SYRTO

SYstemic Risk TOmography
Signals, Measurements, Transmission Channels, and Policy Interventions

Imputing missing values in multivariate data sets with mixed-type variables

Maurizio Carpita, Marica Manisera

SYRTO WORKING PAPER SERIES
Working paper n. 2 | 2015

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 320270.

Imputing missing values in multivariate data sets with mixed-type variables

Maurizio Carpita & Marica Manisera

Department of Economics and Management
University of Brescia
C.da S. Chiara, 50, 25122 Brescia, Italy
email: maurizio.carpita@unibs.it, marica.manisera@unibs.it

Working Paper

(Preliminary version: please do not quote without authors' permission)

This research was funded by a grant from the European Union Seventh Framework Programme `Cooperation - Socio-Economic Sciences and Humanities' (FP7-SSH/2007-2013); `SYstemic Risk TOmography: Signals, Measurements, Transmission Channels, and Policy Interventions' - SYRTO - Project ID: 320270.

Abstract: Missing values arise in several fields and several procedures addressing this issue can be implemented in real data analyses. Recently, an interesting imputation procedure, called *Approximate Bayesian bootstrap with Propensity score and Nearest neighbour* (ABPN), has been proposed to impute missing values in data coming from surveys with Likert-type scales. The aim of this paper is to extend that procedure in order to cope not only with ordinal data, but also with data having having categorical (nominal and/or ordinal), quantitative or mixed nature. Another important generalization is concerned with multiple imputation.

.

Keywords: missing data; nonresponse; imputation; Approximate Bayesian Bootstrap

1. Introduction

Missing values arise in a variety of fields and several procedures addressing this issue have been proposed and discussed in the literature. The quality of the data analysis can be affected by nonresponse: for example, in surveys investigating individuals' opinions or perceptions, for many reasons, individuals may decide not to respond to some questions or they may unintentionally skip some questions. More generally, any data matrix with the structure units × variables can be affected by nonresponse: for example, missing values can be present in an economic data set, referring to several macroeconomic indicators (variables) measured for several countries (units). In this paper the focus is on *variable nonresponse*, occurring when the values of a unit only for some variables are missing (we do not consider *unit nonresponse*).

When faced with missing data, the researcher can ignore the missing data and/or omit subjects with missing data from the study; use procedures based on weighting; use model-based procedures; impute the missing data. Ignoring the missing data and considering only the nonmissing data as well as omitting units with missing data from the study (listwise and pairwise deletions) result in loss of information and can lead to serious biases in case of systematic differences between respondents and nonrespondents (depending on the amount of the missing data; Schafer 1997). Examples of weighting and model-based procedures are the well-known EM and data augmentation algorithms.

A last strategy when dealing with missing data is to use imputation procedures; in this case, the missing values are filled in with one (single imputation) or many (multiple imputation) "plausible" values to create a completed data set that can be analysed with standard techniques, requiring complete data sets (Rubin 1987).

Among the existing imputation procedures, we focus on that proposed by Rubin and Schenker (1986), called *Approximated Bayesian Bootstrap* (ABB) imputation: draw from the observed data a bootstrap sample and from this draw with replacement the imputed values for the missing data. The procedure is repeated two or more times for multiple imputation. Lavori, Dawson, and Shera (1995) proposed an algorithm (implemented in the Windows program SOLAS) that uses the ABB imputation within classes characterized by different levels of missingness defined by the propensity score (Rosembaum and Rubin 1983). We call this variant the *ABB with Propensity score* (ABP) imputation. Starting from this, Carpita and Manisera (2011) proposed an imputation technique refining the ABP with the addition of a "nearest neighbour step" (Chen and Shao 2000) to the "propensity score step" of the ABP imputation procedure. The idea is to draw the donor case in the neighbourhood of the nonrespondent, where the neighbourhood is defined as a subclass of the class resulting from the first grouping based on propensity score and includes cases with response patterns similar to the nonrespondent.

The proposal of Carpita and Manisera (2011) was specifically focused on imputation of missing data in multiple-item scales, where the missing data problem has specific characteristics (Downey and King 1998; Huisman 1999; Sijtsma and van der Ark 2003): firstly, since the items together measure one latent trait, a relationship exists among them. The imputation technique should be able to take such relationship into account and to reproduce it in the completed (or imputed) data set. Secondly, the amount of missing data should be considered with regard to the entire data set: even small percentages of missings per item can be problematic when looking at the entire data set. For example, with listwise deletion even subjects with only one missing response are excluded from the analysis, regardless of whether other items have been responded or not. Moreover, some authors suggest to rate certain response categories (such as any “don’t know” categories or the mid-point neutral responses) as missings, although for other authors the “don’t know” response is a valid response to all extent and should be incorporated into the analysis appropriately (see, for example, Manisera and Zuccolotto, 2014).

The aim of this paper is to generalize the proposal in Carpita and Manisera (2011), i.e. the ABPN imputation technique, in order to be used not only as a procedure for missing data treatment in Likert-type scales but also more generally used for multivariate imputation in data sets with variables having categorical (nominal and/or ordinal), quantitative or mixed nature.

The paper is organized as follows: Section 2 briefly recalls the proposed method of ABPN imputation and Section 3 shows the proposed improvement over ABPN in order to generalize it as a procedure for missing data treatment in multivariate data set with mixed-type variables. Section 4 concludes the paper.

2. The ABPN imputation procedure

Following Lavori et al. (1995) and focusing on the problem of missing data in Likert-type scales, Carpita and Manisera (2011) showed that the ABP imputation procedure can be improved by drawing the donor case in the neighbourhood of the nonrespondent. The neighbourhood is defined as a subclass of the class resulting from the first grouping based on propensity score and includes cases with response patterns similar to the nonrespondent. A “nearest neighbour step” was added (Chen and Shao 2000) to the “propensity score step” of the ABP imputation procedure. This justifies the acronym ABPN, which stands for *ABP and Nearest neighbour* imputation.

In practice, the ABPN imputation procedure combines the *Approximate Bayesian Bootstrap* (ABB) with two classification procedures based on (*i*) the *propensity score* (P) and (*ii*) the *Nearest neighbour* (N).

ABB is a Bayesian-theory based imputation technique that first draws a bootstrap sample from the observed data and then draws with replacement the imputation data from the bootstrap sample. Like the *random hot-deck* procedures (Ford 1983), ABB imputes missing data by sampling a donor case from the observed data. For the links between ABB and Bayesian Bootstrap imputation (Rubin 1981; Rubin and Schenker 1986) see Carpita and Manisera (2011).

Lavori et al. (1995) proposed an imputation procedure, here called A*pproximate* B*ayesian bootstrap with* P*ropensity score* (ABP) imputation, combining ABB with the within-class random imputation using a *propensity score* classifier (Rosembaum and Rubin 1983). The idea is: firstly, to form imputation classes using the propensity scores obtained from a logistic regression with a vector of observed predictors and missing/nonmissing as the dichotomous dependent variable; secondly, to draw a bootstrap sample from each class, and in the end to draw the random imputations within each class.

In the ABPN procedure (Carpita and Manisera, 2011), a special selection of the donor case is added to ABP: this selection is based on the neighbourhood of the *nonrespondent*, defined as a case with at least one missing value over the items. In detail, considering the responses of *n* subjects to *k* items, the ABPN technique is composed of 4 steps.

1. *Logistic regression step* (L step). A logistic regression is performed, with dichotomous dependent variable given by the status *respondent* or *nonrespondent* and predictors given by the variables supposed to be related to the missingness mechanism.

2. Propensity score step (P step). The estimated logistic model is used to calculate the propensity score, that is the predicted probability of being nonrespondent given the vector of observed predictors. Then, all the observations (respondents and nonrespondent cases) are sorted by the propensity score and classified into classes. Therefore, each class includes respondents and nonrespondents with comparable level of propensity to missingness.

3. Nearest neighbour step (N step). Let r and t be the number of complete and incomplete cases, respectively, within each class. For each nonrespondent, a fraction f of cases is selected from among the r complete cases. The subclass of $\{f \cdot r\}$ cases constitutes the donor pool for that nonrespondent, where $\{z\}$ is the nearest integer of z. The nearest neighbours of the nonrespondent to be imputed are selected, according to a similarity criterion evaluated on the response patterns.

4. Approximate Bayesian Bootstrap step (ABB step). From the nearest neighbours (the $\{f \cdot r\}$ complete and most similar cases), a bootstrap sample of $\{f \cdot r\}$ cases is drawn, the donor case is randomly drawn from this sample and finally imputation is performed for the missing answers only.

Therefore, ABPN introduces the N step to ABP, while steps L, P and ABB are the same in ABP and ABPN. Of course, the donor case results from the resampled donor pool (ABB step) that is composed of respondents having similar propensity to missingness (P step) and similar response pattern (N step) with respect to the nonrespondent.

In ABPN, classification plays an important role in the final identification of the donor pool, and the use of the N step adds an important feature to ABP, allowing the selection of the donor from the neighbours, which are subjects with similar response patterns.

The L step requires to collect covariates for the model that may predict missingness: findings indicate that covariates usually related to missingness are gender, age, education, income, occupation (Huisman 1998).

Carpita and Manisera (2011) proposed to apply the logit model to a *balanced sample*, that is a sample in which the number of respondents equals the number of nonrespondents; to use 4 classes of equal size in the P step; to set *f* at 0.25; to replace only the missing values of the nonrespondent with the scores of the donor case corresponding to the missing cells, therefore maintaining the nonmissing values of the nonrespondent into the data set.

Results from a wide simulation study showed that ABPN is a promising imputation procedure for Likert-scaled data: in particular, it always performed better than ABP when the focus in on the effect of the imputation technique on item analysis (scale level) and the ability of the technique to keep the original correlation structure in the imputed data set; these promising results justify further research on the ABPN procedure.

3. A generalization of the ABPN imputation procedure

In this section, we discuss the generalization of the ABPN imputation procedure, in order to make it suitable for missing data treatment in multivariate data set with categorical and numerical variables.

The proposed improvement over ABPN is essentially concerned with the N step, in which the nearest neighbours of the nonrespondent to be imputed are selected, according to a similarity criterion evaluated on the response patterns. In Carpita and Manisera (2011), the focus was on Likert-scaled data, therefore, in the N step similarity was measured by their ranks with the Gower's index (Kaufman and Rousseeuw 1990), appropriate when dealing with ordinal variables. When the focus is on multivariate data sets with variables having mixed nature (nominal categorical, ordinal categorical, numerical), the similarity must be evaluated resorting to appropriate indices.

We can consider different situations. When the variables in the data set are all numerical (i.e., interval-scaled variables resulting from continuous measurements; for example, height, temperature, cost, price, …), similarity in the N step of the ABPN procedure can be computed by a distance function. The most popular choice is the Euclidean distance, but other options can be selected: for example, the city-block or Manhattan distance, the Minkowski distance, which generalizes both the Euclidean and the Manhattan metric, and many others (see, for example, Hartigan, 1975).

When the variables in the data set are all binary -assuming only two values 0 and 1- and symmetric, according to the definition in Gower (1971, p. 858), the most common similarity index is the “simple matching coefficient”. It is given by the percentage of matches between subjects (a match happens when two subjects’ states are both 0 or 1). Some other similarity coefficients for binary data are listed in Kaufman and Rousseeuw (1990, p. 24). When the variables are all binary and asymmetric (the outcomes 0 and 1 are not equally important), similarity should be measured by means of coefficients considering the agreement of two 1s more important than the agreement of two 0s. The most well-known coefficient of this kind is the Jaccard coefficient (Jaccard, 1908). Other indices are reported in Kaufman and Rousseeuw (1990, p. 26) and the references therein.

When the variables in the data set are all categorical nominal variables, the simple matching approach can still be used, in its simplest form (Sokal and Minchener, 1958) or different variants (Kaufman and Rousseeuw, 1990).

The most interesting case, however, is perhaps the situation with mixed variables: a data set composed of different types of variables (symmetric and asymmetric binary, nominal, ordinal, interval and ratio variables). In this situation, we could separate the variables, according to their type and performing the N step and the ABB step separately. Otherwise, the Gower's index is still a good choice for measuring the dissimilarity between two individuals and then identify each individual's neighbourhood.

Finally, the other important generalization that can be easily implemented is that the ABPN imputation procedure can be also applied in the context of multiple imputation (Rubin and Schenker 1986), repeating the whole procedure two or more times. This is feasible for ABPN both in its original version and in the generalized version proposed in this section.

4. Concluding remarks

The ABPN procedure proposed in Carpita and Manisera (2011) for Likert-scaled data is generalized in this paper for multivariate imputation in data sets with mixed-type variables. It is worth noting that the ABPN imputation procedure can be used when variables related to each others: without relationships among the variables, the method itself does not make sense and results are not reliable. The performance of the proposed version of the ABPN procedure could be investigated by a simulation study, comparing ABPN with several other imputation techniques for missing data treatment in Likert-type scales (person mean, corrected item mean, item correlation and two-way imputation; Huisman 2000; van Ginkel, van der Ark, and Sijtsma 2007; Vermunt, van Ginkel, van der Ark and Sijtsma 2008).

Acknowledgements

This research was funded by a grant from the European Union Seventh

Framework Programme (FP7-SSH/2007-2013); `SYstemic Risk TOmography: Signals, Measurements, Transmission Channels, and Policy Interventions' - SYRTO – Project ID: 320270.

References

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

Chen, J., Shao, J. (2000) “Nearest neighbour imputation for survey data,” *Journal of Official Statistics*, 16, 2, 113-131.

Downey, R.G., King, C.V. (1998) “Missing data in Likert ratings: a comparison of replacement methods,” *The Journal of General Psychology*, 125, 175-191.

Ford, B.M. (1983) “An overview of hot-deck procedures,” in *Incomplete data in sample survey*, 2, New York: Academic Press, 185-207.

Hartigan, J. (1975), *Clustering algorithms*, New York: Wiley.

Huisman, M. (1998) “Missing data in behavioral sciences research: investigation of a collection of data sets,” *Kwantitatieve Methoden*, 57, 69-93.

Huisman, M. (1999), *Item nonresponse: occurrence, causes, and imputation of missing answers to test items*, Leiden, The Netherlands: DSWO Press.

Huisman, M. (2000) “Imputation of missing item responses: Some simple techniques,” *Quality & Quantity*, 34, 331-351.

Kaufman, L., Rousseeuw, P.J. (1990), *Finding groups in data*, New York: Wiley.

Lavori, P., Dawson, R., and Shera, D. (1995) “A multiple imputation strategy for clinical trials with truncation of patient data,” *Statistics in Medicine*, 14, 1913-1925.

Manisera, M., Zuccolotto, P. (2014) “Modelling "don't know" responses in rating scales,” *Pattern Recognition Letters*, 45, 226-234.

Rosembaum, P., and Rubin, D.B. (1983) “The central role of the propensity score in observational studies for casual effects,” *Biometrika*, 70, 41-50.

Rubin, D.B. (1981) “The Bayesian Bootstrap,” *The Annals of Statistics*, 9, 130-134.

Rubin, D.B. (1987), *Multiple imputation for nonresponse in surveys*, New York: Wiley.

Rubin, D.B., Schenker, N. (1986) “Multiple imputation for interval from simple random samples with ignorable nonresponse,” *Journal of the American Statistical Association*, *Survey Research Methods*, 81, 366-374.

Schafer, J.L. (1997), *Analysis of incomplete multivariate data*, London: Chapman & Hall.

Sijtsma, K., and van der Ark, L.A. (2003) “Investigation and treatment of missing item scores in test and questionnaire data,” *Multivariate Behavioural Research,* 38, 4, 505-528.

van Ginkel, J.R., van der Ark, L.A., and Sijtsma, K. (2007) “Multiple imputation of item scores in test and questionnaire data, and influence on psychometric results,” *Multivariate Behavioral Research,* 42, 387-414.

Vermunt, J.K., van Ginkel, J.R., van der Ark, L.A., and Sijtsma, K. (2008) “Multiple imputation of incomplete categorical data using latent class analysis,” *Sociological Methodology*, 38, 369-397.