Assessing assumptions for data fusion procedures

Alfonso Piscitelli and Antonio D'Ambrosio

Abstract Data fusion consists of merging information coming from two different surveys. The first one is called reference or donor survey while the second is called punctual or receptor survey. In order to perform data fusion, such two independent surveys must have a block of common variables that is used as a bridge between them. A natural question that arise from data fusion definition is: it is always possible to take a merging of two different surveys coming from two independent samples? This work is about the possibility to evaluate the interrelation structure between the common variables aimed to define some criteria which allows to define them as "similar".

Key words: Data Fusion, Missing data imputation, Box's M-test

1 Introduction

Data fusion involves the imputation of a complete block of missing variables in independent data sets. It consists of matching two already held surveys in order to make it possible to transfer part of the information contained in one survey to a second one. The first survey is called reference survey (donor matrix); the second is called punctual survey (receptor matrix). Data fusion allows us to treat data coming from the two distinct surveys as a whole. With the aim of determining the complete block of unobserved values of a set of variables included in a first survey but not in a second, data fusion can be approached by means of missing data imputation techniques. Missing data of the receptor matrix will be imputed by exploiting

Alfonso Piscitelli
Department of Sociology, University of Naples Federico II e-mail: alfonso.piscitelli@unina.it

Antonio D'Ambrosio
Department of Mathematics and Statistics, University of Naples Federico II e-mail: antdambr@unina.it

information coming from the donor matrix. To perform such an imputation a set of variables in common to both surveys is required. Different methodologies have been proposed in literature for data fusion, and they can be classified in two families (Schulte Nordholt, 1998; Saporta, 2002). A first group, *explicit model-based estimation methods*, relies on finding a *model* for the variables to be imputed in the donor survey and on applying it for the receptor survey (see i.e. Rubin, 1987; Barcena and Tusell, 1999; D'Ambrosio, Aria and Siciliano, 2012). The second group includes the so-called *implicit models for imputation*. In such a case, for each statistical unit of the receptor survey, one or more donor units are selected. The values of the donor units are then imputed to the receivers (see i.e. Baker, Harris and O'Brien, 1989; Aluja-Banet, Daunis-i-Estadella and Pellicer, 2007; Piscitelli, 2008).
Several authors have studied the preliminary assumptions in performing data fusion (see in example D'Orazio, Di Zio and Scanu, 2006; Rassler 2002, 2004).
One of these conditions, in using implicit models in particular in the framework of file grafting, concerns the study of the stability of the relationships among the common variables of both donor and receptor surveys. These conditions have been investigated by Bonnefous et al., (1986) and Aluja-Banet and Thio, (2001) through factorial methods. According to the point of view of the authors, the stability assumption among common variables to the two surveys allows us to define a common space on which to represent the whole information of both data sets. We think that such hypothesis of interrelation structure should be verified. This is mandatory for a *consistent* result of the fusion in terms of missing data imputation. We could work either on the correlation or the covariance matrices of the two independent surveys. Our choice is about the covariance matrices because we are also interested in the scale of the common variables. In other words, the analysis of the interrelation structure between the two independent surveys means evaluating the statistical equality (or "'similarity"' in terms of data fusion) of the covariance (or correlation) structure of the common variables to both the independent surveys. Dealing only with numerical variables, a feasible tool is the Box's M-test (Box, 1950; Box and Draper, 1969).
We state that, in the framework of implicit models, verifying the equality of covariance matrices is a necessary but not sufficient condition because different values in mean between the common variables of the two surveys do not allows to use in the better way the best donor(s). For that reason, MANOVA test must complete the check of preliminary assumptions of data fusion. As Box's M-test is usually used to check homoschedasticity in MANOVA analysis, we choose to use Box's M-test to directly check the equality of covariance matrices as preliminary step.

2 The Box's M-test

We assume two independent surveys named survey A and survey B. Let K be a matrix of dimension $n_1 \times q$ representing the A survey and let Z be a matrix of dimension $n_2 \times j$ representig the B survey. Let p be the number of variables common to both

matrices K and Z. Let X_1 be the $n_1 \times p$ submatrix of K matrix called donor matrix. Let X_2 be the $n_2 \times p$ submatrix of Z matrix called receptor matrix.
We assume that $X_1 \sim N_p(\mu, \Sigma_1)$ and $X_2 \sim N_p(\mu, \Sigma_2)$, with $\Sigma_1 = \Sigma_2 = \Sigma$.
A way to verify the equality of the population covariance matrices is the Box's M-test.
The null hypothesis is

$$H_0 : \Sigma_1 = \Sigma_2 = \Sigma$$

where Σ is the presumed common covariance matrix. A likelyhood ratio statistic for testing the null hypothesis is given by

$$\Lambda = \prod_{i=1}^{2} \left(\frac{|S_i|}{|S_{pool}|} \right)^{\frac{n_i-1}{2}},$$

where S_i is the i^{th} sample covariance matrix, S_{pool} is the *pooled* sample covariance matrix given by

$$S_{pool} = \frac{1}{\sum_{i=1}^{2}(n_i - 1)} \{(n_1 - 1)S_1 + (n_2 - 1)S_2\}.$$

The Box M statistic is based on the χ^2 approximation to the sampling distribution of $-2\ln\Lambda$, which gives

$$M = \left[\sum_{i=1}^{2} (n_i - 1) \right] \ln |S_p| - \sum_{i=1}^{2} [(n_i - 1) \ln |S_i|] \quad (1)$$

Under the null hypothesis, M statistic is distributed as a χ^2 with v degrees of freedom, with $v = 0.5k(k+1)(g-1)$, k is the number of variables and g is the number of groups.

Several experiments on simulated datasets show how the consistence of missing data imputation is higher when the interrelation structure between the donor and receptor matrices is verified.

References

1. Aluja-Banet T., Thio S. (2001), Survey Data Fusion, *Bulletin of Sociological Methodology*, 72, 20-36.
2. Aluja-Banet T., Daunis-i-Estadella J., Pellicer D. (2007), GRAFT, a Complete System for Data Fusion, *Computational statistics and data analysis*, 52, 635-649.
3. Baker K., Harris P., O'Brien J. (1989), Data Fusion: an Appraisal and Experimental Evaluation, *Journal of the Market Research Society*, 31, 153-212.
4. Barcena M.J., Tusell F. (1999), Enlace de encuestas: una propuesta metodologica y aplicacion a la Encuesta de Presupuestos de Tempo, *Q uestiio*, 23, 297-320.

5. Bonnefous, S., Brenot, J., Pages, J.P. (1986). Methode de la greffe et communication entre enquêtes, in Diday E. *et al. (eds), Data Analysis and Informatics IV*, North Holland, 603-617.
6. Box, G.E.P. (1950). Problems with analysis of growth and wear curves, *Biometrics*, 6, 362-389.
7. Box, G.E.P., and Draper, N.R. (1969). *Evolutionary Operation: a statistical method for process improvement*. Wilet, New York.
8. D'Ambrosio, A., Aria, M., Siciliano, R. (2012). Accurate Tree-based Missing Data Imputation and Data Fusion within the Statistical Learning Paradigm, *Journal of Classification*, forthcoming
9. D'Orazio, M., Di Zio, M., Scanu, M. (2006). *Statistical Matching: Theory and Practice*. Whiley, New York.
10. Piscitelli, A. (2008). A double imputation method for Data Fusion, *Quaderni di Statistica*, 10, 35-52.
11. Rassler, S. (2002). *Statistical matching: a frequentist theory, practical applications and alternative Bayesian approaches*, Springer-Verlag, New York.
12. Rassler, S. (2004). Data Fusion: identification problems, validity, and multiple imputation. *Austrian Journal of Statistics*, volume 33, number 1 & 2, 153-171.
13. Rubin, D.B., (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York
14. Saporta, G. (2002), Data Fusion and Data Grafting, *Computational Statistics and Data Analysis*, 38, 465-473.
15. Schutle Nordholt, E. (1998), Imputation: Methods, Simulation Experiments and Practical Examples, *International Statistical Review*, 66, 157-180.