Bayesian Selection of Decomposable Models With Incomplete Data

Paola SEBASTIANI and Marco RAMONI

This article describes a new approach to Bayesian selection of decomposable models with incomplete data. This approach requires the characterization of new ignorability conditions for the missing-data mechanism and the development of new computational methods. Both issues are considered, and solutions are proposed. Theory and methods are assessed in controlled experiments and in the analysis of one real-life incomplete dataset.

KEY WORDS: Bayes factor; Conditional independence; Decomposable model; Ignorability; Missing data; Model selection.

1. INTRODUCTION

A popular approach to Bayesian model selection uses data to update the prior distribution on the set of models into the posterior distribution, and chooses the model with maximum posterior probability (Bernardo and Smith 1994). The key quantity to the "prior to posterior" updating is the *marginal likelihood* of each model, that is, the likelihood function in which the model parameters are averaged out. When the models are decomposable, the adoption of hyper-Markov distributions on the model parameters and of decomposable model prior probabilities reduces the overall model selection process to the search for components with maximum posterior probability. In particular, computation of the marginal likelihood of each component can be done independently. When data are only partially observed, the problem is whether the same model selection process can be applied by ignoring the missing-data mechanism. We give a solution in this situation:

1. We suppose that the model we look for can be decomposed in two components, m_{x_s} and $m_{y|x_c}$, where m_{x_s} is a dependency model for the variable set $\mathcal{X}_s$ and $m_{y|x_c}$ models the dependence of Y on the variables $\mathcal{X}_c$ in $\mathcal{X}_s$. The model $(m_{x_s}, m_{y|x_c})$ encodes the conditional independence of Y and $\mathcal{X}_s$, given $\mathcal{X}_c$.
2. We suppose that Y is partially observed and the probability that a value of Y is missing is a function of the variables $\mathcal{X}_d$ in $\mathcal{X}_s$.

This missing-data mechanism produces data that are *missing at random* MAR and, in particular, *missing completely at random* MCAR when $\mathcal{X}_d$ is empty (Rubin 1976). It is also a special case of the coarsening model introduced by Heitjan and Rubin (1991). In the received theory of missing data, this mechanism is ignorable for Bayesian inference a under the assumption of prior independence of the model parameters and of the missing-data mechanism (Heitjan 1994; Heitjan and Basu 1996; Little and Rubin 1987).

Our first goal is to find weak, sufficient conditions to ensure that this missing-data mechanism is ignorable when selecting a decomposable model $(m_{x_s}, m_{y|x_c})$. We show that the prior independence of m_{x_s} and of the model describing the missing-data mechanism is sufficient to ensure that the selection of m_{x_s} and $m_{y|x_c}$ can be done via two independent processes, and that the selection of m_{x_s} can always ignore the missing-data mechanism. Furthermore, when the probability of deleting the Y values is a function of $\mathcal{X}_d$, we show that the selection of $m_{y|x_c}$ cannot ignore the missing-data mechanism. Only when $\mathcal{X}_d$ is empty is the missing-data mechanism ignorable. The fact that mechanisms producing MAR data may not be ignorable for model selection is not new; Little and Rubin (1987, p. 16) and Rubin (1996) warned against improper modeling when data are MAR rather than MCAR. However, we believe that the formal treatment of the problem that we provide here is new.

The nonignorability of the missing-data mechanism considered in this article is conceptually different from that of mechanisms in which the missingness probability is a function of the unobserved data. In our framework, the probability of deleting the Y values is a function of the variables $\mathcal{X}_d$, which are fully observed. In fact, we can show that the incomplete sample contains the information needed to proceed with proper model selection. To characterize this conceptual difference, we define a missing-data mechanism to be *partially ignorable* when the probability of removing the Y values is a function of the variables $\mathcal{X}_d \subseteq \mathcal{X}_s$. An ignorable missing-data mechanism is a special case of this, in which the set $\mathcal{X}_d$ is empty. Distinguishing between an ignorable and a partially ignorable missing-data mechanism is the first step to ensuring proper selection of $m_{y|x_c}$, and we provide a Bayesian solution to this problem. Our approach builds on suggestions of Little and Rubin (1987) and Baker and Laird (1988), although we only try to discriminate between ignorable and partially ignorable missing-data mechanisms.

When the missing-data mechanism is ignorable, selection of the model component $m_{y|x_c}$ can disregard the missing values. When, on the other hand, the missing-data mechanism is only partially ignorable, exact inference appears to be intractable, and we need approximate methods. A well-established approximate solution to inference with incomplete data is imputation (Rubin 1987, 1996; Schafer 1997), which replaces the missing data with values generated by an imputation model. We

Paola Sebastiani is Assistant Professor, Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003 (E-mail: *sebas@math.umass.edu*). Marco Ramoni is Instructor, Children's Hospital Informatics Program, Harvard University Medical School, Boston, MA 02115 (E-mail: *marco_ramoni@harvard.edu*). This research was partially supported by the ESPRIT programme of the Commission of the European Community under Contract EP29105. The authors thank two anonymous referees, the associate editor, and the editor for their useful comments that helped improve the original manuscript.

show that our framework leads to a natural choice of the imputation model, which, for example, disagrees with *Bayesianly proper multiple imputation* (Schafer 1997). The scheme that we suggest, which we term *ignorable imputation*, accounts for the missing-data mechanism and produces, asymptotically, a proper imputation model as defined by Rubin (1987). We also propose a deterministic method to approximate the exact marginal likelihood that we call *model folding*. Preliminary experiments show that model folding reaches high accuracy at a low computational cost, because the complexity of the model search is not affected by the presence of incomplete cases. Both ignorable imputation and model folding reconstruct a completion of the incomplete data by taking into account the variables $\mathcal{X}_d$ responsible for the missing data. This property is in agreement with the suggestion put forward by Rubin (1976), Little and Rubin (1987), and Heitjan and Rubin (1991) that the variables responsible for the missing data should be kept in the model. However, our approach allows us to also evaluate the likelihoods of models, which do not depend explicitly on these variables.

The structure of the article is as follows. Section 2 provides the necessary background on Bayesian selection of decomposable models with complete data. Section 3 describes ignorability conditions and characterizes ignorable and partially ignorable missing-data mechanisms. Section 4 describes a Bayesian method for discriminating between ignorable and partially ignorable missing-data mechanisms. Section 5 puts together the results of Sections 3 and 4 to describe proper imputation schemes and model folding. Section 6 gives experimental evaluations, and Section 7 discusses an example.

2. MODEL SELECTION WITH COMPLETE DATA

Let (x_s, y) be data generated from variables $\mathcal{X}_s = (X_1, \ldots, X_s)$ and Y via some unknown dependency model m. We suppose that m belongs to the set $\mathcal{M}$ of decomposable models (Lauritzen 1996) describing different types of stochastic dependence between the variables $\mathcal{X}_s$, and independence of Y from $\mathcal{X}_s$, given $\mathcal{X}_c$. In this way m has the two components, m_{x_s} and $m_{y|x_c}$ describing the association among the variables $\mathcal{X}_s$ and the dependence of Y on $\mathcal{X}_c$, for different c. In particular, when we condition on $\mathcal{X}_s$, $m_{y|x_c}$ becomes a regression model for the response variable Y.

By defining $\mathcal{M}_{x_s}$ and $\mathcal{M}_{y|x_s}$ to be the set of models m_{x_s} and $m_{y|x_c}$, $\mathcal{M}$ is the Cartesian product of the two sets $\mathcal{M}_{x_s}$ and $\mathcal{M}_{y|x_s}$, say $\mathcal{M} = \mathcal{M}_{x_s} \otimes \mathcal{M}_{y|x_s}$.

Example 1. Suppose that $\mathcal{X}_s$ is the bivariate variable (X_1, X_2). There are only two possible models describing the stochastic dependence between X_1 and X_2: either X_1 and X_2 are independent or not. The set $\mathcal{M}_{x_s}$ comprises these two models. The set $\mathcal{M}_{y|x_s}$ comprises four models: $m_{y|\varnothing}$ specifies the independence of Y and X_1 and X_2; $m_{y|x_1}$ and $m_{y|x_2}$ specify the dependence of Y on X_1 and the dependence of Y on X_2; and $m_{y|x_1,x_2}$ specifies the dependence of Y on both X_1 and X_2. Clearly, any dependency model for the variables X_1, X_2, and Y is characterized by a pair m_x and $m_{y|x_c}$.

Given the modularity of the models in $\mathcal{M}$, we adopt decomposable prior probabilities (Heckerman, Geiger, and Chickering 1995) on the model space, so that

$$p(m) = p(m_{x_s})p(m_{y|x_c}) \tag{1}$$

for any m in $\mathcal{M}$. Associated with each model m is a vector of parameters θ with prior density $p(\theta|m)$. We suppose that θ consists of the two components θ_{x_s} and $\theta_{y|x_c}$, parameterizing m_{x_s} and $m_{y|x_c}$, and that, given $\theta_{x_s}, \theta_{y|x_c}$, the joint density (probability) of $\mathcal{X}_s$ and Y is

$$p(x_s, y|\theta) = p(x_s|\theta_{x_s})p(y|x_c, \theta_{y|x_c}). \tag{2}$$

We also assume that θ_{x_s} and $\theta_{y|x_c}$ are independent, given m, so that $p(\theta|m)$ factorizes as

$$p(\theta|m) = p(\theta_{x_s}|m_{x_s})p(\theta_{y|x_c}|m_{y|x_c}). \tag{3}$$

Given data (x_s, y), we wish to select the maximum posterior probability model m from the set $\mathcal{M}$. By Bayes's theorem, the posterior probability $p(m|x_s, y)$ is

$$p(m|x_s, y) = \frac{p(m)p(x_s, y|m)}{p(x_s, y)}. \tag{4}$$

The *marginal likelihood* $p(x_s, y|m)$ is computed by averaging out θ from the augmented likelihood $p(x_s, y, \theta|m) = p(x_s, y|\theta)p(\theta|m)$, and it solves the integral $\int_\theta p(x_s, y|\theta) \times p(\theta|m)d\theta = \int_\theta p(x_s|\theta_{x_s})p(y|x_c, \theta_{y|x_c})p(\theta|m)d\theta$. By the assumption of parameter independence in (3), it is easy to show that the marginal likelihood becomes

$$p(x_s, y|m) = p(x_s|m_{x_s})p(y|x_c, m_{y|x_c}). \tag{5}$$

The quantity $p(x_s|m_{x_s})$ is a function of $\mathcal{X}_s$, whereas $p(y|x_c, m_{y|x_c})$ is a function of only Y and $\mathcal{X}_c$. If we now use this factorization of the marginal likelihood and the factorization of the prior probability of a model m in (1), then we can write the posterior probability in (4) as

$$p(m|x_s, y) = \frac{[p(m_{x_s})p(x_s|m_{x_s})] \times [p(m_{y|x_c})p(y|x_c, m_{y|x_c})]}{p(x_s, y)}. \tag{6}$$

This factorization has an important consequence for the model selection process. Because this process searches for the model m with maximum posterior probability $p(m|x_s, y)$, and $p(x_s, y)$ is constant, the solution can be found by searching for the models m_{x_s} and $m_{y|x_c}$ with maximum $p(m_{x_s})p(x_s|m_{x_s})$ and $p(m_{y|x_c})p(y|x_c, m_{y|x_c})$ in the two model spaces $\mathcal{M}_{x_s}$ and $\mathcal{M}_{y|x_s}$. In particular, the variables in $\mathcal{X}_s$ not in $\mathcal{X}_c$, say $\mathcal{X}_s \backslash \mathcal{X}_c$, are irrelevant to computing the marginal likelihood of $m_{y|x_c}$. This result is based on the following assumptions:

- Assumption 1: The prior probability of each model $(m_{x_s}, m_{y|x_c})$ is decomposable.
- Assumption 2: Associated with each model $m = (m_{x_s}, m_{y|x_c})$ is a vector of parameters $\theta = (\theta_{x_s}, \theta_{y|x_c})$, with prior density $p(\theta|m) = p(\theta_{x_s}|m)p(\theta_{y|x_s}|m)$.
- Assumption 3: Conditional on $\theta_{y|x_c}$ and on $\mathcal{X}_c$, Y is independent of $\mathcal{X}_s \backslash \mathcal{X}_c$.
- Assumption 4: The dataset (x_s, y) is complete in the sense that there are not unknown entries.

In the next section we disregard assumption 4 about the absence of unknown entries in the data and identify conditions to ensure that, given the same assumptions 1–3, model selection can proceed by implementing two independent searches in the two model spaces $\mathcal{M}_{x_s}$ and $\mathcal{M}_{y|x_s}$, and the variables $\mathcal{X}_s \backslash \mathcal{X}_c$ continue to be irrelevant for evaluating the marginal likelihood of $m_{y|x_c}$. We conclude this section with an example that we use in the remainder of the article.

Example 2. Suppose that $\mathcal{X}_s$ and Y are categorical variables, and consider the selection of the maximum posterior probability model $m_{y|x_c}$. We assume that, conditional on $m_{y|x_c}$, the distribution of Y is a product of q_c multinomial distributions, where q_c is the number of states x_{ck} of $\mathcal{X}_c$, and $k = 1, \ldots, q_c$. We let θ^c_{kj} denote $p(Y = y_j | x_{ck}, \theta_{y|x_c})$, for $j = 1, \ldots, g$ and for parameters $\theta_{y|x_c} \equiv (\theta^c_k)$, and $\theta^c_k = (\theta^c_{kj})$. We further suppose that for fixed k, the prior distribution of θ^c_k is a Dirichlet $D(\alpha^c_{k1}, \ldots, \alpha^c_{kg})$, and $\theta^c_k \perp \theta^c_h$, for $k \neq h$. The joint distribution of $\theta_{y|x_c}$ is then a product of independent Dirichlet and was termed a hyper-Dirichlet by Dawid and Lauritzen (1993). Because we are considering different models, we assume that as c varies, the hyperparameters $\alpha^c_{k1}, \ldots, \alpha^c_{kg}$ determine the same distribution for the parameters associated with the marginal distribution of Y. Thus we impose the constraint that $\sum_k \alpha^c_{kj} = \sum_k \alpha^{c'}_{kj}$ for any j and $c \neq c'$, which is satisfied when $\alpha^c_{kj} = \alpha/(q_c \times g)$, where $\alpha = \sum_{kj} \alpha^s_{kj}$ is the overall prior precision. Because Dirichlet distributions with $\alpha^c_{kj} = \alpha/(q_c \times g)$ are known as symmetric Dirichlet (Good 1968), we define a hyper-Dirichlet with $\alpha^c_{kj} = \alpha/(q_c \times g)$ as a symmetric hyper-Dirichlet, and as a symmetric hyperbeta when $g = 2$.

When the sample is complete, parameter independence and the use of hyper-Dirichlet distributions let us find explicitly $p(y|x_s, m_{y|x_c})$ as

$$p(y|x_c, m_{y|x_c}) = \prod_k \frac{\Gamma(\alpha^c_{k\cdot})}{\Gamma(\alpha^c_{k\cdot} + n^c_{k\cdot})} \prod_j \frac{\Gamma(\alpha^c_{kj} + n^c_{kj})}{\Gamma(\alpha^c_{kj})}, \qquad (7)$$

where $\Gamma(\cdot)$ is the gamma function, n^c_{kj} is the sample frequency of cases with categories (x_{ck}, y_j) under model $m_{y|x_c}$, $n^c_{k\cdot} = \sum_j n^c_{kj}$, and $\alpha^c_{k\cdot} = \sum_j \alpha^c_{kj}$. Details of the calculations have been given by Kass and Raftery (1995), Raftery (1995), and Heckerman et al. (1995).

3. IGNORABLE MISSING-DATA MECHANISMS

In this section we suppose that some entries of Y are unknown so that the dataset (x_s, y_o) is incomplete. We denote by y_o the observed values of Y and suppose that the incomplete sample is generated according to this scheme. First, a complete sample (x_s, y) of size n is generated by an unknown model m in the set $\mathcal{M}$. We then suppose that there exists a binary variable R, taking on one of two values 1 and 0, with probability $\psi(x_d)$ and $1 - \psi(x_d)$, for any value x_d of $\mathcal{X}_d \subseteq \mathcal{X}_s$. We define $\psi_{r|x_d} = (\psi(x_d))$. The variable R is used to remove entries of y as follows. Given $\psi_{r|x_d}$ and the sample x_s, a sample of n values r is generated from the probability distribution $p(R|x_d, \psi_{r|x_d})$. By definition, the r values are generated independently of y, given x_d and $\psi_{r|x_d}$. Given the r values, the data y are subjected to a deletion process, which removes the entries of Y corresponding to $r_i = 1$, for each sample case i. Now we label the subset of y values that survived the deletion process and the subset of y values removed from the sample by y_o and y_m. We refer to R and $\psi_{r|x_d}$ as the missing-data mechanism and y_m as nonresponse.

Given data (x_s, y_o, r), we wish to select a model $m = (m_{x_s}, m_{y|x_c})$ from the set $\mathcal{M} = \mathcal{M}_{x_s} \otimes \mathcal{M}_{y|x_s}$, and each model $m_{y|x_c}$ in the set $\mathcal{M}_{y|x_s}$ describes the overall dependence of y_o and y_m on $\mathcal{X}_s$. We continue to hold assumptions 1–3 of Section 2 and wish to find conditions on the missing-data mechanism such that model selection can proceed as in Section 2; that is, by means of two independent searches in the model spaces $\mathcal{M}_{x_s}$ and $\mathcal{M}_{y|x_s}$ and by disregarding $\mathcal{X}_s \backslash \mathcal{X}_c$ when computing the marginal likelihood of $m_{y|x_c}$. Formally, this is equivalent to finding conditions such that the posterior probability

$$p(m|x_s, y_o, r, \psi_{r|x_d}) = \frac{p(m_{x_s}, m_{y|x_c}|\psi_{r|x_d}) p(x_s|m_{x_s}, m_{y|x_c}, \psi_{r|x_d}) p(y_o, r|x_s, m_{x_s}, m_{y|x_c}, \psi_{r|x_d})}{p(x_s, y_o, r|\psi_{r|x_d})}$$

simplifies to

$$p(m|x_s, y_o, r, \psi_{r|x_d}) = \frac{[p(m_{x_s}) p(x_s|m_{x_s})] \times [p(m_{y|x_c}) p(y_o|x_c, m_{y|x_c})]}{p(x_s, y_o)}, \qquad (8)$$

and thus is independent of r and $\psi_{r|x_d}$. This factorization being true, the search for the maximum posterior probability model m is the same search described in Section 2, independently of r and $\psi_{r|x_d}$. By adopting Rubin's definition (Rubin 1976) we can say that in this case, the missing-data mechanism is *ignorable* for inference about m.

We first note that, conditional on m_{x_s}, the sample x_s is independent of $m_{y|x_c}$ and $\psi_{r|x_d}$ and, given x_s, the sample y_o, r is independent of m_{x_s}. Hence the marginal likelihood $p(x_s|m_{x_s}, m_{y|x_c}, \psi_{r|x_d})$ is $p(x_s|m_{x_s})$, whereas $p(y_o, r|x_s, m_{x_s}, m_{y|x_c}, \psi_{r|x_d})$ is $p(y_o, r|x_s, m_{y|x_c}, \psi_{r|x_d})$. If we assume that, a priori, $p(m_{x_s}, m_{y|x_c}|\psi_{r|x_d}) = p(m_{x_s})\, p(m_{y|x_c})$, then

$$p(m|x_s, y_o, r, \psi_{r|x_d}) \propto [p(m_{x_s}) p(x_s|m_{x_s})] p(m_{y|x_c}) p(y_o, r|x_s, m_{y|x_c}, \psi_{r|x_d}),$$

so that the search for the maximum posterior probability model m_{x_s} in $\mathcal{M}_{x_s}$ can proceed independently of the search in $\mathcal{M}_{y|x_s}$. The next theorem gives sufficient conditions on the distribution of $R|x_d, \psi_{r|x_d}$ such that (8) holds.

Theorem 1. Under assumptions 1, 2, and 3, a missing-data mechanism is ignorable for inference about $m = (m_{x_s}, m_{y|x_c})$ if the probability $p(r|x_s, y_o, y_m, \psi_{r|x_d}, m_{y|x_c})$ simplifies to $p(r|x_d, \psi_{r|x_d})$ and $p(m|\psi_{r|x_d}) = p(m_{x_s}) p(m_{y|x_c})$, for any m in the model space $\mathcal{M}$.

Proof. Let $m = (m_{x_s}, m_{y|x_c})$ be a model in $\mathcal{M}$. We have shown that $p(m|x_s, y_o, r, \psi_{r|x_d}) \propto p(m_{x_s}) p(x_s|m_{x_s}) p(m_{y|x_c}) \times p(y_o, r|x_s, m_{y|x_c}, \psi_{r|x_d})$, when $p(m|\psi_{r|x_d}) = p(m_{x_s}) p(m_{y|x_c})$. The quantity $p(y_o, r|x_s, m_{y|x_c}, \psi_{r|x_d})$ is computed as $\int_{y_m} p(y_o, y_m, r|x_s, m_{y|x_c}, \psi_{r|x_d}) dy_m$, where y_m denotes the

missing y values. Now write $p(y_o, y_m, r|x_s, m_{y|x_c}, \psi_{r|x_d})$ as the product of the two factors $p(y_o, y_m|x_s, m_{y|x_c}, \psi_{r|x_d})$ and $p(r|x_s, y_o, y_m, m_{y|x_c}, \psi_{r|x_d})$ and note that $p(y_o, y_m|x_s, m_{y|x_c}, \psi_{r|x_d}) = p(y_o, y_m|x_c, m_{y|x_c})$ by definition. By $p(r|x_s, y_o, y_m, \psi_{r|x_d}, m_{y|x_c}) = p(r|x_d, \psi_{r|x_d})$, we have $p(y_o, r|x_s, m_{y|x_c}, \psi_{r|x_d}) = p(r|x_d, \psi_{r|x_d})p(y_o|x_c, m_{y|x_c})$, so that the factorization in (8) is true, thus ensuring ignorability of the missing-data mechanism.

A byproduct of Theorem 1 is that, for inference on a model m with incomplete data (x_s, y_o) under an ignorable missing-data mechanism, the values y_m are also ignorable. This result follows by observing that the integral $\int_{y_m} p(y_o, y_m|x_c, m_{y|x_c})dy_m$ can be computed as $p(y_o|x_c, m_{y|x_c}) \times \int_{y_m} p(y_m|y_o, x_c, m_{y|x_c})dy_m$, and $\int_{y_m} p(y_m|y_o, x_c, m_{y|x_c})dy_m$ is equal to 1. We state this result formally in the next corollary.

Corollary 1. The missing data y_m resulting from an ignorable missing-data mechanism are ignorable for inference about a model m.

Theorem 1 gives a sufficient condition to have an ignorable missing-data mechanism for our model selection problem. We prove that there exists only one class of ignorable missing-data mechanisms.

Fix one model $m_{y|x_c}$ in $\mathcal{M}_{y|x_s}$ and consider the probability $p(r|x_s, y_o, y_m, \psi_{r|x_d}, m_{y|x_c})$. The model $m_{y|x_c}$ identifies the variables in $\mathcal{X}_s$ on which Y depends, so that $(x_s, m_{y|x_c}) = x_c$. By definition, R is independent of Y, given $\mathcal{X}_d$ and $\psi_{r|x_d}$. Hence when $\mathcal{X}_c$ contains $\mathcal{X}_d$, we have that $p(r|x_s, y_o, y_m, \psi_{r|x_d}, m_{y|x_c}) = p(r|x_d, y_o, y_m, \psi_{r|x_d})$, which, by definition, simplifies to $p(r|x_d, \psi_{r|x_d})$. This fact follows by the property that if R is independent of Y given $\mathcal{X}_d$ and if $\mathcal{X}_c$ contains $\mathcal{X}_d$, then R is independent of Y given $\mathcal{X}_c$ (Dawid 1979, 1980).

On the other hand, when $\mathcal{X}_c$ does not fully contain $\mathcal{X}_d$, the independence of R and Y given $\mathcal{X}_d$ does not imply independence of R and Y given $\mathcal{X}_c$. Indeed, we can rewrite $p(r|x_s, y_o, y_m, m_{y|x_c}, \psi_{r|x_d})$ as $p(r|x_c, x_d, y_o, y_m, m_{y|x_c}, \psi_{r|x_d})$, which does not simplify into $p(r|x_d, \psi_{r|x_d})$, so that the marginal likelihood $p(y_o, r|x_s, m_{y|x_c}, \psi_{r|x_d})$ does not factorize as in Equation (8). To characterize ignorable missing-data mechanisms, we need to find conditions on $\mathcal{X}_d$ under which Y and R are independent given $\mathcal{X}_c$, for any $\mathcal{X}_c \subseteq \mathcal{X}_s$. This simplification occurs only when $\mathcal{X}_d = \varnothing$ (Dawid 1979, 1980). We state this result in the next theorem.

Theorem 2. Under the same assumption as in Theorem 1, a missing-data mechanism described by the probability distribution $p(R|x_d, \psi_{r|x_d})$ is ignorable if and only if $\mathcal{X}_d = \varnothing$.

Theorem 2 excludes the ignorability of any missing-data mechanism described by a probability distribution $p(R|x_d, \psi_{r|x_d})$. However, this lack of ignorability is conceptually different from the nonignorability of an informative missing-data mechanism in which the missingness probability is a function of the missing values. We can describe these informative missing-data mechanisms by the probability distribution $p(R|x_d, y, \psi_{r|x_d})$. Furthermore, as discussed earlier, a mechanism described by the probability $p(R|x_d, \psi_{r|x_d})$ has the property that the marginal likelihood $p(y_o, r|x_s, m_{y|x_c}, \psi_{r|x_d})$ can be computed as $p(r|x_d, \psi_{r|x_d})p(y_o|x_c, m_{y|x_c})$ whenever x_d is a subset of x_c. Hence, at least for all models $m_{y|x_c}$ specifying the dependence of Y on $\mathcal{X}_c \supseteq \mathcal{X}_d$, the missing-data mechanism can be ignored in the computation of the marginal likelihood $p(y_o|x_c, m_{y|x_c})$. These observations lead us to give the following definition.

Definition 1. We define a missing-data mechanism described by the probability distribution $p(R|x_d, \psi_{r|x_d})$ as partially ignorable.

The results of Theorem 1 and 2 have some interesting consequences on the received view of missing-data mechanisms as recently described by, for example, Schafer (1997). Because a partially ignorable missing-data mechanism is characterized by a probability distribution $p(R|x_d, \psi_{r|x_d})$, and hence entries of the variable Y are removed independently of the missing values but dependently on the observed values, there follows that data removed with this mechanism are MAR. Our results show that these missing-data mechanisms, although paired with technical conditions on the model space, are not ignorable for model selection. Furthermore, a missing-data mechanism described by a probability distribution $p(R|\psi_{r|x_d})$, so that R is marginally independent of all of the variables in the set $\mathcal{X}_s$, produces MCAR data. Theorem 2 shows that this is the only ignorable missing-data mechanism. This fact is not new; for example, Little and Rubin (1987, p. 16) and, more recently, Rubin (1996) have warned against improper treatment of an incomplete sample when data are MAR rather than MCAR. Our results provide a formal foundation for this intuition.

The results of this section open two issues. The distinction between ignorable and partially ignorable missing-data mechanisms has consequences on the implementation of model selection. For example, as shown in Corollary 1, an ignorable missing-data mechanism lets model selection be carried out by simply disregarding the missing values. The first issue is then to see whether the data (x_s, y_o, r) can be used to discriminate between the two mechanisms. The second issue is how to proceed with model selection when the missing-data mechanism is only partially ignorable. For example, some authors (Heitjan and Rubin 1991; Little and Rubin 1987; Rubin 1976) suggest that the variables responsible for the missing data should be kept in the model, thus limiting the model search to a subset of models. We consider these problems in the next two sections.

4. DISCRIMINATING BETWEEN IGNORABILITY AND PARTIAL IGNORABILITY

In this section we show that the sample (x_s, r) can be used to select one missing-data mechanism, that is, a dependency model of R on $\mathcal{X}_s$, which we call $m_{r|x_d}$. The approach that we describe puts suggestions of Little and Rubin (1987) and Baker and Laird (1988) for fitting log-linear models to assess whether a missing-data mechanism is nonignorable into a Bayesian framework. Here we focus on a subclass of log-linear models and try to assess whether the mechanism is ignorable or partially ignorable with no attempt to detect whether there is a dependency between R and Y.

The data (x_s, r) are a complete sample for R and $\mathcal{X}_s$. Therefore, if we focus on decomposable models $(m_{x_s}, m_{r|x_d})$ and suppose that m_{x_s} and $m_{r|x_d}$ are a priori independent,

Table 1. The Obese Dataset

		Obese			
Age	*Gender*	*Yes*	*No*	*NA*	*% NA*
Young	M	82	463	470	46
	F	81	435	418	45
Old	M	247	900	324	22
	F	272	861	303	21

NOTE: Young children are age under 9 years; old children, above 9 years.

then we can apply the model selection method described in Section 2 to choose in particular a dependency model of R on $\mathcal{X}_s$. The solution is the model $m_{r|x_d}$ with maximum $p(m_{r|x_d})p(r|x_d, m_{r|x_d})$. Computation of $p(r|x_d, m_{r|x_d})$ can be done in closed form when the variables $\mathcal{X}_s$ are all categorical, whereas when the variables $\mathcal{X}_s$ are both continuous and categorical, approximations are needed.

Suppose first that the variables $\mathcal{X}_s$ are all categorical. Conditional on $m_{r|x_d}$ specifying the dependence of R on $\mathcal{X}_d$, the distribution of R is modeled as a product of binomial distributions, with parameters $\psi_j = \psi(x_{dj})$, x_{dj} being one of the q_d states of $\mathcal{X}_d$. We assign independent beta distributions with hyperparameters β^d_{j1} and β^d_{j2} to each parameter ψ_j, so that the overall prior distribution is hyperbeta. Using the result of Example 2, the marginal likelihood is

$$p(r|x_d, m_{r|x_d}) = \prod_{j=1}^{q_d} \frac{\Gamma(\beta^d_{j1}+\beta^d_{j2})}{\Gamma(\beta^d_{j1}+\beta^d_{j2}+n^d_j+m^d_j)} \frac{\Gamma(\beta^d_{j1}+m^d_{j1})\Gamma(\beta^d_{j2}+n^d_j)}{\Gamma(\beta^d_{j1})\Gamma(\beta^d_{j2})},$$

where n^d_j denotes the frequency of cases $(R = 0, x_{dj})$ and m^d_j denotes the frequency of cases $(R = 1, x_{dj})$. The next example applies this procedure to a real dataset.

Example 3. We analyze the missing-data mechanism generating the incomplete dataset in Table 1. The data, reported by Park and Brown (1994), consist of the results of a 1977 survey of schoolchildren in which respondents were asked whether they were obese or not. Our goal is to see whether either a child's age or gender is related to the probability of nonresponse. This dataset was analyzed by Park and Brown (1994) under the assumption of nonignorable nonresponse, because a refusal to answer should depend on the child's obesity. Here we assume that the probability of nonresponse is independent of the child's obesity, given age and gender, because children are "naive" and do not answer if they truly do not know the answer.

We let R denote the missingness variable and use the data on age, gender, and R to discriminate between the four models generating ignorable or partially ignorable missing-data mechanisms. We denote age by X_1 and gender by X_2. The three models in which R depends on X_1 alone ($m_{r|x_1}$), on X_2 alone ($m_{r|x_2}$), or on both X_1 and X_2 ($m_{r|x_1,x_2}$) describe partially ignorable missing-data mechanisms, whereas the model in which R is independent of both X_1 and X_2 ($m_{r|\varnothing}$) produces an ignorable missing-data mechanism. We suppose that the four models are a priori equally likely and, conditional on each model $m_{r|x_d}$, we assign symmetric hyperbeta prior distributions to the parameters ψ_j. The hyperparameters β_{ji} are chosen as $\beta_{ji} = \beta/(2 \times q_j)$, where β is overall prior precision.

Table 2. Marginal Likelihood, in Log-Scale, for the Four Models Producing Partially Ignorable or Ignorable Missing-Data Mechanisms for the Data in Table 1

Model $m_{r\|x_d}$	*Log-marginal likelihood* $\beta = .001$	$\beta = 1.000$	$\beta = 4.000$	
$m_{r\|x_1,x_2}$	−2904.563	−2877.638	−2873.594	Age and gender
$m_{r\|x_2}$	−3035.763	−3022.565	−3020.867	Gender
$m_{r\|x_1}$	−2881.274*	−2868.099*	−2866.471*	Age
$m_{r\|\varnothing}$	−3024.872	−3018.491	−3017.863	None

NOTE: Stars identify the maximum value of the marginal likelihood in each column.

Table 2 gives the marginal likelihood of the four models for three different choices of β. The models are ranked in the same way, and there is little doubt that R depends on age. Therefore, the missing-data mechanism can be ignored in the calculation of the marginal likelihood of $m_{r|x_1}$ and $m_{r|x_1,x_2}$, but cannot be ignored during the calculation of the marginal likelihood of $m_{r|x_2}$ and $m_{r|\varnothing}$. Note that the probability of nonresponse is larger in young children than in old children; this finding supports the hypothesis that nonresponse is due to a genuine "ignorance" rather than a willingness to hide the truth.

The simplicity of the computations in the previous examples is lost when some of the variables in $\mathcal{X}_s$ are continuous. In this case we can still define the vector of parameters $\psi_{r|x_d}$ as $\psi_{r|x_d} = (\psi_{x_d})$, $\psi_{x_d} = p(R = 1|x_d, \psi)$. However, now ψ_{x_d} is some function of $\mathcal{X}_d$, such as the logit function, and involves further parameters Φ. This parameterization does not usually lead to a closed-form solution for the marginal likelihood, and we need either stochastic or deterministic approximations. With large datasets, one can use the Bayesian information criterion (BIC) (Schwarz 1978), which makes an asymptotic approximation of the marginal likelihood $p(r|x_d, m_{r|x_d})$. Further discussion has been given by Kass and Raftery (1995), Raftery (1995), and Wasserman (1999).

5. MODEL SELECTION WITH INCOMPLETE DATA

In this section we provide two solutions to the selection of $m_{y|x_c}$ when the missing-data mechanism is partially ignorable. We begin with a discussion of the computational problems due to partial ignorability.

5.1 Exact Modeling

In Section 3 we defined a missing-data mechanism to be partially ignorable when the probability distribution of the missingness variable R is some function of $\mathcal{X}_s$. Suppose that the approach described in Section 4 selects the subset $\mathcal{X}_d$ on which R depends. We assume that, a priori, m_{x_s}, $m_{y|x_c}$, and $\psi_{r|x_d}$ are independent, so that the search for the maximum posterior probability model m_{x_s} can be carried out independently of the missing-data mechanism. The issue remains the selection of a dependency model $m_{y|x_c}$. We distinguish between the two situations discussed in Section 3:

(a) For any $\mathcal{X}_c$ such that $\mathcal{X}_c \cap \mathcal{X}_d = \mathcal{X}_d$, the missing-data mechanism is ignorable because the posterior probability $p(m_{y|x_c}|x_s, y_o, r, \psi_{r|x_d})$ simplifies into $p(m_{y|x_c}|x_c, y_o)$.

(b) For any $\mathcal{X}_c$ such that $\mathcal{X}_c \cap \mathcal{X}_d \neq \mathcal{X}_d$, the quantity $p(y_o, r|x_s, m_{y|x_c}, \psi_{r|x_d})$ is the solution of $\int_{y_m} p(y_o, y_m, r|x_s, m_{y|x_c}, \psi_{r|x_d})dy_m$. Conditional on $m_{y|x_c}$, we have $x_s \equiv x_c$, so that $\int_{y_m} p(y_o, y_m, r|x_s, m_{y|x_c}, \psi_{r|x_d})dy_m = \int_{y_m} p(y_o, y_m|x_c)p(r|x_c, y_o, y_m, \psi_{r|x_d})dy_m$ and, in the last integral, the probability $p(r|x_c, y_o, y_m, \psi_{r|x_d})$ does not simplify. However, we note that

$$\begin{aligned} p(r|x_c, y_o, y_m, \psi_{r|x_d}) &= \int_{\mathcal{X}_d \setminus \mathcal{X}_c} p(r|x_d, y_o, y_m, \psi_{r|x_d}) \\ &\quad \times p(x_d|x_c, y_o, y_m, \psi_{r|x_d})dx_d \setminus x_c \\ &= \int_{\mathcal{X}_d \setminus \mathcal{X}_c} p(r|x_d, \psi_{r|x_d}) \\ &\quad \times p(x_d|x_c, y_o, y_m)dx_d \setminus x_c, \end{aligned}$$

so that whenever we "expand" the set $\mathcal{X}_c$ to contain $\mathcal{X}_d$, the dependency of R on y_o and y_m disappears. Therefore, to integrate out the missing values, we need to expand model $m_{y|x_c}$ to contain the variables $\mathcal{X}_d \setminus \mathcal{X}_c$.

Exact modeling requires the computation of the observed likelihood $p(y_o, r|x_s, m_{y|x_c}, \psi_{r|x_d})$, by solving the foregoing two integrals. Because there does not seem to be a general closed-form solution, exact integration can be approximated by a stochastic estimation. Alternatively, one may use imputation to fill in the missingdata. The next section focuses on imputation.

5.2 Multiple Imputation

Imputation is a popular method for handling incomplete datasets (Rubin 1987). Gelman, Carlin, Stern, and Rubin (1995), Schafer (1997), and Tanner (1996) have given in-depth descriptions of several imputation schemes in a Bayesian framework. Rubin (1987, 1996) discussed pros and cons of imputation for frequentist and Bayesian inferences. Multiple imputation replaces missing values with quantities generated by an appropriate imputation model. By repeating the process several times, posterior inference can be carried out by averaging the results obtained from the imputed samples. There are several multiple imputation schemes that differ in the choice of the conditional distribution used in the imputation step. For example, Schafer (1997) defined *Bayesianly proper multiple imputation* as a procedure in which missing data are imputed from the predictive distribution of Y conditional on x_c, with density $p(y|x_c, y_o, m_{y|x_c})$. Bayesianly proper multiple imputation is repeated for every model $m_{y|x_c}$ evaluated during the search process and produces a posterior probability of $m_{y|x_c}$ conditional on x_c, y_o, and the imputed data y_{imp}. When the missing-data mechanism is only partially ignorable and $\mathcal{X}_c \cap \mathcal{X}_d \neq \mathcal{X}_d$, Y is not independent of R conditional on $\mathcal{X}_c$, and imputing data from the $p(y|x_c, y_o, m_{y|x_c})$, introduces bias. This fact was also noted by Rubin (1996). As was shown at the beginning of this section, the dependence of Y on R disappears when we condition on $\mathcal{X}_d$. Indeed, we have

$$\begin{aligned} &p(y|x_c, y_o, m_{y|x_c}, r) \\ &\quad = \int_{\mathcal{X}_d \setminus \mathcal{X}_c} p(y|x_d, y_o, r)p(x_d|x_c, y_o)dx_d \setminus x_c, \end{aligned}$$

so that imputing missing data from a mixture of distributions appears to be the proper scheme. A simple alternative is to impute data from $p(y|x_d, y_o, m_{y|x_d})$ conditional on $m_{y|x_d}$. The advantage of this approach is that imputation needs to be done only once, and the imputed samples can be used for inference as if they were complete samples. We call this approach, in which missing data are simulated conditional on $m_{y|x_d}$, *ignorable imputation*. The imputed sample is used to evaluate the posterior probabilities of all models. By repeating this procedure v times, the posterior probability of each model then becomes the average of the posterior probabilities computed from the v imputed samples.

Ignorable imputation has two important features. By requiring imputation from only one model, it results in computational savings when compared to schemes in which the imputation model changes with $m_{y|x_c}$. Furthermore, all posterior probabilities are conditional on the same imputed data, thus providing consistent evaluations of the Bayes factor to assess the strength of evidence of one model $m_{y|x_c}$ versus other models. When $\mathcal{X}_s$ and Y are categorical variables, ignorable imputation can be easily implemented following the *data augmentation* scheme proposed by Tanner and Wong (1987). We use the notation of Example 2 and let n_o and n_m denote the number of observed data, y_o, and the number of missing entries, y_m. Initially, the n_o fully observed data y_o are used to update the prior density of $\theta_{y|x_d}$ into the posterior density, and this is then used to compute $p(y|x_{dk}, y_o, m_{y|x_d})$, where x_{dk} is the first state of $\mathcal{X}_d$ in which the entry of Y is missing. If the prior density of $\theta_{y|x_d}$ is a hyper-Dirichlet, with hyperparameters α^d_{kj}, then the predictive probability of $Y = j$ in x_{dk} is

$$p(Y = j|x_{dk}, y_o, m_{y|x_d}) = \frac{\alpha^d_{kj} + n^d_{kj} + 1}{\alpha^d_{k\cdot} + n^d_{k\cdot} + 1},$$

so that a value for Y is generated from this probability distribution and the outcome is used to update the predictive distribution for the next step. Alternatively, one may simulate the n_m incomplete entries of Y simultaneously, conditional on the observed data y_o. When the size of the observed data is large, it is appealing to replace the imputation step by a deterministic step in which one substitutes missing values with expected ones. This is the intuition of model folding, which is presented in the next section. When $\mathcal{X}_s$ and Y are a mixture of categorical and continuous variables, the posterior density of $\theta_{y|x_d}$ may not be computed in closed form, and implementation of ignorable imputation requires more sophisticated methods. Several methods have been described by Tanner (1996, Chap. 5).

Note that imputing missing data from the minimal model for which the missing-data mechanism is ignorable guarantees the asymptotic propriety of ignorable imputation. Rubin (1987) defined a multiple imputation model to be proper when it produces unbiased estimates of quantities of interest. In our context, ignorable imputation produces an unbiased estimate of the exact marginal likelihood. The same propriety is enjoyed by any schemes requiring imputation from the saturated model or at least a model $m_{y|x_c}$ with $\mathcal{X}_c \supseteq \mathcal{X}_d$. Also, Rubin (1996) suggested including all variables in a multiple imputation model to make it proper for any inference.

5.3 Model Folding

Ignorable imputation replaces missing values by imputed ones to estimate the marginal likelihood. When $\mathcal{X}_s$ and Y are categorical variables, we propose a deterministic approximation of the exact marginal likelihood, which is computed as follows. We use the notation of Example 2 and let $\theta_{y|x_c}$ denote the parameters of $m_{y|x_c}$. The prior distribution of $\theta_{y|x_c}$ is a hyper-Dirichlet, with hyperparameters α^c_{kj}. We let m^c_k denote the frequency of missing data for each state x_{ck} of $\mathcal{X}_c$, and let n^c_{kj} denote the frequency of cases in the sample with $Y = j$. It follows that $n^c_{k\cdot} = \sum_j n^c_{kj}$ is the total number of cases fully observed for each state x_{ck}. The approximation is

$$\prod_k \frac{\Gamma(\alpha^c_{k\cdot})}{\Gamma(\alpha^c_{k\cdot} + n^c_{k\cdot} + m^c_k)} \prod_j \frac{\Gamma(\alpha^c_{kj} + n^c_{kj} + \hat{\phi}^c_{kj} m^c_k)}{\Gamma(\alpha^c_{kj})}, \qquad (9)$$

where $\hat{\phi}^c_{kj}$ is an estimate of $\phi^c_{kj} = p(Y = j|x_{ck}, y_o, R = 1, m_{y|x_c}, \psi_{r|x_d})$, which is the probability of $Y = j$ among nonrespondents in x_{ck}, and hence we condition on $R = 1$. The intuition behind this approximation is to use an estimate of the probability distribution of Y among nonrespondents to distribute the missing data across categories of Y. In this way we complete the incomplete sample by replacing missing values with expected values, so that we can apply the results described in Section 2 to compute the marginal likelihood using the expected completion of the data. The crucial step remains estimation of ϕ^c_{kj}, and we describe this next.

Suppose first that the distribution of R is a function of $\mathcal{X}_s$, and denote $p(R = 1|x_{sk}, \psi_{r|x_s})$ by ψ^s_k. As in Section 4, we assign a hyperbeta prior distribution to $\psi_{r|x_s}$, with hyperparameters β^s_{ki}. The missing-data mechanism is ignorable for the saturated model $m_{y|x_s}$, and missing data can be disregarded to compute the marginal likelihood of $m_{y|x_s}$. Consider now a model $m_{y|x_c}$ specifying the dependence of Y on the subset $\mathcal{X}_c$. The ignorability of the missing-data mechanism for $m_{y|x_s}$ gives a simple way to estimate $\phi^c_{kj} = p(Y = j|x_{ck}, y_o, R = 1, m_{y|x_c}, \psi_{r|x_s})$. By letting $\mathcal{X}_s = \mathcal{X}_c \cup \mathcal{X}_{\bar{c}}$, it is easy to show that

$$\phi^c_{kj} \propto \sum_{x_{\bar{c}k}} p(Y = j|x_{sk}, y_o, R = 1, \psi_{r|x_s}) \times p(R = 1|x_{sk}, y_o, \psi_{r|x_s}) p(x_{\bar{c}k}|x_{ck}).$$

The quantity $p(x_{\bar{c}k}|x_{ck})$ can be estimated from the complete data x_s, conditional on the dependency model m_{x_s} selected for the variable $\mathcal{X}_s$. Because Y is independent of R, given all the observed values and $\psi_{r|x_s}$, $p(Y = j|x_{sk}, y_o, R = 1, \psi_{r|x_s})$ is the posterior probability $p(Y = j|x_{sk}, y_o, m_{y|x_s})$ under the saturated model and is independent of $\psi_{r|x_s}$. Hence to estimate $p(Y = j|x_{sk}, y_o, m_{y|x_s})$, we disregard the missing values and use the posterior mean of θ^s_{kj},

$$\hat{\theta}^s_{kj} = \frac{\alpha^s_{kj} + n^s_{kj}}{\alpha^s_{k\cdot} + n^s_{k\cdot}}.$$

In particular, $\hat{\theta}^s_{kj}$ is the generalized maximum likelihood estimate, as shown by Sebastiani and Ramoni (2000). Similarly, by definition, $p(R = 1|x_{sk}, y_o, \psi_{r|x_s}) = p(R = 1|x_{sk}, \psi_{r|x_s}) = \psi^s_k$, and we estimate it by the posterior mean of ψ^s_k,

$$\hat{\psi}^s_k = \frac{\beta^s_{k1} + m^s_k}{\beta^s_{k1} + \beta^s_{k2} + n^s_{k\cdot} + m^s_k}.$$

Both $\hat{\theta}^s_{kj}$ and $\hat{\psi}^s_k$ are used to estimate $\hat{\phi}^c_{kj} \propto \sum_{x_{\bar{c}}} \hat{\theta}^s_{kj} \hat{\psi}^s_k \times p(x_{\bar{c}k}|x_{ck})$. By plugging $\hat{\phi}^c_{kj}$ into (9), we get an approximation of the marginal likelihood for $m_{y|x_c}$. The approximate marginal likelihood of $m_{y|x_c}$ depends on the "augmented" sample, in which missing data are distributed across categories of Y using the estimates $\hat{\phi}^c_{kj}$. A byproduct of this method is a way to estimate θ^c_{kj} once a model $m_{y|x_c}$ is selected, using the expected completion of the data. This estimate,

$$\hat{\theta}^c_{kj} = \frac{\alpha^c_{kj} + n^c_{kj} + \hat{\phi}^c_{kj} m^c_k}{\alpha^c_{k\cdot} + n^c_{k\cdot} + m^c_k},$$

accounts for the missing-data mechanism via $\hat{\phi}^c_{kj}$. This is in agreement with the fact that the missing-data mechanism is not ignorable for inference on $m_{y|x_c}$. In particular, $\hat{\theta}^c_{kj}$ is the estimate computed with the bound and collapse method of Sebastiani and Ramoni (2000) for informative nonresponse.

The exact marginal likelihood of the saturated model is only a function of the incomplete sample. To make consistent comparisons, we compute the marginal likelihood of $m_{y|x_s}$ on the expected completion of the data by using $\hat{\phi}^s_{kj} = \hat{\theta}^s_{kj}$ to distribute the missing data across categories of Y. This approximation is

$$p(y_o|x_s, m_s) \propto \prod_k \frac{\Gamma(\alpha^s_{k\cdot})}{\Gamma(\alpha^s_{k\cdot} + n^s_{k\cdot} + m^s_k)} \prod_j \frac{\Gamma(\alpha^s_{kj} + n^s_{kj} + \hat{\theta}^s_{kj} m^s_k)}{\Gamma(\alpha^s_{kj})}.$$

The estimates of θ^s_{kj} computed on the expected completion of the data are exactly $\hat{\theta}^s_{kj}$, but the posterior precision of θ^s_{kj} is larger than the exact one. This can yield overconfidence, although the underlying idea of the method that we propose follows the suggestion of Little and Rubin (1987) to use multiple imputation even when the missing data are ignorable.

Suppose now that the distribution of the missingness variable R depends on $\mathcal{X}_d \subset \mathcal{X}_s$. Thus for every model $m_{y|x_c}$ with $\mathcal{X}_c \cap \mathcal{X}_d = \mathcal{X}_d$, the missing-data mechanism is ignorable, but the mechanism is not ignorable for computation of the marginal likelihood of any other model $m_{y|x_c}$ with $\mathcal{X}_c \cap \mathcal{X}_d \neq \mathcal{X}_d$. As before, we compute $\phi^c_{kj} = p(Y = j|x_{ck}, y_o, R = 1, m_{y|x_c}, \psi_{r|x_d})$ as

$$\phi^c_{kj} \propto \sum_{x_{dk} \setminus x_{ck}} p(Y = j|x_{dk}, y_o, R = 1, \psi_{r|x_d}) \times p(R = 1|x_{dk}, y_o, \psi_{r|x_d}) p(x_{dk}|x_{ck}),$$

where $p(Y = j|x_{dk}, y_o, R = 1, \psi_{r|x_d}) = p(Y = j|x_{dk}, y_o, m_{y|x_d})$ for the ignorability of the missing-data mechanism for model $m_{y|x_d}$, so that we estimate the probability $p(Y = j|x_{dk}, y_o, m_{y|x_d})$ by

$$\hat{\theta}^d_{kj} = \frac{\alpha^d_{kj} + n^d_{kj}}{\alpha^d_{k\cdot} + n^d_{k\cdot}}.$$

Furthermore, by definition, $p(R = 1|x_{dk}, y_o, \psi_{r|x_d}) = p(R = 1|x_{dk}, \psi_{r|x_d})$, and we estimate it by the posterior mean $\hat{\psi}^d_k$ of ψ^d_k. It follows that the estimate of ϕ^c_{kj} is $\hat{\phi}^c_{kj} \propto \sum_{x_d \setminus x_c} \hat{\theta}^d_{kj} \hat{\psi}^d_k p(x_{dk}|x_{ck})$. Note that $\hat{\phi}^c_{kj} \equiv \hat{\theta}^c_{kj}$ whenever $\mathcal{X}_c \supset \mathcal{X}_d$, so we can use $\hat{\phi}^c_{kj}$ to compute the expected completion of the data for any model $m_{y|x_c}$.

Once a model $m_{y|x_c}$ is selected, the complete sample computed by model folding can also be used to carry out posterior inference, by using the counts $\alpha^c_{kj} + n^c_{kj} + \hat{\phi}^c_{kj} m^c_k$ as updated hyperparameters of the parameter posterior distribution. This posterior approximate inference was described by Sebastiani and Ramoni (2000), and experimental evaluations show that credible intervals computed using this approximation are extremely accurate compared to those computed using, for example, Gibbs sampling (Spiegelhalter, Thomas, and Best 1996).

Although the method is based on a crude approximation, it appears to perform reasonably well compared to ignorable imputation without requiring the same computation effort. In the next section we provide simulation results to support this claim. We conclude here by noting that the approximation provided by model folding is conditional on the model m_{x_s} selected for the variables $\mathcal{X}_s$ and on the model $m_{r|x_d}$ selected for the missing-data mechanism. In simulation studies that we carried out, the model m_{x_s} chosen for $\mathcal{X}_s$ appeared to have very little effect. In the next section we investigate the effect of the model chosen for the missing-data mechanism. To approximate the marginal likelihood of $m_{y|x_c}$, model folding requires estimating the probability distribution of the nonrespondents using the estimates of the parameters of $m_{y|x_d}$ and of $m_{r|x_d}$. Compared to the complete-data case, the complexity of the model search is increased only with this extra estimation step.

6. REPEATED-SAMPLING PROPERTIES

In this section we evaluate the accuracy of model folding and ignorable imputation in three controlled experiments. The first experiment shows that model folding and ignorable imputation perform almost identically, while ignoring the missing-data mechanism can severely bias the modeling process. The second experiment shows the robustness of both methods when either the missing-data mechanism is supposed to be partially ignorable and the deletion process is informative or the missing-data mechanism is partially ignorable but an inaccurate model for the missing-data mechanism is selected. The last experiment gives some insight in one potential problem of imputation, namely that imputation could bias the model selection toward the imputation model, and shows that ignorable imputation does not appear to suffer this problem.

6.1 Accuracy

Data y, x_1, and x_2 in Table 3 are a random sample generated from the model with both X_1 and X_2 associated with Y and $p(Y = 1|X_1 = 1, X_2 = 1) = .5$, $p(Y = 1|X_1 = 1, X_2 = 2) = .2$, $p(Y = 1|X_1 = 2, X_2 = 1) = .7$, and $p(Y = 1|X_1 = 2, X_2 = 2) = .3$. As shown in Example 1, the model space $\mathcal{M}_{y|x_s}$ consists of $m_{y|\varnothing}$, $m_{y|x_1}$, $m_{y|x_2}$, and $m_{y|x_2,x_3}$. Assuming symmetric hyper-Dirichlet distributions on each model parameter, with $\alpha = 8$, the marginal likelihood of each model

Table 3. Sample Generated From Model $m_{y|x_1,x_2}$ Specifying Dependence of Y on X_1,X_2

	X_1, X_2			
Y	(1,1)	(1,2)	(2,1)	(2,2)
1	52	17	66	36
2	48	83	34	64

can be computed using (7), and the values in log scale are $\log p(y|m_{y|\varnothing}) = -275.1086$, $\log p(y|x_1, m_{y|x_1}) = -255.4863$, $\log p(y|x_2, m_{y|x_2}) = -271.6517$, and $\log p(y|x_1, x_2, m_{y|x_1,x_2}) = -252.9841$. Therefore, $m_{y|x_1,x_2}$ is selected, conditional on the observed data, if all models are a priori equally likely.

The complete sample in Table 3 was then subjected to a random deletion of Y entries with the following process. We defined a binary variable R with probabilities $p(R = 1|X_1 = 1, X_2 = 1) = .2$, $p(R = 1|X_1 = 1, X_2 = 2) = .3$, $p(R = 1|X_1 = 2, X_2 = 1) = .1$, and $p(R = 1|X_1 = 2, X_2 = 2) = .6$. For each sample case, we generated a value of R conditional on the values of X_1 and X_2, and we removed the entry of Y if $R = 1$. Given the dependence of R on both X_1 and X_2, the missing-data mechanism is ignorable for $m_{y|x_1,x_2}$, but it is not for the other models. We repeated this deletion process 100 times, and in each incomplete sample we computed the marginal likelihood of $m_{y|\varnothing}$, $m_{y|x_1}$, $m_{y|x_2}$, and $m_{y|x_1,x_2}$ using model folding with $\alpha_{kj} = 8/(2q_c)$, ignorable imputation, and Bayesianly proper imputation with 10 imputed values for each missing entry, and data deletion, in which missing data were disregarded. Data deletion would be the correct approach for an ignorable missing-data mechanism. In all cases, we assumed that R was a function of both X_1 and X_2 and chose $\beta_{kj} = .125$. Ignorable imputation was implemented by simulating the missing data at once from the predictive distribution of Y, conditional on the observed data and both X_1 and X_2. Values of the marginal likelihood computed with ignorable and Bayesianly proper imputation are averages of the marginal likelihood computed from each imputed sample.

Figure 1 reports the marginal likelihood, in log-scale, of $m_{y|\varnothing}$, $m_{y|x_1}$, $m_{y|x_2}$, and $m_{y|x_1,x_2}$. Assuming uniform probabilities on the model space, in the 100 incomplete samples model folding selected the correct model $m_{y|x_1,x_2}$ in 85 samples and selected $m_{y|x_1}$ in 15 samples. Ignorable imputation selected $m_{y|x_1,x_2}$ in 87 samples and $m_{y|x_1}$ in 13 samples. Thus the error rates of the two methods are equivalent. Bayesianly proper imputation selected the correct model in only 26 samples. With data deletion, $m_{y|x_1,x_2}$ was selected in 58 samples and $m_{y|x_1}$ in 42 samples. The error rate of model folding and ignorable imputation is within the sampling variability; in 100 complete samples generated by the same model, $m_{y|x_1,x_2}$ was selected in 80% of cases.

Figure 1 reveals the reasons for the large error rates of Bayesianly proper imputation and data deletion. Figures 1(a) and 1(b) show two distinct groups of points, the estimates of $\log(p(y|m_{y|\varnothing}))$ and $\log(p(y|x_2, m_{y|x_2}))$ in the lower part of the figure and the estimates of $\log(p(y|x_1, m_{y|x_1}))$ and $\log(p(y|x_1, x_2, m_{y|x_1,x_2}))$ in the top. These two groups maintain the ordering between the posterior probabilities computed from the complete samples. Bayesianly proper

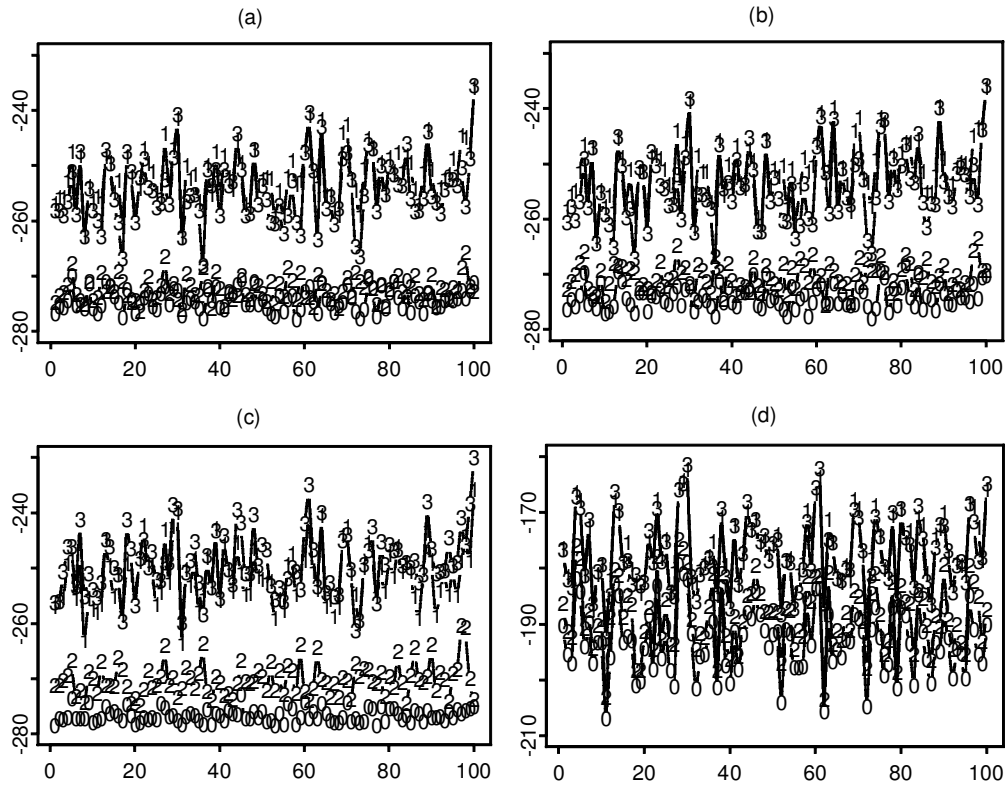

Figure 1. Marginal Likelihood (in log-scale) of $m_{y|\varnothing}(0)$, $m_{y|x_1}(1)$, $m_{y|x_2}(2)$, and $m_{y|x_1,x_2}(3)$, in the 100 Incomplete Samples Generated From the Data in Table 3. (a) Model folding; (b) Ignorable imputation; (c) Bayesianly proper imputation; (d) data deletion.

imputation introduces an evident bias. The estimates of $\log(p(y|x_2, m_{y|x_2}))$ are almost all above the estimates of $\log(p(y|m_{y|\varnothing}))$, and the estimates of $\log(p(y|x_1, m_{y|x_1}))$ are almost all greater than those of $\log(p(y|x_1, x_2, m_{y|x_1,x_2}))$. If missing data are ignored, then there is no longer an evident distinction between the marginal likelihood of the four models. This bias is clear from the summary statistics reported in Table 4.

6.2 Robustness

The accuracy of model folding relies on the ability to use the information about the missing-data mechanism to adjust the estimates computed from the complete cases. However, it is reasonable to wonder whether such accuracy can become a drawback when the missing-data mechanism is supposed to be partially ignorable but the probability of nonresponse is a function of Y. To answer this question, we ran another set of simulations in which the data in Table 3 were deleted with a nonignorable missing-data mechanism. We generated 100 incomplete samples by removing the entries of Y with probability .8, when $Y = 1$ and those with probability .6 when $Y = 2$. The success rate of model folding was 69%, whereas ignorable imputation and data deletion selected the correct model in 71% cases. Bayesianly proper imputation had a success rate of 44% and selected $m_{y|\varnothing}$ in 7 samples, $m_{y|x_2}$ in 36 samples, and $m_{y|x_1,x_2}$ in 13 samples. Both model folding and ignorable imputation are more accurate than Bayesianly proper imputation, although the increased error rates suggests that the inappropriate assumption of ignorability, or partial ignorability, can jeopardize the inference accuracy.

We also ran a small simulation to investigate the robustness of both model folding and ignorable imputation when an inaccurate model for the missing-data mechanism is selected. The data in Table 3 were subjected to the same random deletion of Y entries described in the previous section, to produce 100 incomplete tables. On each incomplete dataset, we ran model

Table 4. Mean Values of the Log-Marginal Likelihood of $m_{y|\varnothing}$, $m_{y|x_1}$, $m_{y|x_2}$, and $m_{y|x_1,x_2}$ in the First Experiment

Method	$m_{y\|\varnothing}$	$m_{y\|x_1}$	$m_{y\|x_2}$	$m_{y\|x_1,x_2}$
Model folding	−275.2661	−255.2350	−271.6484	−252.4831
Ignorable imputation	−275.0549	−255.2079	−271.3363	−252.0717
Bayesianly proper imputation	−277.4527	−249.6824	−270.9706	−251.9317
Data deletion	−193.8340	−178.3028	−190.5554	−177.6876
Complete samples	−274.3278	−255.5745	−271.4870	−252.3128

NOTE: The last row reports average values on 100 simulated complete samples.

Table 5. Sample Generated From Model $m_{y|x_2}$ Specifying Dependence of Y on X_2

	X_1, X_2			
Y	(1,1)	(1,2)	(2,1)	(2,2)
1	101	105	82	79
2	46	41	27	26
3	16	7	39	22

folding and ignorable imputation assuming the three missing-data mechanisms $m_{r|x_1}$, $m_{r|x_2}$, and $m_{r|x_1,x_2}$. When the missing-data mechanism was modeled by $m_{r|x_1}$, model folding selected the correct model $m_{y|x_1,x_2}$ in 86 samples and selected $m_{y|x_2}$ in 14 samples. Similarly, ignorable imputation selected $m_{y|x_1,x_2}$ in 88 samples and $m_{y|x_2}$ in 12 samples. Model folding gave the same results when the model for the missing-data mechanism was $m_{r|x_1,x_2}$, whereas ignorable imputation selected the wrong $m_{y|x_2}$ model in only 10 samples. With the missing-data mechanism modeled by $m_{r|x_2}$, model folding selected $m_{y|x_1,x_2}$ in 81 samples and $m_{y|x_2}$ in 19 samples, whereas ignorable imputation selected $m_{y|x_1,x_2}$ in 88 samples and $m_{y|x_2}$ in 12 samples. This small experiment suggests that both model folding and ignorable imputation are robust to an inaccurate model of the missing-data mechanism. However, using the wrong missing-data mechanism can decrease the accuracy of both methods, particularly that of model folding.

6.3 Bias

The data in Table 5 were randomly generated from the model $m_{y|x_2}$ specifying the dependence of Y on X_2, conditional on a sample generated from the model of association between X_1 and X_2. The data were used to compute the marginal likelihood of the four models $m_{y|\varnothing}$, $m_{y|x_1}$, $m_{y|x_2}$, and $m_{y|x_2,x_3}$ and the logarithm of the Bayes factors of $m_{y|x_2}$ versus $m_{y|\varnothing}$, $m_{y|x_1}$, and $m_{y|x_2,x_3}$ are 7.11, 11.1, and 12.33. In each case we used symmetric hyper-Dirichlet prior distributions with $\alpha = 1$, so that the marginal likelihood of each model is that in (7). Thus under uniform prior probabilities on the model space $\mathcal{M}_{y|x_c}$, $m_{y|x_2}$ is selected conditional on the data. The complete data were subjected to a random deletion of Y entries with the same procedure described in the first experiment. In this case, the distribution of R was randomly chosen to be $p(R = 1|X_1 = 1, X_2 = 1) = .2$, $p(R = 1|X_1 = 1, X_2 = 2) = .4$, $p(R = 1|X_1 = 2, X_2 = 1) = .6$, and $p(R = 1|X_1 = 2, X_2 = 2) = .1$. The deletion process was repeated 50 times, and in each incomplete sample we used model folding, ignorable imputation, Bayesianly proper imputation, and data deletion to compute the Bayes factors of $m_{y|x_2}$ versus $m_{y|\varnothing}$, $m_{y|x_1}$, and $m_{y|x_2.x_3}$. In model folding, incomplete data were distributed using the estimates computed from the saturated model $m_{y|x_1,x_2}$, with $\beta^3_{kj} = .125$ and $\alpha = 1$. Similarly, ignorable imputation replaced missing data by values generated from the predictive distribution of Y, conditional on $m_{y|x_1,x_2}$ and the observed data y_o.

Figure 2 depicts the results. Model folding selected the correct model $m_{y|x_2}$ in 47 samples and selected $m_{y|\varnothing}$ is 3 samples. Ignorable imputation selected $m_{y|x_2}$ in 48 samples and $m_{y|\varnothing}$ in 2 samples. Bayesianly proper imputation selected $m_{y|x_2}$ in 37 samples, $m_{y|\varnothing}$ in 12 samples, and $m_{y|x_1,x_2}$ in 1 sample. Ignoring the missing data performed terribly, leading to selection of $m_{y|\varnothing}$ in all 50 incomplete samples. The average values of the logarithm of the Bayes factor of $m_{y|x_2}$ versus $m_{y|\varnothing}$, $m_{y|x_1}$, and $m_{y|x_2.x_3}$ were 10.79, 7.73, and 11.19 for model folding; 10.17, 8.21, and 11.29 for ignorable imputation; 10.19, 2.73, and 10.29 for Bayesianly proper imputation; and −2.54, .53, and 6.66 for data deletion. The results confirm the accuracy of model folding and ignorable imputation (which perform almost equally), the potential bias of Bayesianly proper imputation, and, in particular, of data deletion. At least in this example, there is no evidence that the model chosen for imputation, (i.e., $m_{r|x_1,x_2}$) should bias model selection.

7. APPLICATION

In this section we model the incomplete dataset given in Table 1. In Example 3 we showed that the missing-data mechanism is only partially ignorable, because the distribution of R is a function of children's age. Assuming independence of age and gender on the missing-data mechanism, the association between these two variables can be modeled regardless of the missing-data mechanism. Using (7) with $\alpha = 8$ and symmetric hyper-Dirichlet, the log-Bayes factor of the model of independence against the model of association is 2.8. Thus age and gender appear to be independent.

Next, we proceed by modeling the incomplete data. We adopt symmetric hyper-Dirichlet prior distributions on the parameters θ^c_{kj} and ψ_k, with precisions $\alpha = 8$ and $\beta = 1$. Model folding returns the values $\log(p(y|m_{y|\varnothing})) = -2447$, $\log(p(y|x_2, m_{y|x_2})) = -2448$, $\log(p(y|x_1, m_{y|x_1})) = -2400$, and $\log(p(y|x_1, x_2, m_{y|x_1,x_2})) = -2410$. Model $m_{y|x_1}$ has the maximum marginal likelihood, followed by the saturated model. Hence, with uniform probabilities on the model space, we select $m_{y|x_1}$, so we conclude that being overweight is related only to the age of the children. Ignorable imputation produces identical conclusions. Disregarding nonresponse gives $\log(p(y_o|m_{y|\varnothing})) = -1695.352$, $\log(p(y_o|x_2, m_{y|x_2})) = -1697.261$, $\log(p(y_o|x_1, m_{y|x_1})) = -1685.701$, and $\log(p(y_o|x_1, x_2, m_{y|x_1,x_2})) = -1690.491$, so that $m_{y|x_1}$ would be selected, but with a different strength of evidence. Bayesianly proper imputation, based on 10 replications, is very unstable and leads to results favoring either model $m_{y|x_2}$ or the saturated model.

Table 7 reports the model folding estimates of the obesity rate among respondents and nonrespondents. The former are calculated by disregarding missing data, and the latter are given by $\hat{\phi}^2_{kj}$. For comparison, the table reports the estimates derived by Park and Brown (1994) under the assumptions that the missing-data mechanism is not ignorable and that both gender and age are significant factors. The results are comparable overall. Note that the estimates of the obesity rate among nonrespondents is slightly lower than the obesity rate among respondents. If the assumption that overweight children tend to hide their status were true, then we would expect the obesity rate to be higher among nonrespondents than among respondents. Hence the nonignorability assumption appears to be doubtful.

We conducted the analysis assuming the missing-data mechanism learned for the variable R. We also analyzed the data by assuming that the variable R depends

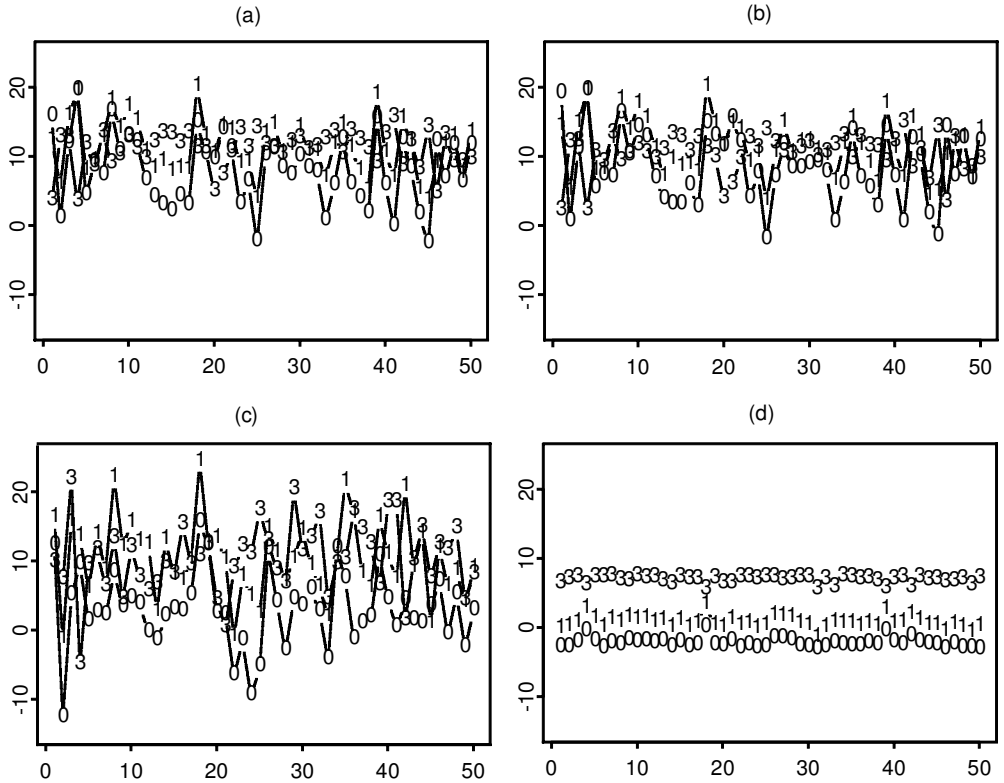

Figure 2. Logarithm of Bayes Factors of $m_{y|x_2}$ Versus $m_{y|\varnothing}(0)$, $m_{y|x_1}(1)$, and $m_{y|x_2,x_3}(3)$ in the 50 Incomplete Samples Generated From Data in Table 5. (a) Model folding; (b) ignorable imputation; (c) Bayesianly proper imputation; (d) data deletion.

on gender and by assuming that the variable R depends on both gender and age. In the first case, model folding estimates the marginal likelihood as $\log(p(y|m_{y|\varnothing})) = -2452$, $\log(p(y|x_2, m_{y|x_2})) = -2452$, $\log(p(y|x_1, m_{y|x_1})) = -2402$, and $\log(p(y|x_1, x_2, m_{y|x_1,x_2})) = -2410$. In the second case, the estimates are $\log(p(y|m_{y|\varnothing})) = -2422$, $\log(p(y|x_2, m_{y|x_2})) = -2424$, $\log(p(y|x_1, m_{y|x_1})) = -2401$, and $\log(p(y|x_1, x_2, m_{y|x_1,x_2})) = -2410$. Ignorable imputation produced comparable results. In both cases, the estimates rank the four models in the same way, and $m_{y|x_1}$ is always the most likely model. This result again suggests a robustness of both model folding and ignorable imputation to misspecification of the missing-data mechanism. Because model folding uses the m_{x_s} model to estimate the probability distribution of the nonrespondents, we also analyzed the data assuming an association between age and gender. The estimates of the marginal likelihood are $\log(p(y|m_{y|\varnothing})) = -2422$, $\log(p(y|x_2, m_{y|x_2})) = -2424$, $\log(p(y|x_1, m_{y|x_1})) = -2405$, and $\log(p(y|x_1, x_2, m_{y|x_1,x_2})) = -2410$. Thus again $m_{y|x_1}$ would be selected, but with a weaker evidence than when compared to $m_{y|\varnothing}$ or $m_{y|x_2}$.

8. CONCLUSIONS

In this article we have provided a new approach to Bayesian selection of decomposable models with incomplete data. We have shown that when only one variable is partially observed and the missingness probability is independent of the variables fully observed in the dataset, the missing-data mechanism is ignorable for Bayesian model selection and missing data can be ignored. When the missingness probability is a function of the variables fully observed, then the missing-data mechanism is only partially ignorable, and we described ignorable imputation and model folding for proper model selection. Both methods reconstruct a complete sample that takes into account the missing-data mechanism, thus following suggestions given by several authors to consider the variables responsible for

Table 6. Distribution of Obese Children Among Respondent and Nonrespondent Using Model Folding Under the Assumption That Only Age Affects the Probability of Being Overweight and the Smoothing Bayesian Method When Both Age and Gender Affects the Obesity Rate

		Model folding		*Not ignorable*	
Age	*Gender*	*Respondent*	*Nonrespondent*	*Respondent*	*Nonrespondent*
Young	Male	.1546	.1539	.1504	.1408
	Female	.1546	.1539	.1572	.1472
Old	Male	.2281	.2238	.2240	.2093
	Female	.2281	.2238	.2403	.2106

the missing data in model selection. Empirical evaluations showed that both methods appear to be very accurate, although further work is needed to make them the standard approach to Bayesian modeling of incomplete data.

The results presented herein are restricted to situations where only one variable is partially observed. The challenge now is to generalize the results presented here to situations in which a whole subset of variables $\mathcal{X}_m$ is partially observed. If we assume that the missingness probabilities can be a function only of the variables $\mathcal{X}_s \setminus \mathcal{X}_m$, which are fully observed, then we can apply a procedure generalizing that discussed in Section 4 to decide whether the missing-data mechanism is ignorable or partially ignorable. Ignorability again will ensure that model selection can be carried out without considering the missing-data mechanism. From a computational standpoint, this can be done by using the first-order approximation of each model marginal likelihood, which depends on estimates of the parameters computed with either the EM algorithm (Dempster, Laird, and Rubin 1977) or Gibbs sampling (Spiegelhalter et al. 1996). An open question is, however, whether model selection can be still carried out by exploiting the decomposability of the models. Partial ignorability will pose further problems, and an interesting hypothesis is to see whether modeling can be carried out by finding, for each model of dependency explored, the minimum model for which the missing-data mechanism is ignorable and then use a generalization of ignorable imputation to reconstruct a completion of the complete sample to estimate the marginal likelihood.

[Received February 1999. Revised March 2001.]

REFERENCES

Baker, S. G., and Laird, N. M. (1988), "Regression Analysis for Categorical Variables With Outcome Subject to Nonignorable Nonresponse," *Journal of the American Statistical Association*, 83, 62–69.

Bernardo, J. M., and Smith, A. F. M. (1994), *Bayesian Theory*, New York: Wiley.

Dawid, A. P. (1979), "Conditional Independence in Statistical Theory" (with discussion). *Journal of the Royal Statistical Society*, Ser. B, 41, 1–31.

——— (1980), "Conditional Independence for Statistical Operation," *The Annals of Statistics*, 8, 598–617.

Dawid, A. P., and Lauritzen, S. L. (1993), "Hyper-Markov Laws in the Statistical Analysis of Decomposable Graphical Models," *The Annals of Statistics*, 21, 1272–1317. Corr. (1995), 23, 1864.

Dempster, A. P., Laird, D., and Rubin, D. (1997), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysis*, London: Chapman and Hall.

Good, I. J. (1968), *The Estimation of Probability: An Essay on Modern Bayesian Methods*. Cambridge, MA: MIT Press.

Heckerman, D., Geiger, D., and Chickering, D. M. (1995), "Learning Bayesian Networks: The Combinations of Knowledge and Statistical Data," *Machine Learning*, 20, 197–243.

Heitjan, D. F. (1994), "Ignorability in General Incomplete-Data Models," *Biometrika*, 81, 701–708.

Heitjan, D. F., and Basu, S. (1996), "Distinguishing 'Missing at Random' and 'Missing Completely at Random'," *The American Statistician*, 50, 207–213.

Heitjan, D. F., and Rubin, D. B. (1991), "Ignorability and coarse data," *The Annals of Statistics*, 19, 2244–2253.

Kass, R. E., and Raftery, A. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Lauritzen, S. L. (1996), *Graphical Models*, Oxford, U.K.: Oxford University Press.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: Wiley.

Park, T., and Brown, M. B. (1994), "Models for Categorical Data With Nonignorable Nonresponse," *Journal of the American Statistical Association*, 89, 44–52.

Raftery, A. E. (1995), "Bayesian Model Selection in Social Research" (with discussion), *Sociological Methodology*, 25, 111–196.

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

——— (1987), *Multiple Imputation for Nonresponse in Survey*, New York: Wiley.

——— (1996), "Multiple Imputation After 18 Years," *Journal of the American Statistical Association*, 91, 473–489.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, London: Chapman and Hall.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Sebastiani, P., and Ramoni, M. (2000), "Bayesian Inference With Missing Data Using Bound and Collapse," *Journal of Computational and Graphical Statistics*, 9, 779–800.

Spiegelhalter, D. J., Thomas, A., and Best, N. G. (1996), "Computation on Bayesian Graphical Models" (with discussion), in *Bayesian Statistics 5*, Oxford, U.K.: Oxford University Press, pp. 407–425.

Tanner, M. A. (1996), *Tools for Statistical Inference* (3rd ed.), New York: Springer.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association*, 82, 528–550.

Wasserman, L. (1999), "Bayesian Model Selection and Model Averaging," *Journal of Mathematical Psychology*, 44, 92–107.