Monte Carlo Methods Appl. **16** (2010), 421–438
DOI 10.1515/MCMA.2010.017

MCMC imputation in autologistic model

Marta Zalewska, Wojciech Niemiro and Bolesław Samoliński

Abstract. We consider statistical inference from incomplete sets of binary data. Our approach is based on the autologistic model, which is very flexible and well suited for medical applications. We propose a Bayesian approach, essentially using Monte Carlo techniques. The method developed in this paper is a special version of Gibbs sampler. We repeat intermittently the following two steps. First, missing values are generated from the predictive distribution. Second, unknown parametes are estimated from the completed data. The Monte Carlo method of computing maximum likelihood estimates due to Geyer and Thompson (1992) is modified to the Bayesian setting and missing data problems. We include results of some small scale simulation experiments. We artificially introduce missing values in a real data set and then use our algorithm to refill missings. The rate of correct imputations is quite satisfactory.

Keywords. Missing data, Markov chain Monte Carlo, medical data, Gibbs sampler, generalized linear models, Bayesian imputation.

2010 Mathematics Subject Classification. 65C05, 62P10, 65C60; 62J12.

1 Introduction

Data from epidemiology, clinical settings and also survey statistics are frequently incomplete. Missing values cause serious problems for statistical inference. Many techniques have been developed to cope with incomplete data sets [6, 8, 9]. Imputation, that is "filling in the gaps in data" is one of possible approaches. There are several methods of imputation, among which so-called Hot-Deck algorithms are perhaps most popular. Alternatives include maximum likelihood and Bayesian methods. The approach in our paper is model based, Bayesian and makes an essential use of Monte Carlo (MC) techniques. We focus on binary data and use a very flexible autologistic model [1].

In this paper, we tackle two closely related problems: *imputation* of missing data and *estimation* of unknown parameters of the autologistic model. Presence of missings makes estimation very difficult. Standard estimation techniques work well only for complete data. The expectation-maximization (EM) algorithm [2]

is probably the most successful method of estimation from incomplete data. But EM is difficult to apply (or even infeasible) in autologistic model because of the computational complexity of the M-step (maximization of the likelihood).

Our algorithm achieves the two above mentioned goals simultaneously. It is a special version of Gibbs sampler (GS). We repeat intermittently the following two steps. First, missing values are generated from the predictive distribution. Second, unknown parametes are estimated from the completed data. The Markov chain Monte Carlo (MCMC) method of computing maximum likelihood estimates due to Geyer and Thompson [4] is modified to the Bayesian setting and missing data problems.

Our chief motivation comes from epidemiological surveys. Medical data often consist of many "cases" which can be regarded as independent binary vectors with highly dependent components. Modelling such data as an independent sample from an autologistic distribution allows us to use asymptotic statistical theory. In this respect, the context in which we apply an autologistic model is quite different from that of spatial statistics (see [5]). In our model we have as many parameters as the pairs of variables and the dimension is moderately large. We point out that the method of Geyer and Thompson is ideally suited to produce samples from an asymptotic approximation to the posterior. In fact we also propose a simplified version of Bayesian estimates, in which likelihood is replaced by pseudo-likelihood. Although this algorithm is only heuristically justified, the results of simulation studies show its practical advantages. Maximum pseudo-likelihood estimates can be very efficiently computed using off-the-shelf methods based on generalized linear models (GLM). Moreover, the sampler needs only a few iterations to approach equilibrium (this conclusion is supported by the results of experiments). This makes our algorithm quite fast.

The paper is organized as follows. In Section 2 we introduce the model which is considered in the sequel. Section 3 describes the main part of our algorithm, namely the estimation step. In Section 4 we complete the description of the algorithm. Section 5 contains the results of some small scale experiments. The aim of this simulation study is to evaluate the performance of the imputation/estimation algorithm proposed in our paper. We use artificial data and also real data from a big allergological survey. Our methodology is the following. We artificially introduce missing values. Then we apply our algorithm to impute missings and simultaneously estimate parameters from incomplete data. Finally we check correctness of our imputations. We also compare estimates computed from incomplete and complete data.

2 Autologistic model

Let $x = (x_1, \ldots, x_d)^\top$ be a random vector with binary components. Assume a Gibbs probability distribution on $\mathcal{X} = \{0, 1\}^d$,

$$p_\beta(x) := \frac{1}{Z(\beta)} \mathrm{e}^{H_\beta(x)}, \tag{2.1}$$

where

$$H_\beta(x) := \sum_{i,j=1}^{d} \beta_{ij} x_i x_j.$$

As usual for this type of models, the norming constant,

$$Z(\beta) := \sum_{x \in \mathcal{X}} \mathrm{e}^{H_\beta(x)},$$

is typically intractable. We say that x has *autologistic* distribution, $x \sim \mathrm{AL}(\beta)$. It depends on a matrix of coefficients $B = (\beta_{ij})$. For identifiability assume that B is symmetric, $\beta_{ij} = \beta_{ji}$. It is more convenient to arrange elements of B in a vector of dimension $d(d+1)/2$, say

$$\beta = (\beta_{11}, \ldots, \beta_{1d}, \beta_{22}, \ldots, \beta_{2d}, \ldots, \beta_{dd})^\top$$

and write ith row of the matrix as

$$\beta_i = (\beta_{i1}, \ldots, \beta_{id})^\top.$$

It is easy to verify that for every $i = 1, \ldots, d$,

$$p_\beta(x_i = 1 \mid x_{-i}) = \frac{\exp\left(\beta_{ii} + \sum_{j \neq i} x_j \beta_{ij}\right)}{1 + \exp\left(\beta_{ii} + \sum_{j \neq i} x_j \beta_{ij}\right)}, \tag{2.2}$$

where $x_{-i} = (x_j, j \neq i)$. This means that the full conditional distributions are the same as in the standard model of logistic regression. For our purposes it is important that

- simulation of x can be easily implemented using Gibbs sampler,
- parameters β can be efficiently estimated using standard GLM methods combined with MCMC techniques.

Simulation via GS is straightforward using (2.2). We proceed to discuss methods of estimation.

3 Estimation

We assume that the set of data X consists of n independent rows, each row having identical auto-logistic distribution

$$X = \begin{pmatrix} x(1)^\top \\ \vdots \\ x(n)^\top \end{pmatrix} = \begin{pmatrix} x_1(1), \ldots, x_d(1) \\ \vdots \\ x_1(n), \ldots, x_d(n) \end{pmatrix},$$

with $x(k) = (x_1(k), \ldots, x_d(k))^\top \sim \mathrm{AL}(\beta)$ for $k = 1, \ldots, n$. In the applications we have in mind, the dimension d is moderately high while the sample size n is large enough to make asymptotic approximations work.

3.1 Maximum pseudo-likelihood via GLM

Let us first discuss a method of estimation based on the idea of maximum pseudo-likelihood and some approximations. Given vector $x \sim \mathrm{AL}(\beta)$, we consider partial log-likelihoods

$$L_i(\beta|x) = L_i(\beta_i|x) := \log p_\beta(x_i|x_{-i}). \tag{3.1}$$

Note that the above expression depends only on β_i but will be regarded as a function of whole $d(d+1)/2$-dimensional β.

Since the rows $x(k)$ of X are i.i.d., the partial log-likelihoods are additive:

$$L_i(\beta|X) = \sum_{k=1}^{n} L_i(\beta|x(k)),$$

where the summands are given by (3.1). Pseudo-log-likelihood is defined as

$$L_{\mathrm{ps}}(\beta|X) = \sum_{i=1}^{d} L_i(\beta|X).$$

Note that each off-diagonal element β_{ij} appears in this sum twice: in $L_i(\beta|X)$ and $L_j(\beta|X)$. Diagonal elements β_{ii} appear once.

If sample size n is large then partial log-likelihoods L_i are approximately quadratic in the neighborhood of the true values of model parameters. Consequently, so is L_{ps}. We start with the following approximation:

$$L_i(\beta_i|X) \simeq -\frac{n}{2}(\beta_i - \hat{\beta}_i)^\top I_i^{-1}(\beta_i - \hat{\beta}_i) + \mathrm{const},$$

where I_i is the partial Fisher information matrix of dimension $d \times d$ (obtained by regarding ith coordinate as response variable and conditioning on the remaining coordinates) and $\hat{\beta}_i$ is the maximizer of the partial likelihood L_i. To rewrite L_i as a quadratic function of β, suppose for a moment that β is rearranged as $\beta^\top = (\beta_i^\top, \beta_{-i}^\top)$, with β_i containing d coordinates of ith row of B and β_{-i}, all remaining $d(d-1)/2$ coordinates. Accordingly, let $\tilde{\beta}_i^\top = (\hat{\beta}_i^\top, 0^\top)$ and define a block matrix

$$K_i = \begin{pmatrix} I_i^{-1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix}.$$

Now we can write

$$\begin{aligned} L_{\mathrm{ps}}(\beta|X) &\simeq -\frac{n}{2}\sum_{i=1}^{d}(\beta-\tilde{\beta}_i)^\top K_i(\beta-\tilde{\beta}_i) + \text{const}, \\ &= -\frac{n}{2}(\beta-\bar{\beta})^\top K(\beta-\bar{\beta}) + \text{const}, \end{aligned} \tag{3.2}$$

where

$$\bar{\beta} = K^{-1}\sum_{i=1}^{d} K_i\tilde{\beta}_i, \quad K = \sum_{i=1}^{d} K_i. \tag{3.3}$$

Standard theory of generalized linear models (for example [7]) ensures that the partial ML estimates are asymptotically normal, if the sample size n goes to infinity: $\hat{\beta}_i \sim_{\text{approx.}} \mathcal{N}(\beta_i, I_i^{-1}/n)$. Equivalently, $\tilde{\beta}_i \sim_{\text{approx.}} \mathcal{N}(\beta, K_i^{-1}/n)$.

Matrices I_i (or equivalently K_i) are unknown but can be consistently estimated from data. Therefore we can use an analogue of formula (3.3) to compute an *estimator*

$$\beta_{\mathrm{ps}}^* = \hat{K}^{-1}\sum_{i=1}^{d}\hat{K}_i\tilde{\beta}_i. \tag{3.4}$$

For our purposes it is sufficient to note that β_{ps}^* is a $\sqrt{n}$-*consistent* estimator. It follows from the fact that β_{ps}^* a weighted sum of asymptotically normal vectors. Note that matrix $\hat{K}$ is nonsingular with probability going to one, because K is nonsingular. In fact, β_{ps}^* is also asymptotically normal, but we will not use this property.

3.2 Maximum likelihood via MCMC

Let us consider a parametric family of Gibbs measures as in (2.1) and write the log-likelihood in the following form:

$$\begin{aligned} L(\beta|x) = \log p_\beta(x) &= H_\beta(x) - \log Z(\beta) \\ &= H_\beta(x) - \log \sum_{x^* \in \mathcal{X}} \mathrm{e}^{H_\beta(x^*)}. \end{aligned}$$

Geyer and Thompson in the influential paper [4] put forward the idea of using Monte Carlo (MC) to approximate the sum in the above formula. For a fixed β^*,

$$\begin{aligned} L(\beta|x) &= H_\beta(x) - \log \sum_{x^* \in \mathcal{X}} \left[\mathrm{e}^{(H_\beta(x^*) - H_{\beta^*}(x^*))} \frac{1}{Z(\beta^*)} \mathrm{e}^{H_{\beta^*}(x^*)} \right] + \log Z(\beta^*) \\ &= H_\beta(x) - \log \mathbb{E} \mathrm{e}^{(H_\beta(x^*) - H_{\beta^*}(x^*))} + \text{const}, \end{aligned}$$

where random variable x^* has the probability distribution p_{β^*}. This change-of-measure identity allows us to apply an importance sampling MC scheme to approximate the ratio $Z(\beta)/Z(\beta^*)$, which is expressed as the expectation in the last display. The method is particularly appealing if the densities p_β form an exponential family, that is if

$$H_\beta(x) = \beta^\top T(x),$$

where $T(x)$ is a vector of sufficient statistics. The formula for the log-likelihood becomes

$$L(\beta|x) = \beta^\top T(x) - \log \mathbb{E} \mathrm{e}^{(\beta - \beta^*)^\top T(x^*)} + \text{const}.$$

Here and in the sequel we adopt the convention that *expectation is computed with respect to $x^* \sim p_{\beta^*}$ while x is kept fixed*. For brevity we will write $T = T(x)$ and $T^* = T(x^*)$. The derivatives of the log-likelihood can be expressed as follows:

$$\begin{aligned} \nabla L(\beta|x) &= T - \frac{\mathbb{E} T^* \mathrm{e}^{(\beta - \beta^*)^\top T^*}}{\mathbb{E} \mathrm{e}^{(\beta - \beta^*)^\top T^*}} = - \frac{\mathbb{E}(T^* - T) \mathrm{e}^{(\beta - \beta^*)^\top (T^* - T)}}{\mathbb{E} \mathrm{e}^{(\beta - \beta^*)^\top (T^* - T)}} \\ \nabla^2 L(\beta|x) &= - \frac{\mathbb{E}(T^* - T)(T^* - T)^\top \mathrm{e}^{(\beta - \beta^*)^\top (T^* - T)}}{\mathbb{E} \mathrm{e}^{(\beta - \beta^*)^\top (T^* - T)}} \\ &\quad + \left[\frac{\mathbb{E}(T^* - T) \mathrm{e}^{(\beta - \beta^*)^\top (T^* - T)}}{\mathbb{E} \mathrm{e}^{(\beta - \beta^*)^\top (T^* - T)}} \right] \left[\frac{\mathbb{E}(T^* - T) \mathrm{e}^{(\beta - \beta^*)^\top (T^* - T)}}{\mathbb{E} \mathrm{e}^{(\beta - \beta^*)^\top (T^* - T)}} \right]^\top. \end{aligned}$$

Now let us consider an i.i.d. sample $X = (x(1), \ldots, x(n))^\top \sim p_\beta$ and suppose that an approximation to the log-likelihood is computed using a Monte Carlo

sample $X^* = (x^*(1), \ldots, x^*(n^*))^\top \sim p_{\beta^*}$. Write $T_k^* = T(x^*(k))$ and $\bar{T} = n^{-1} \sum T(x(k))$. We obtain the MC approximation L^* by replacing, in the formula for L, expectation by an average with respect to the generated sample X^*. The same applies to the derivatives and thus we get the following formulas:

$$
\begin{aligned}
L^*(\beta|X) &= n\beta^\top \bar{T} - n \log \sum_{k=1}^{n^*} \mathrm{e}^{(\beta-\beta^*)^\top T_k^*} + \text{const},\\
\nabla L^*(\beta|X) &= -n \frac{\sum (T_k^* - \bar{T}) \mathrm{e}^{(\beta-\beta^*)^\top (T_k^* - \bar{T})}}{\sum \mathrm{e}^{(\beta-\beta^*)^\top (T_k^* - \bar{T})}}\\
\nabla^2 L^*(\beta|X) &= -n \frac{\sum (T_k^* - \bar{T})(T_k^* - \bar{T})^\top \mathrm{e}^{(\beta-\beta^*)^\top (T_k^* - \bar{T})}}{\sum \mathrm{e}^{(\beta-\beta^*)^\top (T_k^* - \bar{T})}}\\
&\quad + n \left[\frac{\sum (T_k^* - \bar{T}) \mathrm{e}^{(\beta-\beta^*)^\top (T_k^* - \bar{T})}}{\sum \mathrm{e}^{(\beta-\beta^*)^\top (T_k^* - \bar{T})}} \right] \left[\frac{\sum (T_k^* - \bar{T}) \mathrm{e}^{(\beta-\beta^*)^\top (T_k^* - \bar{T})}}{\sum \mathrm{e}^{(\beta-\beta^*)^\top (T_k^* - \bar{T})}} \right]^\top,
\end{aligned}
$$

where every '$\sum$' stands for '$\sum_{k=1}^{n^*}$'. The derivatives in the above formulas are understood with respect to β, with β^* fixed. The formulas simplify if we evaluate derivatives at $\beta = \beta^*$:

$$
\begin{aligned}
\nabla L^*(\beta^*|X) &= -\frac{n}{n^*} \sum (T_k^* - \bar{T})\\
\nabla^2 L^*(\beta^*|X) &= -\frac{n}{n^*} \sum (T_k^* - \bar{T})(T_k^* - \bar{T})^\top\\
&\quad + n \left[\frac{1}{n^*} \sum (T_k^* - \bar{T}) \right] \left[\frac{1}{n^*} \sum (T_k^* - \bar{T}) \right]^\top.
\end{aligned}
\tag{3.5}
$$

Now assume that β^* is a $\sqrt{n}$-consistent estimator of β. An obvious choice in our model is to use the maximum pseudo-likelihood estimator given by (3.4): $\beta^* = \beta^*_{\mathrm{ps}}$. Consider one-step maximum likelihood (ML) estimator computed according to the Newton–Raphson formula

$$
\hat{\beta}_{\text{1-step}} = \beta^* - \nabla^2 L(\beta^*|X)^{-1} \nabla L(\beta^*|X).
$$

It is well known [10, par. 5.7] that $\hat{\beta}_{\text{1-step}}$ has similar asymptotic properties as the genuine ML estimator $\hat{\beta}_{\mathrm{ML}}$. In particular, $\hat{\beta}_{\text{1-step}}$ is asymptotically normal and efficient. We have $\hat{\beta}_{\text{1-step}} - \hat{\beta}_{\mathrm{ML}} = o_{\mathrm{P}}(1/\sqrt{n})$ and therefore

$$
\hat{\beta}_{\text{1-step}} \sim_{\text{approx.}} \mathcal{N}(\beta, I^{-1}/n),
$$

where I is the Fisher information matrix and β denotes the true value of the parameter. Let us stress that to compute $\hat{\beta}_{\text{1-step}}$ exactly, we have to generate a Monte Carlo sample of size $n^* = \infty$. For finite n^* we have to take into account extra variability introduced by random number generation. Let β^*_{MC} be given by

$$\beta^*_{\text{MC}} = \beta^* - \nabla^2 L^*(\beta^*|X)^{-1} \nabla L^*(\beta^*|X), \tag{3.6}$$

where L^* is an MC approximation of L based on generated sample X^* of size n^*. Conditionally, given the real sample X, we have

$$\beta^*_{\text{MC}} \sim_{\text{approx.}} \mathcal{N}(\hat{\beta}_{\text{ML}}, I^{-1}/n^*). \tag{3.7}$$

3.3 Bayesian approach

Now let us equip additionally the autologistic model with Bayesian structure. For simplicity choose a flat, uniform prior: $\pi(\beta) \propto \text{const}$. Then the posterior is proportional to the likelihood: $\pi(\beta|X) \propto p_\beta(X)$. For large n the posterior is approximately normal [10, par. 10.2]

$$\beta|X \sim_{\text{approx.}} \mathcal{N}(\hat{\beta}_{\text{ML}}, I^{-1}/n). \tag{3.8}$$

We can exploit the similarity of (3.7) and (3.8). If we choose $n^* = n$ then the distribution of β^*_{MC} is approximately equal to the posterior. We thus use to advantage the extra random variability of Monte Carlo estimates. Instead of maximizing the likelihood we produce samples from the posterior distribution. It is very convenient for our purposes. Summing up, the proposed sampling scheme is the following:

(i) Compute β^*_{ps} according to formula (3.4) using GLM methodology. In practice we suggest the R functions `glm()` or `glm.fit()`.

(ii) Compute β^*_{MC} according to formula (3.6) with $\beta^* = \beta^*_{\text{ps}}$, using an MC sample of size $n^* = n$.

Although the algorithm described above is very appealing, let us describe a simplified heuristic alternative, which is much faster and not significantly less efficient in practice. To lighten computational burden, let us replace the true likelihood by pseudo-likelihood and use quadratic approximation (3.2). Then distribution $\mathcal{N}(\beta^*_{\text{ps}}, K^{-1}/n)$ takes over the role played by $\mathcal{N}(\hat{\beta}_{\text{ML}}, I^{-1}/n)$ in (3.8) and we arrive at the algorithm with a simplified second step:

(ii)′ Sample β^*_{simpl} from $\mathcal{N}(\beta^*_{\text{ps}}, \hat{K}^{-1}/n)$. Estimator $\hat{K}$ is obtained along with β^*_{simpl} using GLM methods at the first step.

4 Imputation

Now suppose that some elements of the data matrix X are missing:

$$X = (X_{\text{obs}}, X_{\text{mis}}).$$

In the Bayesian setup described at the end of Section 3, we apply the usual scheme of model-based imputations. Start with some initial imputed values of X_{mis}. Then iterate the following two steps of Gibbs sampler:

(i) Generate β from $\pi(\beta|X_{\text{obs}}, X_{\text{mis}})$.

(ii) Generate X_{mis} from $\pi(X_{\text{mis}}|X_{\text{obs}}, \beta)$.

When generating β we can use the normal approximation described in the previous subsection and consequently β^*_{MC}. In our experiments we settled for the simplified version of the algorithm and β^*_{simpl}. Estimators $\hat{\beta}_i$ and matrices $\hat{I}_i$, needed to compute β^*_{ps} and $\hat{K}$, are computed using standard GLM algorithms. Generation of X_{mis}, given β is in principle straightforward. Here we make one or several steps of standard Gibbs sampler on $\mathcal{X}$ based on the formula (2.2) for full conditionals. Of course, the sampler is restricted to X_{mis} and keeps X_{obs} unaltered.

5 Simulation experiments

5.1 Experiments artificial data

Let us report results of a small scale simulation study. First series of experiments was conducted on artificial data. Data sets were generated according to the autologistic model AL(β). The chosen parameters β (arranged in the symmetric matrix B as explained in Section 2) are given in the left upper part of Table 1, "Model parameters". In each single experiment the following steps were performed.

(i) Generation of an i.i.d. sample X ($n = 1000$ binary vectors of dimension $d = 4$). It was done by a GS using formula (2.2).

(ii) Estimation of parameters. We computed β^*_{ps} using formula (3.4) and R function `glm.fit()`. Apart from the estimates themselves, we also computed estimates of their standard errors. Average values are given in the upper right part of Table 1, "Estimated mean errors".

(iii) Introduction of missings, by erasing completely at random a given percent of data (entries of matrix X).

(iv) Application of our algorithm to incomplete data. At the output of the algorithm, we obtained estimates of the parameters (β^*_{simpl}) and also imputed values of missing data.

Since we worked with artificial data, we were able to compare the obtained estimates with the ground truth. We repeated steps (i)–(iv) one hundred times to assess estimation errors. For each of the 100 generated samples (each of size 1000), estimates of parameters β were computed using our method, first for complete and then for incomplete data. Results are summarized in Tables 1–5 below. Monte Carlo approximations of mean values of estimators, "Mean estimates" and their standard errors, "Mean errors of estimates" were computed using 100 repetitions, as explained above. Everything was done for four levels of incompleteness. Table 1 corresponds to complete data.

Inspection of the results shows that the bias of estimates is negligible for complete data and also for data with small percentage of missings. If the percentage of missings is high, the estimates become biased and their standard errors becomes worse, as should be expected. Nevertheless, the estimators work reasonably well. Even for data with 40% missing values, the estimates are still quite meaningful.

Model parameters				Mean errors of estimates			
3	−1	0	1	0.58	0.63	0.34	0.49
−1	0	2	3	0.63	0.66	0.35	0.28
0	2	-2	0	0.34	0.35	0.45	0.26
1	3	0	−1	0.49	0.28	0.26	0.53
Mean estimates				Estimated mean erorrs			
3.18	−1.11	−0.04	0.96	0.58	0.47	0.22	0.33
−1.11	0.08	2.08	3.02	0.47	0.67	0.23	0.19
−0.04	2.08	−2.05	0.01	0.22	0.23	0.43	0.20
0.96	3.02	0.01	−0.94	0.33	0.19	0.20	0.49

Table 1. Estimators of parameters β from complete data.

Mean estimates				Mean errors of estimates			
3.31	−1.26	−0.04	0.96	1.65	1.72	0.4	0.57
−1.26	0.23	2.07	3.01	1.72	1.75	0.43	0.31
−0.04	2.07	−2.04	0.01	0.4	0.43	0.55	0.33
0.96	3.01	0.01	−0.93	0.57	0.31	0.33	0.61

Table 2. Estimators of parameters β from data with 10% missings.

Mean estimates				Mean errors of estimates			
3.92	−1.85	0.06	0.88	3.54	3.58	0.48	0.7
−1.85	0.82	2.04	2.93	3.58	3.57	0.5	0.44
0.06	2.04	−2.07	0.02	0.48	0.5	0.62	0.44
0.88	2.93	0.02	−0.79	0.7	0.44	0.44	0.77

Table 3. Estimators of parameters β from data with 20% missings.

Mean estimates				Mean errors of estimates			
3.68	−1.34	−0.04	0.65	2.86	2.49	0.54	1.89
−1.34	0.46	1.93	2.85	2.49	2.42	0.61	0.45
−0.04	1.93	−1.95	0.08	0.54	0.61	0.82	0.51
0.65	2.85	0.08	−0.59	1.89	0.45	0.51	1.86

Table 4. Estimators of parameters β from data with 30% missings.

Mean estimates				Mean errors of estimates			
3.35	−0.8	−0.01	0.41	3.7	2.39	0.77	2
−0.8	0.06	1.86	2.64	2.39	2.38	0.63	0.7
−0.01	1.86	−1.97	0.17	0.77	0.63	0.88	0.58
0.41	2.64	0.17	−0.29	2	0.7	0.58	2.06

Table 5. Estimators of parameters β from data with 40% missings.

5.2 Experiments on real medical data

In the next series of experiments we used real data, collected in a big epidemiological survey ECAP (prevalence of allergic diseases in Poland 2006–2008). The full data set contains 18617 units (cases, respondents) and 1225 variables (mostly binary, but also numeric). Unit nonresponse in the survey was 4086 cases (18%). There was also significant item nonresponse.

To our experiments we selected a small portion of this large database, containing $n = 2962$ cases and $d = 6$ binary variables. At the present stage of our research we are mostly concerned with examination of developed statistical and computational tools. Therefore we adopted the following approach. We chose a part of data which is *complete*, contains no missings. Then we artificially erased some percentage of values and used our algorithm to refill them.

Figure 1 and Table 6 below indicate the percentage of correctly and wrongly imputed missing values for four different "levels of incompleteness".

The methodology of our experiments was quite analogous to that applied to artificial data. We repeated steps (ii)–(iv) exactly as described in the previous subsection. Of course, the "true values of parameters" were unknown, but estimates computed from complete data could be considered as some sort of ground truth. They are given in Table 7 and compared with those based on incomplete data, given in Tables 8–11. Each of these tables is based on one hundred repetitions. Note that we generated missings at random one hundred times in the same dataset. The tables report "Mean estimates" (averages computed from 100 repetitions) and "Mean errors of estimates" (standard deviations computed from 100 repetitions).

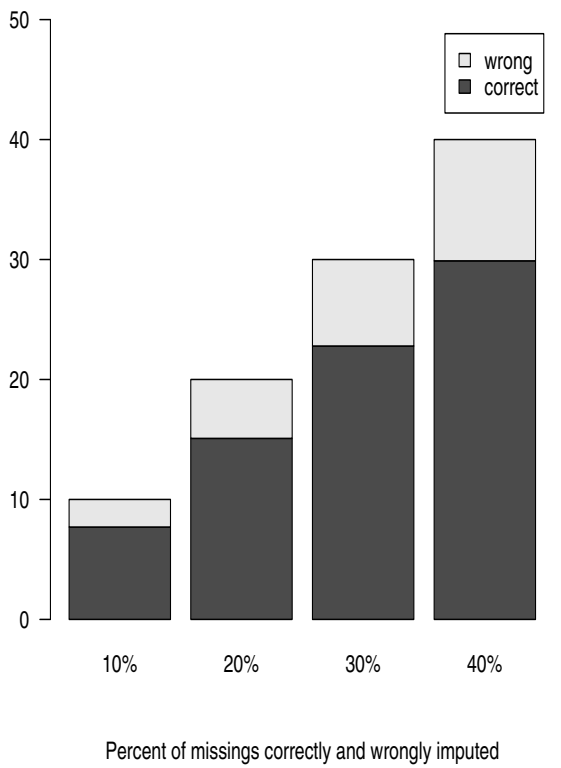

Figure 1. Results of imputations for different percents of missings.

	Percent of missings			
missings	10%	20%	30%	40%
correct	7.7	15.09	22.79	29.89
wrong	2.3	4.91	7.21	10.11

Table 6. Results of imputations for different percents of missings.

Some caution is needed when interpreting "errors". We just computed the square root of mean square differences between estimates from *complete* and *incomplete* data. Put differently, we assess only the "error due to incompleteness". Similarly, comparison of "Mean estimates" in Tables 8–11 with "Estimates" in Table 7 allows us to assess the "bias due to incompleteness".

Figure 2 shows how the "errors due to incompleteness" depend on the percentage of missings. We indicate only 6 components of 21-dimensional vector β. The single horizontal bars corresponding to "0%" correspond to estimates from complete data.

The study reported in this section had a preliminary character. We restricted ourselves to moderate size datasets and used a simplified version of our algorithm. However, the results obtained so far are encouraging. The performance of estimators on incomplete data is satisfactory, even when the percent of missings in significant. Let us emphasize the results concerning real data. We obtained a high rate of successful, correct imputations. This is not only a positive result about our algorithm but also an evidence in favour of using autologistic model in epidemiological applications.

Further research is needed to examine the usefulness of our algorithm for large datasets. Sampling from the posterior distribution, based on Monte Carlo approximation to the likelihood, is a method with sound theoretical justification. However,

Estimates					
−3.02	0.19	1.07	1.64	0.95	0.6
0.19	−5.15	1.92	0.58	1.97	0.28
1.07	1.92	−3.52	0.98	0.98	0.28
1.64	0.58	0.98	−1.02	0.65	0.56
0.95	1.97	0.98	0.65	−2.37	0.7
0.6	0.28	0.28	0.56	0.7	−1.78

Table 7. Estimators of parameters β from complete data.

Mean estimates

−3.02	0.14	1.07	1.64	0.95	0.6
0.14	−5.17	1.91	0.58	2.05	0.27
1.07	1.91	−3.52	0.99	0.98	0.28
1.64	0.58	0.99	−1.02	0.65	0.57
0.95	2.05	0.98	0.65	−2.37	0.7
0.6	0.27	0.28	0.57	0.7	−1.79

Mean errors of estimates

0.06	0.17	0.1	0.09	0.11	0.09
0.17	0.16	0.16	0.19	0.19	0.14
0.1	0.16	0.08	0.1	0.11	0.09
0.09	0.19	0.1	0.03	0.09	0.08
0.11	0.19	0.11	0.09	0.06	0.07
0.09	0.14	0.09	0.08	0.07	0.04

Table 8. Estimators of parameters β from data with 10% missings.

Mean estimates

−3.03	0.17	1.06	1.66	0.95	0.57
0.17	−5.15	1.94	0.62	1.96	0.23
1.06	1.94	−3.53	0.98	0.99	0.28
1.66	0.62	0.98	−1.02	0.66	0.57
0.95	1.96	0.99	0.66	−2.38	0.71
0.57	0.23	0.28	0.57	0.71	−1.79

Mean errors of estimates

0.1	0.26	0.15	0.12	0.13	0.13
0.26	0.25	0.24	0.29	0.29	0.24
0.15	0.24	0.11	0.17	0.16	0.15
0.12	0.29	0.17	0.04	0.13	0.1
0.13	0.29	0.16	0.13	0.07	0.12
0.13	0.24	0.15	0.1	0.12	0.07

Table 9. Estimators of parameters β from data with 20% missings.

Mean estimates

−3	0.2	1.07	1.64	0.92	0.59
0.2	−5.19	1.93	0.58	2.01	0.27
1.07	1.93	−3.51	0.98	0.98	0.27
1.64	0.58	0.98	−1.02	0.68	0.56
0.92	2.01	0.98	0.68	−2.38	0.71
0.59	0.27	0.27	0.56	0.71	−1.79

Mean errors of estimates

0.13	0.31	0.2	0.16	0.22	0.19
0.31	0.28	0.32	0.38	0.35	0.27
0.2	0.32	0.16	0.25	0.22	0.23
0.16	0.38	0.25	0.06	0.17	0.14
0.22	0.35	0.22	0.17	0.09	0.16
0.19	0.27	0.23	0.14	0.16	0.07

Table 10. Estimators of parameters β from data with 30% missings.

Mean estimates

−3.01	0.1	1.04	1.64	1.01	0.57
0.1	−5.21	2	0.66	1.97	0.28
1.04	2	−3.54	1.02	0.99	0.26
1.64	0.66	1.02	−1.02	0.62	0.58
1.01	1.97	0.99	0.62	−2.39	0.7
0.57	0.28	0.26	0.58	0.7	−1.78

Mean errors of estimates

0.49	0.29	0.24	0.28	0.22	
0.49	0.45	0.41	0.51	0.58	0.39
0.29	0.41	0.21	0.28	0.33	0.28
0.24	0.51	0.28	0.07	0.25	0.18
0.28	0.58	0.33	0.25	0.13	0.2
0.22	0.39	0.28	0.18	0.2	0.1

Table 11. Estimators of parameters β from data with 40% missings.

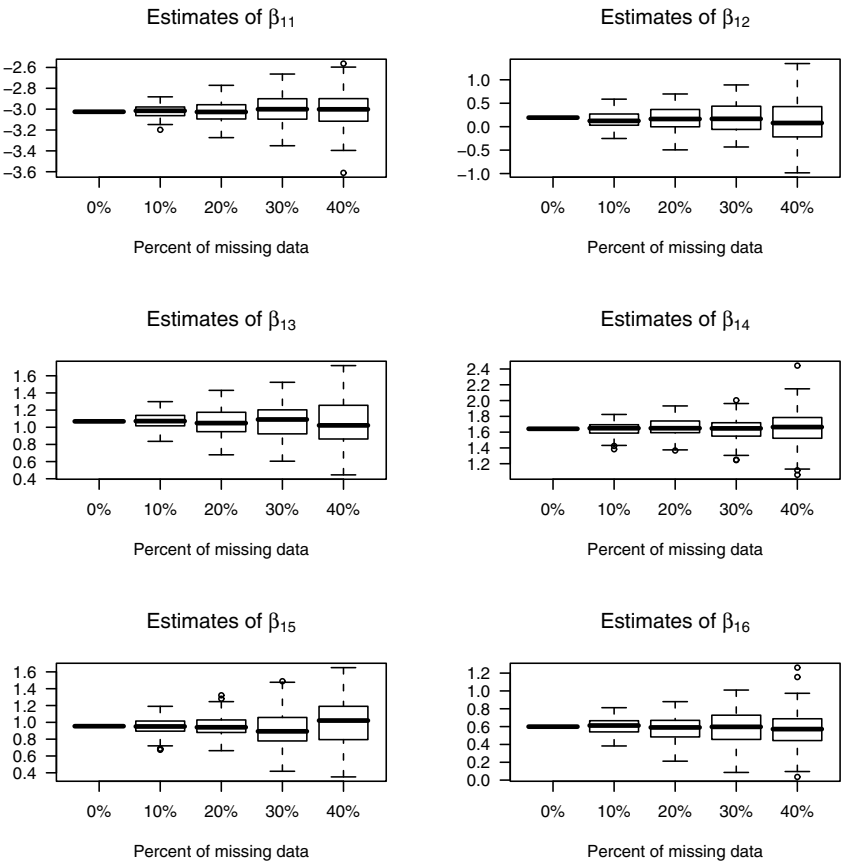

Figure 2. Boxplots of estimates computed from incomplete data.

for big datasets it may turn out to be computationally costly. Our experience shows that the heuristic method based on pseudo-likelihood has comparable statistical properties but is much faster. Finally let us note that several generalizations and extensions of autologistic model are possible. It is not difficult to include covariates (binary or numeric explanatory variables) in the autologistic model. Huffer and Wu in [5] proposed this in an application to spatial statistics. Presence of covariates should improve the imputation of missing values, but also poses new problems, for example if covariates themselves are incomplete.

6 Concluding remarks

The algorithm proposed in our paper aims at estimating the parameters of the autologistic model from incomplete data and, simultaneously, imputing missing data. Note that parameters of the autologistic model have clear and intuitive interpretation. Their estimates are thus valuable for the user. This is particularly true for epidemiological studies, in which estimated parameters reflect interdependence of various risk factors and occurrence of symptoms. On the other hand, imputed values might be of independent interest, if the rate of correct guesses is high.

We consistently use *model-based approach* and Bayesian methodology, as opposed to predominantly heuristic character of many other imputation techniques. Our algorithm is a complex version of Gibbs sampler. The output includes both imputed missing values and estimated parameters, sampled from (approximately) posterior joint distribution. The *reliability* of our method was tested on real epidemiological data with artificially generated missings. We think that this simulation methodology, described in detail in Section 5, is a honest way of evaluating the performance of the method in practice. The results were good and encouraging. Moreover, the experiments performed on both synthetic and real data show efficiency of our algorithm.

As a by-product of our considerations we developed a method of approximating the maximum likelihood or Bayesian estimates, which is more efficient than maximum pseudo-likelihood *even for complete data*.

Further work is needed to develop a similar algorithm for *data with auxiliary variables*. Such a generalization is important for many applications, especially in epidemiology and clinical research.

Acknowledgments. Work partially supported by Polish Ministry of Science and Higher Education Grant No. N N206 356036.

Bibliography

[1] Besag, J., Spatial interaction and the statistical analysis of lattice systems (with discussion), *J. R. Statist. Soc. B* 36 (1974), 192–236.

[2] Dempster, A. P. , Laird, N. M. and Rubin, D. B., Maximum Likelihood from Incomplete Data via the EM Algorithm, *J. R. Statist. Soc. B* 39 (1977), 1–38.

[3] Geyer, C. J., On the convergence of Monte Carlo maximum likelihood calculations, *J. R. Statist. Soc. B* 56 (1994), 261–274.

[4] Geyer, C. J. and Thompson, E. A., Constrained Monte Carlo maximum likelihood for dependent data, *J. R. Statist. Soc. B* 54 (1992), 657–699.

[5] Huffer, F. W. and Wu, H., Markov chain Monte Carlo for autologistic regression models with application to the distribution of plant species, *Biometrics* 54 (1998), 3, 509–524.

[6] Little, R. J. A., Missing-data adjustments in large surveys, *Journal of Business & Economic Statistics* 6 (1988), 287–296.

[7] McCullagh, P. and Nelder, J. A., *Generalized Linear Models*, London, Chapman and Hall, 1989.

[8] Rubin, D. B., Inference and missing data, *Biometrika* 63 (1976), 581–592.

[9] Rubin, D. B., *Multiple Imputation for Nonresponse in Surveys*, New York, Wiley, 1987.

[10] van der Vaart, A. W., *Asymptotic Statistics*, Cambridge University Press, 1998.

Received November 15, 2009; revised September 15, 2010.

Author information

Marta Zalewska, Department of Environmental Hazards Prevention and Allergology, Medical University of Warsaw, Zwirki i Wigury 61, 02-091 Warszawa, Poland.
E-mail: zalewska.marta@gmail.com

Wojciech Niemiro, Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Chopina 12/18, 87-100 Toruń, Institute of Applied Mathematics and Mechanics, University of Warsaw, Banacha 2, 02-097 Warszawa, Poland.
E-mail: wniemiro@gmail.com

Bolesław Samoliński, Department of Environmental Hazards Prevention and Allergology, Medical University of Warsaw, Zwirki i Wigury 61, 02-091 Warszawa, Poland.
E-mail: bsamol@amwaw.edu.pl