KNOWLEDGE MEDIA INSTITUTE

Model Selection and Model Averaging with Missing Data

Marco Ramoni and Paola Sebastiani

KMI-TR-65

July 1998

Model Selection and Model Averaging with Missing Data

Marco Ramoni
Knowledge Media Institute
The Open University

Paola Sebastiani
Department of Actuarial Science and Statistics
City University

Abstract

Missing data can impair the reliability of statistical inference as they may affect the representativity of the sample. Nonetheless, under some conditions guaranteeing that the missing data mechanism is ignorable, reliable conclusions can be still drawn from the incomplete sample. Ignorability conditions are well-understood for parameter estimation but when the inference task involves the computation of the posterior probability of a data model, as required by Bayesian model selection and prediction through model averaging, these conditions are not sufficient. This paper defines new ignorability conditions for model selection and model averaging from incomplete data.

Keywords: Bayesian Learning, Missing Data, Model Selection, Model Averaging, Ignorability.

Reference: KMi Technical Report KMi-TR-65, July 1998.

Address: Marco Ramoni, Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom MK7 6AA. PHONE: +44 (1908) 655721, FAX: +44 (1908) 653169, EMAIL: m.ramoni@open.ac.uk, URL: http://kmi.open.ac.uk/people/marco.

1. Introduction

Missing data can impair the reliability of statistical inference as they may affect the representativity of the sample. When the process yielding missing data is not ignorable, as it sometimes happens in the case of drop-outs from follow-up studies or non responses in sample surveys, the available data may not contain sufficient information to restore the representativity of the sample and external information is required about the missing data mechanism (MDM). This information can be notably large, quite difficult to encode and, sometimes, not readily available. Under some conditions, nonetheless, knowledge of the MDM is not necessary and reliable conclusions can still be drawn from the available data. In these cases, the MDM is said to be *ignorable*.

Ignorability conditions are well-understood for parameter estimation and they have been established by Rubin [3]. However, this paper will show that, when the inference task involves the computation of the posterior probability of a data model, as required by Bayesian model selection and prediction through model averaging, these conditions are not sufficient to guarantee the ignorability of the MDM.

This paper will provide new ignorability conditions for model selection and model averaging. These conditions will introduce a classification of different ignorable MDMs and the paper will provide examples and simulation studies to show the impact of these assumptions on the robustness of model selection and model averaging.

2. Ignorability Conditions

Formally, ignorability of the MDM allows us to simplify the calculation of the likelihood function by dropping the parameters encoding the MDM. Let x be a multivariate $n \times v$ data set, with n cases observed on v variables $X = (X_1, \ldots, X_v)$. Variables can be either continuous or discrete but, for simplicity of notation, this paper will assume that they are continuous. Let $p(x|\theta)$ be the likelihood function and let the prior density of θ be $p(\theta)$.

With complete data, Bayesian inference on θ is based on the posterior density $p(\theta|x) \propto p(\theta)p(x|\theta)$. Suppose now that the data set is incomplete, i.e. some values of $X_1, \ldots, X_v$ are recorded as unknown and they will be denoted as $X_i = ?$. We can decompose x into x_o, x_m, where x_o denotes the observed values, and x_m denotes the subset of x with unknown values. In doing so, we define a variable X_m taking values x_r that are the possible realizations of the unobserved variables in the sample.

According to Rubin's definition [3], the *ignorability* of the MDM implies that inference on θ can be based on the posterior density $p(\theta|x_o) \propto p(\theta)p(x_o|\theta)q(x_o, \theta)$, where $p(x_o|\theta)q(x_o, \theta) = p(x_o|\theta) \int_{x_r} p(x_r|\theta, x_o)dx_r$ and $q(x_o, \theta)$ is an adjusting factor. Thus, missing data are integrated out regardless of the MDM. The ignorability of the MDM requires two conditions to be fulfilled. The first condition states that unreported data are *Missing at Random* (MAR). This condition is formally characterized by associating each variable X_i to

a binary variable R_i. For each case in the sample, the variable R_i takes value 1 if X_i is observed and 0 otherwise. Let r be the observed values of $R_1, ..., R_v$ in the sample and denote the joint probability of r by $p(r|x, \psi)$. Missing data are said to be MAR if $R = (R_1, ..., R_v)$ is independent of X_m, given x_o and ψ:

$$R \perp X_m | x_o, \psi. \tag{1}$$

The second condition requires the independence of the parameters θ and ψ

$$\theta \perp \psi \tag{2}$$

so that the prior density simplifies into $p(\theta, \psi) = p(\theta)p(\psi)$. Under the assumptions (1) and (2) the posterior density of θ, ψ factorizes into the product

$$p(\theta, \psi | x_o, r) \propto [p(\theta)p(x_o|\theta)q(x_o, \theta)][p(\psi)p(r|x_o, \psi)]$$

and inference about θ is based on the marginal density $p(\theta|x_o) \propto p(\theta)p(x_o|\theta)q(x_o, \theta)$ that is not a function of r, so that the MDM becomes ignorable. If $q(x_o, \theta) = 1$, then missing data themselves are ignorable and can be discarded from the analysis. When this simplification is not possible, the MDM is not ignorable.

However, we note that a further assumption is made throughout all these calculations, namely that

$$X_m \perp \psi | x_o, \theta. \tag{3}$$

The conditional independence of X_m and ψ is equivalent to assuming that the MDM is *not informative* with respect to the missing data so that the probabilities of the completions x_r are simply the predictive probabilities given the observed values in the sample. The importance of making this last assumption explicit becomes even clearer when we move from the task of estimating the parameters θ to the task of computing the posterior probability of a model m for the variables X. In this case, the MDM can be either ignorable with respect to all possible models or with respect to just few of them. We shall call the former case *total ignorability* and the latter *partial ignorability*. The posterior probability of m that accounts for the MDM is

$$p(m, \psi | x_o, r) \propto p(m, \psi)p(x_o, r|m, \psi), \tag{4}$$

and the MDM is ignorable if, by integrating ψ out, we obtain a quantity that is not a function of r, say $p(m|x_o) \propto p(m)p(x_o|m)q(x_o, m)$ where, $q(x_o, m)$ is an adjusting factor. In order to characterize the ignorability of the MDM for model selection, it is therefore sufficient to identify under which conditions we have this simplification.

We shall show in the paper that total ignorability of the MDM holds if $R \perp X_m | x_o, \psi$, $M \perp \psi$, and $X_m \perp \psi | x_o, M$, where M is the variable representing the possible models. In

	X_1			
X_2	1	2	?	Total
1	$n(x_{11}\|x_{21})_c$	$n(x_{12}\|x_{21})_c$	$n(?\|x_{21})$	$n(x_{21})$
2	$n(x_{11}\|x_{22})_c$	$n(x_{12}\|x_{22})_c$	$n(?\|x_{22})$	$n(x_{22})$
Total	$n(x_{11})_c$	$n(x_{12})_c$	$n(X_1 =?)$	n

Table 1: The representation of the incomplete contingency table used in the example.

this case, inference on M can be based on $p(m)p(x_o|m)q(x_o, m)$. Partial ignorability holds when we require that $p(x_r|x_o, m, \psi) = p(x_r|x_o, m)$ holds for some model m or, in words, that the MDM is non informative only for some model. This paper will show that enforcing total ignorability can yield inconsistencies and introduce a severe bias in model selection and prediction based either on the best model selected or model averaging. Here, we give only one example. The full paper will provide detailed simulation studies using both exact and imputation-based techniques.

3. An Example

Let X_1 and X_2 be binary variables and consider two possible directed graphical models: *model* m_0 specifies that the two variables are independent and, conditional on m_0, we can parameterize $p(x_{11}|\theta^{(0)}) = \theta_{11}$ and $p(x_{21}|\theta^{(0)}) = \theta_{21}$ where $\theta^{(0)}$ is the parameter vector associated to m_0; *model* m_1 specifies that X_2 is a parent of X_1, so that the induced parameterization is $p(x_{21}|\theta^{(1)}) = \theta_{21}$ and $p(x_1|x_{2j}, \theta^{(1)}) = \theta_{j11}$. Suppose that the complete sample is subject to some random process which deletes part of the entries of the variable X_1 but we continue to have a complete sample for X_2. We model the MDM by associating, with each case in the sample, the value of a binary variable R_2 whose distribution is a function of X_2. Let $p(R_2 = 1|X_2 = i) = \psi_i$, $i = 1, 2$, and $\psi = (\psi_1, \psi_2) \perp \theta^{(m)}|m$. For each case in the sample, we generate a value for $R_2|X_2$ and remove the entry of X_1 if $R_2 = 1$. After the deletion process, the incomplete counts are summarized into Table 1. Since the values of X_2 are always observed and the probability that a value of X_1 is deleted is only dependent on X_2, missing data are MAR.

Let $n(x_{2j})_c$ be $n(x_{11}|x_{2j})_c + n(x_{12}|x_{2j})_c$, $j = 1, 2$. Thus, $n(x_{2j})_c$ is the frequency of cases with both X_1 and X_2 observed. The total number of complete cases is then $n_c = n(x_{21})_c + n(x_{22})_c$.

Denote by $B(\alpha_1, \alpha_2)$ a Beta distribution [1]. Suppose that, given m_0, $\theta_2 \equiv (\theta_{21}, \theta_{22}) \sim B(2, 2)$ and $\theta_1 \equiv (\theta_{11}, \theta_{12}) \sim B(2, 2)$, and they are independent. Given m_1, we assume that $\theta_{j1} \equiv (\theta_{j11}, \theta_{j12}) \sim B(1, 1)$, and they are independent. Thus, a priori, the marginal probabilities of x_{2j}, x_{1k} and $x_{1k}|x_{2j}$ are all uniform and are based on a total prior precision

4. Under the assumption that the MDM is ignorable for computing $p(m_1|x)$, [4] show that the marginal likelihood $p(x_o|m_1)$ can be computed exactly and is:

$$p(x_o|m_1) = \prod_{j=1}^{2} \frac{\Gamma(4)\Gamma(2+n(x_{2j}))}{\Gamma(\alpha+n)\Gamma(2)} \prod_{k=1}^{2} \frac{\Gamma(2)\Gamma(1+n(x_{1k}|x_{2j})_c)}{\Gamma(2+n(x_{2j})_c)\Gamma(1)},$$

and since $q(x_o, m_1) = 1$ missing entries are simply ignored. Consider now the model of independence. If we also assume that the MDM is ignorable for computing $p(m_0|x)$, then we can show that

$$p(x_o|m_o) \propto \prod_{j=1}^{2} \frac{\Gamma(4)\Gamma(2+n(x_{2j}))}{\Gamma(4+n)\Gamma(2)} \prod_{k=1}^{2} \frac{\Gamma(4)\Gamma(2+n(x_{1k})_c)}{\Gamma(4+n_c)\Gamma(2)}.$$

Thus, the assumption that the MDM is totally ignorable implies that missing data are simply ignored and the incomplete sample on X_1 is treated as if it were representative.

However, if we do not consider the sampling variability, the expected frequency of cases removed under the MDM considered are $\psi_j n(x_{1k}|x_{2j})$, where $n(x_{1k}|x_{2j})$ are the counts of the complete sample. The relationship between the complete counts and those in Table 1 is therefore $n(x_{1k}|x_{2j}) = n(x_{1k}|x_{2j})_c + \psi_j n(x_{1k}|x_{2j})$. The marginal counts observed on X_1 are in the relationship $n(x_{1k}) = n(x_{1k})_c + \psi_1 n(x_{1k}|x_{21}) + \psi_2 n(x_{1k}|x_{22})$. The proportions in each conditional distribution are maintained, since $n(x_{11}|x_{2j})_c/n(x_{12}|x_{2j})_c = n(x_{11}|x_{2j})/n(x_{12}|x_{2j})$ and hence the incomplete samples within parent configurations are representative. However, the marginal frequencies of X_1 are no longer in the same original proportion, since:

$$\frac{n(x_{11})_c}{n(x_{12})_c} = \frac{(1-\psi_1)n(x_{11}|x_{21}) + (1-\psi_2)n(x_{11}|x_{22})}{(1-\psi_1)n(x_{12}|x_{21}) + (1-\psi_2)n(x_{12}|x_{22})} = \frac{n(x_{11})}{n(x_{12})}$$

if and only if either $\psi_1 = \psi_2$, or $n(x_{11}|x_{21})n(x_{12}|x_{22}) = (x_{11}|x_{22})n(x_{12}|x_{21})$. Thus, the marginal counts on X_1 are not a representative sample, and enforcing total ignorability of the MDM will have the consequence of introducing bias in the model selection process as shown in the next simulation study.

4. A Simulation Study

Data in Table 2 is a random sample generated from the model that assumes a dependence between X_1 and X_2 and $p(X_1 = 1|X_2 = 1) = 0.57$ and $p(X_1 = 1|X_2 = 2) = 0.50$. We assume that $p(m_0) = p(m_1)$, together with the parameterizations described above, with $\alpha = 4$. It is easily shown that the Bayes factor is $p(x|m_1)/p(x|m_0) = 3.45$ so that, model m_1 would be selected from the complete sample.

Suppose now that some of the occurrence of X_1 are deleted with the same process described in the previous section. We choose $\psi_1 = 0.6$ and $\psi_2 = 0.1$, so that the expected

	X_1		
X_2	1	2	Total
1	577	423	1000
2	510	490	1000
Total	1087	917	2000

Table 2: Data simulated from a model of association with $p(X_1 = 1|X_2 = 1) = 0.57$ and $p(X_1 = 1|X_2 = 2) = 0.50$.

number of cases that would be removed is 700. In 400 repetitions of this deletion process, in which in each incomplete sample the posterior probability of m_1 and m_0 were computed assuming h total ignorability of the MDM, in 37% of cases only the correct model m_1 was chosen. Thus, a random model selection can be better than doing the exact inference! The wrong number of selection is well below what we would expect by considering the sampling variability. In 400 complete samples generated from the model described above, the Bayes factor was greater than 1 in 74 % of cases. Thus, the large error rate under the MAR assumption coupled with the ignorability of the MDM can be due to the incorrect use of the sample information.

In this example, assuming total ignorability has the effect of simply ignoring missing data. In similar cases [2] suggest using multiple imputation as the panacea. Multiple imputation would consist of estimating the posterior probability of each model by imputing missing data conditional on that model. The results of the simulation study reported in the final paper will show that this strategy is nothing more than a "placebo".

5. Conclusions

The example described show the bias effect of enforcing total ignorability of the MDM for model selection. If we assume only partial ignorability, we can suppose that the MDM is ignorable for one particular model and not ignorable for other models where the non ignorability is a consequence of assuming the MDM informative for the computation of $p(x_r|x_o, m, \psi)$. Thus, in the examples shown in this paper, missing data would not be disregarded in computing for instance $p(x_r|x_o, m_0, \psi)$. However, a consequence of this approach would be to compare the posterior probabilities of models conditional on different sample sizes. [5] have shown how to overcome this problems for incomplete sample in which only the response variable is subject to non response. A general solution seems to be an open problem.

References

[1] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory.* Wiley, New York, 1994.

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

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

[4] P. Sebastiani and M. Ramoni. Bayesian inference from data subject to non response using bound and collapse. Technical Report KMi-TR-48, Knowledge Media Institute, The Open University, 1997. Available at http://kmi.open.ac.uk/techreports/KMi-TR-48.

[5] P. Sebastiani and M. Ramoni. Model folding for data subject to nonresponse. Technical Report KMi-TR-64, Knowledge Media Institute, The Open University, 1998.