Some approaches in imputing missing items with Bayesian Networks

Alcuni metodi per l’imputazione di dati mancanti con le Reti Bayesiane

M. Di Zio, G. Sacco, M. Scanu
ISTAT
dizio,sacco,scanu@istat.it

P. Vicard
Università Roma Tre
vicard@uniroma3.it

Riassunto: In questo lavoro vengono poste a confronto due tecniche, basate sulle reti Bayesiane, per l’imputazione di dati mancanti. Il confronto viene condotto tramite un esperimento su dati del censimento della popolazione del Regno Unito.

Keywords: Markov Blanket, graphical structure, item non-response.

1. Imputation and Bayesian Networks

Dealing with data sets affected by partial non-responses is a pervasive problem in statistical surveys. Let us consider a sample where \( k \) variables, \( X = (X_1, X_2, ..., X_k) \), are observed on \( n \) units (individuals, firms, etc.). The data set is an \( n \times k \) matrix where some of the items in the matrix are missing (\( X_{mis} \)), and some are observed (\( X_{obs} \)). A common approach to deal with partial non-response is to replace (i.e. impute) missing items with artificial plausible values, under the assumption that the missing data mechanism is Missing At Random (MAR, see Rubin (1976)). The MAR hypothesis assumes that the probability of an entry to be missing may depend on observed data but not on unobserved data. More formally, let \( R \) be an \( n \times k \) matrix such that the element \( r_{ij} \) is either zero or one when the corresponding \( x_{ij} \) of the matrix \( X \) is observed or not. According to the MAR hypothesis, we have that:

\[
P(R|X_{obs}, X_{mis}; \theta) = P(R|X_{obs}; \theta)
\]

The main reason why the approach based on imputation is so widely used is that it produces a complete data set that can be easily analysed with the usual statistical techniques. A wide variety of imputation techniques has been developed (e.g. see Little and Rubin (2002)). While most of them have desirable properties when studying univariate characteristics, i.e. tend to reproduce the univariate distributions, the preservation of relationships among variables is still a problem to be furtherly studied. Bayesian Networks (BNs, Cowell et al. (1999)) are a well-known tool to study the relationships among variables in complex surveys. Consequently they can be usefully applied for imputation. A first study on the use of BNs for imputation can be found in Di Zio et al. (2004). The imputation algorithm proposed (Bayesian Network imputation technique, hereafter BN imputation technique) can be summarised in the following steps:

**Step 1** - the available variables should be ordered according to their reliability. Reliability is defined in terms of the accuracy, number of missing items, availability of external sources. Without loosing in generality, let \( X_1 \) be the most reliable variable, \( X_2 \) the second most reliable variable, and so on;

**Step 2** - the structure of the BN is estimated from the available data set (i.e. from the data set with missing items). The structure should fulfill the ordering in step 1. Then the

---

1 This work was partially supported by MIUR-cofin2003 and EUREDIT project financial support from the IST programme of the European Union
probability distribution is estimated by means of the EM algorithm;

**Step 3** - each missing item is imputed by a random generation of a value from the distribution of the variable, conditional on its parents. Note that parents’ values may be either observed or imputed.

The results obtained applying the above method are promising (see Di Zio et al. (2004)). It is also clear that further improvements can be achieved. Actually the parents of the variable to impute are just a part of the available information on that variable. Generally speaking the variable to impute is directly related not only with its parents but with its *Markov Blanket*, i.e. its parents, its children and the parents of its children.

The BN imputation technique, described in Di Zio et al. (2004), is modified to account for the Markov Blanket of each variable (Markov Blanket - Bayesian Networks imputation technique, hereafter MBBN imputation technique, Di Zio et al. (2003). In this case, we should impute a missing value in $X_j$ generating a random value from the distribution of $X_j$ given its Markov Blanket. Usually, this distribution may not be explicitly available, especially when some of the variables in the Markov Blanket are missing and consequently should be imputed first. We propose the following algorithm, that makes possible the computation of the Markov Blanket of each variable by the estimation of different BNs.

**Variables order** - The reliability order among the observed variables is still important. In fact we propose to impute $X_1$ first, then $X_2$ and so on. The last variable to impute is $X_k$.

**Estimation of BNs** - For each variable $X_j$, $j = 1, ..., k$, a BN is estimated. As far as $X_j$ is concerned, the estimation of the BN structure is constrained to avoiding edges starting from $X_j$. Let $G_j$ be the estimated graph. As a consequence, the Markov Blanket for $X_j$ in the graph $G_j$ consists only of the $X_j$ parents. Then, the distribution of the BN is estimated by means of the EM algorithm.

**Imputation** - According to the variables order, let us consider the imputation of missing values for $X_1$. We suggest to apply the BN imputation method adopting the BN whose structure is $G_1$. In other words, we impute a missing item for $X_1$ generating a value from the distribution of $X_1$ conditional on its parents (i.e. its Markov Blanket). Note that the parents of $X_1$ may be either observed or imputed. However, once all $X_1$ missing items have been filled in, we suggest to retain just these values and discard the imputations for the other variables, $(X_2, ..., X_k)$, given that they are not imputed according to their Markov Blanket. We consequently obtain a new data set, $D_1$, where $X_1$ is complete. Imputations for $X_2$ missing items are performed on the data set $D_1$ with the BN whose structure is $G_2$. Once each $X_2$ missing item has been filled in, only the imputed values for $X_2$ are retained, obtaining a new data set $D_2$ where $X_1$ and $X_2$ are complete. This procedure is iterated for all the variables. Note that imputations for $X_k$ are performed on a data set $D_{k-1}$ where the other $k - 1$ variables have been completed, and consequently the imputation step only involves the random generation from the distribution of $X_k$ conditional to the (observed or previously completed) values of the variables in the Markov Blanket of $X_k$ in the BN whose structure is $G_k$.

In order to compare the BN and MBNN imputation techniques, we have carried out an experiment on a subset of anonymous individual records and a subset of variables surveyed in the 1991 UK Population Census (SARS). For each sampled household member only 5 variables were analysed: age (cleta, 6 categories), sex (sex), primary economic position (econprim, 10 categories), long term illness (ltill, 2 categories), number of higher education qualifications (qualnum, 3 categories). As in Di Zio et al. (2004), reliability order among these variables is: 1) sex, 2) age, 3) qualnum, 4) econprim
Figure 1: The 5 BN structures for the MBBN imputation technique. White nodes are the nodes that have to be imputed by that BN.

and 5) ltill. The experiment consists of two steps: 1) random generation of artificial missing items in the set of complete SARS data; 2) missing items imputation by means of the two BN-based methods. As far as the first step is concerned, all the variables have been contaminated according to an MCAR mechanism with different expected percentages of missingness: sex 7%, age 10%, qualnum 8%, econprim 20%, ltill 10%. Furthermore, the variables qualnum and ltill have been contaminated with a MAR mechanism following these rules:

\[ P(\text{qualnum} = \text{missing} | \text{age} = i) = (2 \times i + 1)\%, \quad i = 1, \ldots, 6, \]

\[ P(\text{ltill} = \text{missing} | \text{sex} = 1) = 14\% , \quad P(\text{ltill} = \text{missing} | \text{sex} = 2) = 20\%. \]

We compared the performance of the two imputation techniques in terms of the preservation of the joint distribution as well as of micro data. As in Di Zio et al. (2004), the preservation of distributions has been measured on the subset of \( n^\ast \) missing units for variable \( X \) by the following relative index: \( \Delta = \sum_x |f_x - \tilde{f}_x|/2 \), where \( f_x \) denotes the relative frequency of category \( x \) of \( X \) in the true data set of the \( n^\ast \) missing items, and \( \tilde{f} \) is the corresponding frequency after imputation. The multivariate version of \( \Delta \) is straightforward (see Di Zio et al. (2004) for details). As a matter of fact, \( \Delta \) ranges between 0 and 1. To evaluate the preservation of microdata for each variable, the proportion of correct imputations has been measured: \( \xi = \sum I_x(\hat{x})/n^\ast \), where the sum is over the \( n^\ast \) missing units for variable \( X \), and \( I_x(\hat{x}) \) is 1 when the imputed \( (\hat{x}) \) and the true \( (x) \) values coincide, and zero otherwise. The results are reported in tables 1 and 2. For both the imputation techniques, the necessary BNs were learned with the software Hugin.
Table 1: $\Delta$ for the joint distribution when BN and MBBN imputation techniques are applied

<table>
<thead>
<tr>
<th></th>
<th>BN</th>
<th>MBBN</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta$</td>
<td>0.1</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Table 2: Percentages of correctly imputed values ($\xi$) with the two imputation techniques BN and MBBN

<table>
<thead>
<tr>
<th>Variables</th>
<th>BN</th>
<th>MBBN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEX</td>
<td>52.36</td>
<td>61.64</td>
</tr>
<tr>
<td>AGE (cleta)</td>
<td>19.34</td>
<td>26.05</td>
</tr>
<tr>
<td>QUALNUM</td>
<td>78.32</td>
<td>78.82</td>
</tr>
<tr>
<td>ECONPRIM</td>
<td>42.30</td>
<td>44.15</td>
</tr>
<tr>
<td>LTILL</td>
<td>83.52</td>
<td>84.08</td>
</tr>
</tbody>
</table>

(http://www.hugin.com/): the structure of the network was estimated with the PC algorithm (Spirtes et al. (1993)) while the distribution with the EM algorithm (Lauritzen (1995)). Further research will consider the estimation of the structure and the distribution with a maximum likelihood approach (the SEM algorithm, Friedman (1997)). The algorithms for imputation were implemented in an ad hoc C++ code.

These results are extremely encouraging, although this is just a first evaluation. These tables show that the MBBN imputation technique provides a major improvement in the imputation results. Furthermore we remark that, as shown in Di Zio et al. (2004), the BN imputation method is better than the usual hot-deck donor imputation methods. Consequently the MBBN imputation technique should be analyzed more in depth in order to see if it can be considered as an important alternative.

References


