Statistica Neerlandica (2009) Vol. 63, nr. 1, pp. 24–42
doi:10.1111/j.1467-9574.2008.00404.x

The beta-binomial convolution model for 2×2 tables with missing cell counts

Rob Eisinga*

Department of Social Science Research Methods, Radboud University Nijmegen, PO Box 9104, 6500 HE Nijmegen, The Netherlands

This paper considers the beta-binomial convolution model for the analysis of 2×2 tables with missing cell counts. We discuss maximum-likelihood (ML) parameter estimation using the expectation–maximization algorithm and study information loss relative to complete data estimators. We also examine bias of the ML estimators of the beta-binomial convolution. The results are illustrated by two example applications.

Keywords and Phrases: ecological inference, beta-binomial distribution, convolution likelihood, expectation–maximization algorithm, missing information principle, Fisher information loss, bias correction, 2×2 table.

1 Introduction

This paper is concerned with the maximum-likelihood (ML) estimation of cell probabilities for a series of 2×2 tables with unobserved entries using the marginal totals only. This ecological inference problem has attracted ample attention from statisticians and methodologists (e.g. Plackett, 1977; Brown and Payne, 1986; Hamdan and Nasro, 1986; Haber, 1989; Kocherlakota and Kocherlakota, 1992; McCullagh and Nelder, 1992; King, 1997; King, Rosen and Tanner, 1999, 2004; Rosen *et al.*, 2001; Wakefield, 2004; Haneuse and Wakefield, 2008; Imai, Lu and Strauss, 2008). A typical 2×2 ecological inference example concerns a study of the electorate for each of two political parties in two successive elections. Because of the secret ballot nature of voting, the party choices of individual voters and potential changes therein are not known. One observes only the vote totals for the two parties at each election for a number of constituencies. Ecological inference studies take as their objective the estimation of the individual vote choice transitions from the aggregate election results.

In a previous study, we examined ML parameter estimation and Fisher information loss for 2×2 ecological tables assuming that the unobserved cell counts are distributed according to two independent binomial distributions and that the tables

*r.eisinga@maw.ru.nl

Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

are homogeneous with respect to the conditional probabilities (Eisinga, 2008). The current paper relaxes this assumption and discusses the beta-binomial generalization of the binomial. For the beta-binomial distribution, the probability of success parameter π varies for successive observations according to a beta-distribution. The beta-binomial mixture model has several applications in ecological inference analysis, but most studies adopt a Bayesian approach (e.g. King *et al.*, 1999; Wakefield, 2004). We discuss ML parameter estimation with the expectation–maximization (EM) algorithm (Dempster, Laird and Rubin, 1977) and study Fisher information loss and bias of the ML estimators.

The paper is organized as follows. The complete and observed data likelihoods involved in the analysis of 2×2 ecological tables are presented in section 2. Section 3 considers ML parameter estimation using EM and section 4 subsequently examines Fisher information loss because of data aggregation in ecological inference. Parameter bias correction in the beta-binomial case is discussed in section 5. Two empirical studies follow in section 6 and some concluding remarks are offered in section 7.

2 Data likelihoods

As a benchmark, we first consider the complete data situation. For the sth 2×2 table, $s = 1, \ldots, S$, we use the notation for the counts presented in Table 1.

Table 1. Data for table s, $s = 1, \ldots, S$.

	$Y=0$	$Y=1$	Total
$x=0$		y_{0s}	n_{0s}
$x=1$		y_{1s}	n_{1s}
Total	$n_s - y_s$	y_s	n_s

Let y_{0s} be the number of successes in n_{0s} independent observations with binomial probability π_{0s}, and let π_{1s} denote the binomial probability of success for n_{1s} independent observations for which y_{1s} successes were observed. Let

$$y_s = \sum_{j=0,1} y_{js} = y_{0s} + y_{1s} \quad \text{and} \quad n_s = (n_s - y_s) + y_s.$$

The likelihood of the distribution of Y_{0s} and Y_{1s} can be expressed as a product of two binomial distributions

$$\begin{aligned} L_s(\pi_{0s}, \pi_{1s}) &= P(y_{0s}, y_{1s} \mid n_{0s}, n_{1s}, \pi_{0s}, \pi_{1s}) \\ &= \binom{n_{0s}}{y_{0s}} \pi_{0s}^{y_{0s}} (1-\pi_{0s})^{n_{0s}-y_{0s}} \times \binom{n_{1s}}{y_{1s}} \pi_{1s}^{y_{1s}} (1-\pi_{1s})^{n_{1s}-y_{1s}}, \end{aligned}$$

and the overall complete data likelihood for S tables is the product of the table-specific product binomial likelihoods. Many of the statistical procedures for analyzing the product binomial model assume that the success probabilities π_{js} are constant across tables. When this assumption is unlikely to hold, the π_{js} may alternatively be

modeled as conditionally independent draws from a common Beta(a_{j1}, a_{j0}) distribution with probability density function

$$P(\pi_{js} \mid a_{j0}, a_{j1}) = \frac{\Gamma(a_{j0}+a_{j1})}{\Gamma(a_{j0})\Gamma(a_{j1})}\pi_{js}^{a_{j1}-1}(1-\pi_{js})^{a_{j0}-1},$$

where $\Gamma(\cdot)$ is the gamma function, $0<\pi_{js}<1$, and the shape parameters $a_{j(\cdot)}>0$. The joint beta-binomial distribution is then

$$P(y_{js}, \pi_{js} \mid n_{js}, a_{j0}, a_{j1}) = \binom{n_{js}}{y_{js}} \frac{\Gamma(a_{j0}+a_{j1})}{\Gamma(a_{j0})\Gamma(a_{j1})}\pi_{js}^{y_{js}+a_{j1}-1}(1-\pi_{js})^{n_{js}-y_{js}+a_{j0}-1}.$$

Reparameterizing this hierarchical model in terms of the parameters $\mu_j = a_{j1}/(a_{j0}+a_{j1})$ and $\tau_j = a_{j0}+a_{j1}$ and then integrating out the π_{js} gives the unconditional product beta-binomial likelihood contribution by complete data table s

$$\begin{aligned} L_{cs}(\mu_0, \tau_0, \mu_1, \tau_1) &= P(y_{0s}, y_{1s} \mid n_{0s}, n_{1s}, \mu_0, \tau_0, \mu_1, \tau_1) \\ &= \binom{n_{0s}}{y_{0s}} \frac{\Gamma(\tau_0)}{\Gamma(\mu_0\tau_0)\Gamma((1-\mu_0)\tau_0)} \\ &\quad \times \frac{\Gamma(y_{0s}+\mu_0\tau_0)\Gamma(n_{0s}-y_{0s}+(1-\mu_0)\tau_0)}{\Gamma(n_{0s}+\tau_0)} \\ &\quad \times \binom{n_{1s}}{y_{1s}} \frac{\Gamma(\tau_1)}{\Gamma(\mu_1\tau_1)\Gamma((1-\mu_1)\tau_1)} \\ &\quad \times \frac{\Gamma(y_{1s}+\mu_1\tau_1)\Gamma(n_{1s}-y_{1s}+(1-\mu_1)\tau_1)}{\Gamma(n_{1s}+\tau_1)}. \end{aligned}$$

The parameters of interest are now μ_j and τ_j, which are assumed to be constant over tables while allowing variation in the probabilities π_{js}. The expectation and the variance of the beta distribution are (e.g. Mosimann, 1962)

$$\begin{aligned} &E(\pi_{js}) = a_{j1}/(a_{j0}+a_{j1}) = \mu_j \quad \text{and} \\ &\operatorname{var}(\pi_{js}) = a_{j0}a_{j1}/(a_{j0}+a_{j1})^2(1+a_{j0}+a_{j1})^{-1} = \mu_j(1-\mu_j)/(1+\tau_j). \end{aligned}$$

Therefore, μ_j corresponds to the average probability across tables and τ_j corresponds to the similarity (inverse of heterogeneity) in probability among tables, with smaller values implying less similarity. The mean and the variance of Y_j are

$$E(Y_{js}) = n_{js}\mu_j \quad \text{and} \quad \operatorname{var}(Y_{js}) = n_{js}\mu_j(1-\mu_j)[1+(n_{js}-1)\phi_j(1+\phi_j)^{-1}],$$

with $\phi_j = 1/\tau_j$. Hence the beta-binomial distribution reduces to the binomial when $\phi_j = 0$, so that ϕ_j can be thought of as a measure of overdispersion of the beta-binomial distribution relative to binomial variation. An alternative measure of heterogeneity, suggested by Prentice (1986), is the intra-class correlation coefficient $\rho_j = \phi_j/(1+\phi_j) = 1/(1+\tau_j)$, satisfying $0 \le \rho_j \le 1$. (See Poortema, 1999, for a review of alternative ways of modeling overdispersion.)

The logarithm of the product beta-binomial likelihood function for complete data table s is

$$\begin{aligned}\ell_{cs}(\mu_0,\tau_0,\mu_1,\tau_1;y_{0s},y_{1s}) = \sum_{j=0,1} C_{js} &+ \log\Gamma(\tau_j) - \log\Gamma(\mu_j\tau_j) - \log\Gamma((1-\mu_j)\tau_j)\\ &+ \log\Gamma(y_{js}+\mu_j\tau_j) + \log\Gamma(n_{js}-y_{js}+(1-\mu_j)\tau_j)\\ &- \log\Gamma(n_{js}+\tau_j),\end{aligned}$$

[correction added on 1 September 2008, after first online publication: the following phrase has been corrected] where the logarithm of the binomial coefficient

$$C_{js} = \log\Gamma(n_{js}+1) - \log\Gamma(y_{js}+1) - \log\Gamma(n_{js}-y_{js}+1)$$

depends on the observed data but not on the model parameters. The overall unconditional log-likelihood is the sum of the log-likelihoods of the individual tables.

The ML estimates of μ_j and τ_j can be obtained using a variety of optimization techniques. Some of them require the gradients. The analytic first- and second-order derivatives of the complete data log-likelihood with respect to μ_j and τ_j, presented for later use, are

$$\begin{aligned}\frac{\partial\ell_{cs}}{\partial\mu_j} &= [\psi(z_{j1}) - \psi(\bar{z}_{j1}) - \psi(z_{j0}) + \psi(\bar{z}_{j0})]\tau_j,\\ \frac{\partial\ell_{cs}}{\partial\tau_j} &= [\psi(z_{j1}) - \psi(\bar{z}_{j1})]\mu_j + [\psi(z_{j0}) - \psi(\bar{z}_{j0})](1-\mu_j) - \psi(z_j) + \psi(\bar{z}_j),\\ \frac{\partial^2\ell_{cs}}{\partial\mu_j^2} &= [\psi'(z_{j1}) - \psi'(\bar{z}_{j1}) + \psi'(z_{j0}) - \psi'(\bar{z}_{j0})]\tau_j^2,\\ \frac{\partial^2\ell_{cs}}{\partial\tau_j^2} &= [\psi'(z_{j1}) - \psi'(\bar{z}_{j1})]\mu_j^2 + [\psi'(z_{j0}) - \psi'(\bar{z}_{j0})](1-\mu_j)^2 - \psi'(z_j) + \psi'(\bar{z}_j),\\ \frac{\partial^2\ell_{cs}}{\partial\mu_j\partial\tau_j} &= [\psi'(z_{j1}) - \psi'(\bar{z}_{j1})]\mu_j\tau_j - [\psi'(z_{j0}) - \psi'(\bar{z}_{j0})](1-\mu_j)\tau_j + \psi(z_{j1}) - \psi(\bar{z}_{j1})\\ &\quad - \psi(z_{j0}) + \psi(\bar{z}_{j0}),\end{aligned}$$

where $\psi(\cdot) := \partial\log\Gamma(\cdot)/\partial(\cdot)$ and $\psi'(\cdot) := \partial^2\log\Gamma(\cdot)/\partial(\cdot)^2$ are the digamma and trigamma functions, respectively,

$$\begin{aligned}&z_{j1} = y_{js} + \mu_j\tau_j, \quad \bar{z}_{j1} = \mu_j\tau_j, \quad z_{j0} = n_{js} - y_{js} + (1-\mu_j)\tau_j, \quad \bar{z}_{j0} = (1-\mu_j)\tau_j,\\ &z_j = n_{js} + \tau_j \quad \text{and} \quad \bar{z}_j = \tau_j.\end{aligned}$$

The derivatives are easily shown to be equivalent to the specification provided by Morgan (1992), using the recurrence relations for the polygamma functions reported in Appendix A.1.

The ecological inference problem occurs if the cell counts are unobserved and the only data available are the row and column sums of the entries observed as marginal totals. If the row totals are postulated to be fixed and the column totals are taken to be random observations, the distribution of Y_s is a convolution of

two beta-binomial distributions (Böckenholt and Dillon, 2000; Wakefield, 2004), conveniently expressed as

$$L_{os}(\mu_0,\tau_0,\mu_1,\tau_1)=P(y_s\,|\,n_{0s},n_s,\mu_0,\tau_0,\mu_1,\tau_1)$$
$$=\sum_{g_{0s}=y^l_{0s}}^{y^u_{0s}}P(g_{0s},y_s|n_{0s},n_s,\mu_0,\tau_0,\mu_1,\tau_1),$$

where the summation is over all possible values that y_{0s} can take on given the row and column margins, with lower bound $y^l_{0s}=\max[0,y_s-(n_s-n_{0s})]$ and upper bound $y^u_{0s}=\min(n_{0s},y_s)$. Using the summary notation

$$C^g_{js}=\log\Gamma(n_{js}+1)-\log\Gamma(g_{js}+1)-\log\Gamma(n_{js}-g_{js}+1),$$

with $g_{1s}=y_s-g_{0s}$, the convolution log-likelihood of the marginal observation for table s is

$$\ell_{os}(\mu_0,\tau_0,\mu_1,\tau_1;y_s)=\sum_{g_{0s}=y^l_{0s}}^{y^u_{0s}}\Bigg\{\sum_{j=0,1}C^g_{js}+\log\Gamma(\tau_j)-\log\Gamma(\mu_j\tau_j)$$
$$-\log\Gamma((1-\mu_j)\tau_j)+\log\Gamma(g_{js}+\mu_j\tau_j)$$
$$+\log\Gamma(n_{js}-g_{js}+(1-\mu_j)\tau_j)-\log\Gamma(n_{js}+\tau_j)\Bigg\},$$
$$(g_{1s}=y_s-g_{0s}).$$

For a single 2×2 table, the conditional distribution of Y_{0s} given Y_s is an extended beta-hypergeometric likelihood, obtained as

$$L_{ms}(\mu_0,\tau_0,\mu_1,\tau_1)=P(y_{0s}\,|\,y_s,n_{0s},n_s,\mu_0,\tau_0,\mu_1,\tau_1)$$
$$=\frac{P(y_{0s},y_{1s}\,|\,n_{0s},n_s,\mu_0,\tau_0,\mu_1,\tau_1)}{\sum_{g_{0s}=y^l_{0s}}^{y^u_{0s}}P(g_{0s},y_s\,|\,n_{0s},n_s,\mu_0,\tau_0,\mu_1,\tau_1)}.\qquad(1)$$

Let $\boldsymbol{\theta}=(\mu_0,\tau_0,\mu_1,\tau_1)^{\mathrm{T}}$. Taking logarithms of both sides of (1) and rearranging we have

$$\ell_{cs}(\boldsymbol{\theta};y_{0s},y_{1s})=\ell_{os}(\boldsymbol{\theta};y_s)+\ell_{ms}(\boldsymbol{\theta};y_{0s}\,|\,y_s),\qquad(2)$$

where $\ell_{cs}(\boldsymbol{\theta};y_{0s},y_{1s})$ is the complete, $\ell_{ms}(\boldsymbol{\theta};y_{0s}\,|\,y_s)$ the missing and $\ell_{os}(\boldsymbol{\theta};y_s)$ the observed data log-likelihood for table s.

3 Expectation–maximization

Maximum-likelihood estimation helps obtain model parameters for which the observed data are most likely. The EM algorithm circumvents direct consideration of $\ell_{os}(\boldsymbol{\theta};y_s)$ by working with the complete data log-likelihood $\ell_{cs}(\boldsymbol{\theta};y_{0s},y_{1s})$ (Dempster *et al.*, 1977). Define $\boldsymbol{\theta}^{(t)}$ to be the maximizer at iteration t, for $t=0,1,\ldots$. As the

observed data log-likelihood does not depend on Y_{0s}, taking expectations of both sides of (2) with respect to the current (posterior) conditional distribution $P(g_{0s} \mid y_s, \boldsymbol{\theta}^{(t)})$ – hereafter referred to as $E\{\cdot \mid y_s, \boldsymbol{\theta}^{(t)}\}$ – yields

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \sum_{s=1}^{S} \ell_{os}(\boldsymbol{\theta}; y_s) + \sum_{s=1}^{S} E\{\ell_{ms}(\boldsymbol{\theta}; y_{0s} \mid y_s) \mid y_s, \boldsymbol{\theta}^{(t)}\},$$

where

$$\begin{aligned} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) &= \sum_{s} E\{\ell_{cs}(\boldsymbol{\theta}; y_{0s}, y_{1s}) \mid y_s, \boldsymbol{\theta}^{(t)}\} \\ &= \sum_{s} \sum_{g_{0s}} \log\{P(g_{0s} \mid \boldsymbol{\theta}) P(y_s \mid g_{0s}, \boldsymbol{\theta})\} P(g_{0s} \mid y_s, \boldsymbol{\theta}^{(t)}). \end{aligned}$$

The EM algorithm starts from $\boldsymbol{\theta}^{(0)}$ and then alternates between the expectation and maximization steps until a stopping criterion has been met

E-step: compute $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$,
M-step: maximize $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$.

In the M-step, the estimates $\boldsymbol{\theta}^{(t+1)}$ are obtained from $\boldsymbol{\theta}^{(t)}$ as $\boldsymbol{\theta}^{(t+1)} = \arg\max_{\theta} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$. To perform this maximization, the gradient of Q with respect to $\boldsymbol{\theta}$ is equated to zero

$$\frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}} = \sum_{s} \sum_{g_{0s}} \partial \log\{P(g_{0s} \mid \boldsymbol{\theta}) P(y_s \mid g_{0s}, \boldsymbol{\theta})\} / \partial \boldsymbol{\theta} P(g_{0s} \mid y_s, \boldsymbol{\theta}^{(t)}) = 0.$$

Solving for θ_{jk}, with $\theta_{01} = \mu_0 \tau_0$, $\theta_{00} = (1 - \mu_0)\tau_0$, $\theta_{11} = \mu_1 \tau_1$ and $\theta_{10} = (1 - \mu_1)\tau_1$, results in the following iterative M-step scheme

$$\psi(\theta_{jk}^{(t+1)}) = \frac{1}{s} \sum_{s} \sum_{g_{0s}} \left\{ \psi(z_{jk}^{(t)}) - \psi(z_j^{(t)}) + \psi(\bar{z}_j^{(t)}) \right\} P(g_{0s} \mid y_s, \boldsymbol{\theta}^{(t)}), \tag{3}$$

$$\mu_j^{(t+1)} = \psi^{-1}(\theta_{j1}^{(t+1)}) / \tau_j^{(t+1)},$$

$$\tau_j^{(t+1)} = \psi^{-1}(\theta_{j0}^{(t+1)}) + \psi^{-1}(\theta_{j1}^{(t+1)}),$$

where $z_{j1} = g_{js} + \mu_j \tau_j$, $z_{j0} = n_{js} - g_{js} + (1 - \mu_j)\tau_j$, $z_j = n_{js} + \tau_j$, and $\bar{z}_j = \tau_j$. Note that the M-step requires inverting the digamma function. This inversion can be performed efficiently using a Newton update procedure proposed by Minka (2003) and reproduced in Appendix A.2.

Minka (2003) also offers a fixed-point iteration algorithm to obtain parameter estimates for the Dirichlet-multinomial distribution. The idea behind this alternative generalized EM algorithm is to guess an initial $\boldsymbol{\theta}^{(t)}$, find a function that bounds the log-likelihood from below which is tight at $\boldsymbol{\theta}^{(t)}$, and then to optimize this function in closed form to arrive at a new guess $\boldsymbol{\theta}^{(t+1)}$. This approach is also guaranteed

to converge to a stationary point of the likelihood. As described in Appendix A.3, when applied to the current model, the fixed-point algorithm leads to the following convergent iterative scheme for computing the ML estimates

$$\theta_{jl}^{(t+1)} = \theta_{jl}^{(t)} \frac{\sum_s \sum_{g_{0s}} \{\psi(z_{jl}^{(t)}) - \psi(\bar{z}_{jl}^{(t)})\} P(g_{0s} \mid y_s, \boldsymbol{\theta}^{(t)})}{\sum_s \sum_{g_{0s}} \{\psi(z_j^{(t)}) - \psi(\bar{z}_j^{(t)})\} P(g_{0s} \mid y_s, \boldsymbol{\theta}^{(t)})}, \tag{4}$$

where the parameter θ_{jl} is taken to be $\theta_{01} = \mu_0$, $\theta_{00} = \tau_0$, $\theta_{11} = \mu_1$, $\theta_{10} = \tau_1$, and

$$z_{j1} = g_{js} + \mu_j \tau_j, \quad \bar{z}_{j1} = \mu_j \tau_j, \quad z_{j0} = n_{js} - g_{js} + (1 - \mu_j)\tau_j, \quad \bar{z}_{j0} = (1 - \mu_j)\tau_j,$$
$$z_j = n_{js} + \tau_j, \quad \bar{z}_j = \tau_j.$$

Because the bound on the log-likelihood only matches the first-order derivatives, convergence of the EM iterations is linear and the algorithm can thus be slow. Therefore, while EM has an important theoretical advantage in that it provides statistical structure to the current missing data problem, it may be more convenient for parameter estimation to switch to a quadratic convergence method such as Newton–Raphson (NR). The NR general update rule is $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \mathbf{H}^{-1}\mathbf{g}$, where $\mathbf{H}$ is the Hessian matrix of the second-order derivatives of the observed data log-likelihood function and $\mathbf{g}$ the gradient vector of the log-likelihood, all evaluated at $\boldsymbol{\theta}^{(t)}$. The NR algorithm converges faster than EM, especially if there is a large range of possible values for the missing interior cell counts. The observed data derivatives are presented below.

4 Information loss

An important feature of the missing information principle formulated by Orchard and Woodbury (1972) is that the first-order derivative of the observed data log-likelihood with respect to the parameters can be obtained by taking the estimated conditional expectation of the score of the complete data log-likelihood, given the observed data (see also Woodbury, 1971; Laird, 1985; Steel, Beh and Chambers, 2004). That is, partially differentiating both sides of (2) with respect to $\boldsymbol{\theta}_j = (\mu_j, \tau_j)^{\mathrm{T}}$, and then taking expectations of both sides with respect to the conditional distribution of the complete data given the observed data, yields

$$\frac{\partial \ell_{os}}{\partial \theta_j} = E\left\{ \frac{\partial \ell_{cs}}{\partial \theta_j} \,\middle|\, y_s, \boldsymbol{\theta} \right\} = \sum_{g_{0s} = y_{0s}^{l}}^{y_{0s}^{u}} \frac{\partial \ell_{cs}}{\partial \theta_j} P(g_{0s} \mid y_s, \boldsymbol{\theta}).$$

Differentiating the identity (2) twice and then averaging both sides over the conditional complete data distributions generates an expression for the pure and mixed second-order derivatives of the observed data log-likelihood for table s, evaluated at the ML estimates of $\boldsymbol{\theta}$

$$\begin{aligned}
\frac{\partial^2 \ell_{os}}{\partial\theta_j \partial\theta_r} &= E\left\{ \frac{\partial^2 \ell_{cs}}{\partial\theta_j \partial\theta_r} \middle| y_s, \boldsymbol{\theta} \right\} + E\left\{ \frac{\partial^2 \ell_{ms}}{\partial\theta_j \partial\theta_r} \middle| y_s, \boldsymbol{\theta} \right\} \\
&= E\left\{ \frac{\partial^2 \ell_{cs}}{\partial\theta_j \partial\theta_r} \middle| y_s, \boldsymbol{\theta} \right\} + \operatorname{cov}\left\{ \left(\frac{\partial \ell_{cs}}{\partial\theta_j}, \frac{\partial \ell_{cs}}{\partial\theta_r} \right) \middle| y_s, \boldsymbol{\theta} \right\} \\
&= E\left\{ \frac{\partial^2 \ell_{cs}}{\partial\theta_j \partial\theta_r} \middle| y_s, \boldsymbol{\theta} \right\} + E\left\{ \left(\frac{\partial \ell_{cs}}{\partial\theta_j}, \frac{\partial \ell_{cs}}{\partial\theta_r} \right) \middle| y_s, \boldsymbol{\theta} \right\} \\
&\quad - E\left\{ \frac{\partial \ell_{cs}}{\partial\theta_j} \middle| y_s, \boldsymbol{\theta} \right\} E\left\{ \frac{\partial \ell_{cs}}{\partial\theta_r} \middle| y_s, \boldsymbol{\theta} \right\} \\
&= \sum_{g_{0s}} \frac{\partial^2 \ell_{cs}}{\partial\theta_j \partial\theta_r} P(g_{0s}|y_s, \boldsymbol{\theta}) + \sum_{g_{0s}} \frac{\partial \ell_{cs}}{\partial\theta_j} \frac{\partial \ell_{cs}}{\partial\theta_r} P(g_{0s}|y_s, \boldsymbol{\theta}) \\
&\quad - \sum_{g_{0s}} \frac{\partial \ell_{cs}}{\partial\theta_j} P(g_{0s}|y_s, \boldsymbol{\theta}) \sum_{g_{0s}} \frac{\partial \ell_{cs}}{\partial\theta_r} P(g_{0s}|y_s, \boldsymbol{\theta}), \qquad (5)
\end{aligned}$$

with θ_j and θ_r each being replaced by either μ_0, μ_1, τ_0 or τ_1 when the derivatives are with respect to these parameters, and $E\{(\partial \ell_{cs}/\partial\theta_j, \partial \ell_{cs}/\partial\theta_r)|y_s, \boldsymbol{\theta}\} = 0$, if $r = 1 - j$.

Negating the derivatives in (5) and summating over all S tables yields the observed information matrices

$$\mathbf{i}_o(\boldsymbol{\theta}) = \mathbf{i}_c(\boldsymbol{\theta}) - \mathbf{i}_m(\boldsymbol{\theta}),$$

where $\mathbf{i}_o(\boldsymbol{\theta})$ is the observed data observed information matrix, $\mathbf{i}_c(\boldsymbol{\theta})$ the information in the complete data, and $\mathbf{i}_m(\boldsymbol{\theta})$ the information in the unobserved data conditional on the observed. This result, termed 'the MIP', is appealing in that it argues that the observed information equals the complete information minus the missing information.

The observed data Fisher (expected) information is obtained by taking expectations of both sides of (5) over the marginal distribution $P(y_s|\boldsymbol{\theta})$ – denoted as $E_{y_s}\{\cdot\}$ – and multiplying the result by -1. The expected values of the pure and mixed second-order derivatives of the observed data log-likelihood for table s are

$$\begin{aligned}
E_{y_s}\left\{ \frac{\partial^2 \ell_{os}}{\partial\theta_j \partial\theta_r} \right\} &= E_{y_s}\left\{ E\left(\frac{\partial^2 \ell_{cs}}{\partial\theta_j \partial\theta_r} \middle| y_s, \boldsymbol{\theta} \right) \right\} + E_{y_s}\left\{ E\left(\frac{\partial^2 \ell_{ms}}{\partial\theta_j \partial\theta_r} \middle| y_s, \boldsymbol{\theta} \right) \right\} \\
&= \sum_{y_s=0}^{n_s} \left\{ \sum_{g_{0s}=y_{0s}^l}^{y_{0s}^u} \frac{\partial^2 \ell_{cs}}{\partial\theta_j \partial\theta_r} P(g_{0s}|y_s, \boldsymbol{\theta}) + \sum_{g_{0s}} \frac{\partial \ell_{cs}}{\partial\theta_j} \frac{\partial \ell_{cs}}{\partial\theta_r} P(g_{0s}|y_s, \boldsymbol{\theta}) \right. \\
&\quad \left. - \sum_{g_{0s}} \frac{\partial \ell_{cs}}{\partial\theta_j} P(g_{0s}|y_s, \boldsymbol{\theta}) \sum_{g_{0s}} \frac{\partial \ell_{cs}}{\partial\theta_r} P(g_{0s}|y_s, \boldsymbol{\theta}) \right\} P(y_s|\boldsymbol{\theta}),
\end{aligned}$$

where

$$\begin{aligned}
E_{y_s}\{E(\partial^2 \ell_{cs}/\partial\theta_j \partial\theta_j | y_s, \boldsymbol{\theta})\} &= -E_{y_s}\{E((\partial \ell_{cs}/\partial\theta_j, \partial \ell_{cs}/\partial\theta_j)|y_s, \boldsymbol{\theta})\}, \\
E_{y_s}\{E(\partial^2 \ell_{cs}/\partial\theta_j \partial\theta_{1-j} | y_s, \boldsymbol{\theta})\} &= -E_{y_s}\{E((\partial \ell_{cs}/\partial\theta_j, \partial \ell_{cs}/\partial\theta_{1-j})|y_s, \boldsymbol{\theta})\}
\end{aligned}$$

and

$$E_{y_s}\{E((\partial \ell_{cs}/\partial\theta_j, \partial \ell_{cs}/\partial\theta_r)|y_s, \boldsymbol{\theta})\} = 0, \quad \text{if } r = 1 - j.$$

Negating the expected derivatives and summating over tables gives the observed data Fisher information matrix which, upon setting $\boldsymbol{\theta}$ to the ML values, is presented as

$$\mathbf{I}_o(\boldsymbol{\theta}) = \mathbf{I}_c(\boldsymbol{\theta}) - \mathbf{I}_m(\boldsymbol{\theta}).$$

This identity, as shown by Meng and Rubin (1991), may be rewritten as

$$\mathbf{I}_o(\boldsymbol{\theta}) = \mathbf{I}_c(\boldsymbol{\theta})\{\mathrm{I} - \boldsymbol{\Phi}(\boldsymbol{\theta})\}^{\mathrm{T}}, \tag{6}$$

where I is a (4×4) identity matrix and $\boldsymbol{\Phi}(\boldsymbol{\theta}) = \mathbf{I}_m(\boldsymbol{\theta})[\mathbf{I}_o(\boldsymbol{\theta}) + \mathbf{I}_m(\boldsymbol{\theta})]^{-1}$ is the information ratio matrix which measures the proportion of information about $\boldsymbol{\theta}$ that is missing by not also observing y_{0s} in addition to y_s (see also Imai *et al.*, 2008). Hence this result expresses the observed data Fisher information matrix as the complete data Fisher information matrix and a shrinkage matrix that takes account of the loss of information because of the missing cell entries. Inverting both sides of (6) gives the estimates

$$\mathbf{I}_o(\boldsymbol{\theta})^{-1} = \mathbf{I}_c(\boldsymbol{\theta})^{-1}\{\mathrm{I} + \boldsymbol{\Phi}(\boldsymbol{\theta})[\mathrm{I} - \boldsymbol{\Phi}(\boldsymbol{\theta})]^{-1}\}.$$

Thus the observed data (co)variance matrix equals the complete data (co)variance matrix plus an incremental matrix that represents the additional uncertainty in 2×2 ecological tables.

5 Bias of ML estimators

Maximum-likelihood estimators may be biased estimators of the true parameter values. In many practical problems, this bias is of limited consequence as the absolute bias decreases with the sample size (or total Fisher information) and as it is typically small relative to the standard error. In finite samples of limited size (and for estimates with large standard errors), bias may be substantial, however, making it important to study. Cox and Snell (1968) provide an order n^{-1} approximation for the biases of the ML estimators of parameters of any distribution. Let $b(\hat{\theta}_j)$ be the n^{-1} bias of the estimator, with $\hat{\theta}_j$ being either $\hat{\mu}_j$ or $\hat{\tau}_j$. For the beta-binomial convolution model considered here, the bias can be expanded in the form

$$b(\hat{\theta}_j) = \frac{1}{2}\sum \mathbf{I}^{\theta_j\theta_r}\mathbf{I}^{\theta_t\theta_u}\left(\mathbf{K}_{\theta_r\theta_t\theta_u} + 2\mathbf{J}_{\theta_r\theta_t,\theta_u}\right), \tag{7}$$

where the index parameters θ_r, θ_t and θ_u are each replaced by either μ_0, μ_1, τ_0 or τ_1 and the summation is over the resulting 4^3 elements of the sum. The explicit expression for the biases of $\hat{\mu}_j$ and $\hat{\tau}_j$ is presented in Appendix A.4. The superscripts in (7) denote matrix inversion of the observed data expected information matrix $\mathbf{I}$ (the subscript 'o' is suppressed to simplify notation), so that $\mathbf{I}^{\theta_j\theta_r} = (\mathbf{I}^{-1})_{\theta_j\theta_r}$, with

$$(\mathbf{I})_{\theta_j\theta_r} = \sum_s E_{y_s}(-\partial^2 \ell_{os}/\partial\theta_j\partial\theta_r),$$

$$\mathbf{K}_{\theta_r\theta_t\theta_u} = \sum_s E_{y_s}(\partial^3 \ell_{os}/\partial\theta_r\partial\theta_t\partial\theta_u) \quad \text{and}$$

$$\mathbf{J}_{\theta_r\theta_t,\theta_u} = \sum_s E_{y_s}(\partial^2 \ell_{os}/\partial\theta_r\partial\theta_t, \partial\ell_{os}/\partial\theta_u),$$

with the expectations taken over the marginal distribution $P(y_s|\boldsymbol{\theta})$, and θ_r, θ_t and θ_u each being replaced by either μ_0, μ_1, τ_0 or τ_1, when the derivatives are with respect to these parameters.

The expected third-order derivatives of the observed data log-likelihood are

$$E_{y_s}\left(\frac{\partial^3 \ell_{os}}{\partial\theta_r\partial\theta_r\partial\theta_r}\right) = -3E_{y_s}\left(\frac{\partial\ell_{os}}{\partial\theta_r}, \frac{\partial^2\ell_{os}}{\partial\theta_r^2}\right) - E_{y_s}\left(\frac{\partial\ell_{os}}{\partial\theta_r}, \frac{\partial\ell_{os}}{\partial\theta_r}, \frac{\partial\ell_{os}}{\partial\theta_r}\right),$$

$$E_{y_s}\left(\frac{\partial^3 \ell_{os}}{\partial\theta_r\partial\theta_r\partial\theta_t}\right) = -2E_{y_s}\left(\frac{\partial\ell_{os}}{\partial\theta_r}, \frac{\partial^2\ell_{os}}{\partial\theta_r\partial\theta_t}\right) - E_{y_s}\left(\frac{\partial\ell_{os}}{\partial\theta_t}, \frac{\partial^2\ell_{os}}{\partial\theta_r^2}\right) - E_{y_s}\left(\frac{\partial\ell_{os}}{\partial\theta_r}, \frac{\partial\ell_{os}}{\partial\theta_r}, \frac{\partial\ell_{os}}{\partial\theta_t}\right),$$

$$E_{y_s}\left(\frac{\partial^3 \ell_{os}}{\partial\theta_r\partial\theta_t\partial\theta_u}\right) = -E_{y_s}\left(\frac{\partial\ell_{os}}{\partial\theta_r}, \frac{\partial^2\ell_{os}}{\partial\theta_t\partial\theta_u}\right) - E_{y_s}\left(\frac{\partial\ell_{os}}{\partial\theta_t}, \frac{\partial^2\ell_{os}}{\partial\theta_r\partial\theta_u}\right) - E_{y_s}\left(\frac{\partial\ell_{os}}{\partial\theta_u}, \frac{\partial^2\ell_{os}}{\partial\theta_r\partial\theta_t}\right) - E_{y_s}\left(\frac{\partial\ell_{os}}{\partial\theta_r}, \frac{\partial\ell_{os}}{\partial\theta_t}, \frac{\partial\ell_{os}}{\partial\theta_u}\right),$$

where

$$\partial\ell_{os}/\partial\theta_r = E(\partial\ell_{cs}/\partial\theta_r|y_s, \boldsymbol{\theta}),$$
$$\partial^2\ell_{os}/\partial\theta_r^2 = E(\partial^2\ell_{cs}/\partial\theta_r^2|y_s, \boldsymbol{\theta}) + E(\partial{\ell_{cs}}^2/\partial\theta_r|y_s, \boldsymbol{\theta}) - E(\partial\ell_{cs}/\partial\theta_r|y_s, \boldsymbol{\theta})^2,$$

and

$$\partial^2\ell_{os}/\partial\theta_r\partial\theta_t = E(\partial^2\ell_{cs}/\partial\theta_r\partial\theta_t|y_s, \boldsymbol{\theta}) + E((\partial\ell_{cs}/\partial\theta_r, \partial\ell_{cs}/\partial\theta_t)|y_s, \boldsymbol{\theta}) - E(\partial\ell_{cs}/\partial\theta_r|y_s)E(\partial\ell_{cs}/\partial\theta_t|y_s, \boldsymbol{\theta}),$$

with

$$E((\partial\ell_{cs}/\partial\theta_r, \partial\ell_{cs}/\partial\theta_t)|y_s, \boldsymbol{\theta}) = 0, \quad \text{if } t = 1 - j.$$

The bias-corrected ML estimate $\hat{\theta}_j^c$ can be obtained using $\hat{\theta}_j^c = \hat{\theta}_j - b(\hat{\theta}_j)$, where $\hat{\theta}_j$ is the uncorrected estimate.

6 Empirical examples

We present two studies to illustrate selected results of the previous sections.

6.1 Student workload and grades

The first application is a study to assess the relationship between student-reported workload required to complete course units and teacher-awarded course grades show-

ing student's performance. Data were taken from the records and reports of $n_s = 9$ MSc students who in 2006-2007 have evaluated $S = 10$ courses offered by the Radboud University Nijmegen. The data available for analysis consist of the number of students who reported more than nominal workload for course s in an anonymous evaluation questionnaire (n_{0s}) and the number of students who received a grade of eight or higher (y_s) for course s, on a 10-point scale, potentially ranging from 0 (very bad) to 10 (excellent). The marginal totals are $n_0 = (6, 6, 3, 4, 4, 1, 3, 8, 1, 6)$ and $y = (9, 4, 7, 3, 7, 2, 5, 3, 9, 0)$. The cross-classified counts are unavailable as the evaluations were submitted anonymously. Table 2 displays the EM iterates obtained using (3).

Table 2. Expectation–maximization iterates.

t	$\mu_0^{(t)}$	$\tau_0^{(t)}$	$\mu_1^{(t)}$	$\tau_1^{(t)}$	$Q(\boldsymbol{\theta})^{(t)}$	$-2\ell_o(\boldsymbol{\theta})^{(t)}$
0	0.500000	1.000000	0.500000	1.000000	−31.469871	47.868245
1	0.559371	1.004756	0.523584	1.023260	−31.125080	47.258738
2	0.587411	1.007772	0.528429	1.043160	−31.014048	47.115523
3	0.601756	1.008137	0.527058	1.061520	−30.959361	47.064992
4	0.609701	1.006755	0.524061	1.079192	−30.924327	47.037748
5	0.614420	1.004329	0.521030	1.096445	−30.898975	47.018535
10	0.622313	0.986462	0.512183	1.176663	−30.829666	46.957551
50	0.626118	0.873011	0.500493	1.504337	−30.660371	46.840406
100	0.626587	0.834535	0.498147	1.597617	−30.611626	46.832941
250	0.626680	0.825826	0.497640	1.619123	−30.600623	46.832693
400	0.626680	0.825771	0.497637	1.619265	−30.600553	46.832693
434	0.626680	0.825771	0.497637	1.619265	−30.600553	46.832693

The ML estimates $\hat{\mu}_0$ and $\hat{\mu}_1$ – determined by EM as 0.627 and 0.498, respectively – suggest that students who reported above-nominal workload have on average a higher probability of obtaining a grade of eight or higher than those who reported nominal workload or less. However, as indicated by the low values for the $\hat{\tau}_0$ and $\hat{\tau}_1$ estimates, there is a substantial amount of heterogeneity in probability across courses. The estimates $\hat{\mu}_0$ and $\hat{\tau}_0$ imply a beta distribution with shape parameters $\text{Beta}(0.518, 0.308)$ $[= (0.62668 \times 0.825771, (1 - 0.62668) \times 0.825771)]$ and $\hat{\mu}_1$ and $\hat{\tau}_1$ a $\text{Beta}(0.806, 0.814)$ distribution. The probability density function of the two distributions is U-shaped.

The estimated Fisher information about the parameters provided by the observed data and its inverse are

$$\mathbf{I}_o(\hat{\boldsymbol{\theta}})_{\mu_0,\tau_0,\mu_1,\tau_1} = \begin{pmatrix} 41.872 & -0.519 & 21.307 & -0.038 \\ -0.519 & 1.168 & -0.349 & 0.101 \\ 21.307 & -0.349 & 42.960 & -0.046 \\ -0.038 & 0.101 & -0.046 & 0.284 \end{pmatrix},$$

$$\mathbf{I}_o(\hat{\boldsymbol{\theta}})^{-1} = \begin{pmatrix} 0.032 & 0.010 & -0.016 & -0.002 \\ 0.010 & 0.889 & 0.002 & -0.313 \\ -0.016 & 0.002 & 0.031 & 0.002 \\ -0.002 & -0.313 & 0.002 & 3.634 \end{pmatrix}.$$

Had the complete data 2×2 tables been available, the Fisher information matrix and the variance–covariance matrix, evaluated at the observed data ML estimates, would have been

$$\mathbf{I}_c(\hat{\boldsymbol{\theta}})_{\mu_0,\tau_0,\mu_1,\tau_1} = \begin{pmatrix} 66.757 & -1.853 & 0 & 0 \\ -1.853 & 2.984 & 0 & 0 \\ 0 & 0 & 70.916 & 0.015 \\ 0 & 0 & 0.015 & 0.641 \end{pmatrix},$$

$$\mathbf{I}_c(\hat{\boldsymbol{\theta}})^{-1} = \begin{pmatrix} 0.015 & 0.010 & 0 & 0 \\ 0.010 & 0.341 & 0 & 0 \\ 0 & 0 & 0.014 & -0.001 \\ 0 & 0 & -0.001 & 1.560 \end{pmatrix}.$$

Perhaps the most notable difference with respect to the off-block elements is the relatively strong covariance between the estimates $\hat{\mu}_0$ and $\hat{\mu}_1$. The estimated parameter inter-correlation is $\hat{r}_{\mu_0,\mu_1} = -0.501$. As to the block-diagonal elements, the greatest information loss appears to occur for the $\hat{\tau}_j$ parameters. The observed marginal data provide relatively little information about the heterogeneity parameters and their estimated variances are consequently large relative to the estimates themselves. This may be due to the sparsity of the data. If all the observed data marginal counts are increased by a factor of 10 and the parameters are subsequently re-estimated, the ML estimates change to $\hat{\boldsymbol{\theta}}_{\mu_0,\tau_0,\mu_1,\tau_1} = (0.627, 0.375, 0.558, 0.568)$ and the observed data parameter variances to $\hat{\boldsymbol{\sigma}}^2_o = (0.024, 0.107, 0.022, 0.162)$. This finding suggests that $\hat{\tau}_1$ is most affected by the meagre marginal totals. The estimated biases, calculated as $b(\hat{\mu}_0, \hat{\tau}_0, \hat{\mu}_1, \hat{\tau}_1) = (-0.010, 0.609, 0.013, 1.694)$, support this view. Whereas the $\hat{\mu}_j$ parameters are only slightly biased, the estimated biases of the $\hat{\tau}_j$ parameters, $\hat{\tau}_1$ in particular, are severe.

These results may lead one to examine a convolution binomial model for these data rather than a beta-binomial one. Application of the convolution binomial distribution yields similar estimates for $\hat{\mu}_0$ (0.610) and $\hat{\mu}_1$ (0.500), but this model, albeit more parsimonious, has a worse fit to the observed data ($-2\ell_o(\hat{\boldsymbol{\theta}}) = 64.421$). In addition, note from Table 2 that the EM algorithm requires 434 iterations to converge. The NR algorithm helps obtain the same ML estimates in five iterations, using identical starting values and stopping criterion.

6.2 Party registration and race

The second example concerns an examination of the party registration–race data analyzed by Wakefield (2004). The population data are from $S = 64$ counties in the southern US state of Louisiana collected in 1990, for which n_{0s} and n_{1s} are black and white, respectively, and y_s are Republican and $(n_s - y_s)$ are Democratic registration ($N = 1{,}980{,}775$). For these 64 tables the cross-classified counts are available. The observed population-averaged proportions of Republican party registration are 0.035 for black people and 0.254 for white people. The standard deviations

Table 3. ML parameter estimates and biases.

	Beta-binomial convolution ecological data	Product beta-binomial complete data
$\hat{\mu}_0$	0.075	0.031
$\hat{\tau}_0$	58.054	145.837
$\hat{\mu}_1$	0.174	0.195
$\hat{\tau}_1$	18.973	20.704
$b(\hat{\mu}_0)$	-6.0×10^{-4}	-3.9×10^{-7}
$b(\hat{\tau}_0)$	301.117	7.519
$b(\hat{\mu}_1)$	1.7×10^{-4}	-1.1×10^{-5}
$b(\hat{\tau}_1)$	1.666	0.972

– obtained as 0.013 and 0.107, respectively – indicate that there is substantial variation in the proportions of white people with Republican registration across counties; the fractions vary from 0.072 to 0.417. It may also be noted that the grand totals in this application are large (the maximum value of n_s is 217,967) and that the number of possible complete data tables to be processed in computing the marginal expectations is in the order of 31.7 billion [$\approx \sum_{s=1}^{64} \sum_{y_s=0}^{n_s} y_{0s}^u - y_{0s}^l + 1$]. Hence it may take some time for a computer to calculate the desired results.

The NR algorithm was used to estimate both the ML parameters of the beta-binomial convolution model using the marginal totals only, and the ML parameters of the product binomial distribution employing the internal cell counts. The results are presented in Table 3.

The $\hat{\mu}_j$ parameter estimates of the beta-binomial convolution model are relatively close to the corresponding product beta-binomial estimates. The same goes for $\hat{\tau}_1$. A large difference in magnitude is found for the heterogeneity parameter $\hat{\tau}_0$. However, the discrepancy is not as gross as it may seem. The overall shapes of the probability density function of Beta(0.075,58.054) and Beta(0.031,145.837) are rather similar. The latter has a somewhat more extended tail to the right, but they are both very peaked.

The biases of the ML parameters of the beta-binomial convolution model were obtained using (7). The bias formula for the beta-binomial model is presented in Appendix A.5. As can be seen in Table 3, the estimated biases of the $\hat{\mu}_j$ parameters are negligibly small. The absolute biases of $\hat{\tau}_j$ appear large, with the largest bias found for $\hat{\tau}_0$. However, the parameters have a small bias compared with the magnitude of the standard error.

For the beta-binomial convolution model applied to the 2×2 ecological tables, the variance–covariance matrix is estimated as

$$\mathbf{I}_o(\hat{\boldsymbol{\theta}})^{-1}_{\mu_0,\tau_0,\mu_1,\tau_1} = 10^{-2} \begin{pmatrix} 0.131 & -188.404 & -0.055 & -1.025 \\ -188.404 & 1,828,112.399 & 80.291 & -24,460.628 \\ -0.055 & 80.291 & 0.035 & -0.526 \\ -1.025 & -24{,}460.628 & -0.526 & 1692.182 \end{pmatrix},$$

and the variance–covariance matrix which would have been obtained had all within-table counts been observed is

$$\mathbf{I}_c(\hat{\boldsymbol{\theta}})^{-1}_{\mu_0,\tau_0,\mu_1,\tau_1} = 10^{-2}\begin{pmatrix} 0.002 & -1.336 & 0 & 0 \\ -1.336 & 11{,}177.834 & 0 & 0 \\ 0 & 0 & 0.011 & -0.982 \\ 0 & 0 & -0.982 & 1111.071 \end{pmatrix}.$$

This may be compared with the estimated (co)variance matrix for the product beta-binomial model applied to the complete data cross-classified counts

$$\mathbf{I}_{cc}(\hat{\boldsymbol{\theta}})^{-1}_{\mu_0,\tau_0,\mu_1,\tau_1} = 10^{-2}\begin{pmatrix} 0.001 & -1.525 & 0 & 0 \\ -1.525 & 75{,}577.540 & 0 & 0 \\ 0 & 0 & 0.011 & -0.922 \\ 0 & 0 & -0.922 & 1315.850 \end{pmatrix}.$$

Note that the estimated variances for $\hat{\mu}_j$ obtained for the ecological data are close to the variances obtained for complete data cross-classified counts. The covariances between $\hat{\mu}_j$ and $\hat{\tau}_j$ are also roughly similar. The parameter variances for $\hat{\tau}_j$ are very large and fail to correspond.

7 Discussion

This paper examined the beta-binomial convolution model for the analysis of a series of 2×2 tables with missing cell counts. The model is appropriate to use when the totals of one margin are fixed at their observed values, and the other marginal totals are said to be the sum of two independent beta-binomials. We considered ML parameter estimation using the EM algorithm and Fisher information loss and bias of the ML estimators.

When analyzing ecological data some simplifying assumptions or approximations have to be necessarily made. The current paper takes the average probabilities and the heterogeneity parameters to be constant across tables. A modification of the model could be to regress the parameters on relevant covariates, thereby allowing them to vary. Moreover, inference is accomplished by restricting the parameter space to the set of 2×2 contingency tables that have the same fixed row sums as the observed tables. If the assumption that part of the data is fixed is difficult to fulfill, one may consider the adoption of a bivariate beta-binomial distribution with neither margin fixed. The likelihood function then factorizes into a marginal beta-binomial random variable for the row totals and two conditional beta-binomials for the two rows, making implementation of the distributions and ML parameter estimation straightforward. (See Hamdan and Nasro, 1986 and Kocherlakota and Kocherlakota, 1992, for a discussion of the bivariate binomial distribution if the data consist of only marginal information.)

Acknowledgements

Thanks are due to Priscilla Read for providing the MSc education data. This work was supported by the Netherlands Organisation for Scientific Research (NWO),

Division for Social Sciences (no. 480–04–009). The Louisiana party registration–race data are available at the following address: http://www.faculty.washington.edu/jonno/data.html.

Appendix

A.1 Recurrence relations for polygamma functions

The digamma and trigamma functions satisfy the following difference equations

$$[\psi(n+\theta_j\theta_r)-\psi(\theta_j\theta_r)]\theta_j=\sum_{y=1}^{n}\frac{\theta_j}{(\theta_j\theta_r+y-1)},$$

$$[\psi'(n+\theta_j\theta_r)-\psi'(\theta_j\theta_r)]\theta_j^2=\sum_{y=1}^{n}\frac{\theta_j^2}{(\theta_j\theta_r+y-1)^2},$$

$$[\psi'(n+\theta_j\theta_r)-\psi'(\theta_j\theta_r)]\theta_j\theta_r=\sum_{y=1}^{n}\frac{\theta_j\theta_r}{(\theta_j\theta_r+y-1)^2}.$$

A.2 Inversion of the ψ function

The inversion of the digamma function $x=\psi^{-1}(y)$ has to be done iteratively, since no closed-form solution exists. Minka (2003) proposes Newton's method to obtain a highly accurate iterative solution with guaranteed and rapid convergence. The complete Newton iteration to solve $\psi(x)=y$ for x given y is as follows. Set the initial value

$$x^{(0)}=\begin{cases}\exp(y)+0.5 & \text{if } y\geq -2.22\\ -1/(y+\gamma) & \text{if } y<-2.22,\end{cases}$$

where γ is the Euler–Mascheroni constant. Then iterate $x^{(t+1)}=x^{(t)}+\delta^{(t)}$, with $\delta^{(t)}=(y-\psi(x^{(t)}))/\psi'(x^{(t)})$. Less than five Newton updates are adequate to achieve more than ten–digit accurate values.

A.3 Generalized EM update procedure using fixed-point iteration

Using the following inequalities associated with the ratio $\Gamma(x+\alpha)/\Gamma(x)$ to construct a lower bound on the log likelihood

$$\frac{\Gamma(\tau)}{\Gamma(n+\tau)}\geq\frac{\Gamma(\tau^{(t)})\exp((\tau^{(t)}-\tau)d)}{\Gamma(n+\tau^{(t)})},$$

$$\frac{\Gamma(y+\mu\tau)}{\Gamma(\mu\tau)}\geq\frac{\Gamma(y+\mu^{(t)}\tau^{(t)})}{\Gamma(\mu^{(t)}\tau^{(t)})}(\mu^{(t)}\tau^{(t)})^{-e_0}(\mu\tau)^{e_0}\quad\text{if } y\geq 1,$$

$$\frac{\Gamma((n-y)+(1-\mu)\tau)}{\Gamma((1-\mu)\tau)} \geq \frac{\Gamma((n-y)+(1-\mu^{(t)})\tau^{(t)})}{\Gamma((1-\mu^{(t)})\tau^{(t)})}((1-\mu^{(t)})\tau^{(t)})^{-e_1}((1-\mu)\tau)^{e_1}$$
$$\text{if } n-y \geq 1,$$

where

$$d = \psi(n+\tau^{(t)}) - \psi(\tau^{(t)}),$$
$$e_0 = [\psi(y+\mu^{(t)}\tau^{(t)}) - \psi(\mu^{(t)}\tau^{(t)})]\mu^{(t)}\tau^{(t)},$$
$$e_1 = [\psi((n-y)+(1-\mu^{(t)})\tau^{(t)}) - \psi((1-\mu^{(t)})\tau^{(t)})](1-\mu^{(t)})\tau^{(t)},$$

we obtain [correction added on 1 September 2008, after first online publication: the following inequality has been corrected]

$$\sum_s \sum_{g_{0s}} \log \left\{ \prod_{j=0,1} \exp\{C_{js}^g\} \frac{\Gamma(\tau_j)}{\Gamma(n_{js}+\tau_j)} \frac{\Gamma(g_{js}+\mu_j\tau_j)}{\Gamma(\mu_j\tau_j)} \frac{\Gamma((n_{js}-g_{js})+(1-\mu_j)\tau_j)}{\Gamma((1-\mu_j)\tau_j)} \right\}$$
$$P(g_{0s}|y_s, \boldsymbol{\theta}^{(t)})$$
$$\geq \sum_s \sum_{g_{0s}} \log \left\{ \prod_{j=0,1} \exp\{C_{js}^g\} \frac{\Gamma(\tau_j^{(t)})\exp\{(\tau_j^{(t)}-\tau_j)\mathrm{d}_j\}}{\Gamma(n_{js}+\tau_j^{(t)})} \frac{\Gamma(g_{js}+\mu_j^{(t)}\tau_j^{(t)})}{\Gamma(\mu_j^{(t)}\tau_j^{(t)})} \right.$$
$$\times \frac{\Gamma((n_{js}-g_{js})+(1-\mu_j^{(t)})\tau_j^{(t)})}{\Gamma((1-\mu_j^{(t)})\tau_j^{(t)})} (\mu_j^{(t)}\tau_j^{(t)})^{-e_{0j}} (\mu_j\tau_j)^{e_{0j}}$$
$$\left. \times ((1-\mu_j^{(t)})\tau_j^{(t)})^{-e_{1j}} ((1-\mu_j)\tau_j)^{e_{1j}} \right\} P(g_{0s}|y_s, \boldsymbol{\theta}^{(t)}) \equiv Q'(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}).$$

Setting the gradient of Q' wrt $\boldsymbol{\theta}$ to zero and solving for θ_{jl} yields the update step (4).

A.4 Biases of μ_j and τ_j for the beta-binomial convolution model

Using the general expression (7), the bias of $\hat{\mu}_j$ is obtained as

$$\begin{aligned} b(\hat{\mu}_j) = \frac{1}{2}\Big\{ \sum_{k=0,1} & [\mathbf{I}^{\mu_j\mu_j}\mathbf{I}^{\mu_k\mu_k}(\mathbf{K}\mu_j\mu_k\mu_k + 2\mathbf{J}\mu_j\mu_k, \mu_k) \\ & + \mathbf{I}^{\mu_j\mu_j}\mathbf{I}^{\tau_k\tau_k}\left(\mathbf{K}\mu_j\tau_k\tau_k + 2\mathbf{J}\mu_j\tau_k, \tau_k\right) \\ & + \mathbf{I}^{\mu_j\tau_k}\mathbf{I}^{\tau_k\tau_k}\left(\mathbf{K}\tau_k\tau_k\tau_k + 2\mathbf{J}\tau_k\tau_k, \tau_k\right) \\ & + \mathbf{I}^{\mu_j\tau_k}\mathbf{I}^{\mu_{1-j}\mu_{1-j}}\left(\mathbf{K}\tau_k\mu_{1-j}\mu_{1-j} + 2\mathbf{J}\tau_k\mu_{1-j}, \mu_{1-j}\right) \\ & + \mathbf{I}^{\mu_j\tau_k}\mathbf{I}^{\tau_{1-k}\tau_{1-k}}\left(\mathbf{K}\tau_k\tau_{1-k}\tau_{1-k} + 2\mathbf{J}\tau_k\tau_{1-k}, \tau_{1-k}\right) \\ & + \mathbf{I}^{\mu_j\mu_{1-j}}\mathbf{I}^{\tau_k\tau_k}\left(\mathbf{K}\mu_{1-j}\tau_k\tau_k + 2\mathbf{J}\mu_{1-j}\tau_k, \tau_k\right) \\ & + \mathbf{I}^{\mu_j\mu_j}\mathbf{I}^{\mu_j\tau_k}\left(3\mathbf{K}\mu_j\mu_j\tau_k + 2\mathbf{J}\mu_j\mu_j, \tau_k + 4\mathbf{J}\mu_j\tau_k, \mu_j\right) \end{aligned}$$

$$
\begin{aligned}
&+2\mathbf{I}^{\mu_j\mu_j}\mathbf{I}^{\mu_{1-j}\tau_k}\left(\mathbf{K}\mu_j\mu_{1-j}\tau_k+\mathbf{J}\mu_j\mu_{1-j},\tau_k+\mathbf{J}\mu_j\tau_k,\mu_{1-j}\right)\\
&+2\mathbf{I}^{\mu_j\mu_k}\mathbf{I}^{\tau_j\tau_{1-j}}\left(\mathbf{K}\mu_k\tau_j\tau_{1-j}+\mathbf{J}\mu_k\tau_j,\tau_{1-j}+\mathbf{J}\mu_k\tau_{1-j},\tau_j\right)\\
&+2\mathbf{I}^{\mu_j\tau_k}\mathbf{I}^{\mu_j\tau_k}\left(\mathbf{K}\tau_k\mu_j\tau_k+\mathbf{J}\tau_k\mu_j,\tau_k+\mathbf{J}\tau_k\tau_k,\mu_j\right)\\
&+2\mathbf{I}^{\mu_j\tau_j}\mathbf{I}^{\mu_{1-j}\tau_k}\left(\mathbf{K}\tau_j\mu_{1-j}\tau_k+\mathbf{J}\tau_j\mu_{1-j},\tau_k+\mathbf{J}\tau_j\tau_k,\mu_{1-j}\right)\\
&+2\mathbf{I}^{\mu_j\tau_k}\mathbf{I}^{\tau_j\tau_{1-j}}\left(\mathbf{K}\tau_k\tau_j\tau_{1-j}+\mathbf{J}\tau_k\tau_j,\tau_{1-j}+\mathbf{J}\tau_k\tau_{1-j},\tau_j\right)\\
&+2\mathbf{I}^{\mu_j\mu_{1-j}}\mathbf{I}^{\mu_{1-j}\tau_k}\left(\mathbf{K}\mu_{1-j}\mu_{1-j}\tau_k+\mathbf{J}\mu_{1-j}\mu_{1-j},\tau_k+\mathbf{J}\mu_{1-j}\tau_k,\mu_{1-j}\right)\\
&+2\mathbf{I}^{\mu_j\tau_{1-j}}\mathbf{I}^{\mu_{1-j}\tau_k}\left(\mathbf{K}\tau_{1-j}\mu_{1-j}\tau_k+\mathbf{J}\tau_{1-j}\mu_{1-j},\tau_k+\mathbf{J}\tau_{1-j}\tau_k,\mu_{1-j}\right)\\
&+2\mathbf{I}^{\mu_j\tau_k}\mathbf{I}^{\mu_j\mu_{1-j}}\left(2\mathbf{K}\tau_k\mu_j\mu_{1-j}+\mathbf{J}\tau_k\mu_j,\mu_{1-j}+2\mathbf{J}\tau_k\mu_{1-j},\mu_j+\mathbf{J}\mu_j\mu_{1-j},\tau_k\right)]\\
&+\mathbf{I}^{\mu_j\mu_{1-j}}\mathbf{I}^{\mu_{1-j}\mu_{1-j}}\left(\mathbf{K}\mu_{1-j}\mu_{1-j}\mu_{1-j}+2\mathbf{J}\mu_{1-j}\mu_{1-j},\mu_{1-j}\right)\\
&+\mathbf{I}^{\mu_j\mu_j}\mathbf{I}^{\mu_j\mu_{1-j}}\left(3\mathbf{K}\mu_j\mu_j\mu_{1-j}+2\mathbf{J}\mu_j\mu_j,\mu_{1-j}+4\mathbf{J}\mu_j\mu_{1-j},\mu_j\right)\\
&+2\mathbf{I}^{\mu_j\mu_{1-j}}\mathbf{I}^{\mu_j\mu_{1-j}}\left(\mathbf{K}\mu_{1-j}\mu_j\mu_{1-j}+\mathbf{J}\mu_{1-j}\mu_j,\mu_{1-j}+\mathbf{J}\mu_{1-j}\mu_{1-j},\mu_j\right)\\
&+2\mathbf{I}^{\mu_j\tau_j}\mathbf{I}^{\mu_j\tau_{1-j}}\left(2\mathbf{K}\tau_j\mu_j\tau_{1-j}+\mathbf{J}\tau_j\mu_j,\tau_{1-j}+2\mathbf{J}\tau_j\tau_{1-j},\mu_j+\mathbf{J}\mu_j\tau_{1-j},\tau_j\right)\Big\},\quad j=0,1.
\end{aligned}
\tag{A.4}
$$

The expression for $b(\hat{\tau}_j)$ is obtained by interchanging the index parameters μ and τ in the RHS of (A.4).

A.5 Biases of μ_j and τ_j for the beta-binomial model

The bias of the ML estimators for the beta-binomial model is obtained using

$$
\begin{aligned}
b(\hat{\mu}_j)=\frac{1}{2}\{&\mathbf{I}^{\mu_j\mu_j}\mathbf{I}^{\mu_j\mu_j}(\mathbf{K}\mu_j\mu_j\mu_j+2\mathbf{J}\mu_j\mu_j,\mu_j)+\mathbf{I}^{\mu_j\mu_j}\mathbf{I}^{\tau_j\tau_j}(\mathbf{K}\mu_j\tau_j\tau_j+2\mathbf{J}\mu_j\tau_j,\tau_j)\\
&+\mathbf{I}^{\mu_j\tau_j}\mathbf{I}^{\tau_j\tau_j}(\mathbf{K}\tau_j\tau_j\tau_j+2\mathbf{J}\tau_j\tau_j,\tau_j)\\
&+\mathbf{I}^{\mu_j\mu_j}\mathbf{I}^{\mu_j\tau_j}(3\mathbf{K}\mu_j\mu_j\tau_j+2\mathbf{J}\mu_j\mu_j,\tau_j+4\mathbf{J}\mu_j\tau_j,\mu_j)\\
&+2\mathbf{I}^{\mu_j\tau_j}\mathbf{I}^{\mu_j\tau_j}(\mathbf{K}\tau_j\mu_j\tau_j+\mathbf{J}\tau_j\mu_j,\tau_j+\mathbf{J}\tau_j\tau_j,\mu_j)\},\quad j=0,1,
\end{aligned}
\tag{A.5}
$$

where the superscripts denote matrix inversion of the complete data expected information matrix $\mathbf{I}$ (the subscript *'c'* is omitted for notational convenience), so that

$$\mathbf{I}^{\theta_j\theta_r}=(\mathbf{I}^{-1})_{\theta_j\theta_r},$$

with

$$
\begin{aligned}
(\mathbf{I})_{\theta_j\theta_r}&=-\sum_s E_{y_s}(\partial^2\ell_{cs}/\partial\theta_j\partial\theta_r),\\
\mathbf{K}_{\theta_r\theta_t\theta_u}&=\sum_s E_{y_s}(\partial^3\ell_{cs}/\partial\theta_r\partial\theta_t\partial\theta_u)\quad\text{and}\\
\mathbf{J}_{\theta_r\theta_t,\theta_u}&=\sum_s E_{y_s}(\partial^2\ell_{cs}/\partial\theta_r\partial\theta_t,\partial\ell_{cs}/\partial\theta_u),
\end{aligned}
$$

with θ_r, θ_t and θ_u each being replaced by either μ_j or τ_j. The expression for $b(\hat{\tau}_j)$ is obtained by interchanging μ_j and τ_j in the RHS of (A.5).

The third-order derivatives of the complete data log-likelihood wrt $\theta_{r,t,u}(=\mu_j, \tau_j)$ are

$$\frac{\partial^3 \ell_{cs}}{\partial\theta_r\partial\theta_r\partial\theta_r} = -3E_{y_s}\left(\frac{\partial\ell_{cs}}{\partial\theta_r}, \frac{\partial^2\ell_{cs}}{\partial\theta_r^2}\right) - E_{y_s}\left(\frac{\partial\ell_{cs}}{\partial\theta_r}, \frac{\partial\ell_{cs}}{\partial\theta_r}, \frac{\partial\ell_{cs}}{\partial\theta_r}\right),$$

$$\frac{\partial^3 \ell_{cs}}{\partial\theta_r\partial\theta_r\partial\theta_t} = -2E_{y_s}\left(\frac{\partial\ell_{cs}}{\partial\theta_r}, \frac{\partial^2\ell_{cs}}{\partial\theta_r\partial\theta_t}\right) - E_{y_s}\left(\frac{\partial\ell_{cs}}{\partial\theta_t}, \frac{\partial^2\ell_{cs}}{\partial\theta_r^2}\right) - E_{y_s}\left(\frac{\partial\ell_{cs}}{\partial\theta_r}, \frac{\partial\ell_{cs}}{\partial\theta_r}, \frac{\partial\ell_{cs}}{\partial\theta_t}\right),$$

where the expectation is taken over the marginal distribution $P(y_s|\boldsymbol{\theta})$.

References

Böckenholt, U. and W. R. Dillon (2000), Inferring latent brand dependencies, *Journal of Marketing Research* **37**, 72–87.

Brown, Ph. J. and C. D. Payne (1986), Aggregate data, ecological regression and voting transitions, *Journal of the American Statistical Association* **81**, 452–460.

Cox, D. R. and E. J. Snell (1968), A general definition of residuals, *Journal of the Royal Statistical Society, Series B*, **30**, 248–275.

Dempster, A. P., N. M. Laird and D. B. Rubin (1977), Maximum likelihood from incomplete data via the *EM* algorithm, *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Eisinga, R. (2008), Information loss for 2×2 tables with missing cell counts: binomial case, *Statistica Neerlandica* **62**, 239–254.

Haber, M. (1989), Do the marginal totals of a contingency table contain information regarding the table proportions? *Communications in Statistics – Theory and Methods* **18**, 147–156.

Hamdan, M. A. and M. O. Nasro (1986), Maximum likelihood estimation of the parameters of the bivariate binomial distribution, *Communications in Statistics – Theory and Methods* **15**, 747–754.

Haneuse, S. J.-P. A. and J. C. Wakefield (2008), The combination of ecological and case–control data, *Journal of the Royal Statistical Society, Series B* **70**, 73–93.

Imai, K., Y. Lu and A. Strauss (2008), Bayesian and likelihood inference for 2×2 ecological tables: an incomplete-data approach, *Political Analysis* **16**, 41–68.

King, G. (1997), *A solution to the ecological inference problem. Reconstructing individual behavior from aggregate data*, Cambridge University Press, Cambridge, MA.

King, G., O. Rosen and M. Tanner (1999), Binomial-beta hierarchical models for ecological inference, *Sociological Methods and Research* **28**, 61–90.

King, G., O. Rosen and M. Tanner (2004), *Ecological inference. New methodological strategies*, Cambridge University Press, Cambridge, MA.

Kocherlakota, S. and K. Kocherlakota (1992), *Bivariate discrete distributions*, Marcel Dekker, New York.

Laird, N. (1985), Missing information principle, *Encyclopedia of Statistical Sciences* **5**, 548–552.

McCullagh, P. and J. A. Nelder (1992), *Generalized linear models* (2nd edn), Chapman and Hall, London.

Meng, X. L. and D. B. Rubin (1991), Using the EM algorithm to obtain asymptotic variance–covariance matrices: the SEM algorithm, *Journal of the American Statistical Association* **86**, 899–909.

Minka, Th. (2003), *Estimating a Dirichlet distribution* (available at: http://research.microsoft.com/∼minka/papers/dirichlet/; last accessed May 2008).
Morgan, B. J. T. (1992), *Analysis of quantal response data*, Chapman and Hall, London.
Mosimann, J. E. (1962), On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions, *Biometrika* **49**, 65–82.
Orchard, T. and M. A. Woodbury (1972), A missing information principle: theory and applications, *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability* **1**, 697–715.
Plackett, R. L. (1977), The marginal totals of a 2×2 table, *Biometrika* **64**, 37–42.
Poortema, K. (1999), On modelling overdispersion of counts, *Statistica Neerlandica* **53**, 5–20.
Prentice, R. L. (1986), Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors, *Journal of the American Statistical Association* **81**, 321–327.
Rosen, O., W. Jiang, G. King and M. Tanner (2001), Bayesian and frequentist inference for ecological inference: the $r \times c$ case, *Statistica Neerlandica* **55**, 134–156.
Steel, D. G., E. J. Beh and R. L. Chambers (2004), The information in aggregate data, in: G. King, O. Rosen and M. Tanner (eds), *Ecological inference. New methodological strategies*, Cambridge University Press, Cambridge, MA, pp. 51–68.
Wakefield, J. (2004), Ecological inference for 2×2 tables (with discussion), *Journal of the Royal Statistical Society, Series A* **167**, 385–445.
Woodbury, M. A. (1971), Discussion of H.O. Hartley and R. R. Hocking, The analysis of incomplete data, *Biometrics* **27**, 808–813.

Received: April 2008. Revised: May 2008.