KNOWLEDGE MEDIA INSTITUTE

Bayesian Inference with Missing Data Using Bound and Collapse

Paola Sebastiani *Marco Ramoni*

KMI-TR-58

November 1997

Bayesian Inference with Missing Data Using Bound and Collapse

Paola Sebastiani
City University

Marco Ramoni
The Open University

Abstract

Current Bayesian methods to estimate conditional probabilities from samples with missing data pose serious problems of robustness and computational efficiency. This paper introduces a new method, called Bound and Collapse (BC), able to overcome these problems. When no information is available on the pattern of missing data, BC returns *bounds* on the possible estimates consistent with the available information. These bounds can be then collapsed to a point estimate using information about the pattern of missing data, if any. Approximations of the variance and of the posterior distribution are proposed, and their accuracy is compared to approximations based on alternative methods in a real data set of polling data subject to non-response.

Keywords: Bayesian Estimates; Bound and Collapse; Gibbs Sampling; Ignorability; Imputation; Missing Data.

Reference: KMi Technical Report KMi-TR-58, December 1997.

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.

Contents

1 Introduction **2**
1.1 Missing-Data Mechanisms 2
1.2 Bayesian Analysis of Incomplete Contingency Tables 3
1.3 Scope of This Paper 4

2 Background **5**
2.1 Complete Samples 5
2.2 Incomplete Samples 7

3 Bound and Collapse **11**
3.1 Bound 12
3.2 Collapse 13

4 Inference **15**

5 An Application to Polling Data **16**
5.1 Bound 17
5.2 Collapse 19
5.2.1 Missing at Random 19
5.2.2 Non Ignorable Non-response 20

6 Conclusions **23**

References **24**

A Proof of Theorems **25**

Probability is relative, in part to our ignorance, in part to our knowledge.

P.S. DE LAPLACE (1840)

1. Introduction

On 9 April 1992, the Conservative Party won its fourth consecutive British General Election and inflicted a spectacular *débâcle* on the polling industry: on election day, a final poll of four major polling companies suggested a Labour lead of 0.9 %. The Conservative party won by 7.6 %. "An 8.5 % error, the largest ever." (Butler and Kavanagh, 1992, page 135). An inquiry conducted by the Market Research Society found out that one of the major causes of the disaster, accounting for more than the 2 % of the error, laid in the voting intentions not expressed during the polls. Answers like *don't know* and refusal to answer were simply discarded from the samples, thus assuming missing data to be *ignorable*. This assumption was fatal to the polls: follow up surveys revealed that Tories gained disproportionately among those who made up their mind at the last minute and that, in general, Conservative voters were less inclined to declare their voting intention.

1.1 Missing-Data Mechanisms

Incomplete samples challenge the analyst because they hit the very foundations of modern statistics (Copas and Li, 1997): missing data can affect the randomness of a sample, thus removing the grounds for the use of most statistical methods. Our current understanding of the effect of missing data on the representativity of a sample is based on the classification of missing-data mechanisms proposed by Rubin (1976) and further developed by Little and Rubin (1987) and Gelman *et al.* (1995). Consider a sample classified according to the values of two categorical variables X and Y, in which X, the independent variable, is always observed and Y, the response variable, is subject to non-response. We will denote a non-response by $Y =?$, and an incomplete case by $(X = i, Y =?)$. This simple two-variable model may be easily extended to include more general cases, in which a set of always observed variables affects a variable subject to non-response. This scenario is common in various statistical applications, such as prospective studies, controlled experiments, supervised classification in machine learning and data mining tasks and, needless to say, survey samples, where Y may be the voting intention and missing data may be due to genuine uncertainty or to a decision not to reveal the voting intention.

The classification of missing-data mechanisms depends on whether the probability of $Y =?$ depends on the state of Y and/or X. If this probability depends on X but not on Y, then data are said to be *missing at random* (MAR), and observed values of Y are not representative of the complete sample as a whole, but they are so, when considered within categories of X. A special case of this situation is realized when the probability of $Y =?$ is neither dependent on Y nor on X. In this case, data are said to be *missing completely at random* (MCAR) and the observed values of Y are a representative sub-sample of the complete but unknown original sample, since observed and unobserved entries are generated

by the same mechanism. When data are MAR or MCAR, the missing-data mechanism is *ignorable*, in the sense that inference does not depend on it. When the probability of $Y =?$ depends on both X and Y, the missing-data mechanism is said to be *not ignorable* (NI), and the resulting incomplete sample is no longer representative under any respect.

This classification provides a powerful framework to understand the behavior of incomplete samples but, unfortunately, it does leave open the problem of how to handle non responses, since ignorability of the mechanism underlying non responses does not imply that missing data can be simply ignored. The Bayesian approach provides some theoretical tools to handle this problem.

1.2 Bayesian Analysis of Incomplete Contingency Tables

As long as the sample is complete, Bayesian conjugate analysis discloses a simple way to estimate the conditional distributions of Y given X, and the marginal distribution of Y. Given that the sampling model is a multinomial distribution with parameter vector $\boldsymbol{\theta}$, and $\theta_{ij} = p(X = i, Y = j|\boldsymbol{\theta})$, a conjugate prior for $\boldsymbol{\theta}$ is a Dirichlet distribution, and the posterior distribution of $\boldsymbol{\theta}$ is still Dirichlet (Lindley, 1964). From the joint posterior distribution of $\boldsymbol{\theta}$, posterior distributions of the conditional probabilities of Y given X, and of the marginal probabilities of Y can be easily derived. Measures of association between X and Y can also be inferred. Unfortunately, this simple way is precluded when the observed sample contains incomplete cases, because the amount of information about Y and X is unbalanced. An exception is when the missing-data mechanism is MCAR, and the frequencies of complete cases are large. In this case, incomplete cases can be safely ignored and Bayesian conjugate analysis carried out on the complete cases available in the sample.

When the missing-data mechanism is MAR, the posterior distribution of the conditional probabilities of Y given X is still conjugate and therefore estimates of the probabilities of (X, Y) and Y can be easily computed, as well as their posterior variance. However, posterior distributions of the joint probabilities of (X, Y) and of the marginal probabilities of Y do not have simple expressions. A nowaday popular solution is to resort to MCMC methods, such as Gibbs Sampling (Geman and Geman, 1984), and treat missing entries as unknown parameters. The procedure will eventually return a sample from the posterior distribution of quantities of interest, from which empirical estimates and credibility intervals can be computed. A detailed description appears in Gilks and Roberts (1996).

The task becomes even more difficult for NI missing-data mechanisms. In this case, observed data do not carry information about non responses, and exogenous information about the distribution $\boldsymbol{\phi}$ of non-response is needed to be used in an *Imputation-based* analysis: missing data are simulated from $\boldsymbol{\phi}$, and the statistical analysis is based on the filled-in sample. A recent review is in Gelman *et al.* (1995, Ch. 17).

Gibbs Sampling and Imputation-based approaches share three main drawbacks:

1. They both rely on the assumption that information about the missing-data mechanism is available, but, unfortunately, this is not always the case.

2. They provide measures of reliability of the estimates which fail to take into full account missing data, so that sampling variability and extra-variability due to non-responses

are mixed up.

3. The computational cost of both approaches is a function of the number of missing data. Gibbs Sampling treats missing entries as unknown parameters, so that convergence rate of the method will be a decreasing function of the number of missing entries. In Imputation-based methods, the precision of the estimates increases as the number of simulated complete samples does, and the computational cost of each simulation is an increasing function of the number of missing data.

1.3 Scope of This Paper

This paper introduces a new methodological framework, called *Bound and Collapse* (BC), to achieve three main objectives: the definition of an estimation method from incomplete samples robust with respect to the pattern of missing data, the identification of reliability measures able to account for the presence of missing data in a sample, and the development of efficient computational methods to perform these calculations.

The intuition behind BC is that, when no information about the missing-data mechanism is available, an incomplete sample is still able to *bound* the set of possible estimates within an interval defined by extreme distributions. Along this approach, a complete sample is just a special case in which the available information is sufficient to constrain the set of possible estimates to a single point. When information about the mechanism yielding missing data is available, it is encoded in a probabilistic model of non-response and used to select a single estimate. The second step of BC *collapses* each interval computed in the bound step to a single value via a convex combination of the extreme estimates with weights depending on the assumed pattern of missing data. These point estimates can be then used to approximate the posterior distribution of parameters of interest. Different missing-data mechanisms, such as MAR or MCAR, can be easily represented by modeling non-response from complete cases in the sample and, in this case, BC returns a generalized version of the Maximum Likelihood estimates. Under a general missing-data mechanism, BC estimates are the expected Bayesian estimates, conditional on the model for non-response $\boldsymbol{\phi}$.

The fundamental character of BC is to represent explicitly and separately the information conveyed by the sample and the assumptions about the pattern of missing data. This character provides a general estimation method which is independent of any particular missing-data mechanism. Bounds on the possible estimates represent uncertainty due to missing data and the width of intervals computed in the bound step can be regarded as a measure of the quality of information conveyed by the incomplete sample on the estimates. By computing bounds, the uncertainty due to missing data is therefore retained in the analysis, so that sampling variability and uncertainty due to non-responses are independently computed and separately represented. A further advantage of BC is its computational cost: for each conditional distribution, BC reduces the computational complexity of the analysis to one exact updating for each state of the response variable in the bound step, and a convex combination in the collapse step. The computational complexity of BC is therefore only function of the number of states of Y, and its cost is independent of the number of missing data.

Compared to current approaches, in which the missing-data mechanism is modeled and hierarchical log-linear models are used (Little and Rubin, 1987; Park and Brown, 1994; Rubin *et al.*, 1995), BC exploits, in the collapse step, a model for non-responses which is function of the missing-data mechanism, and the resulting inference is conditional on a particular model for non-response. The computational simplicity of BC allows different models for non-responses to be efficiently represented, so that sensitivity of the conclusions to the assumed pattern of missing data can be fast evaluated. Thus, BC provides a general framework for the sensitivity analysis approach to incomplete samples advocated by Kadane (1993) and Kadane and Terrin (1997) in several applications. Furthermore, marginal inference can be easily obtained by averaging out results computed for different models of non-response.

The reminder of this paper is structured as follows. Section 2 will review some theoretical and computational issues relevant to the development and the presentation of BC. BC will be presented in Section 3 and inference methods based on it will be described in Section 4. Features of BC will be illustrated in Section 5, using the polling data from the 1992 British General Elections, and they will be summarized in Section 6.

2. Background

We begin by describing standard Bayesian conjugate analysis of complete multinomial data and then we will report some fundamental results for the analysis of incomplete samples.

2.1 Complete Samples

Let $X_1, ..., X_k$ and Y be categorical random variables. For simplicity of notation, we shall write $X = (X_1, ..., X_k)$, and denote by $X = i$, $(i = 1, \cdots, r)$, and $Y = j$, $(j = 1, \cdots, c)$ the states of the variables X and Y. We assume that (X, Y) has a multinomial distribution with probabilities $\theta_{ij} = p(X = i, Y = j|\boldsymbol{\theta})$, so that $\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \cdots, \theta_{rc}) = (\theta_{ij})$ $(\theta_{ij} > 0$, for all i and j, and $\sum_{ij} \theta_{ij} = 1)$ parameterizes the joint distribution of (X, Y). The standard conjugate prior for $\boldsymbol{\theta}$ is a Dirichlet distribution $D(\boldsymbol{\alpha})$, with $\boldsymbol{\alpha} = (\alpha_{11}, \cdots, \alpha_{rc})$, whose density function is:

$$p(\boldsymbol{\theta}) = \prod_{i=1}^{r} \prod_{j=1}^{c} \frac{\Gamma(\alpha)}{\Gamma(\alpha_{ij})} \theta_{ij}^{\alpha_{ij}-1}$$

with $\alpha_{ij} \geq 0$ for all i and j, and $\alpha = \sum_{ij} \alpha_{ij}$. Thus, the prior expectation $E(\theta_{ij})$ is $p(X = i, Y = j) = \alpha_{ij}/\alpha$ and a measure of the uncertainty about $p(X = i, Y = j)$ is the prior variance $V(\theta_{ij}) = E(\theta_{ij})\{1 - E(\theta_{ij})\}/(\alpha + 1)$. Since $V(\theta_{ij})$ is a decreasing function of α, for fixed $E(\theta_{ij})$, the quantity α is sometimes called *precision*. The parameterization of the joint probability distribution of (X, Y) induces parameterizations of the marginal distributions of X and Y, and of the r and c conditional distributions of $Y|x = i$ and $X|y = j$. The result is known (Fang *et al.*, 1990, page 19) but, since it is crucial for the development of the subsequent theory, it is recalled here without proof.

Theorem 1 *Let* $\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \cdots, \theta_{rc}) \sim D(\boldsymbol{\alpha})$, $\boldsymbol{\alpha} = (\alpha_{11}, \cdots, \alpha_{rc})$, *and for* $i = 1, \cdots, r$, $j = 1, \cdots, c$ *define*

$$\theta_{i+} = \sum_j \theta_{ij},\ \theta_{+j} = \sum_i \theta_{ij},\ \theta_{j|i} = \frac{\theta_{ij}}{\theta_{i+}},\ \theta_{i|j} = \frac{\theta_{ij}}{\theta_{+j}},\ \alpha_{i+} = \sum_j \alpha_{ij},\ \alpha_{+j} = \sum_i \alpha_{ij}.$$

Then

$$\begin{array}{lcl} \boldsymbol{\theta}_I & = & (\theta_{1+}, \cdots, \theta_{r+}) \sim D(\boldsymbol{\alpha}_I), \boldsymbol{\alpha}_I = (\alpha_{1+}, \cdots, \alpha_{r+}), \\ \boldsymbol{\theta}_J & = & (\theta_{+1}, \cdots, \theta_{+c}) \sim D(\boldsymbol{\alpha}_J), \boldsymbol{\alpha}_J = (\alpha_{+1}, \cdots, \alpha_{+c}), \\ \boldsymbol{\theta}_{I|j} & = & (\theta_{1|j}, \cdots, \theta_{r|j}) \sim D(\boldsymbol{\alpha}_{I|j}), \boldsymbol{\alpha}_{I|j} = (\alpha_{1j}, \cdots, \alpha_{rj}), \\ \boldsymbol{\theta}_{J|i} & = & (\theta_{1|i}, \cdots, \theta_{c|i}) \sim D(\boldsymbol{\alpha}_{J|i}), \boldsymbol{\alpha}_{J|i} = (\alpha_{i1}, \cdots, \alpha_{ic}). \end{array}$$

Furthermore $\boldsymbol{\theta}_I$ *and* $\boldsymbol{\theta}_{J|i}$ *are marginally independent, as well as* $\boldsymbol{\theta}_J$ *and* $\boldsymbol{\theta}_{I|j}$.

Suppose that we wish to infer $\boldsymbol{\theta}$ from a random sample $S = \{s_1, \cdots, s_n\}$ of independent cases given $\boldsymbol{\theta}$. Data can be classified in a $r \times c$ contingency table, where n_{ij} denotes the frequency of $(X = i, Y = j)$ in S. Let $\boldsymbol{n}$ be the vector $(n_{11}, \cdots, n_{rc})$. The joint probability of the sample is:

$$p(S|\boldsymbol{\theta}) = \prod_{i=1}^{r} \prod_{j=1}^{c} \theta_{ij}^{n_{ij}},$$

and, by conjugacy, the posterior distribution of $\boldsymbol{\theta}$ is still Dirichlet, with updated hyperparameters $\alpha_{ij} + n_{ij}$:

$$\boldsymbol{\theta}|S \sim D(\alpha_{11} + n_{11}, \cdots, \alpha_{rc} + n_{rc}) \equiv D(\boldsymbol{\alpha} + \boldsymbol{n}).$$

Thus, the posterior precision is $\alpha + n$ and the posterior expectation of θ_{ij} provides a point estimate of the joint probability of $(X = i, Y = j)$:

$$\hat{\theta}_{ij} = E(\theta_{ij}|S) = \frac{\alpha_{ij} + n_{ij}}{\alpha + n}.$$

The posterior variance of θ_{ij}

$$V(\theta_{ij}|S) = \frac{E(\theta_{ij}|S)\{1 - E(\theta_{ij}|S)\}}{\alpha + n + 1}$$

gives a measure of the posterior uncertainty about $\hat{\theta}_{ij}$. A credibility interval for θ_{ij} can be calculated using $\theta_{ij}|S \sim D(\alpha_{ij} + n_{ij}, \alpha + n - \alpha_{ij} - n_{ij})$, (Wilks, 1963, Ch 7.7).

The fact that the joint posterior distribution of $\boldsymbol{\theta}$ is still Dirichlet leads to a simple updating rule for the distributions of $\boldsymbol{\theta}_I$, $\boldsymbol{\theta}_J$, $\boldsymbol{\theta}_{I|j}$ and $\boldsymbol{\theta}_{J|i}$. Denote row and column totals by $n_{i+} = \sum_j n_{ij}$ and $n_{+j} = \sum_j n_{ij}$, and let $\boldsymbol{n}_I = (n_{i+})$, $\boldsymbol{n}_J = (n_{+j})$, $\boldsymbol{n}_{J|i} = (n_{i1}, \cdots, n_{ic})$, and $\boldsymbol{n}_{I|j} = (n_{1j}, \cdots, n_{rj})$. Theorem 1 ensures that $\boldsymbol{\theta}_I|S \sim D(\boldsymbol{\alpha}_I + \boldsymbol{n}_I)$, $\boldsymbol{\theta}_J|S \sim D(\boldsymbol{\alpha}_J + \boldsymbol{n}_J)$, and $\boldsymbol{\theta}_{J|i}|S$ and $\boldsymbol{\theta}_{I|j}|S$ are $D(\boldsymbol{\alpha}_{J|i} + \boldsymbol{n}_{J|i})$ and $D(\boldsymbol{\alpha}_{I|j} + \boldsymbol{n}_{I|j})$. Furthermore, $\boldsymbol{\theta}_I|S$ and $\boldsymbol{\theta}_{J|i}|S$, for all i, are still marginally independent, and so are $\boldsymbol{\theta}_J|S$ and $\boldsymbol{\theta}_{I|j}|S$, for all j. Therefore Bayesian estimates of $p(X = i)$ and $p(Y = j|X = i)$ are

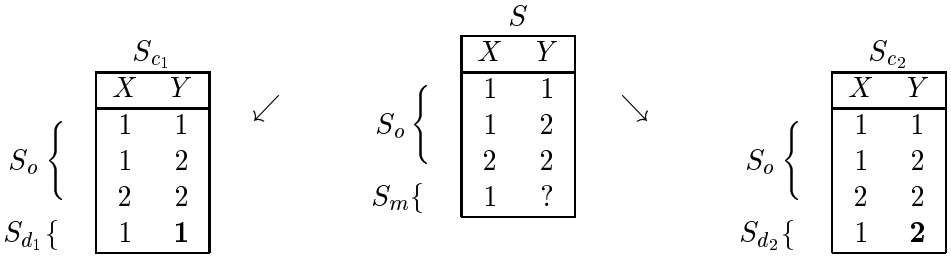

Figure 1: Possible completions of an incomplete sample with two binary variables.

$$\hat{\theta}_{i+} = E(\theta_{i+}|S) = \frac{\alpha_{i+} + n_{i+}}{\alpha + n} \text{ and } \hat{\theta}_{j|i} = E(\theta_{j|i}|S) = \frac{\alpha_{ij} + n_{ij}}{\alpha_{i+} + n_{i+}}.$$

Remark 1 An alternative approach parameterize the marginal distribution of X and the r conditional distributions of $Y|X = i$, and assume independence of the parameters θ_I and $\theta_{J|i}$ for all i. These assumptions are also known as *local* and *global* independence (Spiegelhalter and Lauritzen, 1990) in the field of Bayesian Belief Networks. We merely note here that, as long as the hyperparameters of the prior distributions are chosen consistently, exact Bayesian analysis leads to identical conclusions under both parameterizations (Geiger and Heckerman, 1997).

2.2 Incomplete Samples

Exact Bayesian analysis presents no difficulties as long as the sample is complete. Suppose now that some of the entries on the variable Y are reported as unknown. Let $S = (S_o, S_m)$, where S_o and S_m respectively denote the sample with complete observations and the one with unknown entries on Y, and let S_{c_i} be a possible completion of S_o. Thus $S_{c_i} = (S_o, S_{d_i})$, where S_{d_i} is a possible distribution of the unclassified cases in S_m. Figure 1 shows the two possible completions of S obtained by distributing the incomplete case $(X = 1, Y = ?)$, when $c = 2$. We continue to assume that n_{ij} is the frequency of complete cases and let m_i be the frequency of cases $(X = i, Y = ?)$. Thus, $n = \sum_{ij} n_{ij}$ is the number of cases completely observed, $m = \sum_i m_i$ is the number of cases partially observed, and $n + m$ is now the sample size. Following Little and Rubin (1987), we can represent the incomplete sample in a $r \times (c + 1)$ contingency table, in which the $c + 1$-th column represents the frequency of unknown cases for each category of X. Table 1 gives an example.

By the Total Probability Theorem, the exact posterior distribution of $\boldsymbol{\theta}$ is a mixture of Dirichlet distributions weighted by the probabilities of possible completions of S. Thus, the posterior density is

$$p(\boldsymbol{\theta}|S) = \sum_{c_i} p(\boldsymbol{\theta}|S_{c_i}) p(S_{c_i}|S),$$

with weights $p(S_{c_i}|S) \propto p(S_{d_i}|S_o)$, that can be computed if information about the missing-

	Y			Y
X	1	$\cdots$	c	?
1	n_{11}	$\cdots$	n_{1c}	m_1
$\vdots$	$\vdots$		$\vdots$	$\vdots$
r	n_{r1}	$\cdots$	n_{rc}	m_r

Table 1: An incomplete sample represented in a contingency table.

data mechanism is available. Suppose that this information leads to the formulation of the probability of non-response

$$p(X = i, Y = j|Y =?, \boldsymbol{\theta}, \boldsymbol{\phi}) = \theta_{i+}\phi_{j|i},$$

where $\phi_{j|i} = p(Y = j|X = i, Y =?, \boldsymbol{\phi}, \boldsymbol{\theta})$. We can let $\phi_{j|i}$ depend explicitly on the missing-data mechanism a follows:

$$\begin{aligned} \phi_{j|i} &= p(Y = j|X = i, Y =?, \boldsymbol{\phi}, \boldsymbol{\theta}) \\ &= \frac{p(X = i, Y = j|\boldsymbol{\theta})}{p(X = i|\boldsymbol{\theta})} \frac{p(Y =?|X = i, Y = j, \boldsymbol{\phi}, \boldsymbol{\theta})}{\sum_j p(Y =?|X = i, Y = j, \boldsymbol{\theta}, \boldsymbol{\phi}) p(Y = j|X = i, \boldsymbol{\theta})} \\ &\propto \theta_{j|i} k_{ij}, \end{aligned} \tag{1}$$

where $k_{ij} \propto p(Y =?|X = i, Y = j, \boldsymbol{\phi}, \boldsymbol{\theta})$ describes the process that yields missing data.

A fully Bayesian approach regards $\boldsymbol{\phi}$ as a random vector, with prior density $p(\boldsymbol{\phi})$. Rubin (1976) shows that if the expectation of k_{ij}, conditional on the pattern of missing data and the observed cases, is constant, then the missing-data mechanism is ignorable. In other words, correct inference does not need to take into account the missing-data process. From Rubin's property, it follows that a stronger condition of ignorability of the missing-data mechanism is that the distribution of k_{ij} is independent of S. This stronger sufficient condition for ignorability easily translates into a condition of ignorability in terms of $\boldsymbol{\phi}$.

Once the probabilistic model for non-response is defined, the mixture weights of the exact posterior distribution are found to be:

$$\begin{aligned} p(S_{d_i}|S_o) &= \int p(S_{d_i}|S_o, \boldsymbol{\phi}, \boldsymbol{\theta}) p(\boldsymbol{\phi}, \boldsymbol{\theta}|S_o) d\boldsymbol{\phi} d\boldsymbol{\theta} \\ &= \int p(S_{d_i}|S_o, \boldsymbol{\phi}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|S_o) p(\boldsymbol{\phi}|S_o, \boldsymbol{\theta}) d\boldsymbol{\phi} d\boldsymbol{\theta}. \end{aligned} \tag{2}$$

The assumed missing-data mechanism shapes the probability $p(S_{d_i}|S_o)$ via $\boldsymbol{\phi}$. If the missing-data mechanism is MAR, the probability of an entry being missing does not depend on Y but may depend on X (when this probability is not dependent on X, missing data are MCAR.) We have $p(Y =?|X = i, Y = j, \boldsymbol{\phi}, \boldsymbol{\theta}) = p(Y =?|X = i, \boldsymbol{\phi}, \boldsymbol{\theta})$ and $k_{ij} = 1$ in (1), so that $\phi_{j|i} = \theta_{j|i}$. Ignorability of the missing-data mechanism (Little and Rubin, 1987) thus implies:

$$p(S_{d_i}|S_o) = \int p(S_{d_i}|\boldsymbol{\theta})p(\boldsymbol{\theta}|S_o)d\boldsymbol{\theta}. \tag{3}$$

For NI missing-data mechanisms, the probability of a missing observation on Y is generally function of X and Y, as given in (1), and (2) can be computed once a prior distribution on $\boldsymbol{\phi}$ is specified. Since the posterior distribution is a mixture over all completions of the sample consistent with the available information, the simplicity of conjugate analysis is lost when the sample contains unreported data. However some simplifications are still possible and some of the prior independence are retained. The first simplification concerns the posterior distribution of $\boldsymbol{\theta}_I$ and $\boldsymbol{\theta}_{J|I} = (\boldsymbol{\theta}_{J|1}, \cdots, \boldsymbol{\theta}_{J|r})$, and is given in the next theorem. The result is known: it is mentioned for instance by Forster and Smith (1998). The proof is reported in the Appendix.

Theorem 2 *Let S be an incomplete sample in which n_{ij} is the frequency of observed cases $(X = i, Y = j)$, and m_i is the frequency of cases $(X = i, Y =?)$. If $\boldsymbol{\theta} \sim D(\boldsymbol{\alpha})$, the posterior distribution of $\boldsymbol{\theta}_I$ is $D(\alpha_{1+} + n_{1+} + m_1, \cdots, \alpha_{r+} + n_{r+} + m_r) \equiv D(\boldsymbol{\alpha}_I + \boldsymbol{n}_I + \boldsymbol{m})$, and $\boldsymbol{\theta}_I$ and $\boldsymbol{\theta}_{J|I}$ are independent.*

Marginal independence of $\boldsymbol{\theta}_I$ and $\boldsymbol{\theta}_{J|I}$ is therefore retained in the posterior distribution even when the sample is incomplete, under both MAR and NI assumption. In both cases, the Bayesian estimate of θ_{i+} is

$$\hat{\theta}_{i+} = \frac{\alpha_i + n_i + m_i}{\alpha + n + m}. \tag{4}$$

When the missing-data mechanism is MAR, it can be shown that the joint distribution of $\boldsymbol{\theta}_{J|I}$ simplifies, and incomplete cases need not be taken into account in the inference about the conditional probabilities of $Y|X$. Therefore, missing data are ignorable for $\boldsymbol{\theta}_{J|I}$. The next result is mentioned, without proof, by Spiegelhalter and Lauritzen (1990). Our proof is in the Appendix.

Theorem 3 *Suppose that the missing-data mechanism is* MAR. *Then the distribution of $\boldsymbol{\theta}_{J|I}$ factorizes into a product of independent Dirichlet distributions $D(\boldsymbol{\alpha}_{J|i} + \boldsymbol{n}_{J|i})$.*

Thus, if the missing data mechanism is MAR, incomplete cases are ignorable for the Bayesian estimates of $\theta_{j|i}$, that reduce to:

$$\hat{\theta}_{j|i} = \frac{\alpha_{ij} + n_{ij}}{\alpha_{i+} + n_{i+}}. \tag{5}$$

The results in Theorems 2 and 3 exclude the possibility that the joint posterior distribution of $\boldsymbol{\theta}$ is Dirichlet, when the missing-data mechanism is MAR. The marginal posterior distribution of $\boldsymbol{\theta}_I$ is $D(\boldsymbol{\alpha}_I + \boldsymbol{n}_I + \boldsymbol{m})$. If the joint posterior distribution were a Dirichlet distribution, then $\alpha_{i+} + n_{i+} + m_i$ would be the posterior precision of $\boldsymbol{\theta}_{J|i}$, but this is excluded by Theorem 3 which ensures that the posterior precision of $\boldsymbol{\theta}_{J|i}$ is $\alpha_{i+} + n_{i+}$. The loss of conjugacy can therefore be attributed to the lack of balance in the information about X and Y conveyed by the sample. Conjugacy can be recovered if the missing-data mechanism

is MCAR and the frequencies of complete cases are large. In this case, we can assume that $m_i/m = (\alpha_i + n_i)/(\alpha + n)$ so that

$$\hat{\theta}_{i+} = \frac{\alpha_i + n_i + m_i}{\alpha + n + m} = \frac{\alpha_i + n_i}{\alpha + n}. \tag{6}$$

Thus, incomplete cases become ignorable also for estimation of θ_{i+}, and we can simply disregard incomplete cases, since they will not add any relevant information. Note that this simplification is never possible if data are MAR although the sample with complete cases is large. Information from incomplete cases is never ignorable for estimation of θ_{i+}, since (6) holds if and only if $m_i/m = (\alpha_i + n_i)/(\alpha + n)$.

Under the MAR assumption, it is however possible to compute posterior mean and variance of θ_{ij} and θ_{+j} in closed form. We will use $\theta_{ij} = \theta_{i+}\theta_{j|i}$ and $\theta_{+j} = \sum_{i=1}^{r} \theta_{i+}\theta_{j|i}$. By Theorems 2 and 3, and the results in Wilks (1963, Ch 7.7), we know that $\theta_{i+} \perp \theta_{j|i}$, $\theta_{i+} \sim D(\alpha_{i+} + n_{i+} + m_i, \alpha + n + m - n_{i+} - m_i)$, and $\theta_{j|i} \sim D(\alpha_{ij} + n_{ij}, \alpha_{i+} + n_{i+} - \alpha_{ij} - n_{ij})$. Thus, it is easy to show that:

$$\hat{\theta}_{ij} = E(\theta_{ij}|S) = \frac{\alpha_{ij} + n_{ij} + m_i \frac{\alpha_{ij} + n_{ij}}{\alpha_{i+} + n_{i+}}}{\alpha + n + m} \tag{7}$$

so that incomplete cases are not ignorable for estimation of θ_{ij}, and are distributed across categories of $Y|X = i$ according to the distribution of $\boldsymbol{\theta}_{J|i}$. Note that (7) generalize the MLEs given by Little and Rubin (1987) by adding the flattening constant α_{ij}. Similarly, it can be shown that

$$\hat{\theta}_{+j} = E(\theta_{+j}|S) = \sum_{i=1}^{r} \frac{\alpha_i + n_i + m_i}{\alpha + n + m} \frac{\alpha_{ij} + n_{ij}}{\alpha_i + n_i}. \tag{8}$$

If $m_i/m = (\alpha_i + n_i)/(\alpha + n)$, which holds when the missing-data mechanism is MCAR and the complete sample is large, then $E(\theta_{+j}|S) = (\alpha_{+j} + n_{+j})/(\alpha + n)$. By letting $V(\theta_{ij}|S) = E(\theta_{i+}^2|S)E(\theta_{j|i}^2|S) - E(\theta_{i+}|S)^2 E(\theta_{j|i}|S)^2$, where

$$E(\theta_{i+}^2|S) = \frac{\hat{\theta}_{i+}(1 - \hat{\theta}_{i+})}{\alpha + n + m} + \hat{\theta}_{i+}^2 \tag{9}$$

and

$$E(\theta_{j|i}^2|S) = \frac{\hat{\theta}_{j|i}(1 - \hat{\theta}_{j|i})}{\alpha_i + n_i + km_i} + \hat{\theta}_{j|i}^2, \tag{10}$$

the posterior variance of θ_{ij} can be then written as a function of the Bayesian estimates $\hat{\theta}_{i+}$ and $\hat{\theta}_{j|i}$. The exact posterior variance of θ_{+j} is

$$V(\theta_{+j}|S) = \sum_{i=1}^{r} V(\theta_{ij}|S) + \sum_{i=1}^{r} \sum_{h \neq i=1}^{r} Cov(\theta_{i+}\theta_{j|i}, \theta_{h+}\theta_{j|h}|S),$$

and

	MCAR	MAR	NI
$\boldsymbol{\theta}_{I+}$	CP	CP	CP
$\boldsymbol{\theta}_{J\|i}$	CP	CP	?
$\boldsymbol{\theta}$	CP(a)	EV	?
$\boldsymbol{\theta}_{+J}$	CP(a)	EV	?

Table 2: Inference for various missing-data mechanisms. CP means that the posterior distribution is Dirichlet, and thus conjugate to the prior. CP(a) means that conjugacy is recovered for large samples. EV means that only posterior mean and variance can be computed exactly, and ? means that there is no simple expression for the posterior distribution that is a mixture of Dirichlet distributions.

$$Cov(\theta_{i+}\theta_{j|i}, \theta_{h+}\theta_{j|h}|S) = E(\theta_{j|i}|S)E(\theta_{j|h}|S)Cov(\theta_{i+}, \theta_{h+}|S),$$

where

$$Cov(\theta_{i+}, \theta_{h+}|S) = -\frac{\hat{\theta}_{i+}\hat{\theta}_{h+}}{\alpha + n + m + 1}.$$

However, there is no simple expression for the posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{+J}$, so that inference can be based on approximations using, for instance, MCMC methods or moment-matching approximations.

As far as we know, there are no similar known simplifications for general NI missing-data mechanisms. One exception is the next result that is proved in the Appendix. Let $\boldsymbol{\phi}$ be written as $(\boldsymbol{\phi}_{J|1}, \ldots, \boldsymbol{\phi}_{J|r})$, where $\boldsymbol{\phi}_{J|i} = (\boldsymbol{\phi}_{1|i}, \ldots, \boldsymbol{\phi}_{c|i})$ so that $\boldsymbol{\phi}_{J|i}$ parameterizes the probabilities of non-response within categories of X.

Theorem 4 *Let $\boldsymbol{\phi}_{J|i}$ and $\boldsymbol{\phi}_{J|i'}$ be independent for $i \neq i' = 1, \ldots, r$. Then $\boldsymbol{\theta}_{J|i}|S$ and $\boldsymbol{\theta}_{J|i'}|S$ are independent for $i \neq i' = 1, \ldots, r$.*

Note that the posterior distribution of $\boldsymbol{\theta}_{J|i}$ will typically be a mixture of Dirichlet distributions. This theoretical framework exploits the flexibility of the Bayesian approach to incorporate assumptions on the pattern of missing data. Unfortunately, Table 2 shows that exact inference is possible only in few particular cases, and often it must resort either to simplifying assumptions or to expensive approximate methods.

3. Bound and Collapse

Section 2 described how the information on the pattern of missing data can be used to compute, in principle, the exact posterior distribution of $\boldsymbol{\theta}$. Unfortunately, this information about the pattern of missing data may not be available. In this Section, we will show that, even without this exogenous information, the incomplete sample is still able to induce bounds on the possible estimates consistent with the information available in the sample. When information about the missing-data mechanism becomes available, it can be used to

select a single estimate within the set of possible ones. This is the intuition behind BC. The method first *bounds* the possible estimates consistent with the available data, and then *collapses* the resulting intervals to point estimates via a convex combination of the extreme points, on the basis of information about the missing-data mechanism.

3.1 Bound

Let S be an incomplete sample. We assume that information is complete on X and there are non-reported cases only on Y. We continue to denote by m_i the frequency of cases $(X = i, Y = ?)$. By Theorem 2, the posterior distribution of $\boldsymbol{\theta}_I$ is $D(\boldsymbol{\alpha}_I + \boldsymbol{n}_I + \boldsymbol{m})$, independently of the missing-data mechanism, from which (4) is derived. Uncertainty on the estimate of $p(X = i, Y = j)$ depends on the unreported observations on Y, and hence on the estimate of $p(Y = j|X = i)$. If no missing-data mechanism is specified then, for fixed i, any value

$$\hat{\theta}_{j|i} = \frac{\alpha_{ij} + n_{ij} + m_{ij}}{\alpha_{i+} + n_{i+} + m_i}, \quad \sum_j m_{ij} = m_i \tag{11}$$

is a possible estimate, consistent with the information available in the sample. For fixed i and j, (11) is maximized when $m_{ij} = m_i$, so that the set of possible estimates of $p(Y = j|X = i)$ is bounded above by

$$p^\bullet(j|i) = \frac{\alpha_{ij} + n_{ij} + m_i}{\alpha_{i+} + n_{i+} + m_i}. \tag{12}$$

The maximum probability $p^\bullet(j|i)$ is obtained when all unclassified cases with $X = i$ are assigned to $Y = j$, so that the exact posterior distribution of $\boldsymbol{\theta}_{J|i}$ is $D(\alpha_{i1} + n_{i1}, \ldots, \alpha_{ij} + n_{ij} + m_i, \ldots, \alpha_{ic} + n_{ic})$. This distribution identifies a unique minimum probability of $(Y = l|X = i)$ for all $l \neq j$:

$$p_\bullet(l|i) = \frac{\alpha_{il} + n_{il}}{\alpha_{i+} + n_{i+} + m_i}.$$

This minimum probability is independent of j, from which it is easy to conclude that the possible estimates of $p(Y = j|X = i)$ are bounded as

$$p_\bullet(j|i) = \frac{\alpha_{ij} + n_{ij}}{\alpha_{i+} + n_{i+} + m_i} \leq p(Y = j|X = i|S) \leq \frac{\alpha_{ij} + n_{ij} + m_i}{\alpha_{i+} + n_{i+} + m_i} = p^\bullet(j|i). \tag{13}$$

From (13), we can also derive bounds on the possible estimates of θ_{ij} and θ_{+j}. By independence of θ_{i+} and $\theta_{j|i}$, it is easy to show that $\hat{\theta}_{i+}p_\bullet(j|i) \leq p(X = i, Y = j|S) \leq \hat{\theta}_{i+}p^\bullet(j|i)$. Furthermore, $E(\theta_{+j}|S) = \sum_{i=1}^r \hat{\theta}_{i+}E(\theta_{j|i}|S)$ and this function is maximized when $E(\theta_{j|i}|S) = p^\bullet(j|i)$ for all i, and it is minimized when $E(\theta_{j|i}|S) = p_\bullet(j|i)$ for all i. Therefore bounds on $p(Y = j|S)$ are:

$$p_\bullet(+j) = \sum_i \hat{\theta}_{i+}p_\bullet(j|i) \leq p(Y = j|S) \leq \sum_i \hat{\theta}_{i+}p^\bullet(j|i) = p^\bullet(+j) \tag{14}$$

for all j. Note that, since

$$p^{\bullet}(j|i) - p_{\bullet}(j|i) = \frac{m_i}{\alpha_{i+} + n_{i+} + m_i},$$

the width of the probability interval returned by the bound step is constant for all j, and it is increasing in the number of unclassified cases. Hence, the interval $p^{\bullet}(j|i) - p_{\bullet}(j|i)$ can be regarded as a direct measure of the information conveyed by the incomplete sample on $\theta_{j|i}$ and a faithful representation of the reliability of the estimates as function of the unreported cases for each category of X.

These results are consistent with the Incomplete Dirichlet Model (IDM) proposed by Walley (1996) in order to relax the assumption that the sample space is always known before any analysis can start. This analogy is not surprising when we realize that IDM and the Bound step share the same goal of providing robust estimates when some information is missing: about the sample space in the IDM case, about some entries in the sample in our case. In this second case, however, some external information on the pattern of non-response may be available. The nex Section will show how this information can be used to select a point estimate within probability intervals extracted at the bound step.

3.2 Collapse

For each category of X, incomplete cases induce a set of c extreme distributions corresponding to the most extreme situations in which data are systematically missing on one category of Y. Any assumption about the pattern of missing data will induce a distribution of $\boldsymbol{\theta}_{J|i}|S$ within these extreme distributions. Information about the missing-data mechanism can be therefore used to identify a single distribution within these bounds. This is the key idea of the collapse step.

In this Section, we assume that some external information on the missing-data mechanism is available, from which a probabilistic model for non-response can be deduced:

$$p(Y = j|Y = ?, X = i, \boldsymbol{\phi}, \boldsymbol{\theta}) = \phi_{j|i} \tag{15}$$

where $\sum_j \phi_{j|i} = 1$, all i. This information can be then used to identify a point estimate within the probability interval $[p_{\bullet}(j|i), p^{\bullet}(j|i)]$ via a convex combination of the extreme probabilities:

$$\hat{p}_{j|i} = \phi_{j|i} p^{\bullet}(j|i) + (1 - \phi_{j|i}) p_{\bullet}(j|i) \tag{16}$$

$$= \frac{\alpha_{ij} + n_{ij} + \phi_{j|i} m_i}{\alpha_{i+} + n_{i+} + m_i}. \tag{17}$$

For fixed i, (17) define a probability distribution since $\sum_j \hat{p}_{j|i} = 1$ for all i. Estimates (17) distribute the unclassified cases within categories of X across categories of Y according to the prior information about the pattern of missing data and (17) is the *expected Bayesian estimate* of $\theta_{j|i}$ given the assumed pattern of missing data. Note that (17) can be written as:

$$\hat{p}_{j|i} = \frac{\alpha_{i+} + n_{i+}}{\alpha_{i+} + n_{i+} + m_i} \frac{\alpha_{ij} + n_{ij}}{\alpha_{i+} + n_{i+}} + \frac{m_i}{\alpha_{i+} + n_{i+} + m_i} \phi_{j|i}$$

which is a weighed average of the estimate of $\theta_{j|i}$ obtained in the complete sample S_o and the prior information $\phi_{j|i}$, with weights that depend on m_i. As m_i decreases, then the sample estimate has more weight than $\phi_{j|i}$, and, when $m_i = 0$, $\hat{p}_{j|i} = (\alpha_{ij} + n_{ij})/(\alpha_{i+} + n_{i+})$. Thus, when the sample is complete, (17) is the exact estimate $E(\theta_{j|i}|S)$. As m_i increases then $\hat{p}_{j|i} \to \phi_{j|i}$, so that coherently nothing is learned from an empty sample.

Once $\hat{p}_{j|i}$ and $\hat{\theta}_{i+}$ are known, by independence of $\boldsymbol{\theta}_I$ and $\boldsymbol{\theta}_{J|I}$, the joint probability of $(X = i, Y = j|S)$ can be written as $E(\theta_{i+}|S)E(\theta_{j|i}|S)$, leading to

$$\hat{p}_{ij} = \frac{\alpha_{ij} + n_{ij} + \phi_{j|i} m_i}{\alpha + n + m}. \tag{18}$$

The marginal posterior probability of $(Y = j)$ is then given by:

$$\hat{p}_{+j} = \frac{\alpha_{+j} + n_{+j} + \sum_i \phi_{j|i} m_i}{\alpha + n + m}. \tag{19}$$

If the missing-data mechanism is MAR, there is no need for specifying $\boldsymbol{\phi}$, since we have shown in Section 2 that $\phi_{j|i} = \theta_{j|i}$. Thus, we can use the observed cases in the sample to estimate $\boldsymbol{\phi}$. This situation can be easily structured in BC. If the observed sample is representative of the complete one, $\phi_{j|i}$ can be replaced by the generalization of the MLE computed from S_o:

$$\hat{\phi}_{j|i} = \frac{\alpha_{ij} + n_{ij}}{\alpha_{i+} + n_{i+}}$$

and simple algebra shows that $\hat{p}_{j|i} = \hat{\phi}_{j|i}$. This is (5) in Section 2.

From a computational point of view, BC provides a deterministic method able to reduce the cost of estimating the conditional and marginal distributions of Y to the cost of one exact Bayesian updating and one convex combination for each category of Y in each category of X. The computational complexity of BC is therefore independent of the number of missing data and, being deterministic, BC does not pose convergence rate and detection problems that afflict iterative and stochastic methods currently used for the analysis of incomplete samples. This computational complexity translates into a time saving of several order of magnitudes which increases as the number of missing data scales up. The computational simplicity of BC provides also an alternative way to incorporate uncertainty about $\boldsymbol{\phi}$, that overcomes the limitation of a full Bayesian approach described in Section 2. A specified model $\boldsymbol{\phi}$ for non-response yields point estimates of $\theta_{j|i}$, from which point estimates of θ_{ij} and θ_{+j} are derived. These estimates are conditioned on $\boldsymbol{\phi}$. Uncertainty about $\boldsymbol{\phi}$ can be represented by assuming different values of $\boldsymbol{\phi}$, so that the sensitivity of the conclusions to the assumed pattern of missing data can be easily and fast examined. Furthermore, marginal inference can be obtained by averaging out estimates given different $\boldsymbol{\phi}$.

4. Inference

The BC method computes estimates of the conditional, joint and marginal probabilities of Y, for a given $\boldsymbol{\phi}$. However, complete inference on parameters of interest relies on the knowledge of the posterior distribution of these parameters. In this Section, we explore possible moment-matching approximations of the posterior distributions of $\boldsymbol{\theta}_{J|i}$, from which approximations of the posterior distribution of other parameters can be derived. Let $\hat{\boldsymbol{p}}_{J|i}$ be the vector of BC estimates $\hat{p}_{j|i}$ and $\hat{\boldsymbol{p}}$ be the vector of BC estimates $\hat{p}_{ij}$. If we approximate $\boldsymbol{\theta}_{J|i}|S, \boldsymbol{\phi} \sim D(\hat{\alpha}_{i+}\hat{\boldsymbol{p}}_{J|i})$, then $E(\theta_{j|i}|S) = \hat{p}_{j|i}$, and $\hat{\alpha}_i$ will be the posterior precision on $\boldsymbol{\theta}_{J|i}$, so that

$$\hat{V}(\theta_{j|i}|S) = \frac{\hat{p}_{j|i}(1-\hat{p}_{j|i})}{\hat{\alpha}_i + 1}.$$

We further approximate the posterior precision $\hat{\alpha}_{i+}$ by $\hat{\alpha}_{i+} = \alpha_{i+} + n_{i+} + km_i$, where k is a weight assigned to an incomplete case. If $k = 1$, we weigh an incomplete case as if it were a complete one. This would imply a strong prior confidence on the specified $\boldsymbol{\phi}$. On the other hand, if the missing-data mechanism is believed to be MAR, we can set $k = 0$, so that the incomplete cases are not considered in the analysis, and $D(\hat{\alpha}_{i+}\hat{\boldsymbol{p}}_{J|i})$ becomes the exact posterior distribution of $\boldsymbol{\theta}_{J|i}$ as shown in Section 2. A value $0 < k < 1$ would reflect the confidence in the prior specification of $\boldsymbol{\phi}$.

From the approximate posterior variance of $\theta_{j|i}$, we can further derive approximations of the posterior variance of θ_{ij} and θ_{+j}, that will be used as kernel of moment-matching approximations of their marginal posterior distribution. The posterior variances of θ_{ij} can be computed by letting $\theta_{ij} = \theta_{i+}\theta_{j|i}$. Given that, by Theorem 2, $\boldsymbol{\theta}_I|S \sim D(\boldsymbol{\alpha}_I + \boldsymbol{n}_I + \boldsymbol{m})$ and is independent of $\boldsymbol{\theta}_{J|I}|S$ we find:

$$\hat{V}(\theta_{ij}|S) = E(\theta_{i+}^2|S)\hat{E}(\theta_{j|i}^2|S) - E(\theta_{i+}|S)^2\hat{E}(\theta_{j|i}|S)^2,$$

with $E(\theta_{i+}^2|S)$ computed as in (9), $\hat{E}(\theta_{j|i}^2|S)$ as in (10), with $\hat{\theta}_{j|i}$ replaced by the BC estimate $\hat{p}_{j|i}$, and $\hat{E}(\theta_{j|i}|S) = \hat{p}_{j|i}$. The marginal probability of $Y = j$ is $\theta_{+j} = \sum_{i=1}^r \theta_{ij} = \sum_{i=1}^r \theta_{i+}\theta_{j|i}$ and, for a general NI pattern of missing data, $\theta_{j|i}$ and $\theta_{j|h}$ are dependent. If we assume independence of $\theta_{j|i}$ and $\theta_{j|h}$, the posterior variance of θ_{+j} can be approximated by

$$\hat{V}(\theta_{+j}|S) = \sum_{i=1}^r \hat{V}(\theta_{ij}|S) + \sum_{i=1}^r \sum_{h\neq i=1}^r \hat{Cov}(\theta_{i+}\theta_{j|i}, \theta_{h+}\theta_{j|h}|S)$$

and

$$\hat{Cov}(\theta_{i+}\theta_{j|i}, \theta_{h+}\theta_{j|h}|S) = -\frac{\hat{p}_{j|i}\hat{p}_{j|h}\hat{\theta}_{i+}\hat{\theta}_{h+}}{\alpha + n + m + 1}.$$

Thus, we have an approximation of the posterior variance of θ_{+j} in terms of the BC estimates of $\theta_{j|i}$. The approximate variance is the exact variance when data are MAR. In other cases, the accuracy of this approximation depends on the magnitude of the correlation between $\theta_{j|i}$ and $\theta_{j|h}$. If the probabilities of non-response are functionally independent for different

		Y				
X_1	X_2	1	2	3	4	m_i
1	1	26	8	7	0	11
	2	87	37	30	6	64
	3	66	77	23	8	77
	4	14	25	15	1	12
	5	6	6	2	0	7
2	1	1	1	0	1	2
	2	63	34	32	2	68
	3	102	52	22	4	77
	4	10	32	10	2	38
	5	20	25	8	2	19

Table 3: Data from the British General Election Panel Survey. Voting Intention (Y): 1=Conservative, 2=Labour, 3=Liberal Democrat, 4=Other; Sex (X_1): 1=Male, 2=Female; Social Class (X_2): 1=Professional, 2=Managerial and Technical, 3=Skilled, 4=Semiskilled and Unskilled, 5=Never Worked.

categories of X, then the accuracy of the approximation is only dependent on the accuracy of the BC estimates. Simulation studies have shown that the approximation is very accurate for large samples, and example will be given in Section 5.

The final step is to approximate the marginal posterior distribution of θ_{ij} and θ_{+j}. Consider for instance the marginal distribution of θ_{ij}. A simple choice is to set $\theta_{ij}|S, \boldsymbol{\phi} \sim D(\hat{\alpha}_{ij1}, \hat{\alpha}_{ij2})$, where the hyperparameters $\hat{\alpha}_{ij1}, \hat{\alpha}_{ij2}$ are chosen to match the marginal mean and variance of θ_{ij}. Therefore

$$\hat{\alpha}_{ij1} = \frac{\hat{\theta}_{ij}^2(1-\hat{\theta}_{ij})}{\hat{V}(\theta_{ij}|S)} - \hat{\theta}_{ij} \tag{20}$$

$$\hat{\alpha}_{ij2} = \frac{\hat{\theta}_{ij}(1-\hat{\theta}_{ij})^2}{\hat{V}(\theta_{ij}|S)} + \hat{\theta}_{ij} - 1. \tag{21}$$

An alternative approximation, suitable for large samples, is to set $\theta_{ij}|S, \boldsymbol{\phi} \sim N(\hat{\theta}_{ij}; \hat{V}(\theta_{ij}|S))$. Under the MAR assumption this is the asymptotic approximation of the posterior distribution based on the generalized MLE (Berger, 1985) of the parameters, if the prior hyperparameters were all increased by 1. Similarly, the marginal posterior distribution of θ_{+j} can be approximated by a $D(\hat{\alpha}_{+j1}, \hat{\alpha}_{+j2})$, with hyperparameters chosen to match (19) and $\hat{V}(\theta_{+j}|S)$ given above.

5. An Application to Polling Data

This Section will illustrate the features of BC using the polling data examined by Forster and Smith (1998). Data are extracted from the British General Election Panel Survey

		Y				Width
X_1	X_2	1	2	3	4	
1	1	0.4995	0.1540	0.1348	0.0005	0.2112
		0.7107	0.3652	0.3460	0.2116	
	2	0.3883	0.1652	0.1340	0.0269	0.2856
		0.6739	0.4508	0.4196	0.3125	
	3	0.2629	0.3068	0.0917	0.0320	0.3067
		0.5696	0.6134	0.3983	0.3386	
	4	0.2090	0.3730	0.2239	0.0153	0.1789
		0.3879	0.5518	0.4028	0.1941	
	5	0.2855	0.2855	0.0960	0.0012	0.3318
		0.6173	0.6173	0.4277	0.3329	
2	1	0.2010	0.2010	0.0049	0.2010	0.3921
		0.5931	0.5931	0.3971	0.5931	
	2	0.3165	0.1709	0.1608	0.0102	0.3416
		0.6581	0.5124	0.5024	0.3517	
	3	0.3968	0.2024	0.0857	0.0157	0.2995
		0.6963	0.5018	0.3852	0.3151	
	4	0.1088	0.3477	0.1088	0.0220	0.4125
		0.5214	0.7603	0.5214	0.4346	
	5	0.2702	0.3377	0.1083	0.0273	0.2567
		0.5267	0.5941	0.3647	0.2837	
	$\hat{\theta}_{+j}$	0.3180	0.2391	0.1201	0.0211	0.3017
		0.6197	0.5408	0.4218	0.3228	

Table 4: Bounds on the conditional probabilities estimated by BC.

and are reported in Table 3. The frequencies are classified according to Sex (X_1), Social Class (X_2), and the response variable Voting Intention (Y). The total sample size is 1242 cases, of which 375 cases do not record the Voting Intention. Following Forster and Smith (1998), we assume that Sex and Social Class are associated, and they both affect Voting Intention. We denote $p(Y = j|X_1 = i, X_2 = h, \boldsymbol{\theta})$ by θ_{ihj}. We also assume a Perks prior distribution $\boldsymbol{\theta} \sim D(\boldsymbol{\alpha})$, with $\alpha_{ihj} = 1/40$ (Good, 1968), so that the total prior precision is 1. From the specification of the prior distribution of $\boldsymbol{\theta}$, we can derive prior distributions of the parameters of the model by using Theorem 1. Thus $\boldsymbol{\theta}_{IH+} \sim D(\boldsymbol{\alpha}_{IH+})$ and $\alpha_{ih+} = 1/10$, from which $\boldsymbol{\theta}_{I++} \sim D(\boldsymbol{\alpha}_{I++})$ and $\alpha_{i++} = 1/2$ and $\boldsymbol{\theta}_{+H+} \sim D(\boldsymbol{\alpha}_{+H+})$ and $\alpha_{+h+} = 1/5$. Furthermore $\boldsymbol{\theta}_{J|ih} \sim D(\boldsymbol{\alpha}_{J|ih})$ and $\alpha_{j|ih} = 1/40$.

5.1 Bound

Table 4 reports lower and upper bounds of the estimates of the conditional probabilities computed using (13), when no model for non-response is assumed. The last column is the width of the probability intervals of each conditional distribution. The last two rows give

		Y				
X_1	X_2	1	2	3	4	$\hat{\theta}_{ih+}$
1	1	0.6332	0.1953	0.1709	0.0006	0.0419
		(0.0743)	(0.0611)	(0.0580)	(0.0038)	(0.0057)
	2	0.5436	0.2313	0.1875	0.0376	0.1803
		(0.0392)	(0.0332)	(0.0308)	(0.0150)	(0.0109)
	3	0.3792	0.4424	0.1323	0.0461	0.2020
		(0.0367)	(0.0375)	(0.0256)	(0.0158)	(0.0114)
	4	0.2545	0.4542	0.2727	0.0186	0.0540
		(0.0582)	(0.0665)	(0.0594)	(0.0184)	(0.0064)
	5	0.4273	0.4273	0.1436	0.0018	0.0170
		(0.1273)	(0.1273)	(0.0903)	(0.0108)	(0.0037)
2	1	0.3306	0.3306	0.0081	0.3306	0.0041
		(0.2323)	(0.2323)	(0.0442)	(0.2323)	(0.0018)
	2	0.4807	0.2595	0.2443	0.0154	0.1602
		(0.0435)	(0.0381)	(0.0374)	(0.0108)	(0.0104)
	3	0.5665	0.2889	0.1223	0.0223	0.2068
		(0.0368)	(0.0337)	(0.0243)	(0.0110)	(0.0115)
	4	0.1853	0.5919	0.1853	0.0374	0.0741
		(0.0523)	(0.0660)	(0.0523)	(0.0255)	(0.0074)
	5	0.3634	0.4541	0.1456	0.0368	0.0596
		(0.0642)	(0.0665)	(0.0471)	(0.0251)	(0.0067)

Table 5: BC estimates of conditional probabilities of $Y = j$ given $(X_1 = i, X_2 = h)$. Standard errors are reported in brackets.

the lower and upper bounds on the estimates of the marginal probabilities of Y, computed as in (14).

The width of probability intervals gives a measure of the amount of uncertainty among non respondents, as well as the effect of the variables Sex and Social Class on the amount of non-responses. Intervals are tighter for men than for women, and there seems to be an effect of Social Class on the amount of non-responses that varies according to gender. Therefore, if we accept ignorability of the missing-data mechanism, bounds would support the MAR assumption over MCAR. Semiskilled and unskilled men $(X_1 = 1, X_2 = 4)$ are the least uncertain. Most uncertain are men who never worked $(X_1 = 1, X_2 = 5)$. Among women, semiskilled and unskilled $(X_1 = 2, X_2 = 4)$ are the most uncertain, with a clear preference toward the Labour party. These results could be useful to single out the categories to be addressed more effectively during the political campagne. The uncertainty on the marginal probabilities of Voting Intention shows that, if we assume that non-response are due to genuine indecision, the results of the political election could be a real surprise, but, although a victory of the Labour party cannot be excluded, Tories seem to be far ahead.

5.2 Collapse

Once bounds have been estimated, the *Collapse* step can be used to model different assumptions about the pattern of missing data.

5.2.1 Missing at Random

Suppose that the missing-data mechanism is MAR, so that the BC estimates of the conditional probabilities of $Y = j$ are the exact ones given in (5). Table 5 reports the BC estimates and standard errors of the conditional probabilities of Voting Intention. Estimates of the joint probabilities of $X_1 = i$ and $X_2 = h$ are given in the last column, with their standard errors. From these quantities the joint probabilities of $(X_1 = i, X_2 = h, Y = j)$ can be easily calculated, as well as their exact standard errors. Marginal probabilities of $Y = j$ and standard errors are derived using the approach described in Section 2, and are reported in Table 6. Estimates of θ_{+j} differ from 0.4480, 0.3458, 0.1704, 0.0357 reported by Forster and Smith (1998), and it can be easily verified that the latter are computed under the assumption that Sex and Social Class are marginally independent.

The approximations discussed at the end of Section 4 can be used to find 95% credibility intervals about the predictive probabilities of Voting Intention. Suppose first that the posterior distribution of θ_{+j} is approximated to a Beta to match mean and variance. To evaluate the goodness of the approximation, we have generated a sample of 5,000 observations from the posterior distribution of $\boldsymbol{\theta}_{++J}$ using the Gibbs Sampler implemeted in Bugs5 (Thomas *et al.*, 1992) after a first burn-in of 1,000 observations. The results are in Table 6. Credibility intervals computed via the BC approximation are extremely accurate. 95% credibility intervals obtained by using a Normal approximation are (0.4204;0.4858), (0.3128;0.3764), (0.1466;0.1968) and (0.0192;0.0420). Time used for computing estimates with Bugs5 was 120 sec on a Sparc Ultra, whilst time for computing BC estimates was less than 1 sec. Credibility intervals computed for the joint probabilities are also very similar to those found by Bugs5, even when the frequencies of complete cases are small. For instance, estimates of the joint probabilities of $\theta_{1,5,1}, \theta_{1,5,2}, \theta_{1,5,3}$ and $\theta_{1,5,4}$ are 0.007, 0.007, 0.002,0.00003; 95% credibility intervals computed using the moment-matching approximation described in Section 4 are (0.003;0.013), (0.003;0.013), (0.0003;0.007) and (0.00;0.005). The same credibility intervals computed with Bugs5 are (0.003;0.013), (0.003;0.013), (0.0003;0.007) and (0.00;0.003).

In the follow up survey reported by Forster and Smith (1998), the marginal frequencies of Voting Intention were 0.441, 0.322, 0.210 and 0.0282. Thus, the MAR produces an overestimation of preferences for the Conservative and Labour parties, and an underestimation of preferences for the Liberal Democrat Party. A naive analysis ignoring incomplete cases would yield estimates of θ_{+j} equal to 0.456, 0.343, 0.172, and 0.030, that are not very different from the estimates computed under the MAR assumption. The effect is more evident on the estimates of the conditional probabilities: taking into account incomplete cases yields estimates that are more robust. This was also noted by Little and Rubin (1987).

	Y			
BC	1	2	3	4
$\hat{\theta}_{++j}$	0.4531	0.3446	0.1717	0.0306
s.e	0.0167	0.0162	0.0128	0.0058
95%CI	(0.4182;0.4881)	(0.3121;0.3779)	(0.1470;0.1980)	(0.0202;0.0431)
MCMC	1	2	3	4
$\hat{\theta}_{++j}$	0.4527	0.3447	0.1718	0.0307
s.e	0.0168	0.0157	0.0127	0.0059
95%CI	(0.4206;0.4860)	(0.3141;0.3755)	(0.1476;0.1973)	(0.0202;0.0431)

Table 6: MCMC and BC estimates of the marginal probabilities of Voting Intention.

5.2.2 Non Ignorable Non-response

When respondents do not reveal their voting intention, it is quite sensible to assume that they have a reason for withelding their opinion. It is believed that "nonrespondents to polls before the 1992 British General Election were more heavily pro-Conservative than respondents" (Forster and Smith, 1998). Here we investigate particular assumptions on the missing-data mechanism, including the *Silent Conservative* effect (Butler and Kavanagh, 1992), and we compare the BC estimates with Imputation-based ones.

We first assume that non-respondents are truly uncertain among the three major parties, and we model this assumption by setting

j	1	2	3	4
$\phi_{j\|ih}$	0.32	0.32	0.32	0.04

for all h, i. Thus we assume a uniform pattern of non-response across categories of X_1 and X_2. BC estimates of the conditional probabilities are easily computed by mixing the upper and lower bounds given in Table 4 and are reported in Table 7, together with their standard errors. The latter were computed by assuming a total precision $\hat{\alpha}_{ih} = \alpha_{ih+} + n_{ih+} + m_i$. Estimates of the marginal probabilities θ_{+j}, standard errors and 95% credibility intervals based on the moment-matching approximation to a Beta distribution are reported in Table 8.

The effect of the assumed model for non-response is to balance Voting Intention among categories of X_1, X_2 so that, for instance, professional and managerial men ($X_1 = 1, X_2 = 1, 2$) show a less evident preference for the Conservative party which, in return, would gain votes from semiskilled and unskilled women. The global effect on the marginal probabilities of $Y = j$ is the prediction of a smaller majority of the Conservative party over the Labour and the Liberal Democrat. Note that, since the total number of complete and incomplete cases is considered in the analysis, the posterior estimates have a smaller standard error than the same estimates obtained under the MAR assumption. In order to avoid an overestimation of the precision, a smaller proportion of the incomplete cases can be considered, as described in Section 4.

The accuracy and efficiency of the BC estimates can be compared to Imputation-based

		Y			
X_1	X_2	1	2	3	4
1	1	0.5671	0.2216	0.2024	0.0089
		(0.0140)	(0.0118)	(0.0114)	(0.0027)
	2	0.4797	0.2566	0.2254	0.0383
		(0.0142)	(0.0124)	(0.0118)	(0.0054)
	3	0.3611	0.4049	0.1898	0.0442
		(0.0136)	(0.0139)	(0.0111)	(0.0058)
	4	0.2662	0.4302	0.2811	0.0224
		(0.012)	(0.0140)	(0.0127)	(0.0042)
	5	0.3917	0.3917	0.2021	0.0145
		(0.0138)	(0.0138)	(0.0114)	(0.0034)
2	1	0.3265	0.3265	0.1304	0.2167
		(0.0133)	(0.0133)	(0.0095)	(0.0117)
	2	0.4258	0.2802	0.2701	0.0238
		(0.0140)	(0.0127)	(0.0126)	(0.0043)
	3	0.4927	0.2982	0.1815	0.0276
		(0.0142)	(0.01300)	(0.0109)	(0.0046)
	4	0.2409	0.4797	0.2409	0.0385
		(0.0121)	(0.0142)	(0.0121)	(0.0055)
	5	0.3523	0.4198	0.1906	0.0376
		(0.0135)	(0.0140)	(0.0111)	(0.0054)

Table 7: BC estimates of conditional probabilities of $Y = j$ given $X_1 = i, X_2 = h$, assuming an informative patter of missing data. Standard errors are reported in brackets.

	Y			
BC	1	2	3	4
$\hat{\theta}_{+j}$	0.4145	0.3357	0.2166	0.0332
s.e.	0.0083	0.0075	0.0057	0.0021
95% CI	(0.3983;04308)	(0.3211;0.3505)	(0.2055;0.2279)	(0.0292;0.0374)
Imputation	1	2	3	4
$\hat{\theta}_{+j}$	0.4145	0.3358	0.2165	0.0332
s.e.	0.0074	0.0071	0.0072	0.0031
95% CI	(0.4000;0.4290)	(0.3220;0.3494)	(0.2029;0.2303)	(0.0276;0.0396)

Table 8: BC and Imputation-based estimates of marginal probabilities, standard errors and 95% credibility intervals, assuming an informative pattern of missing data. The probability of non-response is assumed to be $\phi_{j|ih} = 0.32$ for $j = 1, 2, 3$ and all i, h.

	Y			
BC	1	2	3	4
$\hat{\theta}_{+j}$	0.4236	0.3296	0.2045	0.0422
s.e.	0.0084	0.0074	0.0055	0.0024
95% CI	(0.4072;0.4401)	(0.3152;0.3442)	(0.1938;0.2154)	(0.0376;0.0470)
Imputation	1	2	3	4
$\hat{\theta}_{+j}$	0.4236	0.3300	0.2042	0.0423
s.e.	0.0076	0.0073	0.0072	0.0039
95% CI	(0.4080;0.4379)	(0.3156;0.3429)	(0.1901;0.2190)	(0.0348;0.0501)

Table 9: BC and Imputation-based estimates of marginal probabilities, standard errors and 95% credibility intervals, assuming an informative patter of missing data. The probability of non-response is assumed to be $\phi_{j|ih} = 0.35, 0.3, 0.28, 0.07$ for all i, h.

estimates. The incomplete sample was completed 1,000 times by generating the missing entries from the probability of non-response $\boldsymbol{\phi}$. In each completed sample, the exact estimates of the conditional and marginal probabilities $\theta_{j|ih}$ and θ_{+j} were computed by using the standard conjugate analysis described in Section 2. Final estimates and standard errors were then taken as means and standard errors of the 1,000 estimates generated by the simulation. Empirical 95% confidence intervals were also computed by using 2.5% and 97.5% quantiles. The imputation algorithm was implemented in R (Ihaka and Gentleman, 1996), an environment implementing the S Language (Becker *et al.*, 1988) and execution time exceeded 15 minutes against less than 1 sec for BC. Results are in Table 8. The comparison reveals the extreme accuracy of BC estimates. Differences are in the third decimal digit, and the precision of the Imputation-based results is to second decimal digit.

Consider now another pattern of non-responses implementing the *Silent Conservative* effect, that is, the common assumption that non respondents turn out to be more heavily pro-Conservative than respondents. We assume again a uniform pattern of non-responses across categories of X_1 and X_2 and we set

j	1	2	3	4
$\phi_{j\|ih}$	0.35	0.30	0.28	0.07

for all h, i. BC estimates of the marginal probabilities of Y and the corresponding Imputation-based estimates are given in Table 9. If we assume a prior distribution granting a slightly larger preference to the Conservative party:

j	1	2	3	4
$\phi_{j\|ih}$	0.41	0.28	0.28	0.03

for all h, i, the effect on the marginal probability of Voting Intention is given in Table 10. Compared to responses observed in the follow up survey (0.441, 0.322, 0.210 and 0.0282) the estimates turns out to be extremely accurate and support the hypothesis that the missing-data mechanism is NI.

	Y			
BC	1	2	3	4
$\hat{\theta}_{+j}$	0.4417	0.3236	0.2045	0.0302
s.e.	0.0086	0.0073	0.0055	0.0020
95% CI	(0.4248;0.4586)	(0.3094;0.3380)	(0.1938;0.2154)	(0.0264;0.0340)
Imputation	1	2	3	4
$\hat{\theta}_{+j}$	0.4408	0.3233	0.2045	0.0304
s.e.	0.0075	0.0069	0.0069	0.0027
95% CI	(0.4266;0.4556)	(0.3107;0.3365)	(0.1909;0.2182)	(0.0251;0.0360)

Table 10: BC and Imputation-based estimates of marginal probabilities, standard errors and 95% credibility intervals, assuming an informative patter of missing data. The probability of non-response is assumed to be $\phi_{j|ih} = 0.41, 0.28, 0.28, 0.03$ for all i, h.

6. Conclusions

Three main objectives were set at the begginning of this paper: the definition of an estimation method from incomplete samples robust with respect to the pattern of missing data, the identification of realibity measures able to account for the presence of missing data in a sample, and the development of efficient computational methods to perform these calculations. BC provides a methodological framework within which these goals can be achieved.

The basic intution behind BC is that information about the incomplete sample and exogenous knowledge about the pattern of missing data should be kept separated. This assumption naturally produces a two-step method: *Bound* and *Collapse*. The first step extracts from the sample all the available information and returns a set of possible estimates consistent with the incomplete sample. The bound step provides, as a by product, a new measure of the reliability of the estimates with respect to the amount of information actually conveyed by the incomplete sample about each parameter of interest. The second step of BC uses exogenous information available on the missing-data mechanism to select single estimates within the sets defined by the first step. These estimates are weighted averages of estimates computed from the complete sample and probabilities of non-response. When data are MAR, BC returns the exact Bayesian estimates. Under a generic missing-data mechanism, BC estimates are the expected Bayesian estimates given the non-response model $\boldsymbol{\phi}$.

BC provides an estimation procedure able to encode different assumptions about the pattern of missing data and to quickly evaluate the sensitivity of the estimates to different assumptions about the non-response model. Marginal inference can be obtained by averaging out estimates obtained under different non-response models. From a computational point of view, BC provides a deterministic method able to reduce the cost of estimating the conditional and marginal distributions of Y to the cost of one exact Bayesian updating and one convex combination for each category of Y in each category of X. The computational

complexity of BC is therefore independent to the number of missing data and, being deterministic, BC does not pose the problems of convergence rate and detection afflicting iterative and stochastic methods currently used for the analysis of incomplete samples. This computational complexity translates into a time saving of several order of magnitudes, which increases as the number of missing data scales up. Simplicity and efficiency make of this method a powerful tool for the analysis of incomplete samples to foster the application of principled statistical methods to real-world problems.

References

Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988). *The New S Language*. Chapman & Hall, London.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, NY.

Butler, D., and Kavanagh, D. (1992). *The British general election of 1992*. St Martins's Press, New York.

Copas, J. B., and Li, H. G. (1997). Inference in non-random samples (with discussion). *J. R. Statist. Soc.*, *59*, 55–95.

de Laplace, P. S. (1840). *Essai philosophique sur les probabilites* (6th edition). Bachelier, Paris.

Fang, K. T., Kotz, S., and Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions*. Chapman and Hall, London.

Forster, J. J., and Smith, P. W. F. (1998). Model-based inference for categorical survey data subject to non-ignorable non-response. *J. R. Statist. Soc.*, *60*. To appear.

Geiger, D., and Heckerman, D. (1997). A characterization of Dirichlet distributions through local and global independence. *Ann. Statist.*, *25*, 1344–1368.

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

Geman, S., and Geman, D. (1984). Stochastic relaxation, gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Gilks, W. R., and Roberts, G. O. (1996). Strategies for improving MCMC. In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (Eds.), *Markov Chain Monte Carlo in practice*, pp. 89–114. Chapman and Hall, London.

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

Ihaka, R., and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, *5*(3), 299–314.

Kadane, J. B. (1993). Subjective Bayesian analysis for surveys with missing data. *The Statistician*, *42*, 415–426.

Kadane, J. B., and Terrin, N. (1997). Missig data in the forensic context. *J. R. Statist. Soc.*, *160*, 351–357.

Lindley, D. L. (1964). The Bayesian analysis of contingency tables. *Ann. Math. Statist.*, *35*, 1622–1643.

Little, R., and Rubin, D. (1987). *Statistical Analysis with Missing Data.* Wiley, New York, NY.

Park, T., and Brown, M. B. (1994). Models for categorical data with nonignorable nonresponse. *J. Amer. Statist. Assoc.*, *89*, 44–52.

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

Rubin, D. B., Stern, H. S., and Vehovar, V. (1995). Handling "don't know" survey responses: the case of the slovenian plebiscite. *J. Amer. Statist. Assoc.*, *90*, 822–828.

Spiegelhalter, D., and Lauritzen, S. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks*, *20*, 157–224.

Thomas, A., Spiegelhalter, D., and Gilks, W. (1992). Bugs: A program to perform Bayesian inference using Gibbs Sampling. In *Bayesian Statistics 4*, pp. 837–42. Clarendon Press, Oxford.

Walley, P. (1996). Inference from multinomial data: learning about a bag of marbles (with discussion). *J. R. Statist. Soc.*, *58*, 3–57.

Wilks, S. S. (1963). *Mathematical Statistics.* Wiley, New York.

A. Proof of Theorems

This appendix reports the proves of Theorems 2, 3 and 4 stated in Section 2.

Theorem 2 *Let S be an incomplete sample in which n_{ij} is the frequency of observed cases $(X = i, Y = j)$, and m_i is the frequency of cases $(X = i, Y = ?)$. If $\boldsymbol{\theta} \sim D(\boldsymbol{\alpha})$, the posterior distribution of $\boldsymbol{\theta}_I$ is $D(\alpha_{1+} + n_{1+} + m_1, \cdots, \alpha_{r+} + n_{r+} + m_r) \equiv D(\boldsymbol{\alpha}_I + \boldsymbol{n}_I + \boldsymbol{m})$, and $\boldsymbol{\theta}_I$ and $\boldsymbol{\theta}_{J|I}$ are independent.*

Proof. The exact posterior distribution of $\boldsymbol{\theta}$ can be written as

$$p(\boldsymbol{\theta}|S) = \sum_{c_i} w_{c_i} \prod_{ij} \frac{\Gamma(\alpha + n + m)}{\Gamma(\alpha_{ij} + n_{ij} + m_{ij})} \theta_{ij}^{\alpha_{ij}+n_{ij}+m_{ij}-1}$$

where $w_{c_i} = p(S_{c_i}|S)$ and m_{ij} is the frequency of the m_i incomplete cases $(X = i, Y = ?)$ in S_{c_i}. Consider the vector $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}_I, \boldsymbol{\theta}_{J|1}, \cdots, \boldsymbol{\theta}_{J|r})$ whose elements are functions of $\boldsymbol{\theta}$. It is easy to show that the Jacobian of the transformation $\boldsymbol{\theta} \to \tilde{\boldsymbol{\theta}}$ is $\prod_{i=1}^{r} \theta_{i+}^{c-1}$ so that

$$\begin{aligned} p(\tilde{\boldsymbol{\theta}}|S) &= \left(\prod_{i=1}^{r} \theta_{i+}^{c-1}\right) \sum_{c_i} w_{c_i} \prod_{i=1}^{r} \frac{\Gamma(\alpha + n + m)}{\Gamma(\alpha_{i+} + n_{i+} + m_i)} \theta_{i+}^{\alpha_{i+}+n_{i+}+m_i-c} \\ &\quad \times \prod_{j=1}^{c} \frac{\Gamma(\alpha_{i+} + n_{i+} + m_i)}{\Gamma(\alpha_{ij} + n_{ij} + m_{ij})} \theta_{j|i}^{\alpha_{ij}+n_{ij}+m_{ij}-1} \\ &= \prod_{i=1}^{r} \frac{\Gamma(\alpha + n + m)}{\Gamma(\alpha_{i+} + n_{i+} + m_i)} \theta_{i+}^{\alpha_{i+}+n_{i+}+m_i-1} \\ &\quad \times \sum_{c_i} w_{c_i} \prod_{ij} \frac{\Gamma(\alpha_{i+} + n_{i+} + m_i)}{\Gamma(\alpha_{ij} + n_{ij} + m_{ij})} \theta_{j|i}^{\alpha_{ij}+n_{ij}+m_{ij}-1}. \end{aligned}$$

Since the density of $\tilde{\boldsymbol{\theta}}$ factorizes, the independence of $\boldsymbol{\theta}_I$ and $\boldsymbol{\theta}_{J|I}$ is proved. The fact that the distribution of $\boldsymbol{\theta}_I$ is Dirichlet follows easily by integrating $\theta_{j|i}$ out. □

Theorem 3 *Suppose that the missing-data mechanism is* MAR. *Then the distribution of* $\boldsymbol{\theta}_{J|I}$ *factorizes into a product of independent Dirichlet distributions* $D(\boldsymbol{\alpha}_{J|i} + \boldsymbol{n}_{J|i})$.

Proof. Under a general missing-data mechanism, the posterior distribution of $\boldsymbol{\theta}_{J|I}$ is the mixture of Dirichlet distributions, with density:

$$p(\boldsymbol{\theta}_{J|I}|S) = \sum_{c_i} w_{c_i} \prod_{ij} \frac{\Gamma(\alpha_{i+} + n_{i+} + m_i)}{\Gamma(\alpha_{ij} + n_{ij} + m_{ij})} \theta_{j|i}^{\alpha_{ij}+n_{ij}+m_{ij}-1}.$$

Now fix a possible distribution S_{d_i} by assigning m_{ij} cases, out of the m_i partially observed, to $(X = i, Y = j)$, then

$$p(S_{d_i}|\boldsymbol{\theta}) = \frac{\prod_i m_i!}{\prod_{ij} m_{ij}!} \prod_{ij} \theta_{ij}^{m_{ij}}, \quad \sum_j m_{ij} = m_i.$$

Since S_o is complete, the distribution of $\boldsymbol{\theta}|S_o$ is $D(\boldsymbol{\alpha} + \boldsymbol{n})$, and

$$\begin{aligned} p(S_{d_i}|S_o) &= \frac{\prod_i m_i! \Gamma(\alpha + n)}{\prod_{ij} m_{ij}! \Gamma(\alpha_{ij} + n_{ij})} \int \prod_{ij} \theta_{ij}^{\alpha_{ij}+n_{ij}+m_{ij}-1} d\boldsymbol{\theta} \\ &= \frac{\prod_i m_i! \Gamma(\alpha + n) \prod_{ij} \Gamma(\alpha_{ij} + n_{ij} + m_{ij})}{\Gamma(\alpha + n + m) \prod_{ij} m_{ij}! \Gamma(\alpha_{ij} + n_{ij})}, \quad \sum_j m_{ij} = m_i. \end{aligned} \tag{22}$$

From (22) it follows that

$$w_{c_i} \propto \frac{m! \Gamma(\alpha + n) \prod_{ij} \Gamma(\alpha_{ij} + n_{ij} + m_{ij})}{\Gamma(\alpha + n + m) \prod_{ij} m_{ij}! \Gamma(\alpha_{ij} + n_{ij})}.$$

and therefore

$$p(\boldsymbol{\theta}_{J|I}|S) \propto \frac{m!\Gamma(\alpha+n)\prod_i \Gamma(\alpha_{i+}+n_{i+}+m_i)}{\Gamma(\alpha+n+m)\prod_i m_i!\prod_i \Gamma(\alpha_{i+}+n_{i+})} \sum_{c_i} \prod_{ij} \frac{m_i!}{m_{ij}!}\theta_{j|i}^{m_{ij}} \times \prod_{ij} \frac{\Gamma(\alpha_{i+}+n_{i+})}{\Gamma(\alpha_{ij}+n_{ij})}\theta_{j|i}^{\alpha_{ij}+n_{ij}-1}.$$

From the fact that $\sum_{c_i}\prod_{ij} m_i!\theta_{j|i}^{m_{ij}}/m_{ij}! = 1$ the result is proved. □

Theorem 4 *Let $\boldsymbol{\phi}_{J|i}$ and $\boldsymbol{\phi}_{J|i'}$ be independent for $i \neq i' = 1,\ldots,r$. Then $\boldsymbol{\theta}_{J|i}|S$ and $\boldsymbol{\theta}_{J|i'}|S$ are independent for $i \neq i' = 1,\ldots,r$.*

Proof. In the proof of Theorem 2 we have shown that

$$p(\boldsymbol{\theta}_{J|I}|S) = \sum_{c_i} w_{c_i} \prod_{ij} \frac{\Gamma(\alpha_{i+}+n_{i+}+m_i)}{\Gamma(\alpha_{ij}+n_{ij}+m_{ij})}\theta_{j|i}^{\alpha_{ij}+n_{ij}+m_{ij}-1},$$

and $w_{c_i} = p(S_{c_i}|S)$, where S_{c_i} is a possible completion of the incomplete sample. It is straightforward to show that if for the probability of non-response are independent, then w_{c_i} factorizes in a product of r terms, each of which is the probability of completing an incomplete sample for a fixed category i of X. This yields a factorization of the posterior density from which the result follows. □