Statistica Sinica **19** (2009), 251-271

THE NESTED DIRICHLET DISTRIBUTION AND INCOMPLETE CATEGORICAL DATA ANALYSIS

Kai Wang Ng[1], Man-Lai Tang[2], Guo-Liang Tian[1] and Ming Tan[3]

[1] *The University of Hong Kong,* [2]*Hong Kong Baptist University, and* [3] *University of Maryland Greenebaum Cancer Center*

Abstract: The *nested Dirichlet distribution* (NDD) is an important distribution defined on the closed n-dimensional simplex. It includes the classical Dirichlet distribution and is useful in *incomplete categorical data* (ICD) analysis. In this article, we develop the distributional properties of NDD. New large-sample likelihood and small-sample Bayesian approaches for analyzing ICD are proposed and compared with existing likelihood/Bayesian strategies. We show that the new approaches have at least three advantages over existing approaches based on the traditional Dirichlet distribution in both frequentist and conjugate Bayesian inference for ICD. The new methods possess closed-form expressions for both the maximum likelihood and Bayes estimates when the likelihood function is in NDD form; produce computationally efficient EM and data augmentation algorithms when the likelihood is not in NDD form; and provide exact sampling procedures for some special cases. The methodologies are illustrated with simulated and real data.

Key words and phrases: Data augmentation, Dirichlet distribution, EM, incomplete categorical data, matrix rate of convergence, mixing rate of a markov chain, nested Dirichlet distribution.

1. Introduction

The *Dirichlet distribution* (DD) is usually employed as a conjugate prior for the multinomial model in Bayesian analysis of complete contingency tables (Agresti (2002)). Gupta and Richards (1987, 1991, 1992) extended the DD to the Liouville distribution. Fang, Kotz and Ng (1990, Chap. 5) gave an extensive exposition of the Liouville family and its ramifications. Rayens and Srinivasan (1994) studied the generalized Liouville family and considered its application to compositional data analysis (Aitchison (1986)).

For the analysis of *incomplete categorical data* (ICD), Schafer (1997) showed how EM and *data augmentation* (DA) algorithms based on a multinomial model with a Dirichlet prior could be applied under the *missing at random* (MAR) mechanism (Rubin (1976)). Noting that Dirichlet priors do not always provide sufficient flexibility, Dickey (1983) discussed a nested family of distributions that generalize the DD and argued that they are more appropriate for ICD analysis.

Recently, Ng, Tang, Tan and Tian (2008) extended the DD to a new family of grouped Dirichlet distributions and studied its application to ICD analysis. In this article, we discuss the *nested Dirichlet distribution* (NDD) on the closed n-dimensional simplex, which also has important applications in ICD analysis. To our knowledge, NDD was first introduced briefly in Tian, Ng and Geng (2003), in which only one *stochastic representation* (SR) was provided.

The main goal of this article is to investigate the NDD family. Distribution properties such as SR, mixed (or raw) moments and mode are discussed. We examine large-sample likelihood inferences and small-sample Bayesian inferences for ICD based on the NDD. Comparisons between our proposed methods and existing likelihood/Bayesian strategies are presented. We show theoretically that the proposed approaches have at least three advantages over the commonly used approaches based on DD in both frequentist and conjugate Bayesian inference for ICD: when the likelihood takes the NDD form, both the maximum likelihood and Bayes estimates have closed-form expressions in the new approach; when the likelihood is not in NDD form, the corresponding EM and DA algorithms based on our new approaches converge much faster; an exact non-iterative sampling procedure is available for some special cases. The proposed methodologies are illustrated with a hypothetical example and a dental study consisting of ICD. We conclude with a discussion. All technical details are left to the Appendix.

2. The Nested Dirichlet Distribution

2.1. Notations and density function

Let $\mathbf{x} = (x_1, \ldots, x_n)^\top$ denote an n-vector, $||\mathbf{x}|| = \sum_{i=1}^n x_i$ the ℓ_1-norm of $\mathbf{x}$, and $\mathbf{x}_{-n} = (x_1, \ldots, x_{n-1})^\top$. We adopt the following notations throughout this paper.

$\hat{=}$	definition
$\overset{d}{=}$	having the same distribution on both sides
$\mathbb{R}_+^n$	$= \{\mathbf{x} : x_i \geq 0, i = 1, \ldots, n\}$
$\mathbb{T}_n$	$= \{\mathbf{x} : \mathbf{x} \in \mathbb{R}_+^n \text{ and } ||\mathbf{x}|| = 1\},$ the closed simplex
$\mathbb{V}_{n-1}(d)$	$= \{\mathbf{x}_{-n} : \mathbf{x}_{-n} \in \mathbb{R}_+^{n-1} \text{ and } ||\mathbf{x}|| \leq d\},$ the open simplex
$\mathbb{V}_{n-1}$	$= \mathbb{V}_{n-1}(1)$
$\Gamma(a)$	$= \int_0^\infty x^{a-1} e^{-x}\, dx,\ a > 0,$ gamma function
$B_n(a_1, \ldots, a_n)$	$= \prod_{i=1}^n \Gamma(a_i)/\Gamma(\sum_{i=1}^n a_i),$ multivariate beta function
$B(a_1, a_2)$	$= B_2(a_1, a_2),$ beta function
$\mathrm{Beta}(x\|a_1, a_2)$	$= x^{a_1-1}(1-x)^{a_2-1}/B(a_1, a_2),\ 0 \leq x \leq 1,$ beta density
$\mathrm{D}_n(\mathbf{x}\|\mathbf{a})$	$= (\prod_{i=1}^n x_i^{a_i-1})/B_n(\mathbf{a}),\ \mathbf{x} \in \mathbb{T}_n,$ Dirichlet density
$\mathrm{D}_n(\mathbf{a})$	Dirichlet distribution with density $\mathrm{D}_n(\mathbf{x}\|\mathbf{a})$.

An n-vector $\mathbf{x} \in \mathbb{T}_n$ is said to follow an NDD, if the density of $\mathbf{x}_{-n}$ is

$$\mathrm{ND}_{n,n-1}(\mathbf{x}_{-n}|\mathbf{a}, \mathbf{b}) = c^{-1} \cdot (\textstyle\prod_{i=1}^{n} x_i^{a_i - 1}) \cdot \prod_{j=1}^{n-1} (\sum_{k=1}^{j} x_k)^{b_j}, \quad \mathbf{x}_{-n} \in \mathbb{V}_{n-1}, \quad (2.1)$$

where $\mathbf{a} = (a_1, \ldots, a_n)^\top$ is a positive parameter vector, $\mathbf{b} = (b_1, \ldots, b_{n-1})^\top$ is a non-negative parameter vector, $c = \prod_{j=1}^{n-1} B(d_j,\, a_{j+1})$ is the normalizing constant, and

$$d_j \mathrel{\hat{=}} \textstyle\sum_{k=1}^{j} (a_k + b_k). \quad (2.2)$$

We write $\mathbf{x} \sim \mathrm{ND}_{n,n-1}(\mathbf{a}, \mathbf{b})$ on $\mathbb{T}_n$ or $\mathbf{x}_{-n} \sim \mathrm{ND}_{n,n-1}(\mathbf{a}, \mathbf{b})$ on $\mathbb{V}_{n-1}$ accordingly. It is noteworthy that when all $b_j = 0$ the NDD in (2.1) reduces to Dirichlet distribution $\mathrm{D}_n(\mathbf{a})$. One motivation of the NDD density (2.1) comes from the factorization of the likelihood with a monotone pattern for ICD (Rubin (1974) and Little and Rubin (2002, Chap. 13)).

2.2. Two motivating examples

In the first example, the likelihood can be expressed exactly in terms of a NDD (up to a normalizing constant). In the second example, the likelihood can be written as a product of two terms, namely a NDD (up to a normalizing constant) and a product of powers of linear combination of the parameters of interest. Efficient methods for analyzing these data sets are developed in subsequent sections.

Example 1. *Sample surveys with nonresponse.* Let N denote the total number of questionnaires sent out, suppose m individuals respond, and the rest do not. Of the m respondents, there are n_{r+i} individuals whose answers are classified into category i (denoted by $X = i$), $i = 1, \ldots, r$. Denote the nonrespondents by $R = 0$ and the respondents by $R = 1$. Let $\theta_i = \Pr(X = i, R = 0)$ and $\theta_{r+i} = \Pr(X = i, R = 1)$ denote the cell probabilities, $i = 1, \ldots, r$. The parameter of interest is $\Pr(X = i) = \theta_i + \theta_{r+i}$ (Albert and Gupta (1985)). Let $Y_{\text{obs}} = \{(n_{r+1}, \ldots, n_{2r});\ N - m\}$ denote the observed counts and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{2r})^\top$. Under the MAR assumption, the likelihood function for the observed data Y_{obs} is given by

$$L(\boldsymbol{\theta}|Y_{\text{obs}}) \propto (\textstyle\prod_{i=r+1}^{2r} \theta_i^{n_i}) \cdot (\theta_1 + \cdots + \theta_r)^{N-m}, \quad m = \sum_{i=r+1}^{2r} n_i, \quad \boldsymbol{\theta} \in \mathbb{T}_{2r}. \quad (2.3)$$

Obviously, if we treat $\boldsymbol{\theta}$ as a random vector, then $\boldsymbol{\theta} \sim \mathrm{ND}_{2r,2r-1}(\mathbf{a}, \mathbf{b})$ where $\mathbf{a}$ is a $2r \times 1$ vector with $a_i = 1$ for $i = 1, \ldots, r$ and $a_i = n_i + 1$ for $i = r+1, \ldots, 2r$, $\mathbf{b}$ is a $(2r-1) \times 1$ vector with $b_j = 0$ for $j \neq r$ and $b_j = N - m$ for $j = r$.

Example 2. *Dental caries data.* To determine the degree of sensitivity to dental caries, dentists often consider three risk levels: low, medium and high, labeled $X = 1$, $X = 2$ and $X = 3$, respectively. Each subject is assigned a risk level based on the spittle color obtained using a coloration technique. However, some subjects may not be fully categorized due to the inability to distinguish adjacent categories. Paulino and Pereira (1995) analyzed the following data set using Bayesian methods. Of 97 subjects, only 51 were fully categorized, with $n_1 = 14$, $n_2 = 17$, and $n_3 = 20$ subjects being classified as low, medium and high, respectively; total of $n_{12} = 28$ subjects were only classified as low or medium risk, and $n_{23} = 18$ as medium or high risk. The primary objective is the estimation of the cell probabilities. Let $Y_{\text{obs}} = \{(n_1, n_2, n_3); (n_{12}, n_{23})\}$ be the observed frequencies and $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^\top$ the corresponding cell probability vector. Under the assumption of MAR, the observed data likelihood function is

$$L(\boldsymbol{\theta}|Y_{\text{obs}}) \propto \left\{(\textstyle\prod_{i=1}^3 \theta_i^{n_i})\theta_1^0(\theta_1+\theta_2)^{n_{12}}\right\} \cdot (\theta_2+\theta_3)^{n_{23}}, \quad \boldsymbol{\theta} \in \mathbb{T}_3. \tag{2.4}$$

Again, we observe that the first term in (2.4) follows the $\text{ND}_{3,2}(\mathbf{a}, \mathbf{b})$ with $\mathbf{a} = (n_1, n_2, n_3)^\top$ and $\mathbf{b} = (0, n_{12})^\top$, up to a normalizing constant, while the second term is simply a power of a linear combination of $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^\top$.

2.3. Stochastic representation, moments and mode

The first proposition provides a SR of an NDD yielding a simple procedure for generating *independently and identically distributed* (i.i.d.) samples, which in turn play a crucial role in Bayesian analysis for ICD. The result indicates that the NDD can be stochastically represented by a sequence of mutually independent Beta variates. The proof of Proposition 1 is given in the Appendix. As an immediate result of Proposition 1, the second proposition suggests another SR for NDD, and plays an important role in the derivation of marginal and conditional distributions of an NDD. Throughout this paper, $\{d_j\}_{j=1}^{n-1}$ are defined in (2.2).

Proposition 1. *An n-vector* $\mathbf{x} \sim ND_{n,n-1}(\mathbf{a}, \mathbf{b})$ *on* $\mathbb{T}_n$ *if and only if*

$$\begin{cases} x_i \overset{d}{=} (1-y_{i-1}) \cdot \prod_{j=i}^{n-1} y_j, & y_0 \equiv 0, \quad i = 1, \ldots, n-1, \\ x_n \overset{d}{=} 1 - y_{n-1}, \end{cases} \tag{2.5}$$

where $y_j \sim Beta(d_j,\ a_{j+1})$, *and the* $y_1,\ \ldots,\ y_{n-1}$ *are mutually independent.*

Proposition 2. *An n-vector* $\mathbf{x} \sim ND_{n,n-1}(\mathbf{a}, \mathbf{b})$ *on* $\mathbb{T}_n$ *if and only if*

$$\begin{cases} x_1 + \cdots + x_i \overset{d}{=} \prod_{j=i}^{n-1} y_j, & i = 1, \ldots, n-1, \\ x_n \overset{d}{=} 1 - y_{n-1}, \end{cases} \tag{2.6}$$

where $\{y_j\}_{j=1}^{n-1}$ *are defined in Proposition 1.*

Next, in Proposition 3 below, we present the first- and second-order moments of $\mathbf{x}$ by using (2.5). Using (2.6), we give the higher order moments of $\sum_{j=1}^{i} x_j$ in Proposition 4. The corresponding proofs are also given in the Appendix.

Proposition 3. *Let* $\mathbf{x} \sim ND_{n,n-1}(\mathbf{a}, \mathbf{b})$ *on* $\mathbb{T}_n$. *Then*

$$E(x_i) = \frac{a_i}{d_{i-1}+a_i} \prod_{j=i}^{n-1} \frac{d_j}{d_j + a_{j+1}}, \quad i = 1, \ldots, n,$$

$$E(x_i^2) = \frac{a_i(a_i+1)}{(d_{i-1}+a_i)(d_{i-1}+a_i+1)} \prod_{j=i}^{n-1} \frac{d_j(d_j+1)}{(d_j+a_{j+1})(d_j+a_{j+1}+1)}, \quad \textit{and}$$

$$\begin{aligned} E(x_i x_j) = {} & \frac{a_i}{d_{i-1}+a_i} \prod_{k=i}^{j-2} \frac{d_k}{d_k + a_{k+1}} \\ & \cdot \frac{a_j d_{j-1}}{(d_{j-1}+a_j)(d_{j-1}+a_j+1)} \prod_{k=j}^{n-1} \frac{d_k(d_k+1)}{(d_k+a_{k+1})(d_k+a_{k+1}+1)}, \quad i < j. \end{aligned}$$

Proposition 4. *Let* $\mathbf{x} \sim ND_{n,n-1}(\mathbf{a}, \mathbf{b})$ *on* $\mathbb{T}_n$. *For any* $r \geq 0$ *we have*

$$E(\textstyle\sum_{j=1}^{i} x_j)^r = \prod_{j=i}^{n-1} [B(d_j + r,\ a_{j+1})/B(d_j,\ a_{j+1})], \quad i = 1, \ldots, n-1.$$

Finally, Proposition 5 gives a closed-form expression for the mode of an NDD density, implying that explicit MLEs of cell probabilities are available in the frequentist analysis of ICD. The proof is given in the Appendix.

Proposition 5. *The mode of the nested Dirichlet density* (2.1) *is given by*

$$\begin{cases} \hat{x}_n = \dfrac{a_n - 1}{d_{n-1} + a_n - n}, \\ \hat{x}_i \ = \dfrac{(a_i - 1)(1 - \hat{x}_{i+1} - \hat{x}_{i+2} - \cdots - \hat{x}_n)}{d_{i-1} + a_i - i}, \quad i = 2, \ldots, n-1, \\ \hat{x}_1 \ = 1 - \hat{x}_2 - \cdots - \hat{x}_n. \end{cases} \tag{2.7}$$

3. Large-Sample Likelihood Inference

In this section, we consider ICD analyses with large sample sizes. For simplicity, we assume that each subject is classified into one of n categories and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top \in \mathbb{T}_n$ is the corresponding cell probability vector. Let Y_{obs} denote the observed frequencies that consist of two parts: the complete observations (e.g., n_1, n_2, n_3 and n_{12} at (2.4)) and the partial observations (e.g., n_{23} at (2.4)). Under the MAR mechanism, the likelihood function is usually expressed as

$$L(\boldsymbol{\theta}|Y_{\text{obs}}) = \text{ND}_{n,n-1}(\boldsymbol{\theta}|\mathbf{a}, \mathbf{b}) \cdot L^{\text{st}}(\boldsymbol{\theta}|Y_{\text{obs}}), \tag{3.1}$$

where the first term is in the NDD form. For the second term, we consider that $L^{\mathrm{st}}(\boldsymbol{\theta}|Y_{\mathrm{obs}})$ is a constant, or $L^{\mathrm{st}}(\boldsymbol{\theta}|Y_{\mathrm{obs}})$ is not constant, where the superscript "st" represents the second term of (3.1).

3.1. Likelihood with NDD form

If $L^{\mathrm{st}}(\boldsymbol{\theta}|Y_{\mathrm{obs}})$ is a constant, the likelihood function in (3.1) is proportional to the nested Dirichlet distribution $\mathrm{ND}_{n,n-1}(\boldsymbol{\theta}|\mathbf{a},\mathbf{b})$, that is,

$$L(\boldsymbol{\theta}|Y_{\mathrm{obs}}) \propto (\textstyle\prod_{i=1}^n \theta_i^{a_i-1}) \cdot \prod_{k=1}^{n-1}(\sum_{\ell=1}^k \theta_\ell)^{b_k}. \tag{3.2}$$

Recall that (2.3) belongs to this category. From Proposition 5, we immediately obtain the MLE of $\boldsymbol{\theta}$ in closed-form by treating the variates as parameters. The asymptotic variance-covariance matrix of the MLE $\hat{\boldsymbol{\theta}}$ is then given by $I_{\mathrm{obs}}^{-1}(\hat{\boldsymbol{\theta}})$, where $I_{\mathrm{obs}}(\boldsymbol{\theta}) = -\partial^2 \log L(\boldsymbol{\theta}|Y_{\mathrm{obs}})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ is the observed information matrix. From (3.2), it is easy to show that

$$\begin{aligned}
\frac{\partial \log L(\boldsymbol{\theta}|Y_{\mathrm{obs}})}{\partial\theta_i} &= \frac{a_i-1}{\theta_i} - \frac{a_n-1}{\theta_n} + \sum_{k=i}^{n-1} \frac{b_k}{\sum_{\ell=1}^k \theta_\ell}, && i=1,\ldots,n-1,\\
-\frac{\partial^2 \log L(\boldsymbol{\theta}|Y_{\mathrm{obs}})}{\partial\theta_i^2} &= \frac{a_i-1}{\theta_i^2} + \frac{a_n-1}{\theta_n^2} + \sum_{k=i}^{n-1} \psi_k, && i=1,\ldots,n-1,\\
-\frac{\partial^2 \log L(\boldsymbol{\theta}|Y_{\mathrm{obs}})}{\partial\theta_i\partial\theta_j} &= \frac{a_n-1}{\theta_n^2} + \sum_{k=\max(i,j)}^{n-1} \psi_k, && i \neq j,
\end{aligned}$$

where $\psi_k \mathrel{\hat=} b_k/(\sum_{\ell=1}^k \theta_\ell)^2$, $k = 1,\ldots,n-1$. Hence, the observed information matrix can be expressed as

$$I_{\mathrm{obs}}(\boldsymbol{\theta}) = \mathrm{diag}\left(\frac{a_1-1}{\theta_1^2},\ldots,\frac{a_{n-1}-1}{\theta_{n-1}^2}\right) + \frac{a_n-1}{\theta_n^2}\cdot \mathbf{1}_{n-1}\mathbf{1}_{n-1}^\top + A_{n-1}, \tag{3.3}$$

where

$$A_{n-1} = \begin{pmatrix} 1 & 1 & \cdots & 1\\ 0 & 1 & \cdots & 1\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \psi_1 & 0 & \cdots & 0\\ \psi_2 & \psi_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ \psi_{n-1} & \psi_{n-1} & \cdots & \psi_{n-1} \end{pmatrix}. \tag{3.4}$$

3.2. Likelihood beyond NDD form

If $L^{\mathrm{st}}(\boldsymbol{\theta}|Y_{\mathrm{obs}})$ is not constant, it can in general be written as a product of powers of linear functions of $\boldsymbol{\theta}$. We assume $L^{\mathrm{st}}(\boldsymbol{\theta}|Y_{\mathrm{obs}}) = \prod_{j=1}^q(\sum_{i=1}^n \lambda_{ij}\theta_i)^{m_j}$ so that

$$L(\boldsymbol{\theta}|Y_{\mathrm{obs}}) \propto \{(\textstyle\prod_{i=1}^n \theta_i^{a_i-1}) \prod_{k=1}^{n-1}(\sum_{\ell=1}^k \theta_\ell)^{b_k}\} \cdot \prod_{j=1}^q(\sum_{i=1}^n \lambda_{ij}\theta_i)^{m_j}, \tag{3.5}$$

where $\Lambda_{n\times q} = (\lambda_{ij})$ is a known matrix with $\lambda_{ij} = 0$ or 1, and that there exists at least one nonzero entry in each column of Λ. For instance, in (2.4), we have $n = 3$, $q = 1$, $m_1 = n_{23}$, $\lambda_{11} = 0$ and $\lambda_{21} = \lambda_{31} = 1$. Generally speaking, the MLE of $\boldsymbol{\theta}$ based on (3.5) may not be available in closed-form. Here, we propose a new EM algorithm based on NDD rather than the Dirichlet distribution for obtaining the MLE.

For this purpose, we first augment the observed data $Y_{\rm obs}$ with latent data $Y_{\rm mis}^{\rm ND} = \{\mathbf{z}_j\}_{j=1}^q$, where $\mathbf{z}_j = (z_{1j}, \ldots, z_{nj})^\top$ is used to split $(\lambda_{1j}\theta_1 + \cdots + \lambda_{nj}\theta_n)^{m_j}$. When $\lambda_{ij} = 0$, we set $z_{ij} = 0$. The likelihood function for the new augmented-data $Y_{\rm aug}^{\rm ND} = \{Y_{\rm obs}, Y_{\rm mis}^{\rm ND}\}$ (equivalently, the joint density of $Y_{\rm aug}^{\rm ND}$) can be readily shown to be

$$L(\boldsymbol{\theta}|Y_{\rm aug}^{\rm ND}) = f(Y_{\rm aug}^{\rm ND}|\boldsymbol{\theta}) = \mathrm{ND}_{n,n-1}(\boldsymbol{\theta}|\mathbf{a} + Z\mathbf{1}_q, \mathbf{b}), \tag{3.6}$$

where $Z_{n\times q} = (\mathbf{z}_1, \ldots, \mathbf{z}_q)$. Hence, the augmented-data MLEs of $\boldsymbol{\theta}$ (cf. (2.7)) are given by

$$\begin{cases} \hat\theta_n = \dfrac{a_n + \mathbf{z}_{(n)}^\top \mathbf{1}_q - 1}{\sum_{\ell=1}^n (a_\ell + \mathbf{z}_{(\ell)}^\top \mathbf{1}_q - 1) + \sum_{\ell=1}^{n-1} b_\ell}, \\[2ex] \hat\theta_i = \dfrac{(a_i + \mathbf{z}_{(i)}^\top \mathbf{1}_q - 1)(1 - \sum_{j=i+1}^n \hat\theta_j)}{\sum_{\ell=1}^i (a_\ell + \mathbf{z}_{(\ell)}^\top \mathbf{1}_q - 1) + \sum_{\ell=1}^{i-1} b_\ell}, \qquad 2 \le i \le n-1, \\[2ex] \hat\theta_1 = 1 - \sum_{j=2}^n \hat\theta_j, \end{cases} \tag{3.7}$$

where $\mathbf{z}_{(i)}^\top$ denotes the i-th row of the matrix Z. Further, it is easy to show that the conditional predictive distribution is given by

$$f(Y_{\rm mis}^{\rm ND}|Y_{\rm obs}, \boldsymbol{\theta}) = \prod_{j=1}^q f(\mathbf{z}_j|Y_{\rm obs}, \boldsymbol{\theta}), \tag{3.8}$$

where

$$\mathbf{z}_j|(Y_{\rm obs}, \boldsymbol{\theta}) \sim \mathrm{Multinomial}\Big(m_j; \frac{(\lambda_{1j}\theta_1, \ldots, \lambda_{nj}\theta_n)^\top}{\sum_{\ell=1}^n \lambda_{\ell j}\theta_\ell}\Big), \quad 1 \le j \le q.$$

Thus, the E-step of the new EM computes the following conditional expectations

$$E(z_{ij}|Y_{\rm obs}, \boldsymbol{\theta}) = \frac{m_j \lambda_{ij}\theta_i}{\sum_{\ell=1}^n \lambda_{\ell j}\theta_\ell}, \quad 1 \le i \le n, \quad 1 \le j \le q, \tag{3.9}$$

while the M-step updates (3.7) by replacing z_{ij}'s by their conditional expectations. The asymptotic variance-covariance matrix of the MLE $\hat{\boldsymbol{\theta}}$ can be obtained by the method of Louis (1982) or the direct computation of the observed information matrix evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

3.3. Comparison with existing likelihood strategies

When the likelihood (3.1) takes the NDD form in (3.2), the traditional strategy would introduce latent variables to split $\prod_{k=1}^{n-1}(\sum_{\ell=1}^{k}\theta_\ell)^{b_k}$, so that the augmented likelihood is in the form of a Dirichlet density and hence the EM algorithm can be used to obtain the MLE of $\boldsymbol{\theta}$. In this regard, our non-iterative method proposed in Sec. 3.1 is preferable to the EM as one has a closed-form MLE solution of $\boldsymbol{\theta}$.

When the likelihood function is given by (3.5), the traditional strategy would introduce $(n-1)(n-2)/2$ latent variables (denoted by $\{\mathbf{w}_k\}_{k=2}^{n-1}$), when compared with our strategy, where $\mathbf{w}_k = (w_{1k}, \ldots, w_{kk})^\top$ is used to split $(\sum_{\ell=1}^{k}\theta_\ell)^{b_k}$. Thus, the corresponding missing data are denoted by $Y_{\rm mis}^{\rm D} = Y_{\rm mis}^{\rm ND} \cup \{\mathbf{w}_k\}_{k=2}^{n-1}$, so that the likelihood for the augmented data $Y_{\rm aug}^{\rm D} = \{Y_{\rm obs}, Y_{\rm mis}^{\rm D}\}$ (equivalently, the joint pdf of $Y_{\rm aug}^{\rm D}$) is given by

$$L(\boldsymbol{\theta}|Y_{\rm aug}^{\rm D}) = f(Y_{\rm aug}^{\rm D}|\boldsymbol{\theta}) = {\rm D}_n(\boldsymbol{\theta}|s_1, \ldots, s_n), \tag{3.10}$$

$$s_i \hat{=} a_i + \mathbf{z}_{(i)}^\top \mathbf{1}_q + \textstyle\sum_{k=i}^{n-1} w_{ik}, \quad i = 1, \ldots, n-1, \tag{3.11}$$

$$\text{with} \quad w_{11} \hat{=} b_1, \quad \text{and}$$

$$s_n \hat{=} a_n + \mathbf{z}_{(n)}^\top \mathbf{1}_q.$$

Thus, the augmented-data MLEs are

$$\hat{\theta}_i = \frac{s_i - 1}{\sum_{\ell=1}^{n}}(s_\ell - 1), \qquad i = 1, \ldots, n. \tag{3.12}$$

On the other hand, the conditional predictive distributions are given by (3.8) and

$$\mathbf{w}_k|(Y_{\rm obs}, \boldsymbol{\theta}) \sim \text{Multinomial}\Big(b_k; \ \frac{(\theta_1, \ldots, \theta_k)^\top}{\sum_{\ell=1}^{k}\theta_\ell}\Big), \quad k = 2, \ldots, n-1. \tag{3.13}$$

Therefore, the E-step of the traditional EM algorithm computes (3.9) and

$$E(w_{ik}|Y_{\rm obs}, \boldsymbol{\theta}) = \frac{b_k\theta_i}{\sum_{\ell=1}^{k}\theta_\ell}, \quad 2 \le k \le n-1, \quad 1 \le i \le k, \tag{3.14}$$

and the M-step updates (3.12) by replacing the z_{ij}s and w_{ik}s with their conditional expectations.

To compare the traditional EM algorithm with our proposed EM algorithm in Sec. 3.2, we let sequences $\{\boldsymbol{\theta}^{(t)}\}_{t=0}^{\infty}$ be the output of any EM algorithm with augmented data $Y_{\rm aug} = \{Y_{\rm obs}, Y_{\rm mis}\}$ and parameter vector $\boldsymbol{\theta} \in \mathbb{T}_n$. Any EM algorithm implicitly defines a mapping $\boldsymbol{\theta} \to M(\boldsymbol{\theta}) = (M_1(\boldsymbol{\theta}), \ldots, M_n(\boldsymbol{\theta}))^\top$ from $\mathbb{T}_n$ to $\mathbb{T}_n$ such that $\boldsymbol{\theta}^{(t+1)} = M(\boldsymbol{\theta}^{(t)})$. If $\boldsymbol{\theta}^{(t+1)}$ converges to some fixed point $\hat{\boldsymbol{\theta}} \in \mathbb{T}_n$, then $\hat{\boldsymbol{\theta}} = M(\hat{\boldsymbol{\theta}})$. After expanding $M(\boldsymbol{\theta}^{(t)})$ at the neighborhood of $\hat{\boldsymbol{\theta}}$, we have $\boldsymbol{\theta}^{(t+1)} - \hat{\boldsymbol{\theta}} \doteq dM(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}})$. Following Meng (1994), we refer to $dM(\hat{\boldsymbol{\theta}})$, the derivative of $M(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, as the matrix rate of convergence of the sequence $\{\boldsymbol{\theta}^{(t)}\}$. The largest eigenvalue of $dM(\hat{\boldsymbol{\theta}})$, denoted as $\rho\{dM(\hat{\boldsymbol{\theta}})\}$, is called the global rate of convergence of $\{\boldsymbol{\theta}^{(t)}\}$. Furthermore, $S(\hat{\boldsymbol{\theta}}) = I_n - dM(\hat{\boldsymbol{\theta}})$ is called the matrix speed of convergence of $\{\boldsymbol{\theta}^{(t)}\}$, and the smallest eigenvalue $1 - \rho\{dM(\hat{\boldsymbol{\theta}})\}$ of $S(\hat{\boldsymbol{\theta}})$ is known as the global speed of $\{\boldsymbol{\theta}^{(t)}\}$. Under mild regularity conditions, Dempster, Laird and Rubin (1977) showed that $dM(\hat{\boldsymbol{\theta}}) = I_n - I_{\rm aug}^{-1}(\hat{\boldsymbol{\theta}}) I_{\rm obs}(\hat{\boldsymbol{\theta}})$, where

$$I_{\rm aug}(\boldsymbol{\theta}) = E\left[-\frac{\partial^2 \log f(Y_{\rm aug}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \middle| Y_{\rm obs}, \boldsymbol{\theta} \right] \tag{3.15}$$

and $I_{\rm obs}(\boldsymbol{\theta}) = -\partial^2 \log f(Y_{\rm obs}|\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top$ are the expected complete-data information matrix and the observed information matrix, respectively.

To compare two EM algorithms (EM1 and EM2 say) based on the same $Y_{\rm obs}$, but different DA schemes, we need only compare their matrix speeds. Since $I_{\rm obs}(\hat{\boldsymbol{\theta}})$ is independent of DA schemes, it is suffices to compare their $I_{\rm aug}(\hat{\boldsymbol{\theta}})$. Meng and van Dyk (1997) showed that if $I_{\rm aug}^{\rm EM2}(\hat{\boldsymbol{\theta}}) \le I_{\rm aug}^{\rm EM1}(\hat{\boldsymbol{\theta}})$, then the global speed of EM2 is greater than or equal to the global speed of EM1.

Let $c\{B\}$ denote some criterion for measuring the size of the positive definite matrix B. We say EM2 *dominates* EM1 in c-criterion if $c\{I_{\rm aug}^{\rm EM2}(\hat{\boldsymbol{\theta}})\} \le c\{I_{\rm aug}^{\rm EM1}(\hat{\boldsymbol{\theta}})\}$. Furthermore, we say EM2 *uniformly* dominates EM1 in c-criterion if $c\{I_{\rm aug}^{\rm EM2}(\boldsymbol{\theta})\} \le c\{I_{\rm aug}^{\rm EM1}(\boldsymbol{\theta})\}$ for any $\boldsymbol{\theta} \in \mathbb{T}_n$. Besides the largest eigenvalue, the two commonly used criteria for measuring the size of a positive definite matrix are trace and determinant. We have the following result, with the proof given in the Appendix.

Proposition 6. *The EM algorithm given at* (3.7) *and* (3.9) *uniformly dominates the EM algorithm given at* (3.12), (3.9) *and* (3.14) *under the trace criterion, i.e.,* $tr\{I_{\rm aug}^{\rm ND}(\boldsymbol{\theta})\} \le tr\{I_{\rm aug}^{\rm D}(\boldsymbol{\theta})\}$ *for any* $\boldsymbol{\theta} \in \mathbb{T}_n$*, and the strict inequality holds provided that there is at least one* k *such that* $b_k > 0$*. In addition, the new EM dominates the traditional EM under the criterion of largest eigenvalue.*

4. Small-sample Bayesian Inference

When the sample size is small, the asymptotic methods in Sec. 3 are not appropriate and the Bayesian approach is a useful alternative. Furthermore, in situations where some parameters are unidentified (see Sec. 5.1) in frequentist settings, the Bayesian approach may be feasible if an informative prior is assigned. In addition, it is appealing to use a Bayesian approach to specify the whole posterior curve for the parameter of interest.

4.1. Likelihood with NDD form

When $L^{\mathrm{st}}(\boldsymbol{\theta}|Y_{\mathrm{obs}})$ is a constant, the likelihood function is given at (3.2). For Bayesian analysis, the NDD is the natural conjugate prior distribution. Multiplying (3.2) by the prior distribution

$$\boldsymbol{\theta} \sim \mathrm{ND}_{n,n-1}(\mathbf{a}^*,\ \mathbf{b}^*) \tag{4.1}$$

yields the nested Dirichlet posterior distribution

$$\boldsymbol{\theta}|Y_{\mathrm{obs}} \sim \mathrm{ND}_{n,n-1}(\mathbf{a}+\mathbf{a}^*-\mathbf{1}_n,\ \mathbf{b}+\mathbf{b}^*). \tag{4.2}$$

The exact first-order and second-order posterior moments of $\{\theta_i\}$ can be obtained explicitly from Proposition 3. The posterior means are similar to the MLEs. In addition, the posterior samples of $\boldsymbol{\theta}$ in (4.2) can be generated by utilizing (2.5), which only involves the simulation of independent beta variates.

4.2. Likelihood beyond NDD form

When the observed likelihood function is given by (3.5), we propose a new DA algorithm (Tanner and Wong (1987)) based on NDD, rather than the traditional Dirichlet distribution, to generate dependent posterior samples of $\boldsymbol{\theta}$. We take the prior as at (4.1). From (3.6), the complete-data posterior is an NDD, i.e.,

$$f(\boldsymbol{\theta}|Y_{\mathrm{obs}}, Y_{\mathrm{mis}}^{\mathrm{ND}}) = \mathrm{ND}_{n,n-1}(\boldsymbol{\theta}|\mathbf{a}+Z\mathbf{1}_q+\mathbf{a}^*-\mathbf{1}_n,\ \mathbf{b}+\mathbf{b}^*). \tag{4.3}$$

Based on (3.8) and (4.3), the new DA algorithm can be implemented to obtain dependent posterior samples for $\boldsymbol{\theta}$.

Furthermore, when q (cf. Eq.(3.5)) is small (e.g., $q = 1$ or 2), we can adopt the exact sampling approach (Tian, Tan and Ng (2007)) to obtain i.i.d. samples from the posterior distribution $f(\boldsymbol{\theta}|Y_{\mathrm{obs}})$. In fact, from (3.8), (4.3) and the sampling-wise IBF (Tan, Tian and Ng (2003)), we have

$$f(Y_{\mathrm{mis}}^{\mathrm{ND}}|Y_{\mathrm{obs}}) \propto \frac{f(Y_{\mathrm{mis}}^{\mathrm{ND}}|Y_{\mathrm{obs}}, \boldsymbol{\theta}_0)}{f(\boldsymbol{\theta}_0|Y_{\mathrm{obs}}, Y_{\mathrm{mis}}^{\mathrm{ND}})}, \tag{4.4}$$

where $\boldsymbol{\theta}_0$ is an arbitrary point in $\mathbb{T}_n$. Since $Y_{\mathrm{mis}}^{\mathrm{ND}}$ is a discrete random vector assuming finite values on its domain, we can first generate i.i.d. samples $\{Y_{\mathrm{mis}}^{\mathrm{ND}(\ell)}\}_{\ell=1}^{L}$ of $Y_{\mathrm{mis}}^{\mathrm{ND}}$ from the discrete distribution (4.4), and then generate $\boldsymbol{\theta}^{(\ell)} \sim f(\boldsymbol{\theta}|Y_{\mathrm{obs}}, Y_{\mathrm{mis}}^{\mathrm{ND}(\ell)})$. Thus, $\{\boldsymbol{\theta}^{(\ell)}\}_{\ell=1}^{L}$ are i.i.d. samples from the posterior $f(\boldsymbol{\theta}|Y_{\mathrm{obs}})$.

4.3. Comparison with the existing Bayesian strategy

When the likelihood (3.1) has the NDD form given by (3.2), the usual Bayesian strategy introduces latent variables so that the augmented posterior is a Dirichlet distribution, and hence the DA algorithm can be used to obtain *dependent* posterior samples of $\boldsymbol{\theta}$. The proposed non-iterative sampling approach in Sec. 4.1 can more easily obtain i.i.d. posterior samples of $\boldsymbol{\theta}$ than the iterative DA algorithm.

When the likelihood function is given by (3.5), the traditional Bayesian strategy introduces $(n-1)(n-2)/2$ latent variables (denoted by $W = \{\mathbf{w}_k\}_{k=2}^{n-1}$) so that the likelihood function is given by (3.10). If we consider the conjugate Dirichlet distribution $\boldsymbol{\theta} \sim \mathrm{D}_n(\mathbf{a}^*)$ as the prior, then the augmented posterior remains a Dirichlet distribution,

$$f(\boldsymbol{\theta}|Y_{\mathrm{obs}}, Y_{\mathrm{mis}}^{\mathrm{ND}}, W) = \mathrm{D}_n(\boldsymbol{\theta}|\mathbf{s}+\mathbf{a}^*), \quad \mathbf{s} \mathrel{\hat{=}} (s_1, \dots, s_n)^\top, \tag{4.5}$$

where $\{s_i\}_{i=1}^n$ are defined in (3.11). Therefore, the P-step of the traditional DA generates $\boldsymbol{\theta}$ from (4.5), and the I-step independently inputs $Y_{\mathrm{mis}}^{\mathrm{ND}}$ from (3.8) and inputs W from (3.13).

To compare the different DA schemes, we consider the criterion of lag-1 autocorrelation, a common measure for studying the mixing rate of a Markov chain. If the chain from a DA algorithm has reached equilibrium, Liu (1994) shows that for any non-constant scalar-valued function h,

$$\mathrm{Corr}\{h(\boldsymbol{\theta}^{(t)}), h(\boldsymbol{\theta}^{(t+1)})\} = \frac{\mathrm{Var}\,\{E[h(\boldsymbol{\theta})|Y_{\mathrm{aug}}]|Y_{\mathrm{obs}}\}}{\mathrm{Var}\,\{h(\boldsymbol{\theta})|Y_{\mathrm{obs}}\}}.$$

Therefore, the maximum autocorrelation over linear combinations $h(\boldsymbol{\theta}) = \mathbf{x}^\top\boldsymbol{\theta}$ is

$$\sup_{\mathbf{x}\neq\mathbf{0}} \mathrm{Corr}\{\mathbf{x}^\top\boldsymbol{\theta}^{(t)}, \mathbf{x}^\top\boldsymbol{\theta}^{(t+1)}\} = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^\top\mathrm{Var}\,\{E[\boldsymbol{\theta}|Y_{\mathrm{aug}}]|Y_{\mathrm{obs}}\}\mathbf{x}}{\mathbf{x}^\top\mathrm{Var}\,(\boldsymbol{\theta}|Y_{\mathrm{obs}})\mathbf{x}} = \rho\{B\},$$

where $\rho\{B\}$ denotes the spectral radius of $B = I - \{\mathrm{Var}\,(\boldsymbol{\theta}|Y_{\mathrm{obs}})\}^{-1}E[\mathrm{Var}\,(\boldsymbol{\theta}|Y_{\mathrm{aug}})\,|Y_{\mathrm{obs}}]$, the Bayesian fraction of missing information for $\boldsymbol{\theta}$ under $f(Y_{\mathrm{aug}}|\boldsymbol{\theta})$. Thus, to reduce the autocorrelation, we need to maximize $E[\mathrm{Var}\,(\boldsymbol{\theta}|Y_{\mathrm{aug}})|Y_{\mathrm{obs}}]$ over all DA schemes using the positive semi-definite ordering (van Dyk and Meng (2001)). To compare two different DA algorithms, say DA1 and DA2, based on the same observed data Y_{obs}, we only need to compare their $E[\mathrm{Var}\,(\boldsymbol{\theta}|Y_{\mathrm{aug}})|Y_{\mathrm{obs}}]$. We say the algorithm DA1 converges no slower than the DA2 if

$$E[\mathrm{Var}\,(\boldsymbol{\theta}|Y_{\mathrm{aug}}^{\mathrm{DA1}})|Y_{\mathrm{obs}}] \geq E[\mathrm{Var}\,(\boldsymbol{\theta}|Y_{\mathrm{aug}}^{\mathrm{DA2}})|Y_{\mathrm{obs}}].$$

To compare the traditional DA algorithm defined by (4.5), (3.8), and (3.13) with the proposed DA algorithm in Sec. 4.2, we must first choose the same prior distribution, which can be done by setting $\mathbf{b}^* = \mathbf{0}$ in (4.3) or (4.1). Then, the observed posterior distributions corresponding to the two DA algorithms are identical. Since $Y_{\rm aug}^{\rm ND} = \{Y_{\rm obs}, Y_{\rm mis}^{\rm ND}\} \subseteq \{Y_{\rm obs}, Y_{\rm mis}^{\rm ND}, W\} = Y_{\rm aug}^{\rm D}$, we immediately have $\mathrm{Var}\,(\boldsymbol{\theta}|Y_{\rm aug}^{\rm ND}) \geq \mathrm{Var}\,(\boldsymbol{\theta}|Y_{\rm aug}^{\rm D})$, so that $E[\mathrm{Var}\,(\boldsymbol{\theta}|Y_{\rm aug}^{\rm ND})|Y_{\rm obs}] \geq E[\mathrm{Var}\,(\boldsymbol{\theta}|Y_{\rm aug}^{\rm D})|Y_{\rm obs}]$.

Proposition 7. *The DA algorithm defined at* (4.3) *and* (3.8) *converges no slower than the traditional DA algorithm defined at* (4.5), (3.8) *and* (3.13).

Since $Y_{\rm obs}$ does not vary in the sampling process, we can represent the two DA schemes simply as

$$\begin{aligned} &\text{Scheme DA}^{\rm ND}: \boldsymbol{\theta}|Y_{\rm mis}^{\rm ND}, \qquad Y_{\rm mis}^{\rm ND}|\boldsymbol{\theta}. \\ &\text{Scheme DA}^{\rm D}: \ \ \boldsymbol{\theta}|(Y_{\rm mis}^{\rm ND}, W),\ (Y_{\rm mis}^{\rm ND}, W)|\boldsymbol{\theta}. \end{aligned}$$

In Scheme DA$^{\rm ND}$, the two components being iterated are $\boldsymbol{\theta}$ and $Y_{\rm mis}^{\rm ND}$ with W being integrated out, while in Scheme DA$^{\rm D}$, an extra random vector W is introduced. Using Theorem 5.1 of Liu, Wong and Kong (1994), we immediately obtain the following result.

Proposition 8. *Let $F^{\rm ND}$ and $F^{\rm D}$ denote the forward operators of the two DA schemes, and $||F^{\rm ND}||$ and $||F^{\rm D}||$ the corresponding norms. Then* (i) $||F^{\rm ND}|| \leq ||F^{\rm D}||$; (ii) *the spectral radius of Scheme DA$^{\rm ND}$ is less than or equal to that of Scheme DA$^{\rm D}$.*

The former notions of forward operator, norm and spectral radius can be found in Liu, Wong and Kong (1994). Proposition 8 shows that the maximal correlation between $\boldsymbol{\theta}$ and $Y_{\rm mis}^{\rm ND}$ is always smaller than that between $\boldsymbol{\theta}$ and $(Y_{\rm mis}^{\rm ND}, W)$. Furthermore, when q is small, the exact sampling approach proposed in Sec. 4.2 can be used to generate i.i.d. samples from the posterior $f(\boldsymbol{\theta}|Y_{\rm obs})$, avoiding the problems of convergence associated with the iterative DA algorithms.

5. Applications

5.1. Simulated data

For Example 1 in Sec. 2.2, let ϕ and $1-\phi$ denote the probabilities of response and non-response, respectively. Hence, we have $\phi = \Pr(R = 1) = \sum_{i=r+1}^{2r} \theta_i$ and $1 - \phi = \Pr(R = 0) = \sum_{i=1}^{r} \theta_i$. Let $N = 1{,}000$, $r = 5$, and $\phi = 0.65$. By independently drawing 100 m's from Binomial(N, ϕ) and averaging them, we obtained $m = 652$ (m denotes the number of individuals who respond to the survey). We further assume that $\theta_6 = 0.2$, $\theta_7 = 0.12$, $\theta_8 = 0.08$, $\theta_9 = 0.15$, and $\theta_{10} = 0.10$. Independently drawing 100 $(n_6, \ldots, n_{10})$'s from Multinomial$(m; (\theta_6, \ldots, \theta_{10})^\top/0.65)$ and averaging the samples for each component, we obtained $n_6 = 199$, $n_7 = 120$, $n_8 = 81$, $n_9 = 151$, and $n_{10} = 101$.

Table 1. MLEs, SE and Bayesian estimates of parameters for the simulated data.

	True	Frequentist method		Bayesian method		
Parameter	value	MLE	std	mean	std	95% CI
θ_1	-	-	-	0.1428	0.0660	$[0.0280, 0.2741]$
θ_2	-	-	-	0.0727	0.0568	$[0.0025, 0.2114]$
θ_3	-	-	-	0.0538	0.0451	$[0.0019, 0.1680]$
θ_4	-	-	-	0.0447	0.0395	$[0.0013, 0.1498]$
θ_5	-	-	-	0.0390	0.0348	$[0.0011, 0.1311]$
$1-\phi$	0.35	0.348	0.0151	0.3534	0.0150	$[0.3246, 0.3832]$
θ_6	0.20	0.199	0.0126	0.1971	0.0124	$[0.1732, 0.2223]$
θ_7	0.12	0.120	0.0102	0.1192	0.0100	$[0.1002, 0.1397]$
θ_8	0.08	0.081	0.0086	0.0807	0.0084	$[0.0647, 0.0983]$
θ_9	0.15	0.151	0.0113	0.1493	0.0113	$[0.1277, 0.1719]$
θ_{10}	0.10	0.101	0.0095	0.1001	0.0093	$[0.0825, 0.1191]$
ϕ	0.65	0.652	0.0151	0.6466	0.0150	$[0.6167, 0.6753]$

Note: $1-\phi = \Pr(R=0) = \sum_{i=1}^{5}\theta_i$ and $\phi = \Pr(R=1) = \sum_{i=6}^{10}\theta_i$.

Based on the simulated counts $Y_{\rm obs} = \{(n_{r+1}, \ldots, n_{2r}); N-m\}$ and the likelihood (2.3), it is easy to see that the MLE of $\boldsymbol{\theta}$ is exactly the mode of the nested Dirichlet distribution $\mathrm{ND}_{n,n-1}(\mathbf{a}, \mathbf{b})$ with $n=2r$, $\mathbf{a} = (0, \ldots, 0, n_{r+1}, \ldots, n_{2r})^\top + \mathbf{1}_n$ and $\mathbf{b} = (\mathbf{0}_{r-1}^\top, N-m, \mathbf{0}_{r-1}^\top)^\top$. In frequentist settings, we note that $\theta_1, \ldots, \theta_r$ are non-estimable but $1-\phi$ is estimable. From Proposition 5, we obtain $\hat{\theta}_6 = 0.199$, $\hat{\theta}_7 = 0.12$, $\hat{\theta}_8 = 0.081$, $\hat{\theta}_9 = 0.151$, $\hat{\theta}_{10} = 0.101$, and $\hat{\phi} = 0.652$. The corresponding standard errors are listed in the fourth column of Table 1.

On the other hand, in Bayesian settings, $\theta_1, \ldots, \theta_r$ are estimable if an informative prior distribution can be assigned to $\boldsymbol{\theta}$. Here we use the informative prior with $a_1^* = \cdots = a_n^* = 2$ and $b_1^* = \cdots = b_{n-1}^* = 1$ in (4.1). We generated $10,000$ i.i.d. posterior samples of $\boldsymbol{\theta}$ from (4.2). The Bayes means, standard errors and 95% Bayes confidence intervals for $\boldsymbol{\theta}$ and ϕ are given in Table 1.

5.2. Dental caries data

For Example 2 in Sec. 2.2, we are unable to find closed-form MLEs. In this case, we consider the EM algorithm developed in Sec. 3.2 and the exact IBF sampling approach proposed in Sec. 4.2 to handle the likelihood of $\boldsymbol{\theta}$ given at (2.4).

As (2.4) is a special case of (3.5), with $n=3$, $q=1$, $m_1 = n_{23}$, $\lambda_{11} = 0$, and $\lambda_{21} = \lambda_{31} = 1$, we only introduce one latent variable z to split $(\theta_2+\theta_3)^{n_{23}}$, so that (3.6) and (3.8) become $L(\boldsymbol{\theta}|Y_{\rm obs}, z) = \mathrm{ND}_{3,2}(\boldsymbol{\theta}|(n_1+1, n_2+1+z, n_3+1+n_{23}-z)^\top,\ (0, n_{12})^\top)$ and

$$f(z|Y_{\rm obs}, \boldsymbol{\theta}) = \mathrm{Binomial}\Big(z\Big|n_{23},\ \theta_2/(\theta_2+\theta_3)\Big), \quad z = 0, 1, \ldots, n_{23}, \tag{5.1}$$

Table 2. The values of $q_z(\boldsymbol{\theta}_0)$ with $\boldsymbol{\theta}_0 = (1/3, 1/3, 1/3)^\top$ and $p_z = f(z|Y_{\rm obs})$.

z	$q_z(\boldsymbol{\theta}_0)$	p_z	z	$q_z(\boldsymbol{\theta}_0)$	p_z	z	$q_z(\boldsymbol{\theta}_0)$	p_z
0	2.713e+032	3.815e-006	7	8.633e+036	1.214e-001	14	8.300e+035	1.167e-002
1	4.883e+033	6.866e-005	8	1.187e+037	1.669e-001	15	2.213e+035	3.113e-003
2	4.150e+034	5.836e-004	9	1.319e+037	1.855e-001	16	4.150e+034	5.836e-004
3	2.213e+035	3.113e-003	10	1.187e+037	1.669e-001	17	4.883e+033	6.866e-005
4	8.300e+035	1.167e-002	11	8.633e+036	1.214e-001	18	2.713e+032	3.815e-006
5	2.324e+036	3.268e-002	12	5.036e+036	7.082e-002	-		
6	5.036e+036	7.082e-002	13	2.324e+036	3.268e-002	-		

respectively. For the dental caries data, using $\boldsymbol{\theta}^{(0)} = (1/3, 1/3, 1/3)^\top$ as the initial value, the new EM algorithm based on (3.7) and (3.9) converged in seven iterations, with 0.016 seconds CPU time. The resulting MLEs were given by $\hat{\theta}_1 = 0.2393$, $\hat{\theta}_2 = 0.4880$, and $\hat{\theta}_3 = 0.2727$. The corresponding standard errors were 0.0547, 0.0674, and 0.0514, obtained by the direct computation of the observed information matrix evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. However, using the same initial value, the traditional EM based on Dirichlet augmentation (see, (3.12) and (3.14)) converged in 22 iterations, with 0.04 seconds CPU time, which is about 3 (or 2.5) times slower than the new EM in terms of the iteration number (or the computing time). This comparison would be even more impressive if the traditional EM algorithm introduced more than one extra latent variable.

For Bayesian analysis, we adopt the uniform prior, i.e., setting $\mathbf{a}^* = \mathbf{1}_3$ and $\mathbf{b}^* = \mathbf{0}$ in (4.1). Hence, the complete-data posterior (4.3) and the sampling-wise IBF (4.4) become

$$f(\boldsymbol{\theta}|Y_{\rm obs}, z) = \mathrm{ND}_{3,2}\Big(\boldsymbol{\theta}\Big|(n_1{+}1, n_2{+}1{+}z, n_3{+}1{+}n_{23}{-}z)^\top,\ (0, n_{12})^\top\Big) \tag{5.2}$$

and

$$f(z|Y_{\rm obs}) \propto \frac{f(z|Y_{\rm obs}, \boldsymbol{\theta}_0)}{f(\boldsymbol{\theta}_0|Y_{\rm obs}, z)} \mathrel{\hat{=}} q_z(\boldsymbol{\theta}_0), \tag{5.3}$$

respectively. Let $\boldsymbol{\theta}_0 = (1/3, 1/3, 1/3)^\top$. From (5.3), we can calculate $\{q_z(\boldsymbol{\theta}_0)\}_{z=0}^{n_{23}}$. By defining $p_z = f(z|Y_{\rm obs})$ for $z = 0, \ldots, n_{23}$, we have $p_z = q_z(\boldsymbol{\theta}_0)/\sum_{k=0}^{n_{23}} q_k(\boldsymbol{\theta}_0)$, which is independent of $\boldsymbol{\theta}_0$. We list the results in Table 2.

The exact IBF sampling can be conducted as follows: draw $L = 20,000$ independent samples $\{z^{(\ell)}\}_1^L$ of z from the discrete distribution (5.3) with probabilities p_z; generate $\boldsymbol{\theta}^{(\ell)} \sim f(\boldsymbol{\theta}|Y_{\rm obs}, z^{(\ell)})$ at (5.2) for $\ell = 1, \ldots, L$, and $\{\boldsymbol{\theta}^{(\ell)}\}_1^L$ are i.i.d. samples from the observed posterior distribution $f(\boldsymbol{\theta}|Y_{\rm obs})$. For the dental caries data, the Bayes means of $\theta_1, \theta_2, \theta_3$ were given by 0.2457, 0.4784 and 0.2759, with the corresponding Bayes standard errors being 0.0532, 0.0654 and 0.0501. The 95% Bayes interval estimates were $[0.1487, 0.3571]$, $[0.3498, 0.6061]$ and $[0.1832, 0.3785]$, respectively. The computing time was 36.82 seconds. Figure 1 shows the posterior curves of θ_1, θ_2 and θ_3.

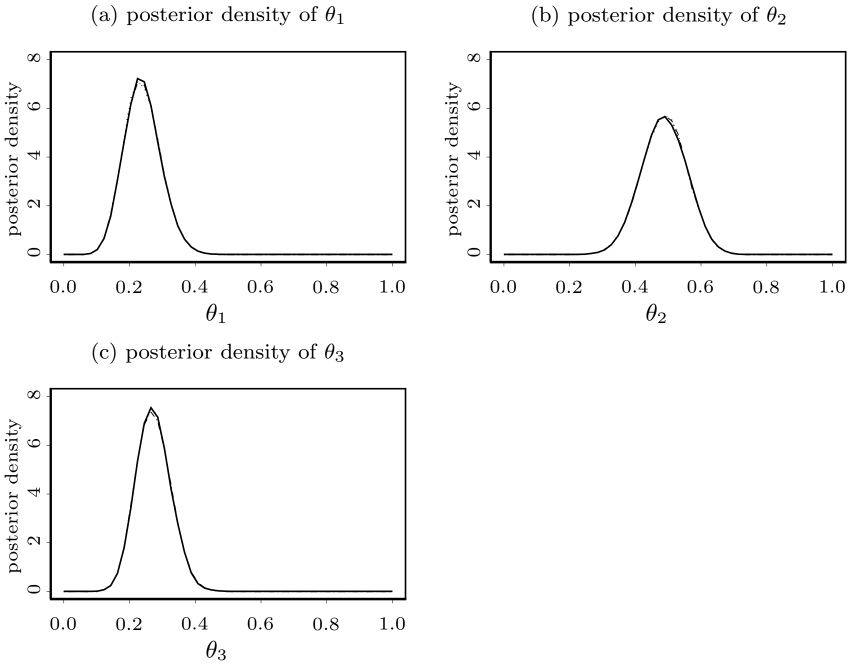

Figure 1. Comparisons of three posterior densities in each plot, obtained by using the exact IBF sampling (solid curve, $L = 20,000$ i.i.d. samples), the DA$^{\rm ND}$ algorithm (dotted curve, a total of $40,000$ and retaining the last $20,000$ samples) and the DA$^{\rm D}$ algorithm (dashed curve, a total of $45,000$ and retaining the last $20,000$ samples). (a) θ_1; (b) θ_2; (c) θ_3.

To compare the new DA$^{\rm ND}$ algorithm defined by (5.1) and (5.2) with the traditional DA$^{\rm D}$ algorithm defined by (5.1),

$$f(w|Y_{\rm obs},\boldsymbol{\theta}) = \text{Binomial}\Big(w\Big|n_{12},\ \frac{\theta_1}{\theta_1+\theta_2}\Big), \quad w = 0,\ldots,n_{12},$$

$$f(\boldsymbol{\theta}|Y_{\rm obs},z,w) = \text{D}_3\Big(\boldsymbol{\theta}\Big|n_1+1+w, n_2+1+n_{12}-w+z, n_3+1+n_{23}-z\Big),$$

we treat the whole posterior curve of $\{\theta_i\}$ obtained from the exact IBF sampling as a benchmark to assess the convergence of the two Markov chains. We ran a single chain of the DA$^{\rm ND}$ (DA$^{\rm D}$) algorithm to produce $40,000$ $(45,000)$ samples, and retained the last $20,000$ samples. The corresponding computing times were 83.562 and 168.922 seconds. Figure 1 shows that the three curves are almost identical, indicating final convergences for the two DA algorithms.

6. Discussion

We extended the Dirichlet distribution to a new family of *nested Dirichlet distribution* (NDD), which allows more flexible parameters and can be readily adopted as the prior distribution in Bayesian ICD analysis. New EM/DA algorithms based on the NDD were proposed and compared theoretically with existing EM/DA algorithms based on the Dirichlet distribution.

The asymmetry of $x_1, \ldots, x_n$ in (2.1) may affect computational efficiency when constructing EM/DA algorithms. For example, for the likelihood (2.4), in Sec. 5.2 we introduced a latent variable z to split $(\theta_2 + \theta_3)^{n_{23}}$. Let the induced EM and DA be denoted by EM^{ND1} and DA^{ND1}, respectively. Alternatively, since (2.4) can be rewritten as $\{\theta_2^{n_2}\theta_3^{n_3}\theta_1^{n_1} \cdot \theta_2^0(\theta_2+\theta_3)^{n_{23}}\} \cdot (\theta_1+\theta_2)^{n_{12}}$, we can introduce a latent variable z' to split $(\theta_1+\theta_2)^{n_{12}}$, resulting in another DA scheme. The corresponding algorithms are denoted by EM^{ND2} and DA^{ND2}. Naturally, we are asked which EM/DA is better? For this simple example, we prefer the first DA scheme if $n_{23} < n_{12}$. For the more general case, it is worthwhile to theoretically compare different DA schemes. In addition, comparing the proposed methods with multiple imputation and importance sampling methods is of research interest.

Throughout this article, we have assumed that the data are *missing at random* (MAR). If this does not hold, our proposed methods are inapplicable. Furthermore, under the assumption of a non-ignorable missing mechanism, the likelihood-based approach is in general not feasible since it leads to large numbers of cells with inestimable probabilities. For a non-ignorable missing mechanism, Tian, Tan and Ng (2007, p.196-197) demonstrated that the grouped Dirichlet distribution (Tang, Ng, Tian and Tan (2007) and Ng et al. (2008)) can be used to analyze ICD only in the Bayesian framework. Finally, the numerical results of Tables 8 and 9 in Tian et al. (2003) show that the cell-probability estimates are quite sensitive to model misspecification. In this connection, it is worthwhile to design a data-driven statistical approach for testing MAR against NMAR.

In practice, high-dimensional (or sparse) categorical data are often analyzed with constrained log-linear models rather than the saturated multinomial model. In Bayesian log-linear model analysis of categorical data, it is convenient to adopt a prior distribution that has the same functional form as the Dirichlet, but which requires the parameters to satisfy the constraints imposed by a log-linear model (Schafer (1997, Chap. 8)). How to incorporate existing models (e.g., logistic and log-linear) with NDD is a challenging topic. This paper may provide a flexible tool for this purpose because the NDD family includes the Dirichlet distribution as a special case. We are now investigating the applications of NDD in modeling. We note that the marginal and conditional distributions of the NDD are rather complicated and this may affect its range of applications. Finally, as in the generalization of the Dirichlet distribution to the Liouville distribution, the extension from NDD to generalized NDD is worthwhile. We provide several S-plus functions as an on-line supplement.

Acknowledgement

We are grateful to the two Co-Editors, an associate editor and two referees for their constructive comments and valuable suggestions. The research of ML Tang was partially supported by a grant FRG/04-05/II-20 from the Hong Kong Baptist University. The researches of GL Tian and M Tan were supported in part by U.S. National Cancer Institute grants CA106767 and CA119758.

Appendix

Proof of Proposition 1. If $\mathbf{x} \sim \mathrm{ND}_{n,n-1}(\mathbf{a}, \mathbf{b})$ on $\mathbb{T}_n$, then the density of $\mathbf{x}_{-n}$ is given by (2.1). We consider two transformations: $y_j = \sum_{i=1}^{j} x_i / \sum_{i=1}^{j+1} x_i$, $j = 1, \ldots, n-2$, $y_{n-1} = \sum_{i=1}^{n-1} x_i$, and

$$u_j = \sum_{i=1}^{j} x_i, \quad j = 1, \ldots, n-1. \tag{A.1}$$

It is easy to obtain

$$u_j = y_j y_{j+1} \cdots y_{n-1}, \quad j = 1, \ldots, n-1. \tag{A.2}$$

From (A.1) and (A.2), the Jacobian $J(\mathbf{u}_{-n} \to \mathbf{x}_{-n}) \,\hat{=}\, |(\partial \mathbf{u}_{-n})/(\partial \mathbf{x}_{-n})| = 1$ and $J(\mathbf{u}_{-n} \to \mathbf{y}_{-n}) = \prod_{j=1}^{n-1} y_j^{j-1}$. Hence, $J(\mathbf{x}_{-n} \to \mathbf{y}_{-n}) = J(\mathbf{x}_{-n} \to \mathbf{u}_{-n}) \cdot J(\mathbf{u}_{-n} \to \mathbf{y}_{-n}) = \prod_{j=1}^{n-1} y_j^{j-1}$, and the joint density of $\mathbf{y}_{-n}$ is given by

$$f(\mathbf{y}_{-n}) = c^{-1} \cdot \prod_{j=1}^{n-1} y_j^{d_j - 1} (1 - y_j)^{a_{j+1} - 1}, \tag{A.3}$$

where $c = \prod_{j=1}^{n-1} B(d_j,\, a_{j+1})$ and $\{d_j\}$ are defined in (2.2). Note that (A.3) has been factored into independent beta distributions. Combining (A.1) and (A.2), we obtain (2.5). Conversely, if (2.5) holds, then the joint density of $\mathbf{y}_{-n}$ is given by (A.3). It is easy to show that the density of $\mathbf{x}_{-n}$ is given by (2.1), i.e., $\mathbf{x} \sim \mathrm{ND}_{n,n-1}(\mathbf{a}, \mathbf{b})$ on $\mathbb{T}_n$.

Proof of Propositions 3 and 4. To derive the 1st- and 2nd-order moments of $\mathbf{x} \sim \mathrm{ND}_{n,n-1}(\mathbf{a}, \mathbf{b})$, we first consider their mixed (or raw) moments. From (2.5), we observe that the independence among $\{y_j\}_{j=1}^{n-1}$ implies

$$E\left(\prod_{i=1}^{n} x_i^{r_i}\right) = \prod_{i=1}^{n} \left[E(1 - y_{i-1})^{r_i} \cdot \prod_{j=i}^{n-1} E(y_j^{r_j}) \right],$$

where $r_1, \ldots, r_n \geq 0$. Utilizing the moments of the beta distribution, the mixed moments of $\mathbf{x}$ are given by

$$E\left(\prod_{i=1}^{n} x_i^{r_i}\right) = \prod_{i=1}^{n}\left[\frac{B(d_{i-1},\ a_i + r_i)}{B(d_{i-1},\ a_i)} \cdot \prod_{j=i}^{n-1} \frac{B(d_j + r_i,\ a_{j+1})}{B(d_j,\ a_{j+1})}\right],$$

where the d_j are defined by (2.2). With $r_i = 1$ and $r_j = 0\,(j \neq i)$, we immediately obtain $E(x_i)$. Similarly, we can obtain $E(x_i^2)$ and $E(x_i x_j)$. To complete the proof of Proposition 3. Using (2.6), for any $r \geq 0$, we have $E(\sum_{j=1}^{i} x_j)^r = \prod_{j=i}^{n-1} E(y_j^r)$, and Proposition 4 follows from the moments of independent beta distributions.

Proof of Proposition 5. The mode of an NDD density is readily obtained. If L is the log-kernel of the density (2.1), we have

$$L = \sum_{i=1}^{n-1}(a_i - 1)\log(x_i) + (a_n - 1)\log(1 - \textstyle\sum_{i=1}^{n-1} x_i) + \displaystyle\sum_{j=1}^{n-1} b_j \log(x_1 + \cdots + x_j).$$

The derivative of L with respect to x_i set to zero yields

$$\frac{a_i - 1}{x_i} - \frac{a_n - 1}{x_n} + \sum_{j=i}^{n-1} \frac{b_j}{x_1 + \cdots + x_j} = 0, \quad i = 1, \ldots, n-1.$$

It is easy to verify that (2.7) is true when $n = 3$. By induction, we obtain (2.7).

Proof of Proposition 6. We first show that $\mathrm{tr}\{I_{\mathrm{aug}}^{\mathrm{ND}}(\boldsymbol{\theta})\} \leq \mathrm{tr}\{I_{\mathrm{aug}}^{\mathrm{D}}(\boldsymbol{\theta})\}$ for any $\boldsymbol{\theta} \in \mathbb{T}_n$. From (3.10), it is easy to obtain

$$\begin{aligned}
-\frac{\partial^2 \log f(Y_{\mathrm{aug}}^{\mathrm{D}}|\boldsymbol{\theta})}{\partial\theta_i^2} &= \frac{s_i - 1}{\theta_i^2} + \frac{s_n - 1}{\theta_n^2}, \quad && i = 1, \ldots, n-1.\\
-\frac{\partial^2 \log f(Y_{\mathrm{aug}}^{\mathrm{D}}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} &= \frac{s_n - 1}{\theta_n^2}, && i \neq j.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
I_{\mathrm{aug}}^{\mathrm{D}}(\boldsymbol{\theta}) &= E\left[-\frac{\partial^2 \log f(Y_{\mathrm{aug}}^{\mathrm{D}}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\top}}\bigg|Y_{\mathrm{obs}}, \boldsymbol{\theta}\right]\\
&= \mathrm{diag}\left(\frac{s_1^* - 1}{\theta_1^2}, \ldots, \frac{s_{n-1}^* - 1}{\theta_{n-1}^2}\right) + \frac{s_n^* - 1}{\theta_n^2} \cdot \mathbf{1}_{n-1}\mathbf{1}_{n-1}^{\top}
\end{aligned}$$

where, for $i = 1, \ldots, n-1$,

$$\begin{aligned}
s_i^* &= E(s_i|Y_{\mathrm{obs}}, \boldsymbol{\theta}) = a_i + E(\mathbf{z}_{(i)}^{\top}\mathbf{1}_q|Y_{\mathrm{obs}}, \boldsymbol{\theta}) + \textstyle\sum_{k=i}^{n-1}[b_k\theta_i/\sum_{\ell=1}^{k}\theta_\ell],\\
s_n^* &= E(s_n|Y_{\mathrm{obs}}, \boldsymbol{\theta}) = a_n + E(\mathbf{z}_{(n)}^{\top}\mathbf{1}_q|Y_{\mathrm{obs}}, \boldsymbol{\theta}).
\end{aligned}$$

On the other hand, from (3.6),

$$L(\boldsymbol{\theta}|Y_{\text{aug}}^{\text{ND}}) = f(Y_{\text{aug}}^{\text{ND}}|\boldsymbol{\theta}) = \text{ND}_{n,n-1}(\boldsymbol{\theta}|\tilde{\mathbf{a}}, \mathbf{b}),$$

where $\tilde{\mathbf{a}} = (\tilde{a}_1, \ldots, \tilde{a}_n)^\top$ with $\tilde{a}_i = a_i + \mathbf{z}_{(i)}^\top \mathbf{1}_q$. Similar to (3.3), we obtain

$$\begin{aligned} I_{\text{aug}}^{\text{ND}}(\boldsymbol{\theta}) &= E\left[-\frac{\partial^2 \log f(Y_{\text{aug}}^{\text{ND}}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \middle| Y_{\text{obs}}, \boldsymbol{\theta} \right] \\ &= \text{diag}\left(\frac{\tilde{a}_1^* - 1}{\theta_1^2}, \ldots, \frac{\tilde{a}_{n-1}^* - 1}{\theta_{n-1}^2} \right) + \frac{\tilde{a}_n^* - 1}{\theta_n^2} \cdot \mathbf{1}_{n-1}\mathbf{1}_{n-1}^\top + A_{n-1}, \end{aligned}$$

where A_{n-1} is given by (3.4), and

$$\tilde{a}_i^* = E(\tilde{a}_i|Y_{\text{obs}}, \boldsymbol{\theta}) = a_i + E(\mathbf{z}_{(i)}^\top \mathbf{1}_q|Y_{\text{obs}}, \boldsymbol{\theta}), \quad i = 1, \ldots, n.$$

Let $h_k = b_k / \sum_{\ell=1}^k \theta_\ell$ for $k = 1, \ldots, n-1$. Then

$$I_{\text{aug}}^{\text{D}}(\boldsymbol{\theta}) - I_{\text{aug}}^{\text{ND}}(\boldsymbol{\theta}) = \text{diag}\Big(\theta_1^{-1} \textstyle\sum_{k=1}^{n-1} h_k, \theta_2^{-1} \sum_{k=2}^{n-1} h_k, \ldots, \theta_{n-1}^{-1} h_{n-1} \Big) - A_{n-1}.$$

For the trace criterion, we have

$$\text{tr}\{I_{\text{aug}}^{\text{D}}(\boldsymbol{\theta})\} - \text{tr}\{I_{\text{aug}}^{\text{ND}}(\boldsymbol{\theta})\} = \sum_{i=1}^{n-1} \sum_{k=i}^{n-1} h_k \left[\frac{1}{\theta_i} - \frac{1}{\sum_{\ell=1}^k \theta_\ell} \right] \geq 0, \quad \forall\, \boldsymbol{\theta} \in \mathbb{T}_n,$$

and strict inequality holds provided there is at least one k such that $b_k > 0$. The first part of Proposition 6 is thus proved.

To prove the second part, we noted that Geng, Wan and Tao (2000) developed a *partial imputation EM* (PIEM) algorithm that imputes partial missing data, and they proved that the convergence speed of the PIEM is faster than the traditional EM algorithm. Since the proposed EM algorithm defined by (3.7) and (3.9) can be viewed as a special PIEM, Theorem 4 of Geng et al. (2000) gives us the second conclusion of Proposition 6.

References

Agresti, A. (2002). *Categorical Data Analysis.* 2nd edition. Wiley, New York.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Chapman & Hall, London.

Albert, J. H. and Gupta, A. K. (1985). Bayesian methods for binomial data with applications to a nonresponse problem. *J. Amer. Statist. Assoc.* **80**, 167-174.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39**, 1-38.

Dickey, J. M. (1983). Multiple hypergeometric functions: Probabilistic interpretations and statistical uses. *J. Amer. Statist. Assoc.* **78**, 628-637.

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

Geng, Z., Wan, K. and Tao, F. (2000). Mixed graphical models with missing data and the partial imputation EM algorithm. *Scand. J. Statist.* **27**, 433-444.

Gupta, R. D. and Richards, D. St. P. (1987). Multivariate Liouville distributions. *J. Multivariate Anal.* **23**, 233-256.

Gupta, R. D. and Richards, D. St. P. (1991). Multivariate Liouville distributions, II. *Probab. Math. Statist.* **12**, 291-309.

Gupta, R. D. and Richards, D. St. P. (1992). Multivariate Liouville distributions, III. *J. Multivariate Anal.* **43**, 29-57.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data.* 2nd edition. Wiley, New York.

Liu, J. S. (1994). Fraction of missing information and convergence rate of data augmentation. In *Computationally Intensive Statistical Methods: Proc.* 26*th Symp. Interface* (Edited by J. Sall and A. Lehmann), 490-497. Interface Foundation of North American, Fairfax Station, Virginia.

Liu, J. S., Wong, W. H. and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81**, 27-40.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44**, 226-233.

Meng, X. L. (1994). On the rate of convergence of the ECM algorithm. *Ann. Statist.* **22**, 326-339.

Meng, X. L. and van Dyk, D. (1997). The EM algorithm — an old folk-song sung to a fast new tune (with discussion). *J. Roy. Statist. Soc. Ser. B* **59**, 511-567.

Ng, K. W., Tang, M. L., Tan, M. and Tian, G. L. (2008). Grouped Dirichlet distribution: A new tool for incomplete categorical data analysis. *J. Multivariate Anal.* **99**, 490-509.

Paulino, C. D. M. and Pereira, C. A. de B. (1995). Bayesian methods for categorical data under informative general censoring. *Biometrika* **82**, 439-446.

Rayens, W. S. and Srinivasan, C. (1994). Dependence properties of generalized Liouville distributions on the simplex. *J. Amer. Statist. Assoc.* **89**, 1465-1470.

Rubin, D. B. (1974). Characterizing the estimation of parameters in incomplete-data problems. *J. Amer. Statist. Assoc.* **69**, 467-474.

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

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

Tan, M., Tian, G. L. and Ng, K. W. (2003). A noniterative sampling method for computing posteriors in the structure of EM-type algorithms. *Statist. Sinica* **13**, 625-639.

Tang, M. L., Ng, K. W., Tian, G. L. and Tan, M. (2007). On improved EM algorithm and confidence interval construction for incomplete $r \times c$ tables. *Computational Statistics and Data Analysis* **51**, 2919-2933.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Comput. Statist. Data Anal.* **82**, 528-540.

Tian, G. L., Ng, K. W. and Geng, Z. (2003). Bayesian computation for contingency tables with incomplete cell-counts. *Statist. Sinica* **13**, 189-206.

Tian, G. L., Tan, M. and Ng, K. W. (2007). An exact noniterative sampling procedure for discrete missing data problems. *Statist. Neerlandica*, **61**, 232-242.

van Dyk, D. and Meng, X. L. (2001). The art of data augmentation (with discussion). *J. Comput. Graph. Statist.* **10**, 1-50.

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, P. R. China.

E-mail: kaing@hku.hk

Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, P. R. China.

E-mail: mltang@math.hkbu.edu.hk

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, P. R. China.

E-mail: gltian@hku.hk

Division of Biostatistics, University of Maryland Greenebaum Cancer Center, MSTF Suite 261, 10 South Pine Street, Baltimore, Maryland 21201, U.S.A.

E-mail: mtan@umm.edu

(Received February 2006; accepted May 2007)