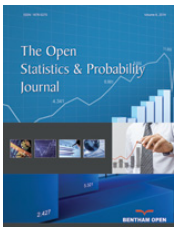
Send Orders for Reprints to reprints@benthamscience.ae

The Open Statistics & Probability Journal, 2017, *08*, 27-38

The Open Statistics & Probability Journal

Content list available at: www.benthamopen.com/TOSPJ/

DOI: 10.2174/1876527001708010027



RESEARCH ARTICLE

Bayesian Inference for Three Bivariate Beta Binomial Models

David Peter Michael Scollnik*

Department of Mathematics and Statistics, University of Calgary, Calgary, Canada

Received: June 02, 2017 Revised: August 16, 2017 Accepted: August 22, 2017

Abstract:

Background:

This paper considers three two-dimensional beta binomial models previously introduced in the literature. These were proposed as candidate models for modelling forms of correlated and overdispersed bivariate count data. However, the first model has a complicated form of bivariate probability mass function involving a generalized hypergeometric function and the remaining two do not have closed forms of probability mass functions and are not amenable to analysis using maximum likelihood. This limited their applicability.

Objective:

In this paper, we will discuss how the Bayesian analyses of these models may go forward using Markov chain Monte Carlo and data augmentation.

Results:

An illustrative example having to do with student achievement in two related university courses is included. Posterior and posterior predictive inferences and predictive information criteria are discussed.

Keywords: Bayesian, Bivariate beta binomial, Data augmentation, MCMC, Negative hypergeometric, OpenBUGS, Overdispersion.

1. INTRODUCTION

The univariate beta binomial model allows for extra binomial variation, *i.e.* overdispersion relative to the binomial model. It is constructed by taking a binomial model and assigning the binomial probability parameter p a beta distribution with parameters α and β. This model's probability mass function is

$$f(x|\alpha,\beta) = \binom{n}{x}\frac{B(\alpha+x,\beta+n-x)}{B(\alpha,\beta)},\ x=0,1,\ldots,n, \tag{1.1}$$

where $\alpha > 0$, $\beta > 0$, and B is the beta function. This beta binomial distribution is known by other names, including the negative hypergeometric distribution and the inverse hypergeometric distribution. A summary of the model's development and early use is given in Johnson, Kotz, and Kemp [1]. To cite a few examples, Skellam [2] applied the model to the association of chromosomes and to traffic clusters and Ishii and Hayakawa [3] used it as a model for the sex composition of families and for the absence of students.

Models that recognise overdispersion in the case of correlated bivariate count data are also naturally useful. The alternative of failing to recognise such overdispersion may lead to faulty statistical inferences and inaccurate conclusions due to an underestimation of the variability of the data. Some examples in the literature involving correlated overdispersed bivariate count data are Bibby and Væth [4], who examined counts of diseased second

* Address correspondence to this author at the Department of Mathematics and Statistics, University of, Calgary, 2500 University Drive NW, Calgary, Alberta, Canada, T2N 1N4; Tel: 001-403-220-5210; E-mail: scollnik@ucalgary.ca

premolars and second molars in Danish children's upper jaws, and Danahar and Hardie [5], who examined the number of bacon and eggs purchases made by households and also the joint readership of two magazines.

Bibby and Væth [4] developed a two-dimensional beta binomial distribution based on the two-dimensional beta distribution introduced in Jones [6], using a construction similar to that in Skellam [2] for the one-dimensional case. Properties of this distribution were given and its estimation and computational aspects were also discussed. A number of additional two-dimensional beta binomial models were also presented. However, these additional models did not have closed forms of probability mass functions and thus were not amenable to analysis using maximum likelihood. This limited their applicability. Such as, they were not considered in detail nor were they applied to any data.

In this paper, we will consider the Bayesian analysis of all three models appearing in Bibby and Væth [4], including the two with intractable forms. These models will be applied to a real data set describing student performances on tests in two related university courses. In Sections 2 and 3, the models under consideration will be presented. In Section 4, we discuss how a Bayesian analysis for any of these models may proceed on the basis of a Markov chain Monte Carlo (MCMC) set-up using data augmentation. The real data set will be analyzed in Section 5, where various approaches to model selection including the use of predictive information criteria will also be illustrated. OpenBUGS code associated with the numerical example appears in the Appendix.

2. THE TWO-DIMENSIONAL BETA BINOMIAL MODEL

Jones [6] introduced the two-dimensional beta distribution described below. Let W_0, W_1, and W_2 be mutually independent random variables such that $W_i \sim \chi^2(2\nu_i)$, *i.e.* $\Gamma(\nu_i, 1)$, for $\nu_i > 0$. Now define

$$B_i = \frac{W_i}{W_i + W_0}, \; i = 1, 2. \tag{2.2}$$

Then the resulting B_i is Beta(ν_i, ν_0)-distributed, for $i = 1, 2$. The two-dimensional beta distribution describes the joint distribution of B_1 and B_2, and its probability density function is given by $\Gamma(\nu)$

$$f_B(b_1, b_2) = \frac{\Gamma(\nu)}{\Gamma(\nu_1)\,\Gamma(\nu_2)\,\Gamma(\nu_0)} \frac{b_1^{\nu_1 - 1}(1 - b_1)^{\nu_2 + \nu_0 - 1} b_2^{\nu_2 - 1}(1 - b_2)^{\nu_1 + \nu_0 - 1}}{(1 - b_1 b_2)^{\nu}} \tag{2.3}$$

for $0 < b_1 < 1$, $0 < b_2 < 1$, and $\nu = \nu_1 + \nu_2 + \nu_0$.

Bibby and Væth [1] developed a two-dimensional beta binomial distribution in terms of the two-dimensional beta distribution in a manner analogous to the one dimensional case.

Let $p = (p_1, p_2)$ be a two-dimensional beta random variable with parameters ν_1, ν_2, and ν_0. Given p, let X_1 and X_2 be two independent binomially distributed random variables, with $X_i | \; p \sim \text{bin}(n_i, p_i)$ for $i = 1, 2$. Then the joint distribution of $X = (X_1, X_2)$ is known as the two-dimensional beta binomial distribution and its probability mass function is given by

$$\begin{aligned} f_X(x_1, x_2) &= \binom{n_1}{x_1}\binom{n_2}{x_2} \frac{\Gamma(\nu)}{\Gamma(\nu_1)\,\Gamma(\nu_2)\,\Gamma(\nu_0)} \\ &\times \frac{\Gamma(x_1 + \nu_1)\,\Gamma(n_1 - x_1 + \nu - \nu_1)}{\Gamma(n_1 + \nu)} \frac{\Gamma(x_2 + \nu_2)\,\Gamma(n_2 - x_2 + \nu - \nu_2)}{\Gamma(n_2 + \nu)} \\ &\times {}_3F_2(\nu, x_1 + \nu_1, x_2 + \nu_2; n_1 + \nu, n_2 + \nu; 1) \end{aligned} \tag{2.4}$$

for $x_1 = 0, 1, \ldots, n_1$ and $x_2 = 0, 1, \ldots, n_2$. This bivariate distribution will be denoted as

$$(X_1, X_2) \sim \text{biv beta binomial}(n_1, n_2; \nu_1, \nu_2, \nu_0). \tag{2.5}$$

The form of the probability mass function given above in (2.4) is significantly complicated by the presence of a

generalized hypergeometric function, denoted by ${}_3F_2$. The definition of this generalized hypergeometric function is given by

$$ {}_3F_2(a_1, a_2, a_3; b_1, b_2; z) = \sum_{k=0}^{\infty} \frac{(a_1)_k (a_2)_k (a_3)_k}{(b_1)_k (b_2)_k} \frac{z^k}{k!} \tag{2.6}$$

where the $(a)_k$ are the Pochhammer symbols

$$(a)_k = \frac{\Gamma(a+k)}{\Gamma(a)} = a(a+1)\cdots(a+k-1)$$

with $(a)_0 = 1$. The series in (2.6) is convergent at the point $z = 1$ if and only if $b_1 + b_2 > a_1 + a_2 + a_3$. This convergence condition is always met in the context of the probability mass function in (2.4). See Bailey [7] for additional details. Bibby and Væth [4] observed that the calculation of the generalized hypergeometric function at the argument 1 is numerically unstable. This presents some problems when estimating the parameters of the two-dimensional beta binomial model using maximum likelihood. The Bayesian method presented later in this paper does not suffer from the same problems, as it circumvents the calculation of the generalized hypergeometric function entirely.

The two-dimensional beta binomial distribution is such that its marginal distributions are univariate beta binomial, that is

$$X_i \sim \text{beta binomial}(n_i; \nu_i, \nu_0) \tag{2.7}$$

for $i = 1, 2$. The marginal mean and variance are given by

$$E(X_i) = \frac{n_i \nu_i}{\nu_i + \nu_0} \tag{2.8}$$

$$Var(X_i) = \frac{n_i \nu_i \nu_0 (n_i + \nu_i + \nu_0)}{(\nu_i + \nu_o)^2 (\nu_i + \nu_0 + 1)}. \tag{2.9}$$

From Bibby and Væth [1], the correlation between the two marginals is given by

$$\begin{aligned} Corr(X_1, X_2) &= \sqrt{\frac{n_1 n_2 \nu_1 \nu_2 (\nu_1 + \nu_0 + 1)(\nu_2 + \nu_0 + 1)}{\nu_0^2 (n_1 + \nu_1 + \nu_0)(n_2 + \nu_2 + \nu_0)}} \\ &\times \{{}_3F_2(1, 1, \nu_0; \nu_1 + \nu_0 + 1, \nu_2 + \nu_0 + 1; 1) - 1\}, \end{aligned} \tag{2.10}$$

and this correlation is always positive with a strictly positive lower bound, that is

$$Corr(X_1, X_2) > \sqrt{\frac{n_1 n_2 \nu_1 \nu_2}{(n_1 + \nu_1 + \nu_0)(\nu_1 + \nu_0 + 1)(n_2 + \nu_2 + \nu_0)(\nu_2 + \nu_0 + 1)}}. \tag{2.11}$$

3. TWO ADDITIONAL TWO-DIMENSIONAL BETA BINOMIAL MODELS

Several other two-dimensional beta binomial distributions were briefly considered in Bibby and Væth [4], two of which are of interest here and will be reviewed below. Whereas the two- dimensional beta binomial distribution of the last section has a positive correlation between the marginals that is positive and bounded away from zero, the models introduced in this section include independent beta binomial marginal distributions as special cases.

The second two-dimensional beta binomial model replaces B_1 and B_2 in (2.2) with

$$B_1 = \frac{U_1}{U_1 + V_1} \tag{3.12}$$

$$B_2 = \frac{U_2}{U_2 + \theta V_1 + (1-\theta)V_2} \tag{3.13}$$

where $U_i \sim \Gamma(v_i, 1)$ and $V_i \sim \Gamma(v_0, 1)$, for $i = 1, 2$, are all mutually independent. Let $p = (p_1, p_2)$ be a two-dimensional random variable defined in accordance with (3.12) and (3.13). Given p, let X_1 and X_2 be two independent binomially distributed random variables as before, with $X_i | p \sim \text{bin}(n_i, p_i)$ for $i = 1, 2$. Then the resulting model for the joint distribution of $X = (X_1, X_2)$ includes the two-dimensional beta binomial distribution (2.5) as a special case when $\theta = 1$, and two independent univariate beta binomial distributions as another when $\theta = 0$. As yet, there appears to be no closed form expression available for this model's probability mass function and it may be that one does not exist.

The third of the two-dimensional beta binomial models uses

$$B_1 = \frac{U_1}{U_1 + V_1 + W} \tag{3.14}$$

$$B_2 = \frac{U_2}{U_2 + V_2 + W} \tag{3.15}$$

where $U_i \sim \Gamma(\mu_i, 1)$ and $V_i \sim \Gamma(v_i, 1)$, for $i = 1, 2$, $W \sim \Gamma(\omega, 1)$, and U_i, V_i, and W are all mutually independent. In this case, the resulting B_i random variables are Beta(μ_i, $v_i + \omega$)- distributed, for $i = 1, 2$. Let $p = (p_1, p_2)$ be defined in accordance with (3.14) and (3.15) and let $X_i | p \sim \text{bin}(n_i, p_i)$, for $i = 1, 2$. Then the resulting model for the joint distribution of $X = (X_1, X_2)$ includes the product of two independent beta binomial distributions as a limiting case when $\omega \to 0$. However, the joint density of B_1 and B_2 involves an integral of a product of two confluent hypergeometric functions leading to an intractable joint probability mass function for (X_1, X_2). Note, the third model's construction bears some similarity to that of the previous two models and so it might casually be described as an extension of the first or second models. However, technically speaking the first and second models are not special cases of (*i.e.* are not nested within) the third.

As noted, the second and third models were introduced and briefly discussed in Bibby and Væth [4]. However, neither model was used in the context of a numerical illustration. Indeed, these additional models do not have closed forms of probability mass functions and are not amenable to analysis using maximum likelihood. However, the Bayesian estimation method presented in the next section for the two-dimensional beta binomial distribution can be easily modified and applied.

4. BAYESIAN ESTIMATION VIA MCMC AND DATA AUGMENTATION

Estimation of the parameters appearing in the two-dimensional beta binomial model (2.5) using maximum likelihood is complicated by the presence of the generalized hypergeometric function in that model's probability mass function. The additional two models presented in the last section pose even greater challenges for this estimation method. A Bayesian estimation method using MCMC and data augmentation can circumvent these challenges.

Consider a random sample of size N from a two-dimensional beta binomial distribution. Denote this sample as $X = (X_1, X_2, \ldots, X_N)$, where $X_i = (X_{i1}, X_{i2})$ is distributed according to the distribution in (2.4). The probability model

$$(X_{i1}, X_{i2}) \sim \text{biv beta binomial}(n_1, n_2; \nu_1, \nu_2, \nu_0) \tag{4.16}$$

for $i = 1, \ldots, N$ and $j = 1, 2$, can be represented with the use of latent variables P and W as

$$X_{ij} \sim \text{binomial}(n_j; P_{ij}) \tag{4.17}$$

$$P_{ij} \leftarrow \frac{W_{ij}}{W_{ij} + W_{i0}} \tag{4.18}$$

$$W_{ik} \sim \chi^2(2\nu_k) \tag{4.19}$$

for $i = 1, \ldots, N, j = 1, 2$, and $k = 0, 1, 2$, in accordance with the development presented in Section 2.

It remains to assign a prior density specification to the unknown parameters v_k , $k = 0, 1, 2$, and then implement a MCMC analysis (given the observed data) of the resulting full probability model based on (4.17) to (4.19). This may be accomplished with the assistance of any of a number of statistical computing packages presently available to implement Gibbs sampling or other more advanced forms of MCMC. For example in this paper, the OpenBUGS package was used. The OpenBUGS package is available at www.openbugs.net.

Illustrative OpenBUGS code corresponding to the example appearing in the next section is provided in the Appendix. With suitable elementary adjustments to the code, the two additional models presented in Section 3 can be analyzed in a likely manner. Some of the foundational papers on Gibbs sampling include Geman and Geman [8] and Gelfand and Smith [9]. Any readers requiring information on how to implement, monitor, and analyse the results of a MCMC simulation are directed to Congdon [10], Gelman *et al.* [11], and the user manual that accompanies (*i.e.* accessible within) OpenBUGS. Manuals and additional resources are also available at www.openbugs.net/w/Documentation.

5. EXAMPLE

This example concerns the number of tests passed by 52 students in each of two distinct but related actuarial science courses at the University of Calgary in the same academic year. The two courses had the same prerequisites and the same instructor. We have applied the three models discussed in this paper to the data. The data are provided in Table **1**.

Table 1. The number of students in each combination of the number of tests passed in two related university courses in actuarial science.

		First course					Total
		0	**1**	**2**	**3**	**4**	
Second course	0	0	0	2	2	0	4
	1	0	0	2	2	1	5
	2	0	0	2	4	6	12
	3	0	0	0	5	26	31
	Total	0	0	6	13	33	52

For each model, a Bayesian method of analysis was adopted and implemented as described in the preceding section. The analysis for each model was performed separately and independent prior densities were assigned to each unknown (top-level) model parameter. Each of the model parameters v_i, μ_i, and ω appearing in any given model was assigned an exponential prior distribution with a common mean of 10. This prior assigns a little over 63.2% and 86.4% of its mass to values below 10 and 20, respectively, and treats an interval of smaller values as being more probable than an interval of larger values with the same interval width, *a priori*. Although this prior is not noninformative, it is relatively diffuse over a large range of parameter values (*e.g.*, values less than 20) containing what we consider to be a reasonable and *a priori* probable subrange (*e.g.*, values less than 10). The parameter θ appearing in the first extended model was assigned a uniform prior on the interval (0, 1). For the problem under consideration the overall choice of prior density specification described above seemed reasonable. However, some other prior density specifications were also tried in conjunction with these models. Although the resulting posterior inferences changed slightly, the relative ranking of the models in terms of their fits and predictive performances (as discussed below) was essentially unchanged.

The MCMC based analysis in OpenBUGS used 4 chains, each was burned for 150,000 iterations, and then allowed to run for an additional 100,000 thus yielding 400,000 kept iterations. Tables **2** to **4** report the posterior means, standard deviations, 95% (symmetric) posterior density intervals, and 95% highest (*i.e.* shortest) posterior density (HPD) intervals for the top-level model parameters assuming the prior density specification mentioned earlier. These posterior summaries are primarily reported for completeness, and the values of the summaries in one table (*e.g.* for v_1 and v_2) are not meant to be compared to those in another. We note that the estimated posterior densities were all unimodal and skewed to varying degrees. The skewness is revealed by comparing a parameter's symmetric posterior density interval to its HPD interval. If a parameter's posterior density was symmetric and unimodal, these two intervals would be the same.

Table 2. Posterior summaries for parameters in the first model.

Parameter	Mean	Std. Dev.	95% Interval	95% HPD Interval
v_1	5.716	2.526	(2.383, 11.98)	(1.843, 10.75)
v_2	2.728	1.201	(1.175, 5.683)	(0.937, 5.105)
v_0	0.707	0.295	(0.323, 1.433)	(0.259, 1.291)

Table 3. Posterior summaries for parameters in the second model.

Parameter	Mean	Std. Dev.	95% Interval	95% HPD Interval
v_1	5.879	3.073	(2.239, 13.52)	(1.695, 11.61)
v_2	2.919	1.477	(1.208, 6.617)	(0.939, 5.66)
v_0	0.719	0.3703	(0.289, 1.625)	(0.221, 1.39)
θ	0.862	0.140	(0.471, 0.997)	(0.572, 1.00)

Table 4. Posterior summaries for parameters in the third model.

Parameter	Mean	Std. Dev.	95% Interval	95% HPD Interval
μ_1	8.594	5.350	(2.876, 23.18)	(2.07, 20.79)
μ_2	3.977	2.247	(1.455, 9.722)	(0.988, 8.756)
v_1	0.304	0.606	(0.017, 2.020)	(0.011, 1.644)
v_2	0.257	0.340	(0.035, 1.146)	(0.0165, 0.925)
ω	0.810	0.430	(0.274, 1.907)	(0.163, 1.722)

Recall that the data set contains 52 observations. As part of the MCMC based analysis, a predictive replicated sample of size 52 was generated from the model at each iteration. Summaries of these predictive replicated samples are reported in Tables **5** to **7**. Specifically, these tables report the estimated predicted expected value and standard deviation of the number of students (assuming a cohort size of 52 students) in each combination of the number of tests passed in the two related actuarial science courses, under each model. One may observe that each model's predicted row and column totals are (for the most part) generally in agreement with those for the original data in Table **1**.

Table 5. The predicted expected value (and standard deviation) of the number of students in each combination of the number of tests passed in the two related courses under a Bayesian analysis using the first model.

		First course					
		0	1	2	3	4	Total
Second course	0	0.21 (0.49)	0.45 (0.70)	0.75 (0.90)	1.09 (1.10)	1.32 (1.31)	3.82 (2.42)
	1	0.18 (0.44)	0.47 (0.70)	0.94 (0.98)	1.66 (1.32)	2.48 (1.66)	5.74 (2.48)
	2	0.15 (0.40)	0.44 (0.68)	1.07 (1.07)	2.48 (1.66)	5.68 (2.54)	9.82 (3.37)
	3	0.09 (0.31)	0.31 (0.57)	0.96 (1.03)	3.44 (2.00)	27.83 (4.94)	32.62 (4.6)
	Total	0.64 (0.91)	1.66 (1.41)	3.73 (2.03)	8.67 (3.15)	37.30 (4.13)	52

Table 6. The predicted expected value (and standard deviation) of the number of students in each combination of the number of tests passed in the two related courses under a Bayesian analysis using the second model.

		First course					
		0	1	2	3	4	Total
Second course	0	0.05 (0.23)	0.18 (0.44)	0.40 (0.66)	0.68 (0.87)	0.84 (1.02)	2.14 (1.77)
	1	0.08 (0.29)	0.33 (0.59)	0.89 (0.97)	1.88 (1.42)	2.97 (1.87)	6.14 (2.67)
	2	0.08 (0.29)	0.39 (0.65)	1.33 (1.20)	3.66 (2.01)	8.51 (3.01)	13.97 (3.73)
	3	0.05 (0.23)	0.28 (0.56)	1.23 (1.18)	4.84 (2.36)	23.34 (5.05)	29.75 (4.84)
	Total	0.26 (0.56)	1.18 (1.22)	3.85 (2.14)	11.06 (3.49)	35.65 (4.52)	52

Table 7. The predicted expected value (and standard deviation) of the number of students in each combination of the number of tests passed in the two related courses under a Bayesian analysis using the third model.

		First course					
		0	1	2	3	4	Total
Second course	0	0.03 (0.18)	0.12 (0.37)	0.33 (0.60)	0.65 (0.86)	0.89 (1.06)	2.03 (1.73)
	1	0.05 (0.23)	0.26 (0.53)	0.84 (0.95)	1.99 (1.46)	3.29 (1.96)	6.44 (2.74)
	2	0.06 (0.25)	0.35 (0.61)	1.36 (1.22)	4.02 (2.10)	8.80 (3.07)	14.59 (3.82)
	3	0.04 (0.21)	0.31 (0.58)	1.47 (1.31)	5.74 (2.63)	21.38 (4.95)	28.94 (4.86)
	Total	0.18 (0.46)	1.05 (1.14)	4.00 (2.20)	12.41 (3.68)	34.36 (4.64)	52

Table **8** summarizes the means and standard deviations of the number of tests passed in each course, along with the correlation between the two. The first rows of numerical values in this table are the empirical values associated with the observed data. These are followed by the Bayesian predictive values associated with the three models under consideration. In this example, it appears that while all three models generally replicate the empirical means and standard deviations, the first model does a much better job at replicating the correlation than either the second or third.

Table 8. The means and standard deviations of the number of tests passed in each of the two courses, along with the correlation between the two courses. The first line gives the observed values corresponding to Table 1; the estimated predictive values associated with the first, second, and third models follow.

	First course		**Second course**		**Correlation**
	Mean	Std. Dev.	Mean	Std. Dev.	
Observed	3.519	0.699	2.346	0.947	0.667
First model	3.545	0.855	2.370	0.946	0.412
Second model	3.551	0.769	2.371	0.847	0.284
Third model	3.533	0.752	2.355	0.843	0.246

When the goal is to pick a model with the best out-of-sample predictive power then selection can be made on the basis of the deviance information criterion (DIC), which is a combined measure of goodness of fit and model complexity. The DIC and its calculation are discussed in Spiegelhalter *et al.* [12]. See also Chapter 7 in Gelman *et al.* [11]. The DIC is implemented in OpenBUGS and according to the OpenBUGS User Manual, "the model with the smallest DIC is estimated to be the model that would best predict a replicate dataset of the same structure as that currently observed". The values of the DIC corresponding to the models represented in Tables **5** to **7** are 162.9, 164.7, and 169.1, respectively. So based on this criterion, the first of the three models under consideration is the preferred model.

Although the DIC is conveniently incorporated in OpenBUGS, it is not without its problems. For instance, it can produce negative estimates of the effective number of parameters in its evaluation of a model's complexity and it is not defined for singular models. See Celeux *et al.* [13], Gelman *et al.* [14], Plummer [15], and Spiegelhalter *et al.* [16]. The WAIC (Watanabe- Akaike or widely applicable information criterion; Watanabe [17]) is another measure of a model's predictive accuracy. This criterion is fully Bayesian and works for singular models and may be viewed as an improvement on the DIC (however, the WAIC is not without its own difficulties; see Gelman *et al.* [14]). Unfortunately, the WAIC is not directly calculated by OpenBUGS, but it is not too difficult to evaluate it using output from OpenBUGS. We did so, using the definition for the version of WAIC found in Gelman *et al.* [14] The values of the WAIC corresponding to the models represented in Tables **5** to **7** are 173.5, 177.4, and 185.7, respectively. Therefore, of the models under consideration, the first again exhibits the best predictive accuracy for the data in Table **1**, this time according to the WAIC. We note that other predictive information criteria for Bayesian models do exist, several of which are reviewed by Gelman *et al.* [14].

Another approach to checking model fit involves focusing attention on a particular test quantity or discrepancy measure of interest. This test quantity may be a function of the known and unknown parameters as well as of the data. Let such a test quantity be denoted as $T(D, \varphi)$, with D denoting the data and φ the model parameters. If D^{pred} denotes a predicted replicated data set generated from the model, then the predictive Bayesian p-value is defined as

$$\Pr(\, T(D^{pred}, \phi) \geq T(D, \phi) \mid D \,), \tag{5.20}$$

where the probability is taken over the posterior distribution of φ and the posterior predictive distribution of D^{pred}. Extreme values of $T(D^{pred}, \varphi)$ relative to $T(D, \varphi)$ are evidence of discrepancy between the model and the data. Thus, a Bayesian *p*-value near 0 or 1 provides evidence of model discrepancy. See Gelman *et al.* [11] for a more detailed discussion of Bayesian *p*-values.

In the present context of an example involving correlated bivariate data, a sensible and meaningful discrepancy measure is the sample correlation. In the original data set, the sample correlation $r = T(D, \varphi)$ is equal to 0.667. Recall, as part of our MCMC based analysis, a predictive replicated sample of size 52 was generated from the model under study at each iteration. For each model, we monitored the posterior predictive distribution of the sample correlation for the replicated data, *i.e.* $r^{pred} = T(D^{pred}, \varphi)$, and monitored its relation to the sample correlation of the original data. The results are presented in Table **9**, and lend further support to the conclusion that the first of the three models under consideration is a better fit to the data than either of the two extended models.

Table 9. Posterior predictive summaries for the sample correlation coefficient r^{pred} for the replicated data associated with each of the two-dimensional beta binomial models.

Model	Mean	Std. Dev.	95% Interval	Bayesian *p*-value
First model	0.412	0.152	(0.097, 0.688)	0.037
Second model	0.284	0.163	(−0.043, 0.589)	0.005
Third model	0.246	0.159	(−0.070, 0.547)	0.002

One final check of predictive model performance was performed. For each predictive replicated sample, the sum of squared deviations between the predicted and observed cell counts was calculated over the original non-empty cells. Denote this statistic as SS^{pred}. When comparing models, smaller values of this statistic are indicative of a better fit. The estimated posterior predictive summaries for SS^{pred} are reported in Table **10**. Once again, the first of the three models under consideration comes out on top.

Table 10. Posterior predictive summaries for SS^{pred} for the replicated data asso- ciated with each of the two-dimensional beta binomial models.

Model	Mean	Std. Dev.	95% Interval
First model	61.11	43.92	(15, 182)
Second model	76.56	58.89	(17, 238)
Third model	96.1	73.03	(19, 291)

CONCLUSION

This paper considered three two-dimensional beta binomial models. Two of these models do not have closed forms of probability mass functions and are not amenable to analysis using maximum likelihood. Instead, a Bayesian analysis of each model was implemented using MCMC with data augmentation.

In the example contained within this paper, the first of the two-dimensional beta binomial models was the best performing model of the three considered. Of course, this does not necessarily mean that it will always perform better than the other two. However, as yet we have not run across an actual data set for which the first model did not perform at least as well as one of the others.

APPENDIX

This BUGS code can be used with OpenBUGS to implement a Bayesian analysis of the two- dimensional beta binomial model presented in Bibby and Væth [4]. The manner of variable indexing and data formatting used in this code is such that tables containing cells with zero counts may be conveniently analysed. Only cells with non-zero counts are read in as data.

```
model{
   # Define the model including the prior.
   for( cell in 1:cells ) {
      for( k in 1:N[cell,3] ) {
         for( l in 1:2 ) {
            x[cell,k,l] <- N[cell,l] - 1
            x[cell,k,l] ~ dbin(p[cell,k,l],n[l])
            p[cell,k,l] <- w[cell,k,l] / (w[cell,k,l] + w[cell,k,3])
         }
         for( l in 1:3 ) {
            w[cell,k,l] ~ dchisqr(df.w[l])
         }
      }
   }
   for( l in 1:3 ) {
      df.w[l] <- 2 * nu.w[l]
      nu.w[l] ~ dexp(0.1)
   }

# Replicated sample for posterior predictive inference.
for( rep in 1:nobs ) {
   for( k in 1:2 ) {
      xp[rep,k] ~ dbin(pp[rep,k],n[k])
      pp[rep,k] <- wp[rep,k] / (wp[rep,k] + wp[rep,3])
   }
   for( l in 1:3 ) {
      wp[rep,l] ~ dchisqr(nu.w[l])
   }
}
for( i in 1:n[1] + 1 ) {
```

```
   for( j in 1:n[2] + 1 ) {
      for( rep in 1:nobs ) {
         np[rep,i,j] <- equals(i,xp[rep,1] + 1) * equals(j,xp[rep,2] + 1)
      }
      nc[i,j] <- sum(np[1:nobs,i,j])
   }
}
for( i in 1:n[1] + 1 ) {
   col[i] <- sum(nc[i,1:n[2] + 1])
}
for( j in 1:n[2] + 1 ) {
   row[j] <- sum(nc[1:n[1] + 1,j])
}

# Posterior predictive model check.
r.obs <- 0.6672033

r.pred <- ( inprod( xp[,1],xp[,2] ) -
          nobs * mean( xp[,1] ) * mean( xp[,2] ) ) /
          ( ( nobs - 1 ) * sd( xp[,1] ) * sd( xp[,2] ) )
post.pred <- step( r.pred - r.obs )

# Another posterior predictive model check statistic.
for( cell in 1:cells ) {
   nc2[cell] <- pow(N[cell,3] - ncsub[cell],2)
   ncsub[cell] <- nc[N[cell,1],N[cell,2]]
}
# nc2[cells + 1] is referred to as SS.pred in the main body of the paper.
nc2[cells + 1] <- sum(nc2[1:cells])
ncsub[cells + 1] <- sum(ncsub[1:cells])
```

```
    nc2[cells + 2] <- nc2[cells + 1] + pow(nobs - ncsub[cells + 1],2)
}


list( cells = 10, n = c(4,3), nobs = 52, N=structure(
.Data = c(
3, 1, 2,
4, 1, 2,
3, 2, 2,
4, 2, 2,
5, 2, 1,
3, 3, 2,
4, 3, 4,
5, 3, 6,
4, 4, 5,
5, 4, 26
),
.Dim=c(10,3)
)
)
```

CONSENT FOR PUBLICATION

Not applicable.

CONFLICT OF INTEREST

The author declares no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

This research was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

[1] N.L. Johnson, A.W. Kemp, and S. Kotz, *Univariate Discrete Distributions.,* 3rd ed John Wiley & Sons Inc.: New York, 2005. [http://dx.doi.org/10.1002/0471715816]

[2] J.G. Skellam, "A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials", *J. R. Stat. Soc. B,* vol. 10, pp. 257-261, 1948.

[3] G. Ishii, and R. Hayakawa, "On the compound binomial distribution", *Annals of the Institute of Statistical Mathematics, Tokyo,* vol. 12, pp.

69-80, 1960.
[http://dx.doi.org/10.1007/BF01577666]

[4] B.M. Bibby, and M. Væth, "The two-dimensional beta binomial distribution", *Stat. Probab. Lett.,* vol. 81, pp. 884-891, 2011.
[http://dx.doi.org/10.1016/j.spl.2010.12.019]

[5] P.J. Danaher, and B.G. Hardie, "Bacon with your eggs? Applications of a new bivariate beta-binomial distribution", *Am. Stat.,* vol. 59, no. 4, pp. 282-286, 2005.
[http://dx.doi.org/10.1198/000313005X70939]

[6] M. Jones, "Multivariate t and beta distributions associated with the multivariate t dis- tribution", *Metrika,* vol. 54, pp. 215-231, 2001.
[http://dx.doi.org/10.1007/s184-002-8365-4]

[7] W.N. Bailey, "Generalized Hypergeometric Series", In: *Cambridge Tracts in Mathematics and Mathematical Physics,* vol. 32. Cambridge University Press, 1935.

[8] S. Geman, and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images", *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 6, no. 6, pp. 721-741, 1984.
[http://dx.doi.org/10.1109/TPAMI.1984.4767596] [PMID: 22499653]

[9] A.E. Gelfand, and A.F. Smith, "Sampling-based approaches to calculating marginal densities", *J. Am. Stat. Assoc.,* vol. 85, no. 410, pp. 398-409, 1990.
[http://dx.doi.org/10.1080/01621459.1990.10476213]

[10] P. Congdon, *Bayesian Statistical Modelling.,* 2nd ed John Wiley & Sons Ltd: West Sussex, 2006.
[http://dx.doi.org/10.1002/9780470035948]

[11] *A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, D.B. Rubin, Bayesian Data Analysis.,* 3rd ed Chapman & Hall / CRC: New York, 2014.

[12] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. van der Linde, "Bayesian measures of model complexity and fit (with discussion)", *J. R. Stat. Soc. B,* vol. 64, pp. 583-640, 2002.
[http://dx.doi.org/10.1111/1467-9868.00353]

[13] G. Celeux, F. Forbes, C. Robert, and D. Titterington, "Deviance information criteria for missing data models", *Bayesian Anal.,* vol. 1, pp. 651-706, 2006.
[http://dx.doi.org/10.1214/06-BA122]

[14] A. Gelman, J. Hwang, and A. Vehtari, "Understanding predictive information criteria for Bayesian models", *Stat. Comput.,* vol. 24, pp. 997-1016, 2014.
[http://dx.doi.org/10.1007/s11222-013-9416-2]

[15] M. Plummer, "Penalized loss functions for Bayesian model comparison", *Biostatistics,* vol. 9, no. 3, pp. 523-539, 2008.
[http://dx.doi.org/10.1093/biostatistics/kxm049] [PMID: 18209015]

[16] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. van der Linde, "The deviance information criterion: 12 years on", *J. R. Stat. Soc. B,* vol. 76, pp. 485-493, 2014.
[http://dx.doi.org/10.1111/rssb.12062]

[17] S. Watanabe, "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory", *J. Mach. Learn. Res.,* vol. 11, pp. 3571-3591, 2010.

© 2017 David Peter Michael Scollnik.

This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Public License (CC-BY 4.0), a copy of which is available at: https://creativecommons.org/licenses/by/4.0/legalcode. This license permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.