KMi
Knowledge Media Institute

# Model Folding for Data Subject to Nonresponse

*Paola Sebastiani and Marco Ramoni*

**KMI-TR-66**

**July 1998**

The Open University

# Model Folding for Data Subject to Nonresponse

**Paola Sebastiani**
Department of Actuarial Science and Statistics
City University

**Marco Ramoni**
Knowledge Media Institute
The Open University

## Abstract

This paper presents a new model selection method, called *Model Folding*, for regression models with partially classified categorical data in which only the dependent variable is subject to non response.

**Keywords:** Bayesian Learning, Missing Data, Model Selection, Model Folding.

**Reference:** KMi Technical Report KMi-TR-66, July 1998.

**Address:** Paola Sebastiani, Department of Actuarial Science and Statistics, City University, Northampton Square, London EC1V 0HB, United Kingdom. PHONE: +44 (171) 477-8959, FAX: +44 (171) 477-8838, EMAIL: p.sebastiani@city.ac.uk, URL: http://www.city.ac.uk/~sn303.

## 1. Introduction

This paper presents a method for model selection in the specific context of partially classified categorical data $X_1, \ldots, X_v, Y$ in which only the response variable $Y$ is subject to non response. We assume that the objective of inference is to select the subset of variables in $X_1, \ldots, X_v$ that have an effect on $Y$ and assume marginal independence of $X_1, \ldots, X_v$. This choice allows us to represent a model in which $Y$ depends on a subset of variables $X_1, .., X_s$ by a directed graph, with arrows pointing from $X_1, .., X_s$ to $Y$. The set $\mathcal{M}$ of possible models is finite, and can be represented by a lattice with $v+1$ levels. We shall denote each model by $M_{is}$, where $is$ is one of the possible combination of $s$ indices out of $1, 2, \ldots, v$. We adopt a Bayesian approach, and denote by $p(M_{is})$ the prior probability of $M_{is}$. Then, data $y$ are used to compute the posterior probability of $M_{is}$: $p(M_{is}|y) \propto p(M_{is})p(y|M_{is})$, and the model with the largest posterior probability is selected as best model given the sample information. We assume that, given $M_{is}$, the distribution of $Y$ is a product of multinomial distributions with parameters $\theta_k^{(is)}$, where $p(Y = y_j|\pi_k^{(is)}, \theta^{(is)}) = \theta_{kj}^{(is)}$, corresponding to the combinations $\pi_k^{(is)}$ of categories of variables with indexes in $is$. The prior distribution of the parameters $\theta^{(is)} = \{\theta_k^{(is)}\}$ is assumed to be a product of Dirichlet distributions [4], so that parameters associated to different conditional distributions $Y|\pi_k^{(is)}$ are mutually independent. Since we will be considering different models, we shall further assume that the hyper-parameters $\alpha_{k1}^{(is)}, \ldots, \alpha_{kc}^{(is)}$ are chosen so that they yield the same distribution for the parameters associated to the marginal distribution of $Y$.

When the sample is complete, these assumptions allow us to find explicitly $p(y|M_{is})$ as

$$p(y|M_{is}) \propto p(M_{is}) \prod_k \frac{\Gamma(\alpha_{k\cdot}^{(is)})}{\Gamma(\alpha_{k\cdot}^{(is)} + n_{k\cdot}^{(is)})} \prod_j \frac{\Gamma(\alpha_{kj}^{(is)} + n_{kj}^{(is)})}{\Gamma(\alpha_{kj}^{(is)})}$$

where $n_{kj}^{(is)}$ is the sample frequency of cases with categories $\pi_k^{(is)}, y_j$ under model $M_{is}$, and $n_{k\cdot}^{(is)} = \sum_j n_{kj}^{(is)}$. Similarly, $\alpha_{k\cdot}^{(is)} = \sum_j \alpha_{kj}^{(is)}$.

When the sample is incomplete and some of the $Y$ values are reported as unknown, information about the MDM is needed for carrying out inference. Let $y_o$ denote the subset of $y$ with observed values, $y_m$ be the subset with missing entries and $Y_m$ be the variable taking as values the possible realization $y_r$ of $y_m$. Following Rubin's approach [3], the MDM can be explicitly represented by associating with each case in the sample an indicator variable $R$ taking value 1 when $Y$ is not observed and 0 otherwise. Let $\psi$ be the parameter associated to the distribution of $R$. Ramoni and Sebastiani [2] define the MDM as *totally ignorable* if it is ignorable for every model in the set $\mathcal{M}$. They also show that sufficient conditions for total ignorability are that $R \perp Y_m|y_o, \psi$, i.e. data are missing at random (MAR) and that $\theta^{(is)} \perp \psi$ and $Y_m \perp \psi|(Y_o, X_1, \ldots, X_v, \theta^{(is)})$ for every model in $\mathcal{M}$. A less restrictive assumption is to require *partial ignorability* that is to assume that $\theta^{(is)} \perp \psi$, and $Y_m \perp \psi|(Y_o, X_1, \ldots, X_v, \theta^{(is)})$ hold only for some model in $\mathcal{M}$. The consequence of assuming either total or partial ignorability results in the computation of the likelihood: for the present problem, assuming total ignorability is simply equivalent to disregarding incomplete cases from the sample with the consequence that non representative samples are treated as if they were indeed representative. Partial ignorability, on the other hand, has the effect of leading to the comparisons of posterior probabilities of rival models that are conditional on samples of different sizes. Next section will present a model selection method to resolve this dilemma.

## 2. Model Folding

The approach we suggest consists of using the sample information to compute an estimate of the parameters for both the sampling model and the MDM assuming that the incomplete samples for the saturated model (in which $X_1, \ldots, X_v$ affect $Y$) are representative. Posterior probabilities of nested models are derived from the estimates of the parameters of the saturated model and of the MDM and hence we call the method *model folding* (MF).

Suppose that data are collected in an augmented contingency table, with rows corresponding to the possible combinations $\pi_k^{(v)}$ of categories of $X_1, ..., X_v$. Columns correspond to the $c$ categories of $Y$, and the $c+1$th column reports the frequencies of unclassified cases within each $\pi_x^{(v)}$. The unclassified cases will be denoted by $m_k^{(v)}$. The other frequencies will be denoted by $n_{kj}^{(v)}$ so that, in the cell $k, j$, $n_{kj}^{(v)}$ is the frequency of cases in the sample with $Y = j$ and $X_1, ..., X_v = \pi_k^{(v)}$. We further denote by $\alpha$ the overall prior precision, by $n$ the size of the complete sample and by $m$ the number of missing data on $Y$. We continue to assume that data are missing at random and that the MDM is ignorable for the saturated model. Thus, $p(R = 1|y_m, y_o, x, \psi)$ is a function of $\pi_k^{(v)}$. We shall denote $p(R = 1|\pi_k^{(v)})$ by $\psi_k^{(v)}$, and assume as prior distributions on $\psi_k^{(v)}$ Dirichlet distributions with hyper parameters $\beta_{k1}^{(v)}, \beta_{k2}^{(v)}$. The possible models of which we wish to evaluate the posterior probabilities are displayed in a lattice. The strategy we propose is top-down, and can be divided into $v + 1$ steps. Each step corresponds to the evaluation of posterior probabilities of models within a level of the lattice. We start from the saturated model, and estimate, for each $\pi_k^v$,

$$\hat{\theta}_{kj}^{(v)} = \frac{\alpha_{kj}^{(v)} + n_{kj}^{(v)}}{\alpha_{k\cdot}^{(v)} + n_{k\cdot}^{(v)}}.$$

These estimates can then be used to evaluate the posterior probability of the saturated model, conditional on the complete sample in which missing data are distributed as predicted values. Thus

$$p(M_v|y,x) \propto p(M_v) \prod_k \frac{\Gamma(\alpha_{k\cdot}^{(v)})}{\Gamma(\alpha_{k\cdot}^{(v)} + n_{k\cdot}^{(v)} + m_k^{(v)})} \prod_j \frac{\Gamma(\alpha_{kj}^{(v)} + n_{kj}^{(v)} + \hat{\theta}_{kj}^{(v)} m_k^{(v)})}{\Gamma(\alpha_{kj}^{(v)})}.$$

Next step consists of computing posterior probability of the $v$ models with $v-1$ explanatory variables. For each of these models, we need to estimate the probabilities of completions from the conditional probabilities $\theta_{kj}^{(v)}$ estimated in the saturated model, and from the estimates of $\psi$. Let $h$ be a combination of $v-1$ indices and consider the model $M_h$. We denote by $n_{kj}^{(h)}$ the frequency of cases with $\pi_k^{(h)}, Y = j$, and by $\alpha_{kj}^{(h)}$ the prior hyper-parameters of the distribution of $\theta_k^{(h)}$. In order to approximate the posterior probability of $M_h$, we need an estimate of the probabilities of completions $\phi_{kj}^{(h)} = p(y_r = j|\pi_k^{(h)}, y_o, M_h, \psi)$. These probabilities are estimated from the estimates of the parameters $\theta_{kj}^{(v)}$ and from the estimates of $\psi_{kj}$, by application of the Total Probability Theorem. Hence, up to a normalizing constant, we estimate $\phi_{kj}^{(h)}$ by $\hat{\phi}_{kj}^{(h)} \propto \sum_l (\alpha_{l\cdot}^{(v)} + n_{l\cdot}^{(v)} + m_l^{(v)}) \hat{\theta}_{lj}^{(v)} \hat{\psi}_{lj}$, where the summation is extended over all categories $\pi_{l,j}^{(v)}$ of the saturated model containing $\pi_{k,j}^{(h)}$, and the quantity $\hat{\psi}_{lj}$ is estimated as

$$\hat{\psi}_{lj} = \frac{\beta_{l1} + m_h^{(v)}}{\beta_{l\cdot} + n_{l\cdot}^{(v)} + m_l^{(v)}}.$$

The quantities $\hat{\phi}_{kj}^{(h)}$ are then used to distribute the incomplete cases across categories of $Y$ as $\hat{\phi}_{kj}^{(h)} m_k^{(h)}$, and hence an estimate of the posterior probability of $M_h$ is

$$p(M_h|y,x) \propto p(M_h) \prod_k \frac{\Gamma(\alpha_{k\cdot}^{(h)})}{\Gamma(\alpha_{k\cdot}^{(h)} + n_{k\cdot}^{(h)} + m_k^{(h)})} \prod_j \frac{\Gamma(\alpha_{kj}^{(h)} + n_{kj}^{(h)} + \hat{\phi}_{kj}^{(h)} m_k^{(h)})}{\Gamma(\alpha_{kj}^{(h)})}.$$

By repeating the same procedure for every model $M_h$, $h \in co(v, v-1)$, we estimate the posterior probabilities of the $v$ models with $v-1$ explanatory variables, and we can then move to evaluate posterior probabilities of model with $v-2$ explanatory variables and so forth. As we move down the lattice of models, the operation of marginalization and estimation of the probabilities of completions can be carried out using counts and estimates collected in the previous step. Note that, if the probability of $Y$ being missing is only a function of a subset of the explanatory variables, then the probabilities of the completions turn out to be the predictive probabilities, adjusted to take into account prior information about the MDM. In particular, if missing data are MCAR, then the effect of carrying on the MDM in the estimation steps become negligible, and the effect of MF is only to enlarge the incomplete samples so as to base the comparisons of posterior probabilities conditional on samples having the same size.

The intuition behind MF goes somehow forward sustaining the idea of using multiple imputation when missing data are MAR [1]. As far as we know, the fact that there are several ways of imputing missing data, according to the assumptions made about the MDM, has gone unnoticed. We shall define *Global imputation* a stochastic version of our approach: missing data are simulated from the saturated model using the predictive distribution conditional on the observed data, and the completed sample is used to evaluate the posterior probabilities of all possible models nested within the saturated one. The effect of using global imputation instead of MF would be to limit unwanted bias. This version of imputation differs from the imputation-based approach suggested for instance by [1], in which missing data are imputed from the predictive distributions conditional on different models. We shall name the latter *Local imputation.* Imputing missing data conditional on different models has a bias effect as shown in the next example.

## 3. A Simulation Study

Data below are a random sample generated from the saturated model $M_{12}$ and $p(Y = 1|(1,1)) = 0.5$, $p(Y = 1|(1,2)) = 0.2$, $p(Y = 1|(2,1)) = 0.7$, $p(Y = 1|(2,2)) = 0.3$.

| | $X_1, X_2$ | | | |
|---|---|---|---|---|
| $Y$ | 1,1 | 1,2 | 2,1 | 2,2 |
| 1 | 52 | 17 | 66 | 36 |
| 2 | 48 | 83 | 34 | 64 |

Four models $M_0$, $M_1$, $M_2$ and $M_{12}$ can be considered. Assuming uniform prior probabilities, their posterior probabilities (in log scale) are easily found to be: $\log p(M_0|y,x) \propto -275.1086$ , $\log p(M_1|y,x) \propto -255.4863$, $\log p(M_2|y,x) \propto -271.6517$, $\log p(M_{12}|y,x) \propto -252.9841$ and $M_{12}$ would be selected, conditional on the observed data. The complete sample was then subject to a random deletion of $Y$ entries with the following process. The values of $Y$ were associated with the values of a binary variable $R$ whose distributions are $p(R = 1|(1,1)) = 0.2$, $p(R = 1|(1,2)) = 0.3$, $p(R = 1|(2,1)) = 0.1$ and $p(R = 1|(2,2)) = 0.6$. The complete sample was then parsed, and for each case in the sample a value of $R$ was generated from the distribution of $R$ conditional on $X_1, X_2$. If the value of $R$ turned out to be 1, the entry of $Y$ was removed. Clearly, missing data are MAR, since the probability of $Y$ being missing is at most a function of the observed cases in the sample. This deletion process was repeated 100 times, and in each incomplete sample, posterior probabilities of the four models $M_0$, $M_1$, $M_2$ and $M_{12}$ were computed using MF, global and local imputation based on 10 imputed values for each missing entry, and disregarding missing data. Figure 1 reports the estimates of the posterior probabilities of $M_0, M_1, M_2, M_{12}$ that are denoted by 0,1,2,and 3. The estimates of the posterior probabilities computed with global and local imputation are obtained by evaluating the posterior probabilities from the completed samples and then by averaging

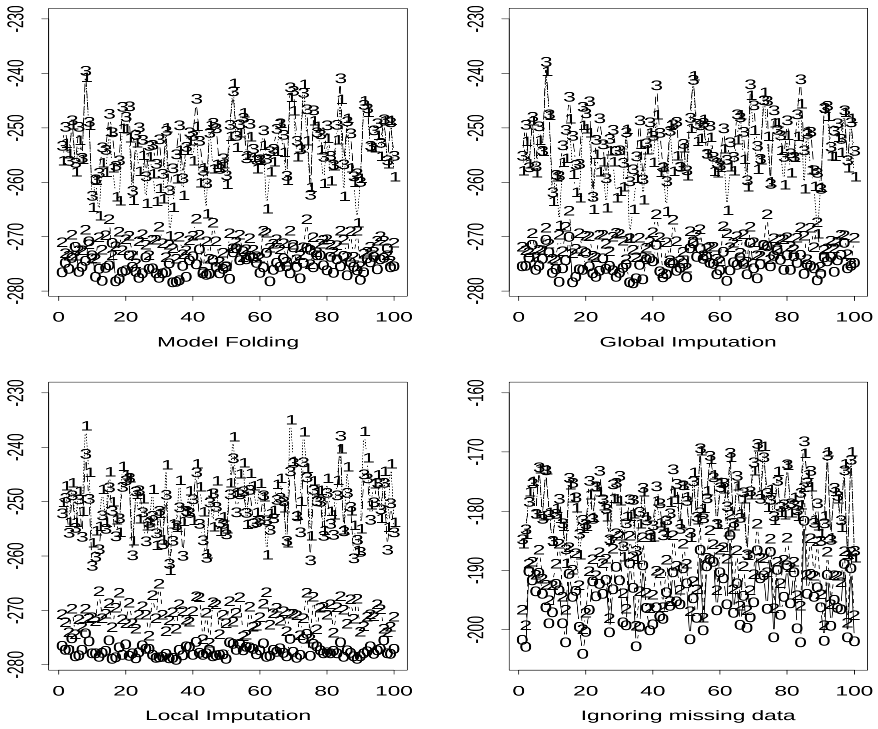

**Figure 1**: Posterior probabilities of $M_0$ (0), $M_1$ (1), $M_2$ (2) and $M_{12}$ (3) from 100 incomplete samples.

| | $\log p(M_0\|y,x)$ | $\log p(M_1\|y,x)$ | $\log p(M_2\|y,x)$ | $\log p(M_{12}\|y,x)$ |
|---|---|---|---|---|
| Model Folding | -275.2661 | -255.2350 | -271.6484 | -252.4831 |
| Global Imputation | -275.0549 | -255.2079 | -271.3363 | -252.0717 |
| Local Imputation | -277.4527 | -249.6824 | -270.9706 | -251.9317 |
| Data Deletion | -193.8340 | -178.3028 | -190.5554 | -177.6876 |

**Table 1**: Mean values of the posterior probabilities of $M_0, M_1, M_2, M_{12}$.

out the results. In the 100 incomplete samples, MF selects $M_1$ in 16 samples, $M_{12}$ in 84 samples. When global imputation is used, the correct model is selected from 15 samples. Thus the error rates of the two methods are equivalent. If local imputation is adopted, the error rate rises to 81%. Removing incomplete cases leads to selecting model $M_1$ from 36 incomplete samples, and model $M_{12}$ in the remaining 64. The error rate incurred by MF and global imputation is within sampling variability: in 100 complete samples, model $M_1$ was selected in 20% of cases.

Figure 1 reveals the reasons behind the large error rates incurred under the assumption of total ignorability, which is assumed by local imputation and data deletion. In the first two plots of Figure 1, there are two distinct patterns of points, the estimates of $p(M_0|y,x)$ and $p(M_2|y,x)$ in the lower part of the figure, and $p(M_1|y,x)$ and $p(M_{12}|y,x)$ in the top. These two patterns reproduce the ordering between the posterior probabilities computed from the complete samples when MF or global imputation are used and the accuracy is shown in the first two rows of Table 1.

However, when local imputation is used, and the MDM is assumed to be totally ignorable, there is an evident bias in the estimates. The estimates of $p(M_2|y,x)$ are almost all above the estimates of $p(M_0|y,x)$. A similar result is for the estimates of $p(M_1|y,x)$ that are almost all greater than the estimates of $p(M_{12}|y,x)$. If missing data are simply ignored there is no longer an evident distinction between the posterior probabilities of the four models. Hence, the assumption of total ignorability coupled with imputation has the effect of maximizing the bias that would be incurred in simply disregarding missing data. This result confirms the conjecture put forward in the previous section. This bias is clear from the summary statistics reported in the second two rows of Table 1.

## References

[1] R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data.* Wiley, New York, NY, 1987.

[2] M. Ramoni and P. Sebastiani. Model selection and model averaging with missing data. Technical Report KMi-TR-63, Knowledge Media Institute, The Open University, 1998.

[3] D.B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

[4] S. S. Wilks. *Mathematical Statistics.* Wiley, New York, 1963.