be themselves interval-valued (Zuccolotto, 2008). The interval-valued average satisfaction index deriving from projections on the h-th factorial axes of the N_g subjects belonging to group G_g can be computed as

$$\bar{y}_{h,G_g} = \left[\frac{1}{N_g} \sum_{i \in G_g} \min(\mathbf{y}_{hi}); \frac{1}{N_g} \sum_{i \in G_g} \max(\mathbf{y}_{hi}) \right]. \tag{1}$$

If a group contains only complete observations, the extremes of its average composite index are coincident.

The functioning of InI can be made more comprehensible by means of an illustrative example on simulated data and a real data case study.

2.1. An illustrative example

In a customer satisfaction survey $N = 105$ subjects have expressed satisfaction judgments about $p = 9$ items, in a Likert scale ranging from 1 to 5, with higher scores reflecting more satisfaction. The subjects are divided into 4 groups. A variable number of missing values affect 8 subjects, all belonging to group 1, with $M = 4$. The data matrix is given by

$$\mathbf{X} = \begin{bmatrix} miss & 5 & 4 & 3 & 2 & 5 & 5 & 5 & 4 \\ 3 & 2 & 3 & 3 & 1 & miss & 4 & 2 & 1 \\ miss & miss & 3 & 4 & 5 & 3 & 4 & 5 & 5 \\ 3 & 3 & 3 & 3 & 4 & miss & miss & 5 & 5 \\ 5 & 5 & miss & miss & miss & 5 & 3 & 2 & 3 \\ 4 & 3 & 5 & 4 & 4 & 4 & miss & miss & miss \\ 2 & 1 & 1 & 1 & miss & miss & 1 & 1 & 1 \\ 4 & 3 & 3 & miss & miss & miss & miss & 2 & 2 \\ 4 & 5 & 5 & 5 & 5 & 5 & 5 & 4 & 4 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 & 1 & 3 & 4 & 4 & 3 \end{bmatrix},$$

and the matrix $\mathbf{X}(InI)$ is obtained replacing each missing with the interval $[1; 5]$.

Each subject is expanded into a matrix with $2^M = 16$ rows containing the vertices of its hyperrectangle replicated 2^{4-m_i} times. For example, subject 1 is a 1-dimensional rectangle ($m_1 = 1$), actually a segment in the 9-dimensional space with two vertices given by the vectors $(1\ 5\ 4\ 3\ 2\ 5\ 5\ 5\ 4)'$ and $(5\ 5\ 4\ 3\ 2\ 5\ 5\ 5\ 4)'$. Its expanded matrix $\mathbf{X1}$ is obtained replicating 8 times the two vertices of its (degenerate) hyperrectangle

$$\mathbf{X1}_{16\times 9} = \begin{bmatrix} 1 & 5 & 4 & 3 & 2 & 5 & 5 & 5 & 4 \\ 5 & 5 & 4 & 3 & 2 & 5 & 5 & 5 & 4 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 5 & 4 & 3 & 2 & 5 & 5 & 5 & 4 \\ 5 & 5 & 4 & 3 & 2 & 5 & 5 & 5 & 4 \end{bmatrix}.$$

Subject 3 is a 2-dimensional rectangle ($m_3 = 2$) with 4 vertices given by the vectors $(1\ 1\ 4\ 3\ 2\ 5\ 5\ 5\ 4)'$, $(5\ 5\ 4\ 3\ 2\ 5\ 5\ 5\ 4)'$, $(1\ 5\ 4\ 3\ 2\ 5\ 5\ 5\ 4)'$ and $(5\ 1\ 4\ 3\ 2\ 5\ 5\ 5\ 4)'$. Its expanded matrix $\mathbf{X3}$ is obtained replicating 4 times its 4 vertices.

$$\mathbf{X3}_{16\times 9} = \begin{bmatrix} 1 & 1 & 3 & 4 & 5 & 3 & 4 & 5 & 5 \\ 5 & 5 & 3 & 4 & 5 & 3 & 4 & 5 & 5 \\ 1 & 5 & 3 & 4 & 5 & 3 & 4 & 5 & 5 \\ 5 & 1 & 3 & 4 & 5 & 3 & 4 & 5 & 5 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 3 & 4 & 5 & 3 & 4 & 5 & 5 \\ 5 & 5 & 3 & 4 & 5 & 3 & 4 & 5 & 5 \\ 1 & 5 & 3 & 4 & 5 & 3 & 4 & 5 & 5 \\ 5 & 1 & 3 & 4 & 5 & 3 & 4 & 5 & 5 \end{bmatrix}.$$

Similarly, subjects 5 and 6 are 3-dimensional rectangles (cubes in the 9-dimensional space) and their 8 vertices are replicated 2 times, subject 8 is a 4-dimensional hyperrectangles with 16 vertices and finally subjects with no missing are expanded by replicating 16 times their complete observation. Subsequently, the matrix $\mathbf{\Xi}(InI)$ has dimension (1680×9) and NL-PCA is applied.

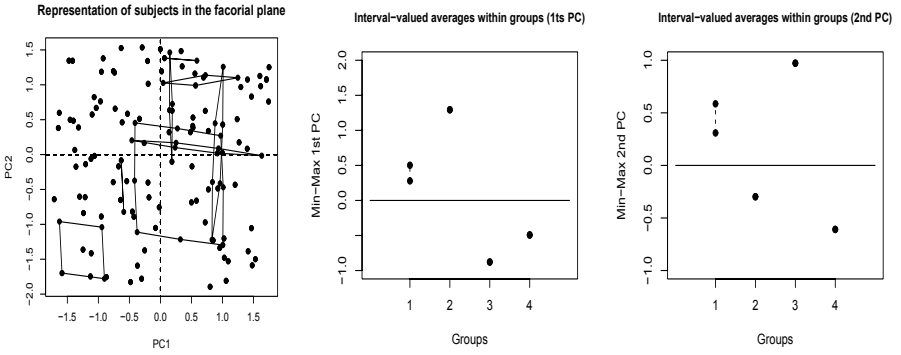

Figure 1. Main results of NL-PCA: projections on the first two factorial axes (left), interval-valued average satisfaction indexes within groups (middle and right)

The 2-dimensional representation accounts for 69.6% explained variance. The first two PCs are highly correlated with items 1 through 5 and 6 through 9, respectively. The projection of subjects in the factorial plane (Figure 1, left) shows 8 irregularly shaped polygons and 97 points. About the polygons we can remark the following main observations: (1) subjects with $m_i = 1$ are represented by segments, (2) the area of the polygons enlarges when m_i increases (the largest polygon corresponds to subject 8), (3) the shape of the polygons depends on where the missing are located. For example, subjects with missing values on the items 1 through 5 have a large base and a narrow height, thus exhibiting their uncertainty about only the 1st PC.

The composite satisfaction indexes given by the first two PCs can be averaged within groups, as specified in (1). Since the incomplete observations are gathered in group 1, its average index is interval-valued, differently to those of other groups which result single-valued (Figure 1, middle and right).

2.2. A real data case study

The data set analyzed in this section derives from ICSI2007, a wide survey about the social service sector, carried out by a team of researchers from six Italian universities (Bergamo, Brescia, Milano, Napoli Federico II, Reggio Calabria, Trento). It involved 320 social cooperatives, selected through a stratified sampling from 24 strata described by three variables: Type (A or B)[1], Area of Italy (North-East, North-West, Center or South), Number of employed workers (15 or less, from 16 to 49, 50 or more). For each social cooperative a sample of different types of workers (with a total of 4134 paid workers, 289 voluntary workers and 287 managers) was requested to fill in a specific questionnaire, asking for information about personal and professional characteristics, work-related attitudes, job satisfaction, relationships with end users, colleagues and superiors. Further

[1] Social cooperatives of type A manage social-assistance and educational services whereas social cooperatives of type B run activities focused on training and job finding for disadvantaged people.

details on the survey and on its main results are in Borzaga(2007) and in Carpita (2009).

In this analysis we focus on the job satisfaction (JS) of the $N = 1084$ paid workers whose job consists specifically in supplying the service (administrative tasks, technical assistance, cleaning services, ... are thus excluded) and assert to have *always* direct relations with end users and their families. Since it is likely that workers with greatly different tasks have different perceptions about their job, with this choice only workers directly involved in the mission of the cooperative are examined.

In the paid workers questionnaire of ICSI2007, facet JS judgements are asked for by means of 23 items (Table 1) where respondents indicated, on a scale ranging from 1 to 7, the level of satisfaction with respect to each aspect. Higher scores reflect more satisfaction and the median score 4 explicitly indicates neutrality ("neither satisfied nor unsatisfied"). The first 12 items are relative to *extrinsic* aspects of JS, concerned with remunerative, organizational and environmental matters, whereas the remaining items deal with *intrinsic* JS, that is with the relational and motivational dimension.

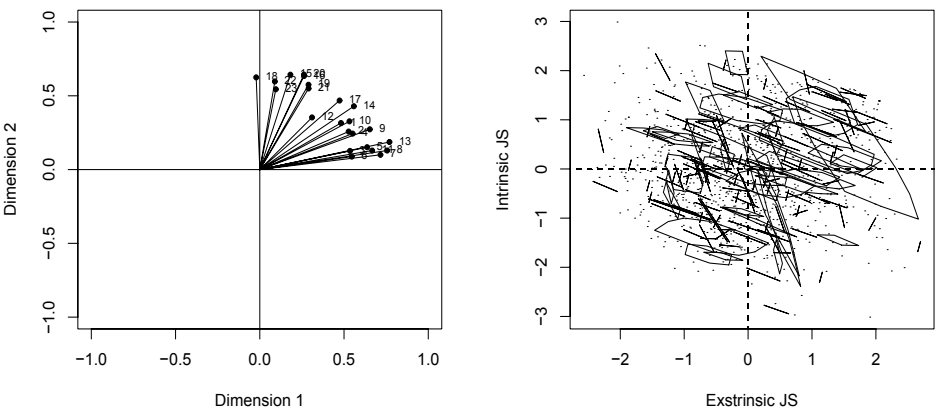

Figure 2. Main results of NL-PCA: varimax rotated factor loadings plot (left), projections of the subjects in the factorial plane (right)

Table 1. Items of job satisfaction

Id	**How satisfied are you with...**	**Short name**
1	... the working hours schedule?	HOURS
2	... the working hours flexibility?	FLEX
3	... the job stability?	STAB
4	... the workplace environment?	ENVIR
5	... the social security protection and benefits?	WELFARE
6	... your total pay (including possible fringe benefits)?	PAY
7	... your involvement in the cooperative decisions?	INVOLV
8	... the transparency in your relations with the cooperative?	TRANSP
9	... your vocational training and professional growth?	GROWTH
10	... your decisional and operative independence?	INDEP
11	... your achieved and prospective career promotions?	CAREER
12	... the consistence with your education and vocational training?	CONSIST
13	... the recognition by the cooperative of your work?	COOP-RECOG
14	... your personal fulfillment?	FULFIL
15	... the relations with your colleagues?	COLLEAG
16	... the relations within the team?	TEAM
17	... the relations with your superiors?	SUPER
18	... your relations with end users and their families?	USERS
19	... the variety and creativity of your work?	VARIETY
20	... the recognition by colleagues of your work?	COLL-RECOG
21	... the social recognition?	SOCIAL-RECOG
22	... the end users recognition?	USER-RECOG
23	... the usefulness of your work for end users and their families?	USEFULNESS

In the analyzed dataset there are about 15% incomplete observations (163 subjects out of the 1084 respondents), with a total of $\sum_{i=1}^{N} m_i = 248$ missing values (about 1% of the 1084×23 data matrix) and a maximum number of missing per row of $M = 5$. The expanded data matrix $\boldsymbol{\Xi}(InI)$ has dimension 34688×23. Due to the ordinal scale of the Likert-type variables, NL-PCA is used in order to obtain composite satisfaction indexes. The inspection of the varimax rotated factor loadings plot (Figure 2, left) shows that extrinsic and intrinsic items tend to be separately correlated to the 1st and the 2nd PC, respectively. The first two dimensions

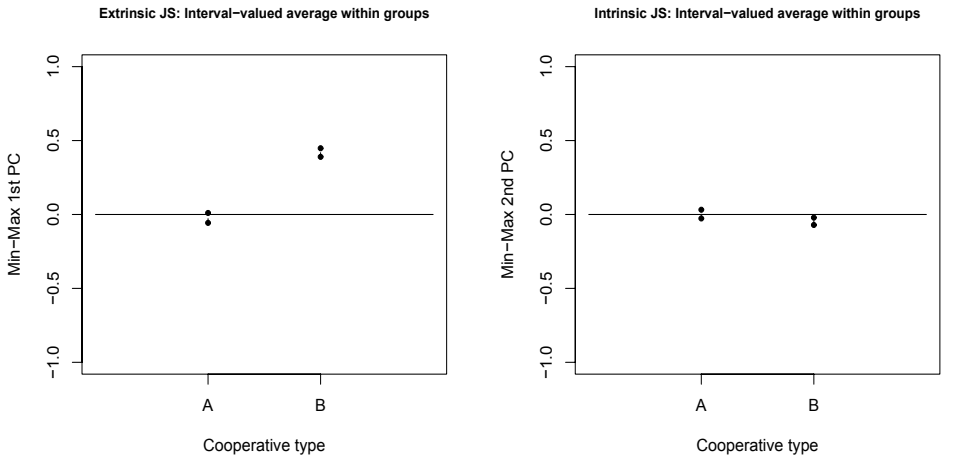

Figure 3. Interval-valued averages of composite satisfaction indices; groups defined according to the categorical variable Type of cooperative

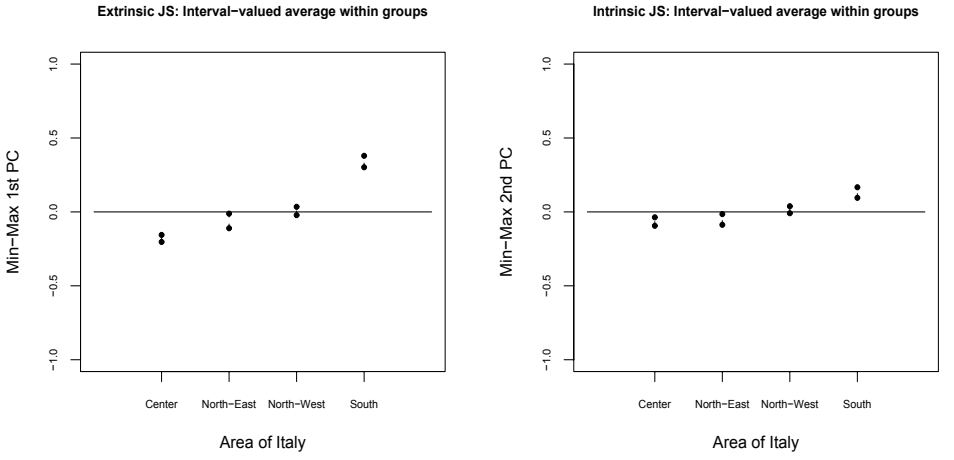

Figure 4. Interval-valued averages of composite satisfaction indices; groups defined according to the categorical variable Area of Italy

altogether account for a 41.7% explained variance. In the right part of Figure 2 the projections in the factorial plane are represented. Due to the high number of observations, the graphic is rather difficult to interpret and

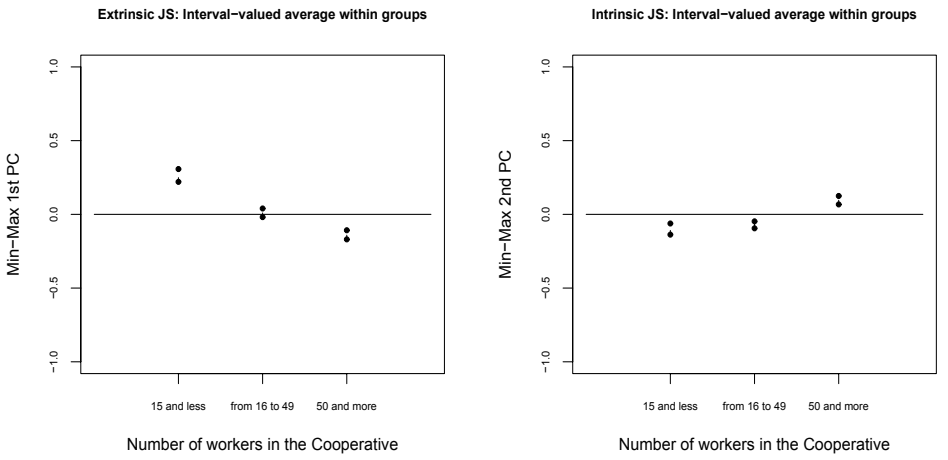

Figure 5. Interval-valued averages of composite satisfaction indices; groups defined according to the categorical variable Number of employed workers

interval-valued averages within groups are much more insightful (Figures 3 through 5). More specifically we observe that, in spite of the presence of missing values, all the intervals are quite narrow and the differences among groups, when present, can be easily appreciated.

3. A simulation study

With the method of NL-PCA two main strategies for missing values treatment are usually employed: (1) *passive*, consisting in setting weights in the loss function equal to zero (De Leeuw and Van Rijckevorsel, 1980), (2) *active*, consisting in imputating the mode of the variable where the missing is located. The terminology is the same used in the SPSS package CatPCA (Meulman and Heiser, 1999). The first method is the default option of the R package `homals - version 0.9-7` (De Leeuw and Mair, 2007), which has been used for the computations presented in this paper.

In this section a simulation study is carried out in order to compare the performance of InI with the strategies passive and active. A dataset

without missing values has been generated with $N = 100$ subjects and $p = 10$ Likert scaled items with $k = 5$. After applying NL-PCA, the first two PCs account for 68.7% explained variance and are highly correlated with items 1 through 5 and 6 through 10, respectively. A categorical dichotomous variable has been defined[2] after the dimensionality reduction so as to divide subjects into two groups G_1 and G_2 characterized by average projections appreciably different on the 1st PC (-0.75 and 0.99, respectively), and quite similar in the 2nd PC (-0.06 and 0.07).

In order to investigate different degrees of missing values contamination, three series of simulations have been carried out: with 50 (5% of the 100×10 data matrix), 100 (10%) and 200 (20%) missing values, with $R = 500$ repetitions for each series. Hereafter the three series of simulations will be denoted with L, M, H (low, medium, high contamination). The missing have been generated with a MCAR mechanism (Missing Completely At Random - Allison, 2002), drawing from a uniform distribution the position ij in the data matrix $\mathbf{X}$.

The results have been evaluated in terms of computability (does NL-PCA reach a solution?), variance accounted for (VAF) by the two dimensional representation, bias and EQM in the estimation of loadings and average projections within groups. In addition the interval-valued indexes (1) are evaluated in terms of the frequency of cases they include the real group average and the frequency of cases they are disjoint.

Table 2. Frequency of cases (%) `homals` *iterations reach convergence*

passive			active			InI		
L	M	H	L	M	H	L	M	H
61	57.6	50.5	100	100	100	100	100	100

About the points mentioned above, we can draw the following remarks:

- *Computability:* the default option of the R package `homals -version 0.9-7` seems to have serious problems in handling mis-

[2] The grouping variable assumes value 0 (group G_1) for each subject with negative object score in the 1st PC and 1 otherwise.

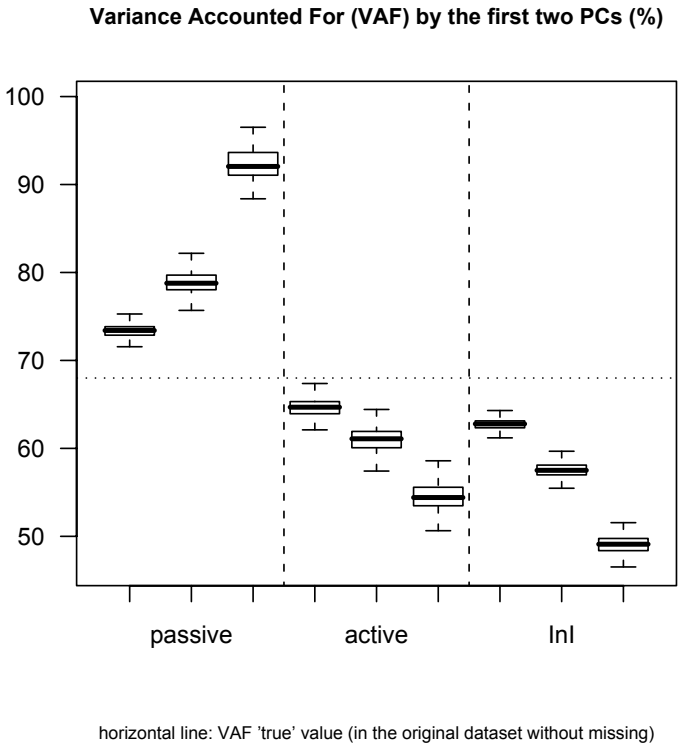

Figure 6. VAF boxplots (the three boxplots corresponding to each technique refer to the three series of simulations: L, M, H)

sing values, especially with heavy contamination. The passive strategy fails to reach convergence in a great number of cases (Table 2) and this makes this procedure potentially impossible to use with many datasets. It is not clear if it is due to anomalies generated in the loss function by the presence of the zero weights or if it is a bug of the R package. The problem has never occurred with the active strategy and with InI.

- *VAF:* a dataset with missing values contains less information than the corresponding complete dataset. So one would reasonably expect a decrease in the VAF of the two-dimensional representation.

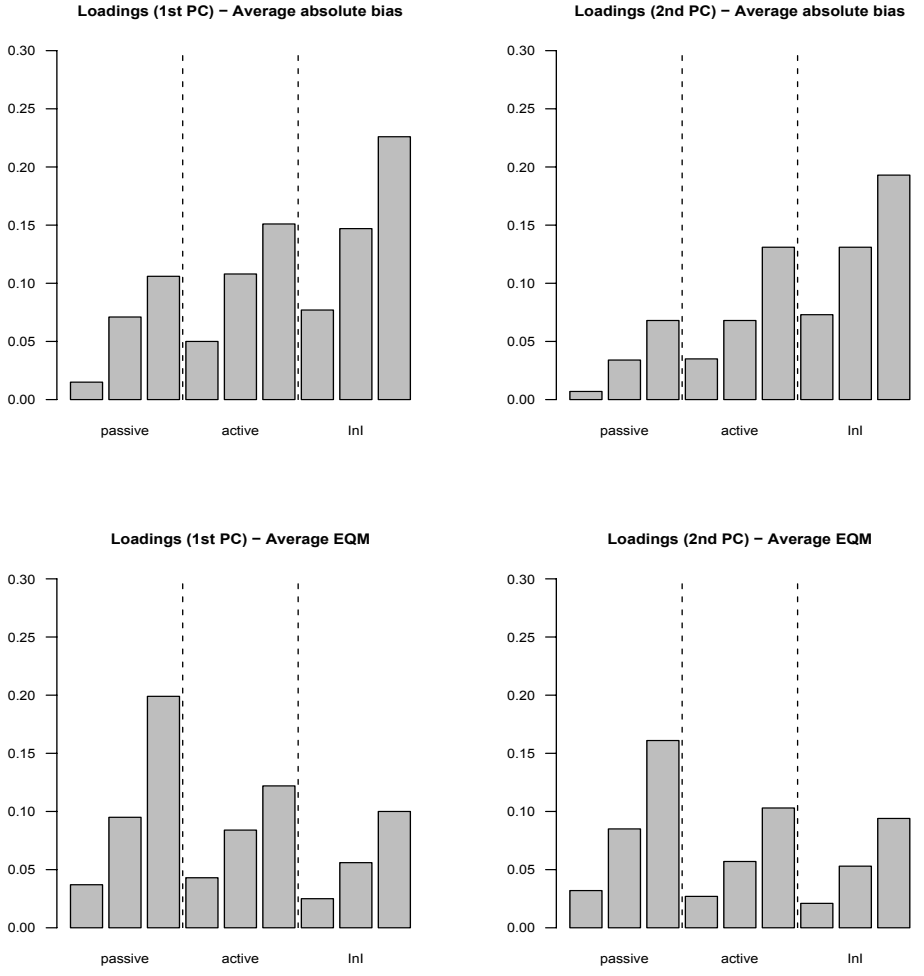

Figure 7. Bias and EQM of loadings estimation (the three bars corresponding to each technique refer to the three series of simulations: L, M, H)

This is exactly what happens with the active strategy and InI, where we observe a greater decrease in the dataset where the contamination is more severe (Figure 6). Quite the opposite, with the passive

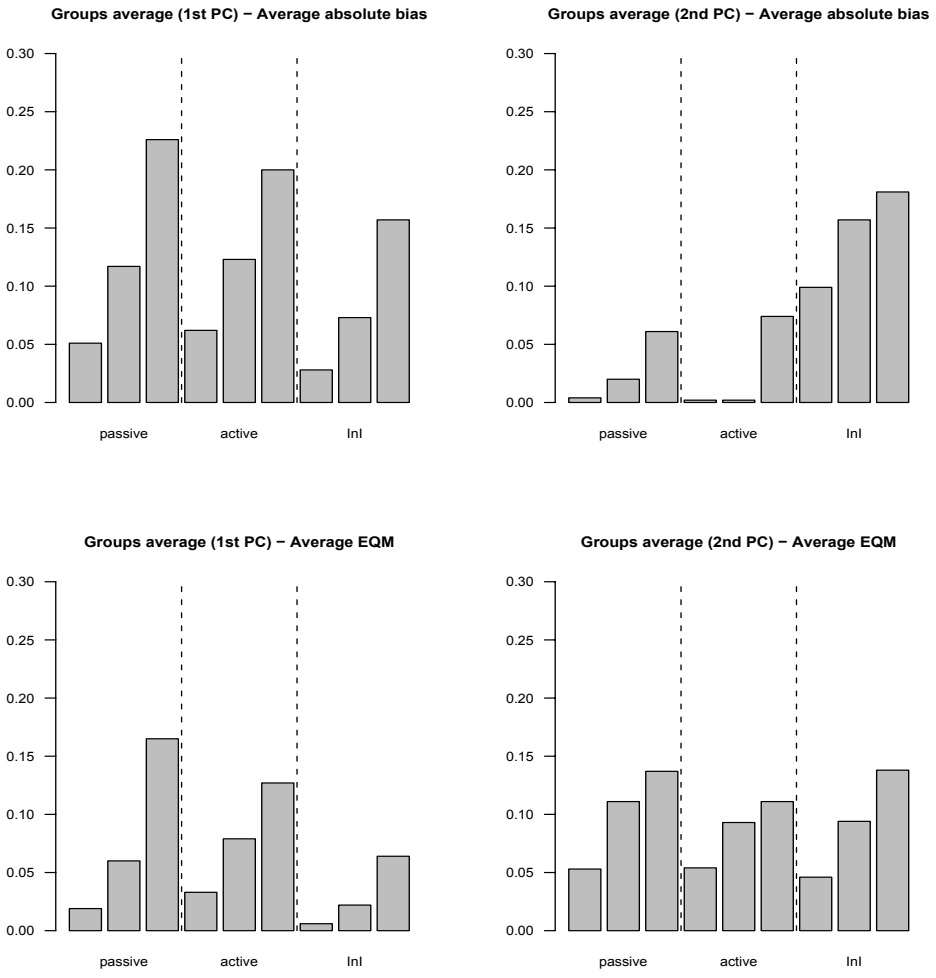

Figure 8. Bias and EQM of groups averages estimation (the three bars corresponding to each technique refer to the three series of simulations: L, M, H)

strategy the dimensionality reduction seems to be improved by the presence of missing values. This surprising odd result is probably due to the fact that with the passive strategy missing data do not

Table 3. Results about the interval-valued indexes provided by InI

frequency of cases (%) ...	L	M	H
... the interval $\bar{y}_{1,G_1}$ includes the real value	97	96.8	97
... the interval $\bar{y}_{1,G_2}$ includes the real value	94.8	94.8	95.5
... the intervals $\bar{y}_{1,G_1}$ and $\bar{y}_{1,G_2}$ are disjoint	93.8	94.8	95.5
... the interval $\bar{y}_{2,G_1}$ includes the real value	37.6	51.4	84
... the interval $\bar{y}_{2,G_2}$ includes the real value	30.2	35.8	53.5
... the intervals $\bar{y}_{2,G_1}$ and $\bar{y}_{2,G_2}$ are disjoint	11.4	3.4	0

enter in the computation, which is carried out on the basis of available data. Thus a prudent care in the evaluation of VAF is strongly recommended when the passive strategy is employed.

- *Bias and EQM of loadings estimation:* with each technique $p = 10$ loadings for each PC are computed. For a given technique and at a given level of missing values contamination, let $b_{h,j} = av_R\{l_{h,j} - \hat{l}_{h,j}\}$ be the average over the $R = 500$ repetitions of the difference between the loading of h-th PC for j-th variable and its estimated value by means of the incomplete data matrix. The top of Figure 7 displays the average absolute bias for the 1st and the 2nd PC, given by $p^{-1}\sum_{j=1}^{p}|b_{1,j}|$ and $p^{-1}\sum_{j=1}^{p}|b_{2,j}|$, respectively. Similarly, let $EQM_{h,j} = av_R\{(l_{h,j} - \hat{l}_{h,j})^2\}$ be the average of the quadratic difference between the loading and its estimate. The bottom of Figure 7 displays the average EQM, $p^{-1}\sum_{j=1}^{p} EQM_{1,j}$ and $p^{-1}\sum_{j=1}^{p} EQM_{2,j}$, for the three techniques, with the three different levels of contamination. As a whole, we notice that InI tends to outperform the other techniques from the point of view of EQM, while the passive strategy has the best performance in terms of bias.

- *Bias and EQM of groups averages estimation:* the same computations described for evaluating bias and EQM of loadings estimates have been carried out for the estimates of the two groups average projections in the 1st and in the 2nd PC. In order to compare the InI technique with the others from this point of view, the midpoints of

intervals $\bar{y}_{1,G_1}$, $\bar{y}_{1,G_2}$, $\bar{y}_{2,G_1}$, $\bar{y}_{2,G_2}$ have been considered, even if this operation is somehow inconsistent with the InI philosophy. Results are shown in Figure 8 and we can remark nearly the same conclusions as for loadings estimation.

- *Results about the interval-valued indexes provided by InI:* for the InI technique a further analysis is possible, thanks to the interval-valued nature of the groups average projections estimates. Table 3 displays the frequency of cases the intervals include the real group average and the frequency of cases they are disjoint. From this point of view we observe a very good performance in the 1st PC, where the two groups are appreciably different.

4. Final remarks

In this paper Interval Imputation, a recently proposed technique for missing values treatment, is used in the context of satisfaction measurement, together with nonlinear PCA. Its functioning is explained by means of an illustrative example and a real data case study. The performance of InI is compared to the traditional strategies used in this context, by means of a simulation study. The most interesting results show that the passive strategy, although capable of offering good estimates of loadings and groups averages in terms of bias, is affected by computational problems and seriously overestimates the explained variance of the dimensionality reduced representation. InI, on its hand, outperforms the other strategies from the point of view of EQM and offers interesting advantages in terms of the graphical interpretation of the subjects with missing values. In addition it provides interval-valued estimates of groups averages.

Further research should formalize the evidences obtained with simulations studies and explore the performance of InI with other data analysis techniques, such as for example cluster analysis.

References

Allison P.D. (2002), *Missing data*, Sage Publications, Thousand Oaks.

Borzaga C. (editor) (2007), Quando le risorse umane fanno la differenza: il modello imprenditoriale delle cooperative sociali. Primi risultati di ICSI 2007: la nuova Indagine sulle Cooperative Sociali Italiane, *Impresa Sociale*, 3.

Carpita M. (editor) (2009), *La qualità del lavoro nelle cooperative sociali, misure e modelli statistici*, Franco Angeli, Milano.

Cazes P., Chouakria A., Diady E. and Schektman Y. (1997), Extension de l'analyse en composantes principales à des donnèes de type intervalle, *Revue de Statistique Appliquèe*, 45, 5–24.

De Leeuw J., Mair P. (2007), Homogeneity analysis in R: the package homals, Preprint available at http://gifi.stat.ucla.edu/homalsR.pdf.

De Leeuw J., Van Rijckevorsel J., (1980), HOMALS and PRINCALS - Some generalizations of principal components analysis, in E. Diday *et al.* (eds.), *Data analysis and informatics*, North-Holland, Amsterdam.

Diday E. (1987), *Introduction l'approche symbolique en Analyse des Donnes*, Premiere Journees Symbolique-Numerique, Universitè de Paris IX Dauphine.

Gifi A. (1990), *Nonlinear multivariate analysis*, J.Wiley, Chichester.

Giordani P., Kiers H.A.L. (2004), Three-way component analysis of interval-valued data, *Journal of Chemometrics*, 18, 253–264.

Godwa K.C., Diday E., Nagabhushan P. (1995), Dimensionality reduction of symbolic data, *Pattern Recognition Letters*, 16, 219–223.

Lauro N.C., Palumbo F. (2000), Principal component analysis of interval data: a symbolic data analysis approach, *Computational Statistics*, 15, 1, 73–87.

Meulman J.J., Heiser W.J., SPSS (1999), *SPSS Categories 10.0*, SPSS Inc., Chicago.

Rodríguez O., Diday E., Winsberg S. (2000), Generalization of the principal components analysis to histogram data, *Proc. PKDD2000*, Lyon, France.

Zuccolotto P. (2008), A symbolic data approach for missing values treatment in principal component analysis, *Statistica & Applicazioni*, 6, 153–180.

Zuccolotto P. (2009), Principal Component Analysis with Interval Imputed missing values, *Rapporti di ricerca del Dipartimento di Metodi Quantitativi*, Università di Brescia, n. 335.