Bayesian Analysis (2008) **3**, Number 3, pp. 541–554

# Bayesian methods for categorical data under informative censoring

Thomas J. Jiang* and James M. Dickey†

**Abstract.** Bayesian methods are presented for categorical sampling when some observations are censored (i.e., suffer missing distinctions between categories). Such problems have been researched over the years, as they can be important in applications. However, previous work has assumed strong restrictions, such as truthful reporting, noninformative censoring, etc. Here, we attempt to remove such restrictions. In particular, we remove two of the three restrictions imposed by Dickey, Jiang, and Kadane (1987). We provide Bayesian methods for cases more general than those considered by Paulino and de B. Pereira (1992, 1995), and others. Thus, it will no longer be necessary to make unrealistic assumptions commonly employed regarding the censoring model. A theorem of Identifiability-by-Conditioning is provided, allowing familiar improper prior densities. By this theorem, we obtain identical Bayesian updating results by imposing constraints on either prior, likelihood, or posterior directly. Several computational procedures are suggested, and an example is used to illustrate methods.

**Keywords:** Bayesian inference, generalized Dirichlet distributions, informative censoring, multiple hypergeometric functions

## 1 Introduction

Bayesian treatments of categorical sampling with censored, or partially-classified, data were given by Karson and Wrobleski (1970), Antelman (1972), Kaufman and King (1973), Albert and Gupta (1983), Gunel (1984), Smith and Gunel (1984), Smith, Choi, and Gunel (1985), Albert (1985), Kadane (1985), and Gibbons and Greenberg (1989). These all dealt with 2 × 2 contingency tables that have information missing regarding row or column variables. Dickey, Jiang, and Kadane (1987) extended consideration to the general multinomial. All these studies were restricted to noninformatively-censored categorical data. (For treatments from the frequentist viewpoint, see e.g., Hartley (1958), Chen and Fienberg (1974, 1976), Dempster, Laird, and Rubin (1977), Little and Rubin (1987).

Bayesian treatments for informatively censored data can be found in Basu and de B. Pereira (1982), Paulino and de B. Pereira (1992, 1995), Walker (1996), and Tian, Ng, and Geng (2003), among others. In particular, Paulino and de B. Pereira (1995), Walker (1996), and Tian, Ng, and Geng (2003) all considered the general censored data problem with truthful reports. Paulino and de B. Pereira (1995) and Tian, Ng, and Geng (2003) both gave

[*] Department of Mathematical Sciences, National Chengchi University, Wen-Shan, Taipei, Taiwan, mailto:jiangt@nccu.edu.tw

[†] School of Statistics, University of Minnesota, Minneapolis, MN, mailto:dickey@stat.umn.edu

DOI:10.1214/08-BA321

posterior distributions and posterior expectations under a restricted Dirichlet prior distribution family, while Walker (1996) used MAP (maximum a posteriori) methods to make inference.

In practice, it is likely that reported data fail to match the true categories, and the pattern of censoring, itself, has information regarding parameters of interest. For example, suppose that each person's income falls into one of five categories. An individual, whose income is in the highest category, may report his/her income as being in the second highest category. This is an example of nontruthful-reporting. In addition, to discourage a refusal to respond, an individual may be allowed to report a set-union of two or more categories. However, then, an individual who is actually in the highest category, may be more likely to report himself/herself as being in the top two categories than another individual who is actually in the second highest category. This would, then, be an example of informatively censored reporting.

In the present paper, we consider an unrestricted Dirichlet prior distribution family and allow sample data having non-truthful reports, both of which contexts are more general than considered by Paulino and de B. Pereira (1992, 1995), and others. We then offer new methods as a breakthrough in the analysis of categorical data in the general context of informative censoring, both the new theory, as such, and methods for its use in practice. Hence, it will no longer be necessary to make the possibly unrealistic assumptions commonly employed regarding the censoring model. Of course, statistical identifiability will still be needed to have sample information about the totality of unknown parameters, but the available sets of restrictions need not be limited to the old choices of noninformative censoring, truthful reporting, and the like. Several computational procedures are suggested, and an example is then used to illustrate methods.

## 2 Sampling Process and Bayesian Inference

### 2.1 Multiple Bernoulli Sampling Process with Informative censoring

In a sequence of $n$ ($n$ prespecified) multiple-Bernoulli trials having $I$ categories, let $Y_1, \ldots, Y_n$ denote the first, second, ..., $n$-th trial variable. With $\theta_i$ denoting the probability that a trial outcome lies in the $i$-th category for $i = 1, \ldots, I$, write for the $k$-th trial,

$$\Pr(Y_k = i) = \theta_i, \qquad \text{for } k = 1, \ldots, n.$$

Then $\theta_+ = 1$, where $\theta_+ = \sum_{i=1}^{I} \theta_i$. (Throughout this paper, a variable or a parameter with "+" in a subscript represents the sum over all possible such subscript values; for example, $a_{i+} = \sum_{j=1}^{J} a_{ij}$, when the possible values of $j$, for this $i$, are $1, \ldots, J$.) For notational convenience, we use the first $I$ positive integers as the possible values for the categorical variable.

The Dirichlet distributions are the conjugate prior family for samples from such a multiple-Bernoulli distribution. The random vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_I)$ is said to have the Dirichlet distribution $D(\boldsymbol{b})$, denoted by $\boldsymbol{\theta} \sim D(\boldsymbol{b})$, with parameter vector $\boldsymbol{b} =$

$(b_1, \ldots, b_I)$, each $b_i > 0$, if $\boldsymbol{\theta}$ has the following joint density in any $I-1$ of its coordinates. For all $\boldsymbol{\theta}$ in the probability simplex $\{\boldsymbol{\theta} \mid \text{each } \theta_i > 0,\ \theta_+ = 1\}$,

$$f(\boldsymbol{\theta}; \boldsymbol{b}) \equiv B(\boldsymbol{b})^{-1} \prod_{i=1}^{I} \theta_i^{b_i - 1}, \tag{1}$$

where $B(\boldsymbol{b}) = \left[\prod_{i=1}^{I} \Gamma(b_i)\right] \Big/ \Gamma(b_+)$.

The prior general moment, for a Dirichlet prior distribution, $\boldsymbol{\theta} \sim D(\boldsymbol{b})$, is

$$g(\boldsymbol{c}; \boldsymbol{b}) \equiv \underset{\boldsymbol{\theta}|\boldsymbol{b}}{E}(\prod_{i=1}^{I} \theta_i^{c_i}) = B(\boldsymbol{b} + \boldsymbol{c})/B(\boldsymbol{b}).$$

The predictive distribution is then the Dirichlet-multiple-Bernoulli, with mass function $\Pr(\boldsymbol{Y} = \boldsymbol{y}) = g(\boldsymbol{x}; \boldsymbol{b})$, where $\boldsymbol{x} = (x_1, \ldots, x_I)$ is the vector of frequency counts of the outcomes vector $\boldsymbol{y} = (y_1, \ldots, y_n)$ in each of $I$ categories. That is, $x_i$ denotes the number of $y_j$'s equal to $i$, and $\sum_{i=1}^{I} x_i = n$.

Consider, first, the situation when all the data is fully and truthfully categorized. Based on the first trial outcome $Y_1 = y_1$, the posterior distribution, starting from the conjugate prior (1), is again a Dirichlet distribution, with updated parameters,

$$\boldsymbol{\theta}|Y_1 = y_1 \ \sim \ D(\boldsymbol{b} + \boldsymbol{\delta}_{y_1}), \tag{2}$$

where $\boldsymbol{\delta}_y$ denotes an $I$-coordinate vector with value 1 for its $y$-th coordinate, and 0 otherwise. The posterior Dirichlet density is then $f(\boldsymbol{\theta}; \boldsymbol{b}+\boldsymbol{\delta}_{y_1})$ and the posterior moment is $g(\boldsymbol{c}; \boldsymbol{b}+\boldsymbol{\delta}_{y_1})$. Before receiving outcome $y_2$, we now treat (2) as the prior distribution. The posterior distribution, after $Y_2 = y_2$, is then

$$\boldsymbol{\theta}|Y_1 = y_1, Y_2 = y_2 \ \sim \ D(\boldsymbol{b} + \boldsymbol{\delta}_{y_1} + \boldsymbol{\delta}_{y_2}). \tag{3}$$

This process continues until we have received all $n$ outcomes $y_1, \ldots, y_n$. The eventual posterior distribution is

$$\begin{aligned} \boldsymbol{\theta}|y_1, \ldots, y_n \quad &\sim \quad D(\boldsymbol{b} + \boldsymbol{\delta}_{y_1} + \boldsymbol{\delta}_{y_2} + \cdots + \boldsymbol{\delta}_{y_n}) \\ &\sim \quad D(\boldsymbol{b} + \boldsymbol{x}). \end{aligned} \tag{4}$$

The corresponding posterior density is $f(\boldsymbol{\theta}; \boldsymbol{b} + \boldsymbol{x})$ and the posterior moment has the closed form, $g(\boldsymbol{c}; \boldsymbol{b} + \boldsymbol{x})$.

It is likely, in practice, that some of the outcomes may not be reported completely and truthfully. We shall use a random variable $R_k$ for the $k$-th report, say, the report of the $k$-th respondent (or subject), where $k = 1, \ldots, n$. Here, the outcome value $r_k$ of random variable $R_k$ is a set of categories, a non-empty subset of $\{1, \ldots, I\}$. For example, suppose the first respondent, who is actually in the second category, reports as being either in the third or the fourth category. Then, in this example, $r_1 = \{3, 4\}$, and this first report is a non-truthful report, since 2, the true category of the first subject, is not in the reported category set $r_1 = \{3, 4\}$. Assume that there are only $J$ ($J \leq 2^I - 1$) different category sets considered possible, that is, available for reporting. We use $j = 1, \ldots, J$ to index these sets of categories. Further, let $\lambda_{i,r}$ and $\lambda_{ij}$ (respectively, with and without a comma) denote the conditional probabilities that a respondent, who is actually from the $i$-th category, reports in the category set $r$, or the $j$-th category set. Here, for each $i$, $\sum_{\text{all } r} \lambda_{i,r} = \sum_{j=1}^{J} \lambda_{ij} = 1$. Let $\Lambda$ be these probabilities $\lambda_{ij}$ arranged in matrix form. Then, $\Lambda$ is an $I \times J$ conditional-probability matrix. In $n$ trials, the probability of receiving reports $R_1 = r_1, \ldots, R_n = r_n$ is then the product of independent marginal probabilities, each of the form,

$$\Pr(R_k = r_k \mid \boldsymbol{\theta}, \Lambda) = \sum_{i=1}^{I} \lambda_{i,r_k} \theta_i, \tag{5}$$

for each $k = 1, \ldots, n$.

In this paper, we treat the matrix $\Lambda$ as an unknown parameter and consider the most general case for the likelihood function with factors (5) in the sense that we allow a datum to be either informatively or noninformatively censored. If we are sure that the censoring mechanism, itself, is noninformative, we could include restrictions on (5). For example, we could assume an equality of conditional probabilities $\lambda_{1,r_k} = \cdots = \lambda_{I,r_k}$. (Then, the $k$-th report $r_k$ would be wholly uninformative.) On the other hand, if we are sure it is not possible for the $k$-th report $r_k$ to have come from the 1st category, we could assume that $\lambda_{1,r_k} = 0$ in (5).

## 2.2 Bayesian Updating under Restrictions

The following theorem says that the posterior distribution based on joint prior and constrained likelihood function is equivalent to that based on constrained prior and full likelihood function, if both approaches use the same linear constraint function. It is also equivalent, then, to the constrained posterior based on joint prior and full likelihood function, for the same constraint function.

*Theorem* 1 *(Identifiability by Conditioning).*

Suppose there exist joint prior density, likelihood function, and joint posterior density. Consider the following three methods to obtain a constrained posterior distribution for a linear constraint. Conditional distributions involved are conditional on the linear constraint.

(1) Take conditional prior from the joint prior to get constrained prior. Then use Bayes formula to yield constrained posterior.

(2) Take joint prior and constrained likelihood function. Then use Bayes formula to yield constrained posterior.

(3) Take joint prior and likelihood function. Use Bayes formula to yield joint posterior. Then take conditional distribution from the joint posterior to obtain a constrained posterior.

Then

1. The conditional prior (or posterior) density is proportional to the constrained unconditional prior (or posterior) density.
2. If methods (1), (2) and (3) use the same linear function for the constraint, they yield the same constrained posterior distribution.

*Proof.* Purely for notational convenience, we will assume here that the vector referred to by the symbol $\boldsymbol{\theta}$ contains all the unknown parameters. We consider a vector of linear constraints,

$$\boldsymbol{\ell}(\boldsymbol{\theta}) = \sum_i \boldsymbol{\ell}_i \theta_i = \boldsymbol{c}.$$

1. The conditional prior (or posterior) density of the remaining, or linearly redefined variables $\tilde{\boldsymbol{\theta}}$, is just given by the familiar expression for a conditional density in terms of the joint density:

   $$h(\tilde{\boldsymbol{\theta}}|\boldsymbol{\ell} = \boldsymbol{c}) = \frac{f[\boldsymbol{\theta}(\tilde{\boldsymbol{\theta}}, \boldsymbol{\ell})]J}{f_1(\boldsymbol{\ell} = \boldsymbol{c})},$$

   where $J$ is the constant Jacobian and, at $\boldsymbol{\ell} = \boldsymbol{c}$, the marginal density in the denominator just acts as a normalizing constant. So the conditional density, given the linear constraints $\boldsymbol{\ell} = \boldsymbol{c}$, is proportional to the constrained joint density.

2. After a linear change of variable in the prior and the likelihood, obtain a posterior density proportional to the product,

   $$f[\boldsymbol{\theta}(\tilde{\boldsymbol{\theta}}, \boldsymbol{\ell})]J \prod_{k=1}^{n} \left[ \sum_{i=1}^{I} \lambda_{i,r_k} \theta_i(\tilde{\boldsymbol{\theta}}, \boldsymbol{\ell}) \right].$$

   This yields the same constraints on the product, hence, the same posterior density of $\tilde{\boldsymbol{\theta}}$, whether the constraints are first applied to the prior or to the likelihood.

□

The above Identifiability-by-Conditioning Theorem gives us the freedom to apply a constraint function to prior, likelihood, or posterior. Therefore, the approach to take to a problem of unidentifiability can be decided, based on which method is easier to apply or more natural in context.

Dickey, Jiang, and Kadane (1987) developed Bayesian inferences with the following three assumptions concerning the censoring process $\lambda_{i,r}$:

(i) Truthful reporting:

$$\lambda_{i,r} = 0 \text{ when } i \notin r.$$

(ii) Every possible report outcome $R_k = r$ differentially noninformative among the categories within $r$:

$$\lambda_{i,r} = \lambda_{i',r}, \text{ whenever both } i \in r \text{ and } i' \in r.$$

(iii) Prior independence assumed between the parameter arrays $\boldsymbol{\theta}$ and $\Lambda$.

In this paper, we drop assumptions (i) and (ii) and assume (iii) with $I+1$ independent Dirichlet prior distributions,

$$\boldsymbol{\theta} \sim D(\boldsymbol{a}) \text{ and } \boldsymbol{\lambda}_{i*} \sim D(\boldsymbol{b}_{i*}),$$

where $i = 1, \ldots, I$. (In this paper, a variable or a parameter with "$*$" notation in a subscript indicates a vector with components having all the possible such subscript values.) Then, the joint prior probability density function of $\boldsymbol{\theta}$ and $\Lambda$ is proportional to

$$\left(\prod_{i=1}^{I} \theta_i^{a_i-1}\right)\left[\prod_{i=1}^{I}\left(\prod_{j=1}^{J} \lambda_{ij}^{b_{ij}-1}\right)\right]. \tag{6}$$

As an example, consider the structure with all subsets included as possible, available for reports, $\boldsymbol{\lambda}_{i*} = (\lambda_{i,\{1\}}, \lambda_{i,\{2\}}, \ldots, \lambda_{i,\{I\}}, \lambda_{i,\{1,2\}}, \ldots, \lambda_{i,\{1,\ldots,I\}}) = (\lambda_{i1}, \lambda_{i2} \ldots, \lambda_{iI}, \ldots, \lambda_{iJ})$, where $J = 2^I - 1$. This can be considered the most general case, in the following sense. We understand that a zero Dirichlet parameter coordinate, $a_i = 0$ or $b_{ij} = 0$, corresponds to a singular prior distribution in which $\theta_i = 0$ or $\lambda_{ij} = 0$, with probability one, respectively. Then if the experimental data don't disagree with such prior-presumed singularities, they would be preserved in the posterior distribution. If the data disagree, a partially noninformative posterior density, with nonzero arguments, would apply automatically.

Even if the prior is not directly expressed by (6) (i.e. a constrained prior not in the joint prior family), using the Identifiability-by-Conditioning Theorem, we can find, first, the general posterior distribution based on (5) and (6), and then use a suitable method, e.g. transformation of variables, to obtain the constrained posterior distribution. Therefore, our discussion in this paper will focus on the general prior (6) and general likelihood (5).

Consider the inference following a report only of the first trial (first respondent). The posterior probability density function resulting from (5) for $k = 1$ (first trial) and the joint prior (6) is then proportional to

$$\begin{aligned}&\left(\prod_{i=1}^{I} \theta_i^{a_i-1}\right)\left[\prod_{i=1}^{I}\left(\prod_{j=1}^{J} \lambda_{ij}^{b_{ij}-1}\right)\right]\left(\sum_{i=1}^{I} \lambda_{i,r_1}\theta_i\right)\\ &\quad = \sum_{m=1}^{I}\left\{\left(\prod_{i=1}^{I} \theta_i^{a_i+\delta_i^m-1}\right)\left[\prod_{i=1}^{I}\left(\prod_{j=1}^{J} \lambda_{ij}^{b_{ij}+\delta_{ij}^{mr_1}-1}\right)\right]\right\},\end{aligned} \tag{7}$$

where $\delta_i^m$ is 1 for $i = m$, and is 0 otherwise, and $\delta_{ij}^{mr_1}$ is 1 for $i = m$ and $j = r_1$, and is 0 otherwise. This posterior density can be expressed as

$$\sum_{m=1}^{I} \frac{A_m}{A_+} \left\{ \left( \prod_{i=1}^{I} \theta_i^{a_i+\delta_i^m-1} \right) \left[ \prod_{i=1}^{I} \left( \prod_{j=1}^{J} \lambda_{ij}^{b_{ij}+\delta_{ij}^{mr_1}-1} \right) \right] \right\} \Big/ A_m, \tag{8}$$

where $A_m = B(\boldsymbol{a} + \boldsymbol{\delta}^m) \prod_{i=1}^{I} B(\boldsymbol{b}_{i*} + \boldsymbol{\delta}_{i*}^{mr_1})$, $\boldsymbol{a} = (a_1, \dots, a_I)$, $\boldsymbol{\delta}^m = (\delta_1^m, \dots, \delta_I^m)$, $\boldsymbol{b}_{i*} = (b_{i1}, \dots, b_{iJ})$, $\boldsymbol{\delta}_{i*}^{mr_1} = (\delta_{i1}^{mr_1}, \dots, \delta_{iJ}^{mr_1})$, and $A_+ = \sum_{m=1}^{I} A_m$.

Here the posterior density (8) is a weighted average of $I$ products of Dirichlet densities. The weight $A_m/A_+$ in the $m^{\text{th}}$ term is proportional to $(a_m * b_{mr_1}/b_{m+})$. The posterior moment is the similarly weighted average of $I$ ratios of $A$'s, each of the form $A'_m/A_m$, where $A'_m$ is, like $A_m$, a product of $B$'s. For example, the posterior mean of $\theta_1$ is $\sum_{m=1}^{I} (A_m/A_+) \, (A_m^{(1)}/A_m)$, where $A_m^{(1)} = B(\boldsymbol{a}^{(1)} + \boldsymbol{\delta}^m) \prod_{i=1}^{I} B(\boldsymbol{b}_{i*} + \boldsymbol{\delta}_{i*}^{mr_1})$, and $\boldsymbol{a}^{(1)} = (a_1 + 1, a_2, \dots, a_I)$.

If we receive, now, the further report $R_2 = r_2$, the updated posterior probability density function is proportional to the product of (8) and (5) for $k = 2$. This would give the new posterior p.d.f. as a mixture of $I^2$ products of Dirichlet densities. The posterior moment is now the weighted average of $I^2$ ratios of $A$'s. As the number of reports received increases, the number of ratios of $A$'s increases dramatically. This would make the computation of posterior moments unfeasible. In the next section, we shall give posterior probability density functions and suggest possible uses of alternative computational methods for posterior moments even when the sample size is not small.

## 3 Posterior distribution and computational methods

With the general prior (6) and general likelihood function (5), the posterior probability density function, after receiving reports $R_1 = r_1, \dots, R_n = r_n$, is then proportional to

$$\left( \prod_{i=1}^{I} \theta_i^{a_i-1} \right) \left[ \prod_{i=1}^{I} \left( \prod_{j=1}^{J} \lambda_{ij}^{b_{ij}-1} \right) \right] \left[ \prod_{k=1}^{n} \left( \sum_{i=1}^{I} \lambda_{i,r_k} \theta_i \right) \right] . \tag{9}$$

Let $n_j = \sum_{k=1}^{n} \delta_j(r_k)$ for $1 \le j \le J$, where $\delta_j(r_k)$ is 1 if $r_k = j$, and is 0 otherwise. Hence, $\sum_{j=1}^{J} n_j = n$. Define $w_{ij} = \theta_i \lambda_{ij}$, for $i = 1, \dots, I$, and $j = 1, \dots, J$. Then the posterior probability density function corresponding to (9) can be reexpressed in terms of $w_{ij}$'s. The kernel of the density can be shown to be

$$\left[ \prod_{i=1}^{I} \left( \prod_{j=1}^{J} w_{ij}^{b_{ij}-1} \right) \right] \left[ \prod_{i=1}^{I} w_{i+}^{a_i - b_{i+}} \right] \left[ \prod_{j=1}^{J} w_{+j}^{n_j} \right] , \tag{10}$$

since the Jacobian is $\{\prod_{i=1}^{I} w_{i+}^{1-J}\}$.

Now, let $K = I \cdot J$, the $1 \times K$ row vector $\boldsymbol{b} = [\boldsymbol{b}_{1*}, \boldsymbol{b}_{2*}, \dots, \boldsymbol{b}_{I*}]$, the $K \times I$ matrix $G^{(1)} = [\boldsymbol{g}_{*1}^{(1)}, \boldsymbol{g}_{*2}^{(1)}, \dots, \boldsymbol{g}_{*I}^{(1)}]$, and the $K \times J$ matrix $G^{(2)} = [\boldsymbol{g}_{*1}^{(2)}, \boldsymbol{g}_{*2}^{(2)}, \dots, \boldsymbol{g}_{*J}^{(2)}]$, where, for each $i$ and $1 \leq i \leq I$, $\boldsymbol{b}_{i*} = (b_{i1}, b_{i2}, \ldots, b_{iJ})$ is a $1 \times J$ row vector and $\boldsymbol{g}_{*i}^{(1)}$ is a $K \times 1$ column vector having 1 on each of the $J \times (i-1) + 1^{st}$, $J \times (i-1) + 2^{nd}$, ..., and $J \times (i-1) + J^{th}$ components and having 0 otherwise, and, for each $j$ and $1 \leq j \leq J$, $\boldsymbol{g}_{*j}^{(2)}$ is a $K \times 1$ column vector having 1 on each of the $j^{th}$, $J + j^{th}$, ..., $(I-1) \cdot J + j^{th}$ components and having 0 otherwise. The above posterior distribution is a generalized Dirichlet distribution, as defined by Dickey (1983) and Dickey, Jiang, and Kadane (1987). Its density can be expressed as

$$B^{-1}\left[\boldsymbol{b}\right] \cdot \mathcal{R}^{-1}\left(\boldsymbol{b}, G, -\boldsymbol{e}\right) \cdot \left[\prod_{i=1}^{I}\left(\prod_{j=1}^{J} w_{ij}^{b_{ij}-1}\right)\right]\left[\prod_{i=1}^{I} w_{i+}^{a_i - b_{i+}}\right]\left[\prod_{j=1}^{J} w_{+j}^{n_j}\right] , \tag{11}$$

where $\boldsymbol{e} = (a_1 - b_{1+}, a_2 - b_{2+}, \ldots, a_I - b_{I+}, n_1, n_2, \ldots, n_J)$, $G = [G^{(1)}|G^{(2)}]$ is a $K \times (I+J)$ matrix, and $\mathcal{R}$ is a Carlson (1977) multiple hypergeometric function. (See Dickey (1983) for probabilistic interpretations and statistical uses of $\mathcal{R}$.) Therefore, the posterior moment is

$$E\left(\prod_{i=1}^{I}\prod_{j=1}^{J} w_{ij}^{d_{ij}}\right) = \frac{B[\boldsymbol{b}+\boldsymbol{d}] \cdot \mathcal{R}(\boldsymbol{b}+\boldsymbol{d}, G, -\boldsymbol{e})}{B[\boldsymbol{b}] \cdot \mathcal{R}(\boldsymbol{b}, G, -\boldsymbol{e})} , \tag{12}$$

which is proportional to a ratio of Carlson functions $\mathcal{R}$. Jiang, Kadane, and Dickey (1992) give computational methods for ratios of $\mathcal{R}$.

We note that the prior distributions that Paulino and de B. Pereira (1995) and Tian, Ng, and Geng (2003) consider are cases where $a_i = b_{i+}$, for all $i = 1, \ldots, I$. In such a special case, (10) can be reexpressed as

$$\left[\prod_{i=1}^{I}\left(\prod_{j=1}^{J} w_{ij}^{b_{ij}-1}\right)\right]\left[\prod_{j=1}^{J} w_{+j}^{n_j}\right] . \tag{13}$$

Using the definition of Jiang, Kadane, and Dickey (1992, p.235), the matrix parameter in Carlson's $\mathcal{R}$ for the normalizing constant of (13) is a 1-level nested-partition indicator. Theorem 2 of Jiang, Kadane, and Dickey (1992) gives a closed form of Carlson's $\mathcal{R}$ function for any level of nested-partition indicator matrix. Hence the normalizing constant of (13) can be expressed in closed form, and so the posterior mean and standard deviation of $\theta_i$ can be easily calculated.

An application area expert (or experts) may not feel comfortable, however, with this restriction when he/she assesses the prior. For example, an expert may have the prior equivalent of 10 observations on first category, but may have only 6 observations on how they might be reported. In this case, he/she would have $a_1 = 10$ and $b_{1+} = 6$. The prior family should be large enough to allow a choice accurately expressing the real predata expert uncertainty concerning $\theta$ and $\Lambda$ (see, e.g., Dickey and Jiang (1998, p.651)). So, in this paper, we make no such restriction on the $a's$ and $b's$.

The computation of posterior moments, the moments of (11) may be done by the "expansion method" or by the Monte Carlo method, given by Jiang, Kadane, and Dickey

(1992). If the sample size is not small, however, the "expansion method" is not feasible. Although the Monte Carlo method and even the Gibbs sampler can give good approximation to the posterior estimates, we note that the quasi-Bayes method, given by Jiang and Dickey (2007), which is analogous to the methods given by Makov and Smith (1977), Smith and Makov (1978) and Titterington, Smith, and Makov (1985, Chapter 6), provides a more efficient approximate computational method for our problems. Specifically, the quasi-Bayes method provides a simpler and easier algorithm for coding and also requires much less CPU time for better accuracy. See Jiang and Dickey (2007) for details. Further, Jiang and Dickey (2007) show that the quasi-Bayes method gives the same posterior means as those given by Paulino and de B. Pereira (1992, 1995) under their restricted priors.

## 4 Example

In this section, to illustrate methods, we use the Monte Carlo, the quasi-Bayes, and the Gibbs sampler approaches to reanalyze the data from Paulino and de B. Pereira (1995). The problem is concerned with the determination of the degree of sensitivity to dental caries, categorized in three risk levels: low, medium, and high. We label the low, medium, and high levels by 1, 2, and 3, respectively, and assume the possible reported category sets are $\{1\}$, $\{2\}$, $\{3\}$, $\{1, 2\}$, and $\{2, 3\}$, labelled as $1, 2, 3, 4$, and 5, respectively. Therefore, $I = 3$ and $J = 5$ in this example. Using our notation, the count data are $n_1 = 14$, $n_2 = 17$, $n_3 = 20$, $n_4 = 28$ and $n_5 = 18$. Two different priors (Table 1) are used here. The first prior, which was given by Paulino and de B. Pereira (1995), is the case where $a_i = b_{i+}$, for all $i = 1, 2$, and 3. In the second prior, $a_i \neq b_{i+}$, for any $i$.

At least two possible scenarios could make the above second prior highly possible.

Scenario 1:
An expert gives his/her prior knowledge on $\theta$ that there are 8, 6, and 8 equivalent sample subjects in low, medium, and high risk levels, respectively. However, this expert is not able to translate his/her prior knowledge on $[\lambda_{ij}]$ into equivalent sample evidence, but he/she does think that three independent Dirichlet distributions would adequately represent his/her knowledge on $\lambda_{i*}$, $i = 1, 2, 3$. He/she then gives the means and the variances of these three distributions. Finally, equating these to the formulas for the mean and the variance of a Dirichlet distribution, one can solve for the parameters of Dirichlet distributions. These parameters $[b_{ij}]$ are given in the bottom right of Table 1. The entire parameters for the second prior are given in the bottom of Table 1.

Scenario 2:
A partial prior is given by an expert who gives his/her initial knowledge that is reported in category sets $\{1\}$, $\{2\}$, $\{3\}$, $\{1, 2\}$, $\{2, 3\}$ of each risk level, as shown in the bottom right of Table 1. At the next stage, he/she thinks that the equivalent 8, 6, and 8, instead of 4, 3, and 4, sample subjects would be best adequate to represent his/her prior knowledge on $\theta$. Therefore, this expert's true prior information is given in the bottom of Table 1.

The old methods fail for both of these two scenarios, since $a_i \neq b_{i+}$ for at least one $i$.

The posterior means and standard deviations, using the Monte Carlo method, quasi-Bayes method, and the Gibbs sampler, are given in Table 2. Under the first prior, the quasi-Bayes method gives the exact posterior means, which are the same as those given by Paulino and de B. Pereira (1995). Under the second prior, Table 2 shows that the posterior mean of $\theta_i$ lies between the value under the first prior and $a_i/a_+$, for all $i$. These results are consistent with what one would anticipate, since the $b_{ij}$'s are the same for both priors and each $a_i$ is increased in the second prior. We note that the CPU time for the quasi-Bayes method here is practically 0 seconds, while those for the Monte Carlo method and the Gibbs sampler are about 14.4 seconds and 1.5 seconds, respectively. In addition, by comparing the posterior means obtained using the Monte Carlo method, the Gibbs sampler is seen to be less accurate than our quasi-Bayes method.

## 5 Conclusions

Methods here generalize Dickey, Jiang, and Kadane (1987)'s fully Bayesian methods by removing two of their three conditions. Our methods also generalize the cases considered by Paulino and de B. Pereira (1992, 1995), and others. From now on, it will no longer be necessary to assume such restrictions, when not realistic in regard to the censoring model. Our example illustrates that the new methods are convenient and computationally feasible.

## Appendix

Table 1: Prior information

| prior | $\boldsymbol{a}'$ | $[b_{ij}]$ |
|---|---|---|
| 1 | $\begin{pmatrix} 4 \\ 3 \\ 4 \end{pmatrix}$ | $\begin{bmatrix} 3 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 3 & 0 & 1 \end{bmatrix}$ |
| 2 | $\begin{pmatrix} 8 \\ 6 \\ 8 \end{pmatrix}$ | $\begin{bmatrix} 3 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 3 & 0 & 1 \end{bmatrix}$ |

Table 2: Posterior estimates of $\boldsymbol{\theta}$

| prior | method | $E(\theta_1)$ | $E(\theta_2)$ | $E(\theta_3)$ | CPU time (Intel(R) Xeon(TM) 3.20GHz) |
|---|---|---|---|---|---|
| 1 | m | 0.2965 | 0.3988 | 0.3047 | 13.87500 sec |
| | | (0.0900) | (0.1064) | (0.0678) | |
| | q | 0.2963 | 0.3981 | 0.3056 | $\doteq 0$ sec |
| | | (0.0437) | (0.0469) | (0.0441) | |
| | g | 0.3114 | 0.3833 | 0.3053 | 1.546875 sec |
| | | (0.0962) | (0.1103) | (0.0707) | |
| 2 | m | 0.3286 | 0.3507 | 0.3208 | 14.42188 sec |
| | | (0.0772) | (0.0880) | (0.0612) | |
| | q | 0.3101 | 0.3738 | 0.3161 | $\doteq 0$ sec |
| | | (0.0422) | (0.0422) | (0.0424) | |
| | g | 0.3356 | 0.3460 | 0.3185 | 1.515625 sec |
| | | (0.0811) | (0.0883) | (0.0639) | |

Note:

1. Entries in parentheses are the (approximate) SD's.
2. Methods m, q, and g are Monte Carlo, quasi-Bayes, and Gibbs sampler, respectively.

## References

Albert, J. H. (1985). "Bayesian Estimation Methods for Incomplete Two-way Contingency Tables Using Prior Belief of Association." In *Bayesian Statistics* 2, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Amsterdam: North-Holland, 589–602. 541

Albert, J. H. and Gupta, A. K. (1983). "Bayesian Estimation Methods for 2 × 2 Contingency Tables Using Mixtures of Dirichlet Distributions." *Journal of the American Ststistical Association*, 78: 708–717. 541

Antelman, G. R. (1972). "Interrelated Bernoulli Processes." *Journal of the American Statistical Association*, 67: 831–841. 541

Basu, D. and de B. Pereira, C. A. (1982). "On the Bayesian Analysis of Categorical Data: The Problem of Nonresponse." *Journal of Statistical Planning and Inference*, 6: 345–362. 541

Carlson, B. C. (1977). *Special Functions of Applied Mathematics.* New York: Academic Press. 548

Chen, T. and Fienberg, S. E. (1974). "Two-Dimensional Contingency Tables With Both Completely and Partially Cross-Classified Data." *Biometrics*, 30: 629–642. 541

— (1976). “The Analysis of Contingency Tables With Incompletely Classified Data.” *Biometrics*, 32: 133–144. 541

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). “Maximum Likelihood From Incomplete Data via the EM Algorithm (with discussion).” *Journal of the Royal Statistical Society B*, 39: 1–38. 541

Dickey, J. M. (1983). “Multiple Hypergeometric Functions: Probabilistic Interpretations and Statistical Uses.” *Journal of the American Statistical Association*, 78: 628–637. 548

Dickey, J. M., Jiang, J. M., and Kadane, J. B. (1987)., “Bayesian Methods for Censored Categorical Data.” *Journal of the American Statistical Association*, 82: 773–781. 541, 545, 548, 550

Dickey, J. M. and Jiang, T. J. (1998). “Filtered-Variate Prior Distributions for Histogram Smoothing.” *Journal of the American Statistical Association*, 93: 651–662.

Gibbons, P. C. and Greenberg, E. (1989). Bayesian Analysis of Contingency Tables With Partially Categorized Data. Typescript, Washington University, St. Louis, Missouri 63130. 541

Gunel, E. (1984)., “A Bayesian Analysis of the Multinomial Model for a Dichotomous Response With Nonrespondents.” *Communications in Statistics—Theory and Methods*, 13: 737–751. 541

Hartley, H. O. (1958). “Maximum Likelihood Estimation From Incomplete Data.” *Biometrics*, 14: 174–194. 541

Jiang, T. J. and Dickey, J. M. (2007). “Quasi-Bayes Methods for Categorical Data Under Informative Censoring.” To appear. 549

Jiang, T. J., Kadane, J. B., and Dickey, J. M. (1992). “Computation of Carlson’s Multiple Hypergeometric Function $\mathcal{R}$ for Bayesian Applications.” *Journal of Computational and Graphical Statistics*, 1: 231–251. 548

Kadane, J. B. (1985). “Is Victimization Chronic? A Bayesian Analysis of Multinomial Missing Data.” *Journal of Econometrics*, 29: 47–67. 541

Karson, M. J. and Wrobleski, W. J. (1970). “A Bayesian Analysis of Binomial Data With a Partially Informative Category.” In *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 532–534. 541

Kaufman, G. M. and King, B. (1973). “A Bayesian Analysis of Nonresponse in Dichotomous Processes.” *Journal of the American Statistical Association*, 68: 670–678. 541

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, New York: John Wiley & Sons. 541

Makov, U. E. and Smith, A. F. M. (1977). "A Quasi-Bayes Unsupervised Learning Procedure for Priors." *IEEE Trans. Inf. Theory*, IT-23, 761–764. 549

Paulino, C. D. M. and de B. Pereira, C. A. (1992). "Bayesian Analysis of Categorical Data Informatively Censored." *Communications in Statistics—Theory and Methods*, 21: 2689–2705. 541, 542, 549, 550

— (1995). "Bayesian Methods for Categorical Data Under Informative General Censoring." *Biometrika*, 82: 439–446. 541, 542, 548, 549, 550

Smith, P. J., Choi, S. C., and Gunel, E. (1985). "Bayesian Analysis of a 2 × 2 Contingency Table With Both Completely and Partially Cross-Classified Data." *Journal of Educational Statistics*, 10: 31–43. 541

Smith, P. J. and Gunel, E. (1984). "Practical Bayesian Approaches to the Analysis of 2 × 2 Contingency Table With Incompletely Categorized Data." *Communications in Statistics—Theory and Methods*, 13: 1941–1963. 541

Smith, A. F. M. and Makov, U. E. (1978). "A Quasi-Bayes Sequential Procedure for Mixtures." *Journal of the Royal Statistical Society B*, 40: 106–112. 549

Tian, G.-L., Ng, K. W., and Geng, Z. (2003). "Bayesian Computation for Contingency Tables with Incomplete Cell-Counts." *Statistica Sinica*, 13: 189–206. 541, 548

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*, New York: John Wiley & Sons.

Walker, S. (1996). "A Bayesian Maximum a Posteriori Algorithm for Categorical Data Under Informative General Censoring." *The Statistician*, 45: 293–298.

**Acknowledgments**

This research was supported in part by the National Science Council, Taiwan. The authors thank the Editor and two referees for their comments.