Data Fusion and Data Pruning

Gilbert Saporta
CNAM, 292 rue Saint Martin, F75141 Paris cedex03, saporta@cnam.fr

Abstract : Data fusion is concerned with the problem of merging data bases coming from different sources into a single data base when variables are absent or missing in some files. After a survey of the main techniques we present some new approaches, in particular one based on homogeneity analysis and future directions. We insist on validation problems and caveats.

Keywords:data fusion, missing values, homogeneity analysis, imputation.

1. Introduction

Data fusion and data pruning are concerned with combining files and informations coming from different sources. In this respect, these techniques participate to data mining and knowledge discovery approaches : the problem here is not to extract information from a single data base, but to merge different data bases collected in sample surveys, administrative sources, socio-economic data at an aggregated level etc., each separate base may be composed of different statistical units or different levels.

1.1 Data Fusion

In data fusion the goal is to obtain a single data-base where all the variables have been completed for the union of units.The resulting base may be analysed afterwards with data mining tools.Basically the problem may be formalised in terms of two data files : the first file contains observations for a whole set of $p+q$ variables measured on n_0 units, the second file contains observations of only a subset of p variables for n_1 units. In some cases, n_0 is small compared to n_1 . If X stands for the common variables, we have the following scheme :

X_0	**Y_0**
X_1	**?**

The problem, here is to fill the blank part of the table : it is a special kind of missing data estimation or imputation, where a lot of variables is missing because they have not been collected.

Data fusion originates from market studies (Baker and al. 1989), especially in media and consumption surveys, where it is often impossible to ask to the same sample all the items when there are too many questions. In order to reduce the burden of respondents, and thus avoid bias, one proceeds with two different independent samples, where the questions of interest are splitted in two parts, with a common set of descriptors (socio-demographic variables).

There is an increasing interest in data fusion due to the availability of multiple sources, in various fields : let us quote studies on customers behaviour where one has from one hand a complete file of transactions, and on the other hand a sample survey about satisfaction. Using multiple and incomplete sources for automatic detection has many military applications, see the Fusion2000 congress (http://www.onera.fr/fusion2000) and the new « Information Fusion »journal.

1.2 Data Pruning

Data pruning is a close methodology where one does not try to estimate missing data, but to paste the results of a survey S1 upon the reference space of a survey S0, in a similar way as procrustes analysis.

Data pruning has been developped in the context of multivariate descriptive analysis (PCA, Correspondence analysis) and its goal is to add points coming from S1, in the graphical displays coming from S0 (Bonnefous and al,1986).

The scheme is here :

$\mathbf{X}_0$	$\mathbf{Y}_0$	
$\mathbf{X}_1$		$\mathbf{Y}_1$

Where $\mathbf{X}_0$ and $\mathbf{X}_1$ are data tables with the same common variables, and $\mathbf{Y}_0$ and $\mathbf{Y}_1$ are tables with specific variables.

Technically, data pruning which could be better called « output pruning » is nothing else than a specific way of dealing with supplementary information. If all the variables are numerical, it consists in the following steps if we want to display results of S1 , in the plots of a PCA of S0.

- Perform a PCA of **($\mathbf{X}_0$ $\mathbf{Y}_0$),** retain k components and regress the principal components $\mathbf{C}_0$ onto the common variables $\mathbf{X}_0$. This lead to approximation formulas $\hat{C}_0 = X_0 b_O$ for reconstructing principal components with a subset of variables.
- Position the units of S1 in the principal plane of S0, $C_1 = X_1 b_O$
- Position the variables of Y_1 by computing correlations between Y_1 and C_1

We have thus a double use of supplementary points (supplementary variables are positionned ttrough supplementary points) combined with an approximation of principal components. For good performances, it is necessary that X_0 and Y_0 are highly correlated , otherwise we could not predict the principal components of S0, and that X_1 and Y_1 be also correlated.

2. Models and methods for data fusion

Since the file $(\mathbf{X_0}, \mathbf{Y_0})$ is used to predict the unknown $\mathbf{Y}$ part of the second file, the first file will be called donor-file and the second one the recipient-file. Data fusion being a (very) special case of missing value estimation, several classical techniques (Little, Rubin 1987) may be applied.

2.1 Explicit model based estimation

Each missing value could be estimated thanks to classical techniques, such as regression or the general linear model for numerical Y, or logistic regression if Y is categorical : each variable of $\mathbf{Y_0}$ is modelled with $\mathbf{X_0}$ as predictors, and the model is applied to the recipient file. These techniques, though simple, suffers from at least two drawbacks : estimations are made variable by variable, not taking into account their correlations, and may lead to inconsistent results :there is no guarantee that incoherent results should be avoided (like age « under 20 » and occupation « retired »).

Maximum likelihood estimation may also be applied : in order to get likely estimations one models the joint distribution of all variables and maximises the likelihood of the incomplete sample, EM algorithm being frequently used. But ML estimation does not prevent from incoherent estimations.

An other drawback of estimation techniques is the following : two units having the same values of the predictors will have the same estimate of their Y variable, hence a loss of variability. Multiple imputation techniques, based on a bayesian framework (Rubin 1987) allows to simulate the posterior distribution of the missing values by imputing each data with several values according to one or more estimation models. One can recover correct variances with multiple imputation. However, these techniques are very complex and time consuming for large data sets.

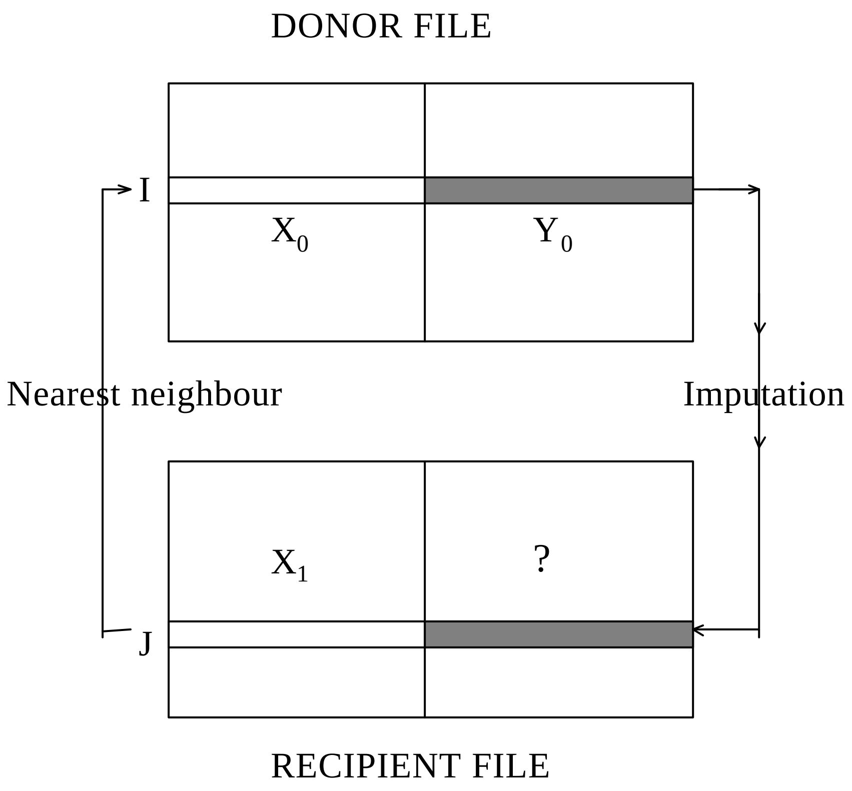
To sum up, explicit estimation techniques seem fitted more to sparse missing values than to the estimation of blocks of thousands of missing data like in data fusion.

2 .2 Imputation with implicit models : nearest neighbours, hotdeck etc.

Much more simpler techniques than the previous ones are based on the principle which consists in giving to the Y variables of a receiver the whole vector of variables of a donor : copy and paste !

Let *i* be a receiver : the basic idea is to look for a donor *j* having a close profile with the X variables : a double if all the variables are identical (which is possible with categorical predictors) or a nearest neighbour such as an appropriate distance d(*i*,*j*) in the $\mathbb{R}^p$ space of common variables is minimal. This method avoids incoherent estimations since the copied values belong to real observations. Furthermore, to avoid loss of variability, one may use a penalty function such that the same donor cannot be used too many times.

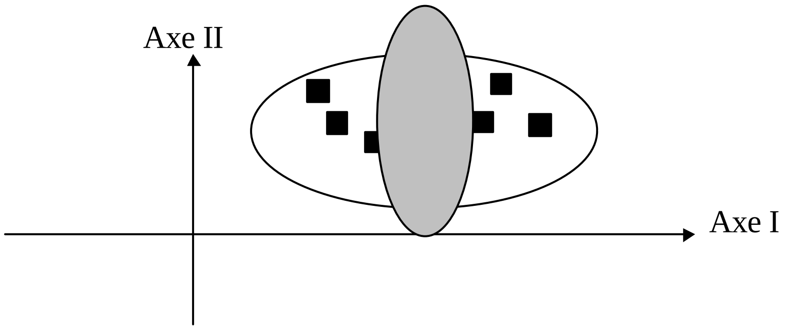
Figure 1 : Imputation scheme



Defining a distance may not be straightforward for categorical variables, which is by far the most common case in market studies. Private institutes use the following technique called « fusion with factorial reference space » or FFR :

- The first step is to perform a Multiple Correspondence Analysis with all units, donors plus receivers, and the *p* common variables, or a subset of p_1« critical variables ». A few significant components are retained, in order to filter the data and to avoid problems due to peculiar situations. Afgterwards we use a classical euclidean distance in the space of factorial coordinates.
- The second step consists in determining the neighbourhood of a receiver, ie potential donors, according to a distance, or to a preset number of neighbours.
- The last step consists in choosing among the potential donors those who fill some condition according to prespecified variables such as, age, gender etc., while avoiding to use too often the same donor.

Figure 2 : donor choice



2.3 Data fusion by maximising internal consistency

Proposed by Van Buuren & Van Rijckevorsel (1992), Co (1997) and Saporta &Co (1999) have studied a technique for categorical data based upon the « dutch » presentation of Multiple Correspondence Analysis called homogeneity analysis, see (De Leeuw 1973) and (Gifi 1990). This presentation enables MCA with missing data (Meulman 1982) but the goal, here is to estimate missing data and not to manage without them.
MCA of a disjunctive table **X=(G1|G2|…|Gm)** may be viewed as the minimisation of a loss function :

$$s\,(\mathrm{X},\mathrm{Y}) = \frac{1}{m}\sum_{j=1}^{m}(\mathrm{X} - \mathrm{G}_j^{'}\mathrm{Y}_j)^{'}(\mathrm{X} - \mathrm{G}_j^{'}\mathrm{Y}_j) \qquad (1)$$

where $\mathbf{Y_j}$ is the matrix of coordinates of the categories .
The essential idea when there are missing data, is to assign categories in order to minimise the loss function, in other words, to get maximal eigenvalues for the completed table.

Formally it consists in minimising over $\mathbf{Y}_\mathrm{j}$ and $\mathbf{G_j}$*****:

$$s\,(\mathrm{X};\mathrm{Y}_1,\ldots,\mathrm{Y}_\mathrm{m},\mathrm{G}_1^*,\ldots,\mathrm{G}_\mathrm{m}^*) = \sum_{j\in\Omega}\left\|\mathrm{X} - \mathrm{G}_\mathrm{j}\mathrm{Y}_j\right\|^2 + \sum_{j\notin\Omega}\left\|\mathrm{X} - \mathrm{G}_\mathrm{j}^*\mathrm{Y}_j\right\|^2 \qquad (2)$$

where Ω is the set of variables with no missing data and $\mathbf{G_j}$***** the indicator matrix of variables with missing data to be completed.

The following simple example drawn from (Van Buuren S. & Van Rijckevorsel J.L.A., 1992) shows how it works for a one-dimensional MCA with 3 variables each having 3 categories:

Table 1 : MCA with missing data

Unit	Income	Age	Car
1	*x*	young	am
2	medium	medium	am
3	*y*	old	jap
4	low	young	jap
5	medium	young	am
6	high	old	am
7	low	young	jap
8	high	medium	am
9	high	*z*	am
10	low	young	am

X,y,z are missing categories, so there exists 27 possibilities of imputing values.

Table 2 : Results of the 27 MCA

x	*y*	*z*	l_1	*x*	*y*	*z*	l_1	*x*	*y*	*z*	l_1
l	l	y	.70104	m	l	y	.63594	h	l	y	.61671
l	l	m	.77590	m	l	m	.72943	h	l	m	.66458
l	l	o	.76956	m	l	o	.72636	h	l	o	.65907
l	m	y	.78043	m	m	y	.70106	h	m	y	.70106
l	m	m	.84394	m	m	m	.77839	h	m	m	.74342
l	m	o	.84394	m	m	o	.84394	h	m	o	.74342
l	h	y	.78321	m	h	y	.73319	h	h	y	.68827
l	h	m	.84907	m	h	m	.80643	h	h	m	.74193
l	**h**	**o**	*.84964	m	h	o	.80949	h	h	o	.74198

The optimal solution is x = « low », y = « high », z = « old », in terms of maximal homogeneity. It is in some respect the most likely solution for an unidimensional model.

Of course an exhaustive search is impossible for real examples, where one has thousand of data, and heuristics or iterative algorithms are necessary. Co (1997) has developped a modification of the original algorithm of Van Buuren & Van Rijckevorsel, which can handles files of several hundreds units.

Data fusion by maximising internal consistency does not avoid the drawback of reducing the variability, since two units with the same non-missing variables will get the same imputations of the missing variables, but multiple imputation is possible. The choice of the number of retained axes needs also to be precised, and a unidimensional solution does not seem very realistic.

3.Validation

How can we assess the quality of data fusion techniques ? Since there is generally no model for the data, the only way is to use empirical validations were known data are hidden and their estimations compared to the true values: cross validation, bootstrap etc. see Comyn 1999.

3.1 Assessment and conditions of validity

Which are the quality indicators ? Recovery of the values at an individual level seems appealing but too severe in most cases. Users are not generally interested in individual predictions and may be satisfied with predictions which are correct in the average for groups of units. But it is not enough to recover marginal distributions or mean values, since a random sampling could do this adequately !
The main problem is to conserve the covariance structure, or for categorical data to have some correct cross-tabulations between variables of interest.
We have seen before that there are two classes of techniques : fusion with data matching using donors (copy-paste) and in a broad sense regression or estimation methods.The first class is generally not optimal in terms of individual estimation, but is efficient in keeping covariance structure and avoids incoherences. The opposite is true for the second class of methods.
In all circumstances, in order to have satisfactory results, it is necessary that there exist :

- A large enough number of common variables
- High correlations between the block of common variables and variables to be imputed.
- A common structure between the donor file and the receiver file. We mean that the distributions of common or critical variables in the donor file, should be close to their distribution in the receiver file. Otherwise the results would be biased.

3.2 An example (Saporta, Co 1999)

We used a classical data set of SPAD software : a sociological survey with 992 interviews, splitted randomly into 2 files of 800 for the donor file and 192 for the receiver file. There were 7 categorical variables :
4 common variables:

Q1 - age categories(5 levels),
Q2 - town size (5 levels),
Q3 - bedtime (7 levels),
Q4 - school leaving age (5 levels) .

3 opinion variables Y to impute on the receiver file:

Q5 - Family is the only place where one feels good ? (Yes, No),
Q6 - Highest education grade (7 levels),
Q7 - Frequency of TV watching (4 levels).

A first attempt was to compare at the individual level (number of cases correctly classified for Q5, Q6 and Q7) and then for the marginal distributions two methods :fusion with maximising internal consistency (MIC) and fusion with factorial reference space or FFR, in contrast with a random assignment .
The results are a good illustration of paragraph 3.1 : MIC, like model-based methods , or methods where variables are estimated at the unit level performs better than FFR, bur FFR is better in terms of margins and cross margins. Due to a lack of space, cross tabulations are omitted.

Table 3 Individual performances

Method	***Correct classifications***
Random	*49%*
MIC	*54%*
FFR	*47%*

Table 4 Marginal performances

Q5	***True margins***	***MIC***	***FFR***
1	*136*	*136*	*125*
2	*56*	*56*	*67*
Q6	***True margins***	***MIC***	***FFR***
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
Q7	***True margins***	***MIC***	***FFR***
1	100	118	100
2	36	18	43
3	37	29	31
4	19	27	18

As we saw earlier, fusion with MIC estimates missing data like a regression model : the estimated value is the most likely in terms of homogeneity and is unique for a defined pattern. FFR may give several different imputations for the same pattern : for instance, here with X=3421, we get for Y 6 different estiamations: 232,121,123,122,114,212 which may better represent the natural variability of responses.

4 Concluding remarks

They will be of two kinds : technical and ethical.

From the methodological point of view, data fusion is a problem of « mass » missing data, and statisticians may be interested in developping and validating new methods : in addition to the above mentioned techniques, one could suggest new directions using for instance non linear learning algorithms (neural networks). It is clear that data fusion fills a need which is frequently met by practitioners and data managers who want to provide to their final user a single full data file.One has of course to be very careful when using « data » which are actually estimates and not observations : they should never be used at an individual level. A perverse consequence of data fusion techniques may result in less effort to collect data, since we may invent them scientifically…

An other danger of data fusion is about confidentiality and privacy of data : many countries have laws about protection of personal data, which regulates the possibility of linking non-anonymous files. With data fusion, we are in the situation where informations which have not been requested are estimated and added without the knowledge of the individuals ! There is here a paradoxical situation when one thinks of the amount of researches (done mainly by National Statistical Institutes) for developping algorithms preserving confidentiality when statistical data bases are disseminated. Data pruning is free from most of these suspicions, since it does not aim to provide individual estimates.

References:

Aluja-Banet T., Morineau A., Rius R. (1997), La greffe de fichiers et ses conditions d'application. Méthode et exemple. in *Enquêtes et sondages,* G.Brossier, A.M.Dussaix (Eds), Dunod, Paris, 94-102

Baker K., Harris P., O'Brien J. (1989), Data fusion: An appraisal and experimental evaluation, *Journal of the Market Research Society,* 31, 153-212

Bonnefous S., Brenot J., Pagès J.P. (1986), Méthode de la greffe et communications entre enquêtes, in *Data Analysis and Informatics vol 4,* E.Diday (ed), North-Holland, Amsterdam, 603-617

Buuren S.V. & Van Rijckevorsel L.A. (1992), Imputation of missing categorical data by maximizing internal consistency, *Psychometrika,*57, 567-580.

Co V. (1997), *Méthodes statistiques et informatiques pour le traitement des données manquantes*. Ph.D., CNAM, Paris.

Comyn M. (1999), *Modélisation et validation des rapprochements et fusions de fichiers d'enquêtes*. Ph.D., ENST, Paris.

De Leeuw J. (1973), *Canonical analysis of categorical data.* Dswo, Leiden.

Gifi A. (1990), *Nonlinear multivariate analysis,* Wiley , New-York.

Lejeune M. (1995), De l'usage des fusions de données dans les études de marché, *Proceedings 50th Session of ISI-Beijing,* Tome LVI, 923-935

Little R.J.A., Rubin D.B. (1987), *Statistical analysis with missing data,* Wiley, New-York.

Meulman J. (1982), *Homogeneity analysis of incomplete data*, Dswo, Leiden.

Rubin D.B. (1987), *Mutiple imputation for nonresponse in survey*s, Wiley, New-York.

Saporta G, Co V.(1999), Fusion de fichiers: une nouvelle méthode basée sur l'analyse homogène, in *Enquêtes et sondages,* G.Brossier, A.M.Dussaix (Eds), Dunod, Paris, 81-93