Random walk methods for imputation of categorical data under edit restrictions

Mark P. J. van der Loo

Dpt. of Methodology, Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague, The Netherlands

m.vanderloo@cbs.nl

Keywords: categorical data, imputation, markov chain monte carlo, imputation variance

1 Introduction

Raw survey data frequently suffers from missing values and inconsistencies. To improve raw data quality, statistical institutes often estimate missing values and replace erratic values. There are a number of well-known algorithms which perform this task by treating raw data record by record. Examples include hot-deck or cold-deck imputation, nearest neighbour imputation and regression imputation. Imputation can be seen as a part of the estimation procedure for a statistic, and it is often difficult to estimate the amount of variance introduced by imputation.

The method presented in this paper differs from the mentioned examples in that data is not treated record by record. Instead, the set of all possible imputed datasets is considered. For categorical data, this set can be represented as a set of contingency tables. The set of valid tables is finite, although large, and the number of elements is limited by the number of respondents, the information in partially filled in records and edit restrictions. It is possible to define a probability measure on this set, parameterised by for example complete or historical data. Using random walk algorithms inspired by te work of Diaconis and Sturmfels (1998), it is possible to find a maximal probability solution or to draw random solutions. The latter facilitates the analysis of imputation variability under a probability model. For an extended description see also van der Loo (2009).

2 Contingency tables and missing items

A contingency table contains all information about a dataset, except the identity of the respondents. Contingency tables are constructed by counting records with similar data, and can always be represented as a d-dimensional nonnegative discrete vector $\mathbf{y}$. Here, d is the number of ways a complete record can be filled, which is the product of the number of possible categories for every variable in the record. The coefficients of $\mathbf{y}$, labeled y_t with $0 \leq t \leq d-1$, represent the number of respondents with the t^{th} answer combination. The marginals of $\mathbf{y}$ can also be represented as a discrete nonnegative vector $\mathbf{b}$. The map sending $\mathbf{y}$ to its marginals can be written as a matrix $\mathbf{A}$ so that

$$\mathbf{A}\mathbf{y} = \mathbf{b}.$$

Every edit restriction on a categorical data record can be written as an invalid value combination. Edit restrictions can therefore be represented as a set of indices $\mathfrak{I}$ so that x_t must be zero whenever $t \in \mathfrak{I}$.

Consider a dataset with n records, of which n^{full} are fully completed and n^{part} records have missing values. The complete contingency table $\mathbf{y}$ can be written as

$$\mathbf{y} = \mathbf{y}^{\mathsf{full}} + \mathbf{x},$$

where $\mathbf{y}^{\mathsf{full}}$ is the contingency table corresponding to the n^{full} completed records and $\mathbf{x}$ the unknown contingency table corresponding to the n^{part} records which contain mising items. The imputation problem consists of finding and estimate $\hat{\mathbf{x}}$ so that the full table $\mathbf{y}$ can be estimated as $\hat{\mathbf{y}} = \mathbf{y}^{\mathsf{full}} + \hat{\mathbf{x}}$.

Altough $\mathbf{x}$ is not known, the n^{part} incomplete records give some information on the marginals of $\mathbf{x}$. Consider for example a two-question survey asking the respondent about educational level and gender. A person who only fills in gender=male contributes to the marginal that is constructed by summing over educational levels, but not to the marginal that is obtained by summing over genders. We can write

$$\mathbf{A}\mathbf{x} = \mathbf{b}^{\mathsf{part}} + \mathbf{b}^{\mathsf{miss}},$$

where $\mathbf{b}^{\mathsf{part}}$ is the part of the marginals that is obtained from the partially filled records and $\mathbf{b}^{\mathsf{miss}}$ is the unknown part. Since all entries of $\mathbf{b}^{\mathsf{miss}}$ are nonnegative by definition, we have

$$\mathbf{A}\mathbf{x} \geq \mathbf{b}^{\mathsf{part}},$$

where $\geq$ is interpreted elementwise.

The set Ω_X of all possible instances of $\mathbf{x}$ can be written as

$$\Omega_X = \{\mathbf{x} \in \mathbb{Z}^d_{\geq 0} \mid \sum_t x_t = n^{\mathsf{part}} \wedge \mathbf{A}\mathbf{x} \geq \mathbf{b}^{\mathsf{part}} \wedge x_t = 0 \text{ when } t \in \mathfrak{I}\}.$$

Here, the first restriction ensures that entries of $\mathbf{x}$ sum to the number of partially filled records, the second that no filled in data is altered, and the third that structural zeros (edit restrictions) are obeyed. We can regard $\mathbf{x}$ an instance of a random variable X, distributed according to

$$p(\mathbf{x}) = \left\{ \begin{array}{l} |\Omega_X|^{-1}\mathcal{M}(\mathbf{x}|\boldsymbol{\theta}) \text{ if } \mathbf{x} \in \Omega_X \\ 0 \text{ otherwise,} \end{array} \right.$$

where $|\Omega_X|$ is the number of elements in Ω_X and $\mathcal{M}$ is the multinomial distribution given by

$$\mathcal{M}(\mathbf{x}|\boldsymbol{\theta}) = n^{\mathsf{part}}! \prod_{t=0}^{d-1} \frac{\theta_t^{x_t}}{x_t!}.$$

Here, $\boldsymbol{\theta}$ is the vector of cell probabilities which can be estimated from for example historical data or the n^{full} complete records. Determining $|\Omega_X|$ in general is an unsolved counting problem, but it is clear that the number of elements in practical problems grows rapidly with d and n^{part}. In imputation problems occurring in practice, the number of elements in Ω_X easily becomes too large to represent in computer memory. In the following section two algorithms are given which can produce random walks on Ω_X which enable one to draw randomized samples according to $p(\mathbf{x})$ or to approximate one of the local maxima of $p(\mathbf{x})$

3 Random walks

As stated before, drawing a sample from Ω_X directly is practically impossible because of the size of Ω_X. The solution is to generate a startvalue $\mathbf{x}(0)$ in Ω_X and to generate a random walk from there by taking basic random steps. After sufficient steps, say N, the instance $\mathbf{x}(N)$ can be considered a sample from Ω_X according to $p(\mathbf{x})$. The following algorithm, based on the Metropolis-Hastings sampler returns the N^{th} step of a random walk.

```
Generate some $\mathbf{x} \in \Omega_X$;
$\tau := 0$;
while $\tau < N$ do
    Draw $i$ and $j$ uniformly and without replacement from $\{0, 1, \ldots, d-1\}\backslash\mathfrak{I}$;
    $\mathbf{v} := \mathbf{0}$; $v_i := 1$; $v_j := -1$;
    if $\mathbf{x} + \mathbf{v} \geq \mathbf{0} \wedge \mathbf{A}(\mathbf{x} + \mathbf{v}) \geq \mathbf{b}^{\mathsf{part}}$ then
        $\tau := \tau + 1$;
        With probability $\min\{(p(\mathbf{x} + \mathbf{v})/p(\mathbf{x}), 1\}$ do $\mathbf{x} := \mathbf{x} + \mathbf{v}$;
```

Here, the backslash indicates set difference. The basic step which is set here, involves moving a record from one possible entry in the contingency table to another one. The fact that this algorithm generates a random sample from Ω_X is a consequence of theorem 3.1 of Diaconis and Sturmfels (1998). It is also explicitly shown in van der Loo (2009). In the latter paper, it is also shown how larger steps can be taken so that less steps are necessary.

The following algorithm tries to take N random steps in the direction of a mode of $p(\mathbf{x})$.

```
Generate some $x$ in $\Omega_X$;
$\tau := 0$;
while $\tau < N$ do
    $\tau := \tau + 1$;
    Draw $i$ and $j$ uniformly and without replacement from $\{0, 1, \ldots, d-1\}\backslash\mathfrak{I}$;
    $\mathbf{v} := \mathbf{0}$; $v_i := 1$; $v_j := -1$;
    if $\mathbf{x} + \mathbf{v} \geq \mathbf{0} \wedge \mathbf{A}(\mathbf{x} + \mathbf{v}) \geq \mathbf{b}^{\mathsf{part}} \wedge p(\mathbf{x} + \mathbf{v}) > p(\mathbf{x})$ then
        $\mathbf{x} := \mathbf{x} + \mathbf{v}$;
    Else if $\mathbf{x} - \mathbf{v} \geq \mathbf{0} \wedge \mathbf{A}(\mathbf{x} - \mathbf{v}) \geq \mathbf{b}^{\mathsf{part}} \wedge p(\mathbf{x} - \mathbf{v}) > p(\mathbf{x})$ then
        $\mathbf{x} := \mathbf{x} - \mathbf{v}$;
```

This algorithm draws a step and walks in the direction of higher probability if possible.

4 Implementation and numerical examples

Random walk algorithms were implemented in C and coupled to the R statistical package to facilitate analysis. Care was taken to optimize the code for speed as much as possible. The current version of the software is capable of treating datasets with maximum one missing value per record (the code gets more complex for general missing data patterns since it involves checking more partial marginals).

Numerical tests were performed using two different categorical datasets with more than 30 000 records, yielding a 8 × 7 × 2 × 5 and a 7 × 5 × 2 × 5 × 7 contingency table. In both

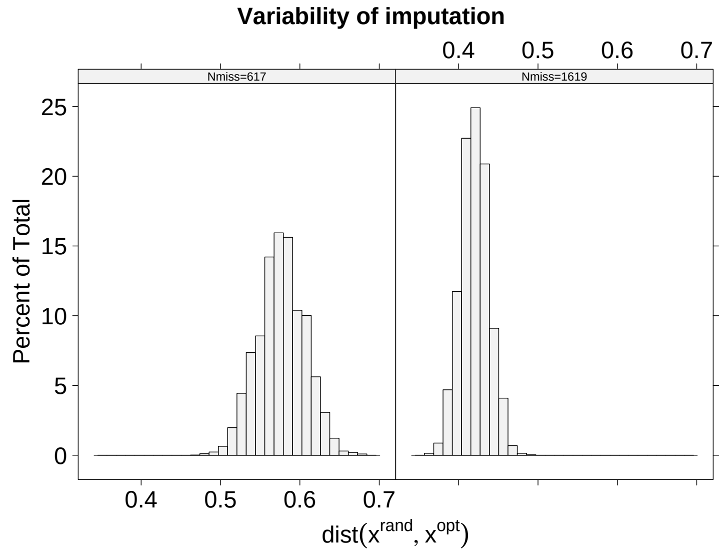

Figure I: Distribution of scaled euclidian distance between the most probable value $\mathbf{x}^{\mathsf{opt}}$ under $p(\mathbf{x})$ and a random value $\mathbf{x}^{\mathsf{rand}}$.

datasets 10 or 25% of the records were provided with missing values which were imputed using the algorithms described in the previous section. The values of θ_t were estimated from the complete data.

The tests show that it takes about $N = 10^5$ steps for the first algorithm to converge to a random drawing, the calculation taking on the order of 1 second on a AMD64 laptop running at 1.8GHz. Finding an optimum value takes about 10^4 steps. To demonstrate the power of the method, in Fig. I shows distributions of the distance between the optimum table found with the second algorithm and tables drawn with the first algorithm. This distribution can be seen as a measure of variance of imputation under the assumed model $p(\mathbf{x})$.

5 Conclusions and outlook

The algorithms described here offer a generic method to numerically investigate the contribution of imputation variability. They also seem very promising as generic methods for imputing categorical datasets randomly or optimally. Various extensions of the software described above are either planned or already implemented, see also van der Loo (2009). The resulting software could be a valuable tool for statistical researchers.

References

P. Diaconis and B. Sturmfels. Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics*, 26:363, 1998.

M. P. J. van der Loo. Algebraic algorithms for stochastic imputation of item nonresponse with edit restrictions. Technical report, Statistics Netherlands, 2009. In press.