Data Fusion: a new method based on Homogeneity Analysis

Gilbert Saporta, Vila Co

DATA FUSION: A NEW METHOD BASED ON HOMOGENEITY ANALYSIS

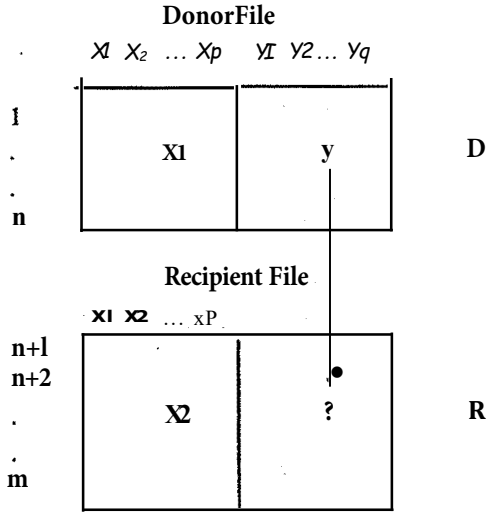
Gilbert Saporta, Vila Co

SUMMARY
Data fusion consists in estimating blocks of rmssmg variables for a set of observations. After a short overview of the problem, we present an application of the criterium of homogeneity analysis for the case of categorica] data.
Ke-ywords: data fusion, correspondence analysis, homogeneity analysis, missing data.

I. DATA FUSION: AN OVERVIEW

Data fusion is concemed by the following problem: we have two sets of units; on the first set p+q variables are measured, whilst on the second set only the first p variables are known and some kind of inference is needed for the missing variables.
Data fusion is frequently encountered in marketing, see Lejeune[1], especially in media-market when for instance:
- Questions are not answered in a survey, or questions are not proposed since the questionnaire is too long.
- Data information comes from different sources.

The aim of data fusion consists in an optimal use of existing infonnation to recover the missing infonnation. It is a large scale application of some particular missing values estimation techniques, see Little & Rubin[2]. Its principle is to estimate the values of the missing variables by means of answered variables, according to the relationships between the two blocks of variables.
The recipient file, is the file with missing variables, the missing values are estimated by using the infonnation in the other data set, which is defined as the *donar file.*



Classical methods,consists in imputing missing data: that is to say replacing for each unit of the recipient file the missing vector y by the values of an unit of the donor file. Missing variables are not estimated one by one but globally: advantages are simplicity and absence of incoherent separate estimations since the imputed values are real values of an individual. Drawbacks are a loss of variability.

Eligible donors are choosed among the nearest neighbours of the recipient according to some of the common variables (the active variables).

For categorical variables, a classical method, referred here as FRF (fusion with factorial referential), used by several french institutes is described by Santini[3]:

A multiple correspondence analysis is performed on the active variables of the union of all the observations of the two files. The first k axis are retained which allow to put the observations in a R^k space and calculate distances between them.Then for each recipient, we choose the donors in the neighborhood of the recipient. Final donors are choosen according to two conditions:
1) Among the potential donors, we choose those which most resemble to the recipient on descriptive variables like age, sex and social status etc.
2) We avoid to use too frequently the same donors by means of a penalty function since a drawback of classical methods is that they cannot create different responses from those of the donor file, for they copy a entire block of the donor. Thus they may create artificial correlations.

Results of data fusion are measured according to two levels:

Global level - the distributions of the answered variables and those of the transferred ones should be close; as well as similar correlations.

Individual level - it measures the fit between estimated and real data incross validation studies.

We propose in this paper a method which is relatively robust, adapted to more general problems without requiring the data to follow a statistical model, and can accept differences in level of structure of population and various proportions of size between the two files. The method we propose is completely different of the classical ones: based on an optimal criterion, maximizing the internal consistency (homogeneity of the data), it does not copy whole blocks of responses, and may give values different from the existing ones in the donor file.

II. HOMOGENEITY ANALYSIS USED FOR DATA FUSION

Homogeneity analysis (Gifi[4]) is a presentation of correspondence analysis in terms of maximizing a measure of internal consistency, which is convenient for dealing with missing data.

The basic idea is similar to factor analysis: if the variables measure more or less the same property, it is possible to replace the different variables by a unique synthetic variable without loosing too much information. In order to evaluate the success of the substitution, a criterion of homogeneity and a loss function are defined. Maximizing

the homogeneity of variables (internal consistency) leads to homogeneity analysis which is similar to multiple correspondence analysis.

II.1 Optimal quantification, correspondence analysis and homogeneity analysis for complete categorical data.

Let G_j be the indicator matrix of variable j: $G_j = (g_{il})_{n \times j_k}$ and D_j the diagonal matrix of frequencies.

Let y_j be a quantification of categories of variable j (i.e scores given to categories) and $G_j y_j$ the quantification of variable j.

$X = \frac{1}{m} \sum_{j=1}^{m} G_j y_j$. is the average quantification of units (the unique " factor "). In the case of perfect consistency of the data, we have :

$$X = G_1 y_1 = G_2 y_2 = \cdots = G_m y_m$$

Thus the homogeneity loss function :

$$\sigma(X,Y) = \frac{1}{m} \sum_{j=1}^{m} (X - G_j y_j)'(X - G_j y_j)$$

is equal to zero in the case of perfect homogeneity. Its minimum value is equal to $1 - \frac{\lambda_1^2}{m}$ where λ_1^2 is the largest eigenvalue of the correlation matrix of the transformed data.

Minimizing $\sigma(X,Y)$ on X and Y (Meulman[5]) gives :

$$y_j = D_j^{-1} G'_j X' \text{ and } X = \frac{1}{m} \sum_{i=1}^{m} G_i y_i$$

which are the transition equations of multiple correspondence analysis.

II.2 Quantification of incomplete categorical data and homogeneous imputation by maximizing the internal consistency

Following Meulman, homogeneity analysis with missing data consists in minimizing the modified loss function:

$$\sigma(X,Y) = \frac{1}{m} \sum_{j=1}^{m} (X - G_j y_j)' M_j (X - G_j y_j)$$

$M_j = (m_{ij})_{nxn}$ diagonal matrice , $\begin{cases} m_{ij} = 0 \\ m_{ii} = \text{sum of elements of the row i of } G_j \end{cases}$

which gives :

$$\begin{cases} Y = D^{-1} G' X \\ X = \frac{1}{m} GY \\ \quad \text{with } U' M_0 X = 0 \text{ and } X' M_0 X = I \end{cases}$$

$$M_0 = \frac{1}{m} \sum_{j=1}^{m} M_j$$

Buuren S.V. & Van Rijckevorsel[6] have proposed a method for estimating missing data by maximizing internal consistency.
Let the loss function be:

$$\sigma(x; y_1, \cdots, y_m, G_1^*, \cdots, G_m^*) = \sum_{j \in \Omega} (X - G_j y_j)^2 + \sum_{j \notin \Omega} (X - G_j^* y_j)^2$$

Ω represents the set of all observed variables - common predictive variables.
G_j* is the incomplete indicator matrix of variable j - transferred variables;
if variable j is missing on unit i, then the ith row of G_j* is full of zeros.

The maximal homogeneity among the imputed values (most consistent imputation) can be found by minimizing σ over X, y_1 , y_2 , ..., y_m and $G_1, G_2, \cdots, G_m$

III. VALIDATIONS AND RESULTS

Validation is performed in comparing real data with their estimation. There are two levels : individual and marginal.:
For individual estimation the fit is given by the difference between the observed error rate, and its expectation in a random allocation (Lejeune&Lebart[7])
The marginal level consists in comparing distributions of imputed variables to the real distributions.

Simulations performed on real and simulated data sets (Co[8]) proved that imputation with homogeneity analysis gives better results than other methods for the individual level. For global level however, a method like FRF better preserves the structure of dependencies between the imputed variables: the explanation being that FRF, which copies entire blocks of data respects the correlation and avoids incoherent answers.
For instance, in the classical sample data set " enquête 1000 " from Spad software, 3 variables have been predicted for 192 observations with 4 common variables. Our method has been compared with FRF : the good classification rate is about 54% for several simulations, compared to 47% with FRF. Comparisons of the actual marginal distribution and the estimated ones shows differences between Homogeneity Analysis and FRF:

Q5	Actual Margin	Homog. An.	FRF
1	136	136	125
2	56	56	67

Q6	Actual Margin	Homog. An	FRF
1	36	6	49
2	70	114	65
3	35	16	27
4	29	23	33
5	4	33	1
6	18	33	15
7	0	0	2

0 7	Actual Mamn	Homo2.An.	ERF
I	100	118	100
2	36	18	43
3	37	29	31
4	19	27	18

V. CONCLUSIONS

Data fusion based on homogeneity analysis could be recommended for estimating individuai data. though it is demanding in computation. It is frequently more relevant than imputation by full blocks of the questionnaire. It has the advantage of maximizing a criterion (internal consistency). Further developments are nevertheless needed in order to recover more variability for maximizing internal consistency gives a unique imputation for the pattern of the X variables which does not reflect the real variability.

REFERENCES

I. M.Lejeune, 'De l'usage des fusions de données dans les études de marché', *50th Session o fthe International Statistica/ lnstitute, Tome LVI,* 923-935, Beijing, 1995.

2. R.Little & D.Rubin, *Statistica! analysis with missing data,* Wiley, New York, 1987.

3. G.Santini, 'An experiment to validate fusionned files obtained by the referential factorial method', *ESOMAR,* Helsinki, Finland, 1986.

4. A.Gifi, *Nonlinear multivariate analysis,* Wiley, Chichester, 1990.

5. J.Meulman, *Homogeneity analysis o fincomplete data,* Dswo Press, Leiden, 1982.

6. S.V.Buuren & J.L.A.Van Rijckevorsel, 'Imputation of missing categorica! data by maximizing internal consistency', *Psychometrika,* 57, 567-580, 1992.

7. L.Lebart & M.Lejeune, 'Assessment of data fusions and injections', *Encuentro lnternational AIMC sobre lnvestigacion de Medios,* Madrid, 1996.

8. V.Co, *Méthodes statistiques et informatiques pour le traitement des données manquantes,* Thèse de doctorat, Cnam, Paris, 1997.