UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Extra-Binomial Variability: A Bayesian Approach

Jorge Alberto Achcar
Juliano José Guimarães Junqueira

Nº 63

NOTAS

São Carlos - SP

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação
ISSN 0103-2577

Extra-Binomial Variability: A Bayesian Approach

Jorge Alberto Achcar
Juliano José Guimarães Junqueira

Nº 63

NOTAS

Série Estatística

São Carlos – SP
Out./2000

EXTRA-BINOMIAL VARIABILITY: A BAYESIAN APPROACH

Jorge Alberto Achcar
Juliano José Guimarães Junqueira

Depto (Estatística)
ICMC / USP – São Carlos
Caixa Postal 668
13560-970, São Carlos - SP, Brazil

Abstract

In this paper, we present a Bayesian analysis for extra-Binomial variability models introduced in the literature. We also consider the use of mixture distributions to model extra-Binomial variability and the introduction of covariates. Considering Gibbs Sampling with Metropolis-Hastings algorithms, we obtain Monte Carlo estimates for the posterior quantities of interest. The methodology is illustrated in two real data sets.

Keywords and phases: extra-Binomial variability, Bayesian analysis, MCMC methods.

1. Introduction

In many applications of binary data, the usual assumption of Binomial Distribution for the counts of sucesses y out of the number of total trials n, the observed variation is greater than expected under the ordinary Binomial assumption. This variation is called extra-Binomial variation (see for example, Skellan, 1948; Altham, 1978; Rudolfer, 1990; or Kupper and Haseman, 1978).

Let Y_i, $i = 1,...,N$ be a random variable denoting observed counts of success associated with $n_i - y_i$ failures. The simple Binomial distribution assumes,

$$P(Y_i = y_i / n_i, p_i) = \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i} \tag{1}$$

with mean $n_i p_i$ and variance $n_i p_i (1 - p_i)$.

Different forms to model extra-Binomial variation are introduced in the literature. Considering continuous random variation, we could assume that the extra-Binomial variability comes from the probability distribution on the probability p_i.

In this case (see for example, Griffiths, 1973; or Crowder, 1978), we assume a conjugate Beta distribution for p_i with parameters α e β,

$$g(p_i / \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p_i^{\alpha - 1} (1 - p_i)^{\beta - 1} \tag{2}$$

Thus, the probability density function for Y_i given n_i, α and β is a Beta-Binomial distribution,

$$P(Y_i = y_i / n_i, \alpha, \beta) = \binom{n_i}{y_i} \frac{B(y_i + \alpha, n_i + \beta - y_i)}{B(\alpha, \beta)} \tag{3}$$

where $B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and Y_i has mean $n_i\xi$ and variance $\frac{n_i\xi(1-\xi)(n_i+\omega^{-1})}{(1+\omega^{-1})}$

with $\xi=\alpha(\alpha+\beta)^{-1}$ and $\omega=(\alpha+\beta)^{-1}$.

The parametrization ξ and ω has been used by many authors (see for example, Willians, 1982; Tarone, 1979; or Hsiao, 1994).

In parametrization ξ and ω, the probability distribution for Y_i given n_i,α and β is,

$$P(Y_i=y_i/n_i,\alpha,\beta)=\binom{n_i}{y_i}\frac{\prod_{j=0}^{y_i-1}(\xi+\omega j)\prod_{j=0}^{n_i+y_i-1}(1-\xi+\omega j)}{\prod_{j=0}^{n_i-1}(1+\omega j)} \tag{4}$$

Observe that when $\omega\to 0$, (4) tends to the Binomial distribution.

Other form to model extra-Binomial variation is to assume the correlated-Binomial model (see Altham, 1978; or Kupper and Haseman, 1978) in which the extra-variation comes from the positive or negative pairwise correlation, called intra-class correlation among binary variables $U_{i1},.....,U_{in_i}$ of Y_i where $Y_i=U_{i1}+U_{i2}+....+U_{in_i}$.

That is,

$$U_{ij}/p\sim\text{Bernoulli}(p) \tag{5}$$

and the correlation δ, where δ is assumed homogeneous across $Y_1,.....,Y_N$ is

$$\delta=corr(U_{ij},U_{ik})=\frac{P(U_{ij}=1,U_{ik}=1)-p^2}{p(1-p)} \tag{6}$$

where $j\neq k$; $j,k=1,...,n_i$ and $\frac{-1}{n_i-1}<\delta<1$.

Thus,

$$P(Y_i = y_i / n_i, \delta, p) = \binom{n_i}{y_i} p^{y_i}(1-p)^{n_i - y_i}\left\{1 + \frac{\delta}{2p(1-p)}\left[(y_i - n_i p)^2 + y_i(2p-1) - n_i p^2\right]\right\} \quad (7)$$

with mean $n_i p$ and variance $n_i p(1-p)[1+(n_i-1)\delta]$.

Considering discrete random variation, we could assume that the random variable Y_i has a finite mixture of Binomial Distributions,

$$P\left(Y_i = y_i / n_i, \underset{\sim}{\theta}, \underset{\sim}{\lambda}\right) = \sum_{j=1}^{J} \lambda_j b(n_i, \theta_j) \quad (8)$$

where $b(n_i, \theta_j)$ denotes a Binomial distribution with mean $n_i\theta_j$ and variance $n_i\theta_j(1-\theta_j)$, $\underset{\sim}{\theta} = (\theta_1, ..., \theta_J)$, $\underset{\sim}{\lambda} = (\lambda_1, ..., \lambda_J)$ and $\sum_{j=1}^{J} \lambda_j = 1$.

Under some conditions in the parameters of model (8), we have identifiability of the parameters (see for example, Titterington, Smith and Markov, 1985).

In the presence of covariates, we could consider logistic regression for modeling.

In this paper, assuming the presence or not of covariates, we consider a Bayesian approach to analyse data with extra-Binomial variability using Markov Chain Monte Carlo (MCMC) methods as the Gibbs Sampling algorithm (see for example, Gelfand and Smith, 1990) or the Metropolis-Hastings algorithm (see for example, Smith and Roberts, 1993).

We also consider some existing Bayesian criteria to discriminate the proposed models.

2. A Bayesian Analysis for the Extra-Binomial Variability Models

Assuming the Beta-Binomial model (3), let us consider the prior distributions for α and β given by,

$$\begin{aligned} &\alpha \sim \Gamma(a_1, b_1); \quad a_1, b_1 \text{ Known}, \\ &\beta \sim \Gamma(a_2, b_2); \quad a_2, b_2 \text{ Known}, \end{aligned} \quad (9)$$

where $\Gamma(a,b)$ denotes a gamma distribution with mean a/b and variance a/b^2.

Assuming prior independence among the parameters, the joint prior distribution for α and β is given by

$$\pi(\alpha,\beta) \propto \alpha^{a_1-1}e^{-b_1\alpha}\beta^{a_2-1}e^{-b_2\beta} \tag{10}$$

where $\alpha > 0$ and $\beta > 0$.

The conditional probability function for Y_i given n_i, a_1, b_1, a_2 and b_2 is given by,

$$P(Y_i = y_i / n_i, a_1, b_1, a_2, b_2) = c\int_0^\infty \int_0^\infty \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_i)\Gamma(\beta+n_i-y_i)}{\Gamma(\alpha+\beta+n_i)} \alpha^{a_1-1}e^{-b_1\alpha}\beta^{a_2-1}e^{-b_2\beta} d\alpha d\beta \tag{11}$$

where $c = \binom{n_i}{y_i} \frac{b_1^{a_1}}{\Gamma(a_1)} \frac{b_2^{a_2}}{\Gamma(a_2)}$

Observe that we could use Laplace's method (see for example, Tierney and Kadane, 1986) to solve the integral in (11).

Other possibility is to consider the use of MCMC methods to get the conditional probability function for Y_i, $i = 1,...,N$ in Bayesian analysis. With the prior distribution (9) for α and β assuming prior independence, the joint posterior distribution for α and β is given by,

$$\pi\left(\alpha,\beta / \underset{\sim}{n}, \underset{\sim}{y}\right) \propto \alpha^{a_1-1}e^{-b_1\alpha}\beta^{a_2-1}e^{-b_2\beta}\left\{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right\}^N \prod_{i=1}^N \left[\frac{\Gamma(\alpha+y_i)\Gamma(\beta+n_i-y_i)}{\Gamma(\alpha+\beta+n_i)}\right] \tag{12}$$

where $\underset{\sim}{n} = (n_1,...,n_N)$ and $\underset{\sim}{y} = (y_1,...,y_N)$.

The conditional posterior distributions for the Gibbs Sampling algorithm are given by,

(i) $\pi\left(\alpha/,\beta,\underset{\sim}{n},\underset{\sim}{y}\right) \propto \alpha^{a_1-1}e^{-b_1\alpha}\psi_1(\alpha,\beta)$

where,

$$\psi_1(\alpha,\beta) = \exp\left\{N\ln[\Gamma(\alpha+\beta)] - N\ln[\Gamma(\alpha)] + \sum_{i=1}^{N}\ln[\Gamma(\alpha+y_i)] - \sum_{i=1}^{N}\ln[\Gamma(\alpha+\beta+n_i)]\right\} \tag{13}$$

(ii) $\pi\left(\beta/,\alpha,\underset{\sim}{n},\underset{\sim}{y}\right) \propto \beta^{a_2-1}e^{-b_2\beta}\psi_2(\alpha,\beta)$

where,

$$\psi_2(\alpha,\beta) = \exp\left\{N\ln[\Gamma(\alpha+\beta)] - N\ln[\Gamma(\beta)] + \sum_{i=1}^{N}\ln[\Gamma(\beta+n_i-y_i)] - \sum_{i=1}^{N}\ln[\Gamma(\alpha+\beta+n_i)]\right\}$$

Observe that we need to use the Metropolis-Hastings algorithm to generate the variables α and β.

The predictive density for Y_i is given by

$$f(y_i/n_i,a_1,b_1,a_2,b_2) = \int\int f(y_i/n_i,\alpha,\beta)\pi\left(\alpha,\beta/\underset{\sim}{n},\underset{\sim}{y}\right)\partial\alpha\partial\beta \tag{14}$$

where $f(y_i/n_i,\alpha,\beta)$ is the Beta-Binomial distribution (3) and $\pi\left(\alpha,\beta/\underset{\sim}{n},\underset{\sim}{y}\right)$ is the joint posterior distribution for α and β (12).

A Monte Carlo estimate for (14) based on S generated Gibbs samples is given by,

$$\hat{f}(y_i/n_i,a_1,b_1,a_2,b_2) = \frac{1}{S}\sum_{s=1}^{S}\binom{n_i}{y_i}\frac{\Gamma(\alpha^{(s)}+\beta^{(s)})}{\Gamma(\alpha^{(s)})\Gamma(\beta^{(s)})}\frac{\Gamma(\alpha^{(s)}+y_i)\Gamma(\beta^{(s)}+n_i-y_i)}{\Gamma(\alpha^{(s)}+\beta^{(s)}+n_i)} \tag{15}$$

Assuming the correlated Binomial model with probability function (7), let us consider the following prior distributions for the parameters δ and p,

$$\begin{aligned} &\text{(i)}\ \ \delta \sim U[a,b];\ \ a,b \text{ Known;} \\ &\text{(ii)}\ \ p \sim Beta[c,d];\ \ c,d \text{ Known;} \end{aligned} \tag{16}$$

where $U(a,b)$ denote an uniform distribution in the interval (a,b).

Assuming prior independence among the parameters, the joint posterior distribution for δ and p is given by,

$$\pi\left(\delta, p / \underset{\sim}{y}\right) \propto p^{c+\sum_{i=1}^{N} y_i - 1} (1-p)^{d+\sum_{i=1}^{N}(n_i - y_i)-1} \prod_{i=1}^{N} \left\{1 + \frac{\delta}{2p(1-p)}\left[(y_i - n_i p)^2 + y_i(2p-1) - n_i p^2\right]\right\} \tag{17}$$

The conditional posterior distributions for the Gibbs Sampling algorithm are given by,

$$\begin{aligned} &\text{(i)}\ \ \pi\left(\delta / p, \underset{\sim}{y}\right) \propto \psi(\delta, p)\ ; \\ &\text{(ii)}\ \ \pi\left(p / \delta, \underset{\sim}{y}\right) \propto p^{c+\sum_{i=1}^{N} y_i - 1} (1-p)^{d+\sum_{i=1}^{N}(n_i - y_i)-1} \psi(\delta, p)\ ; \end{aligned} \tag{18}$$

where $a < \delta < b$; $0 < p < 1$ and

$$\psi(\delta, p) \propto \prod_{i=1}^{N} \left\{1 + \frac{\delta}{2p(1-p)}\left[(y_i - n_i p)^2 + y_i(2p-1) - n_i p^2\right]\right\}$$

Observe that we need to use the Metropolis-Hastings algorithm to generate samples of δ and p.

A Monte Carlo estimate for the predictive density of Y_i is given by,

$$\hat{f}(y_i/n_i,a,b,c,d)=\frac{1}{S}\sum_{s=1}^{S}f(y_i/n_i,\delta^{(s)},p^{(s)}) \tag{19}$$

where $f(y_i/n_i,\delta,p)$ denotes the probability function (7) , and S is the number of generated Gibbs samples for δ and p of the joint posterior distribution (17).

3. Use of Mixture of Binomial Distributions

Let us assume that the extra-Binomial variability could be modeled by a mixture of Binomial distributions (8).

The likelihood function for $\underset{\sim}{\theta}=(\theta_1,...,\theta_J)$ and $\underset{\sim}{\lambda}=(\lambda_1,....,\lambda_J)$ where $\sum_{j=1}^{J}\lambda_j=1$ is given by,

$$L(\underset{\sim}{\theta},\underset{\sim}{\lambda})=\prod_{i=1}^{N}\sum_{j=1}^{J}\left[\lambda_j\binom{n_i}{y_i}\theta_j^{y_i}(1-\theta_j)^{n_i-y_i}\right] \tag{20}$$

Assuming prior independence between $\underset{\sim}{\theta}$ and $\underset{\sim}{\lambda}$, the joint posterior distribution for the parameters is given gy,

$$\pi(\underset{\sim}{\theta},\underset{\sim}{\lambda}/\underset{\sim}{n},\underset{\sim}{y})=\pi(\underset{\sim}{\theta})\pi(\underset{\sim}{\lambda})L(\underset{\sim}{\theta},\underset{\sim}{\lambda}) \tag{21}$$

where $\pi(\underset{\sim}{\theta})$ and $\pi(\underset{\sim}{\lambda})$ are the prior distributions for $\underset{\sim}{\theta}$ and $\underset{\sim}{\lambda}$ and $L(\underset{\sim}{\theta},\underset{\sim}{\lambda})$ is given in (20).

To simplify the joint posterior distribution and the full conditional distributions for the Gibbs Sampling algorithm we introduce latent variables (see for example, Tanner and Wong, 1987) given by $\underset{\sim}{v_i}=(v_{i1},....,v_{iJ})$ with a multinomial distribution $\text{Mult}(1;h_{i1},....,h_{iJ})$ with cell probabilities,

$$h_{ij} = \frac{\lambda_j \binom{n_i}{y_i} \theta_j^{y_i} (1-\theta_j)^{n_i - y_i}}{\sum_{j=1}^{J} \lambda_j \binom{n_i}{y_i} \theta_j^{y_i} (1-\theta_j)^{n_i - y_i}} \tag{22}$$

That is,

$$\pi\left(\underset{\sim}{v}_1, \ldots\ldots, \underset{\sim}{v}_N / \underset{\sim}{\theta}, \underset{\sim}{\lambda}, \underset{\sim}{n}, \underset{\sim}{y}\right) \propto \frac{\prod_{i=1}^{N} \prod_{j=1}^{J} \left[\lambda_j \binom{n_i}{y_i} \theta_j^{y_i} (1-\theta_j)^{n_i - y_i} \right]^{v_{ij}}}{\prod_{i=1}^{N} \sum_{j=1}^{J} \lambda_j \binom{n_i}{y_i} \theta_j^{y_i} (1-\theta_j)^{n_i - y_i}} \tag{23}$$

Combining (23) with (21), we get,

$$\pi\left(\underset{\sim}{\theta}, \underset{\sim}{\lambda} / \underset{\sim}{n}, \underset{\sim}{y}, \underset{\sim}{v}\right) \propto \pi\left(\underset{\sim}{\theta}\right) \pi\left(\underset{\sim}{\lambda}\right) \prod_{i=1}^{N} \prod_{j=1}^{J} \left[\lambda_j \binom{n_i}{y_i} \theta_j^{y_i} (1-\theta_j)^{n_i - y_i} \right]^{v_{ij}} \tag{24}$$

For the special case of $J = 2$ binomial componentes in (8), and assuming prior independence among the parameters, consider the following prior densities for θ_1, θ_2 and λ_1:

$$\begin{aligned} &\text{(i) } \theta_j \sim Beta(a_j; b_j) \text{ ; } a_j, b_j \text{ Known;} \\ &\\ &\text{(ii) } \lambda_1 \sim Beta(c; d) \text{ ; } c, d \text{ Known;} \end{aligned} \tag{25}$$

for $j = 1,2$ and $\theta_1 < \theta_2$.

The joint posterior distribution for θ_1, θ_2 and λ_1 is given by,

$$\pi\left(\theta_1,\theta_2,\lambda_1 / \underset{\sim}{n},\underset{\sim}{y},\underset{\sim}{v}\right) \propto \lambda_1^{c+\sum_{i=1}^{N} v_{i1}-1}(1-\lambda_1)^{d+\sum_{i=1}^{N} v_{i2}(n_i-y_i)-1}\prod_{j=1}^{2}\theta_j^{a_j+\sum_{i=1}^{N} v_{ij}y_i-1}\left(1-\theta_j\right)^{b_j+\sum_{i=1}^{N} v_{ij}(n_i-y_i)-1} \tag{26}$$

where $v_{i1}+v_{i2}=1$.

The full conditional posterior distributions for θ_1, θ_2 and λ_1 are given by,

$$\begin{aligned}
&\text{(i)}\ \ \theta_1/\theta_2,\lambda_1,\underset{\sim}{n},\underset{\sim}{y},\underset{\sim}{v} \sim Beta\left(a_1+\sum_{i=1}^{N} v_{i1}y_i; b_1+\sum_{i=1}^{N} v_{i1}(n_i-y_i)\right);\\
&\text{(ii)}\ \ \theta_2/\theta_1,\lambda_1,\underset{\sim}{n},\underset{\sim}{y},\underset{\sim}{v} \sim Beta\left(a_2+\sum_{i=1}^{N} v_{i2}y_i; b_2+\sum_{i=1}^{N} v_{i2}(n_i-y_i)\right);\\
&\text{(iii)}\ \ \lambda_1/\theta_1,\theta_2,\underset{\sim}{n},\underset{\sim}{y},\underset{\sim}{v} \sim Beta\left(c+\sum_{i=1}^{N} v_{i1}; d+\sum_{i=1}^{N} v_{i2}(n_i-y_i)\right).
\end{aligned} \tag{27}$$

To generate samples of the joint posterior distribution (26) we follow the steps:

(i) Start with the initial values $\underset{\sim}{\theta}^{(0)}$ and $\underset{\sim}{\lambda}^{(0)}$;

(ii) Generate $\underset{\sim}{v}^{(1)}=\left(\underset{\sim 1}{v}^{(1)},....,\underset{\sim N}{v}^{(1)}\right)$ from a Bernoulli distribution with cell probabilities (22);

(iii) Generate a sample of θ_1,θ_2 and λ_1 from the conditional distributions (27).

4. Extra-Binomial Variability in the presence of covariates

If explanatory variables are present, usually we consider logistic regression for modeling. The residual variation, however, can be greater than what is expected, which implies the existence of extra-Binomial variability.

In the presence of a vector $\underset{\sim}{x}=(x_1,....,x_k)$ of covariates, the logistic regression model is given by,

$$y_i/n_i, y_i, \underset{\sim}{x_i} \sim b(n_i,p_i) \tag{28}$$

where $p_i = \dfrac{e^{\underset{\sim}{\beta}' \underset{\sim}{x}_i}}{1+e^{\underset{\sim}{\beta}' \underset{\sim}{x}_i}}$, $\underset{\sim}{\beta}' \underset{\sim}{x}_i = \beta_0 + \sum_{l=1}^{k} \beta_l x_{li}$, $i = 1,...,N$

Observe that we are assuming the same values for the covariates $\underset{\sim}{x}_i = (x_{1i},...,x_{ki})$ for the n_i observations in the Binomial distribution $b(n_i, p_i)$, $i = 1,....,N$.

The residual variation, however may be greater than what can be attributed to Binomial random variation even when all covariates are fitted in the model.

To model extra-Binomial variation in the presence of covariates, we assume the correlated-Binomial model (7), that is,

$$P(Y_i = y_i / n_i, \underset{\sim}{x}_i, \underset{\sim}{\beta}, \delta) = \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i} A_i\left(n_i, y_i, \underset{\sim}{x}_i, \underset{\sim}{\beta}, \delta\right) \tag{29}$$

where $A_i\left(n_i, y_i, \underset{\sim}{x}_i, \underset{\sim}{\beta}, \delta\right) = 1 + \dfrac{\delta}{2p_i(1-p_i)}\left[(y_i - n_i p_i)^2 + y_i(2p_i - 1) - n_i p_i^2\right]$,

$$p_i = \frac{e^{\underset{\sim}{\beta}' \underset{\sim}{x}_i}}{1+e^{\underset{\sim}{\beta}' \underset{\sim}{x}_i}}, \quad \underset{\sim}{\beta}' = (\beta_0, \beta_1, ..., \beta_k), \quad i = 1,2,....,N.$$

Observe that in model (29), we are assuming that the intra-class correlation among binary variables $U_{i1},....,U_{1n_i}$ of $Y_i = \sum_{l=1}^{n_i} U_{il}$ is homogeneous independent of the value of the covariate $\underset{\sim}{x}_i$, $i = 1,...,N$.

Assuming prior independence among the parameters, we could consider the following prior distributions,

(i) $\beta_0 \sim N(\mu_0, \sigma_0^2)$; μ_0, σ_0^2 Known;

(ii) $\beta_l \sim N(\mu_l, \sigma_l^2)$; μ_l, σ_l^2 Known; (30)

(iii) $\delta \sim U(a,b)$; a, b Known;

where $l = 1,2,...,k$ and $N(\mu,\sigma^2)$ denotes a Normal distribution with mean μ and variance σ^2.

Assuming prior (30), we use MCMC methods to simulate Gibbs Samples for the joint posterior distribution of δ and $\underset{\sim}{\beta}$.

Another possibility is to consider different intra-class correlations δ_i for each Binomial distribution $b(n_i,p_i)$, with covariate $\underset{\sim}{x}_i$, $i = 1,...,N$.

In this case, the likelihood function for $\underset{\sim}{\delta} = (\delta_1,...,\delta_N)$ and $\underset{\sim}{\beta} = (\beta_1,....,\beta_k)$ is given by

$$L\left(\underset{\sim}{\delta},\underset{\sim}{\beta}\right) \propto \frac{\exp\left\{\sum_{i=1}^{N} y_i \underset{\sim}{\beta}' \underset{\sim}{x}_i\right\}}{\prod_{i=1}^{N}\left(1+e^{\underset{\sim}{\beta}' \underset{\sim}{x}_i}\right)^{n_i}} \prod_{i=1}^{N} A_i\left(n_i,y_i,\underset{\sim}{x}_i,\underset{\sim}{\beta},\delta_i\right) \tag{31}$$

where $A_i\left(n_i,y_i,\underset{\sim}{x}_i,\underset{\sim}{\beta},\delta_i\right) = 1+\frac{\delta_i}{2p_i(1-p_i)}\left[(y_i-n_ip_i)^2 + y_i(2p_i-1) - n_ip_i^2\right]$ and p_i is the logistic regression model.

For a Bayesian analysis of the model, we can consider the same prior distribution for $\underset{\sim}{\beta}$ given in (30) and consider uniform prior distributions $U(a,b)$ for δ_i, $i = 1,...,N$.

5. Some Examples

5.1. An example with Genetic Data

Let us consider a genetic data set introduced by Skellan (1948). The data concern the secondary association of cromosomes in Brassica, a kind of cabbage and cauliflower.

There are $N = 337$ nuclei observed. In each nucleus there are pairs of bivalents during meiosis. Each pair may or may not show association between bivalentes. The observed frequencies with 0,1,2 and 3 associated pairs in these nuclei are 32,103,122 and

80, respectively. If we assume the simple Binomial distribution $Y_i \sim B(3,p)$, $i = 1,..,N$, we are assuming that the probability of association is the same for any pair any nucleus.

The maximum likelihood estimator for p is $\hat{p} = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} n_i} = \frac{587}{1011} = 0.5806$.

However, the observed variation ($\sum_{i=1}^{N} \frac{\left(y_i - n\hat{p}\right)^2}{N} = 0.86$ where $n = 3$ and $\hat{p} = 0.58$) is larger than the expected Binomial variation $n\hat{p}\left(1-\hat{p}\right) = 0.73$, which suggests that the probability of association is different form nucleus to nucleus. Considering $\hat{p} = 0.58$, the product of predictive probability for $Y_i = 0,1,2$ or 3 is given by $A1 = \prod_{i=1}^{N} \binom{n_i}{y_i} \hat{p}^{y_i} \left(1-\hat{p}\right)^{n_i - y_i} = 5.5278 \text{ x } 10^{-192}$.

Assuming the Beta-Binomial distribution (3) and the prior distributions (9) for α and β, with $a_1 = 170$, $b_1 = 28$, $a_2 = 170$ and $b_2 = 39$, we use Laplace's method (see for example, Tierney and Kadane, 1986) to obtain $A2 = \prod_{i=1}^{N} \hat{f}(y_i / n_i, a_1, b_1, a_2, b_2) =$ 1.4327 x 10^{-190} where $\hat{f}(y_i / n_i, a_1, b_1, a_2, b_2)$ denotes the Laplace's approximation for the conditional probability function (11). We observe that A_2 is larger than A_1, that is, the Beta-Binomial model (3) with prior distributions (9) gives better fit for the genetic data.

Considering S=500 generated Gibbs samples for the joint posterior distribution (12) and using the Monte Carlo estimate (15) for the predictive density $\hat{f}(y_i / n_i, a_1, b_1, a_2, b_2)$ in (14), we find $A3 = \prod_{i=1}^{N} \hat{f}(y_i / n_i, a_1, b_1, a_2, b_2) = 1.9684 \text{ x } 10^{-190}$.

Assuming the correlated-Binomial model (7) and the prior distribution (16) for the parameters δ and p, with $a = 0$, $b = 2$, $c = 6$ and $d = 4.3$ we simulated 5 separate Gibbs chains of size 2000 for each parameter from the conditional posterior distributions

(18). For each chain we discarded the first 1000 iterations (" burn-in-samples") and we considered the 10^{th}, 20^{th}, ... iterations. We monitored the convergence of the Gibbs samples using the Gelman and Rubin (1992) method, which utilizes the analysis of variance technique to determine if further iterations and needed.

In table 1, we have the posterior summaries of interest. We also have in table 1, the estimated potential scale reductions $\hat{R}$ (see Gelman and Rubin ,1992) for all parameters. In this case, the number of iterations considered was sufficient for approximate convergence ($\sqrt{\hat{R}} < 1.1$ for all the parameters).

Parameter	Mean	SD	95% Credible Interval	$\hat{R}$
δ	0.0928	0.0334	(0.0305 ; 0.1599)	0.9993
p	0.5802	0.0168	(0.5418 ; 0.6091)	1.0057

Table 1. Posterior summaries (correlated-Binomial distribution)

Considering the $S = 500$ generated Gibbs samples, we find the Monte Carlo estimates (19) for the predictive density of Y_i, $i = 1,...,N$, and $A4 = \prod_{i=1}^{N} \hat{f}(y_i / n_i, a, b, c, d) =$ 2.083 x 10^{-190}, that is, a larger value than $A_1 = 5.5278$ x 10^{-192} considering the standard Binomial distribution.

Considering a mixture of $J = 2$ Binomial distributions (20) with prior distributions (25) with $a_1 = 2.144$, $b_1 = 6.288$, $a_2 = 4.03$, $b_2 = 1.01$, $c = 2$ and $d = 3$, we generated 5 separate Gibbs chains of size 2000 from the conditional posterior distributions (27). Also discarding the first 1000 iterations for each chain, we considered the 10^{th}, 20^{th}, ... iterations. In table 2, we have the posterior summaries.

Parameter	Mean	95% Credible Interval	$\hat{R}$
θ_1	0.4136	(0.2324 ; 0.5269)	1.0017
θ_2	0.7286	(0.6187 ; 0.8720)	1.0097
λ_1	0.4724	(0.1295 ; 0.7833)	1.0076

Table 2. Posterior summaries (mixture of $J = 2$ Binomial distributions)

Considering the $S = 500$ generated samples, we get $A_5 = \prod_{i=1}^{N} \hat{f}(y_i / n_i, a_1, a_2, b_1, b_2, c, d) = 2.0231 \times 10^{-190}$ where $\hat{f}(y_i / n_i, a_1, a_2, b_1, b_2, c, d)$ is the Monte Carlo estimate of the predictive density for Y_i given by,

$$\hat{f}(y_i / n_i, a_1, a_2, b_1, b_2, c, d) = \frac{1}{500} \sum_{s=1}^{500} \left[\lambda_1^{(s)} \binom{n_i}{y_i} \theta_1^{(s)y_i} \left(1 - \theta_1^{(s)}\right)^{n_i - y_i} + \left(1 - \lambda_1^{(s)}\right) \binom{n_i}{y_i} \theta_2^{(s)y_i} \left(1 - \theta_2^{(s)}\right)^{n_i - y_i} \right] \tag{32}$$

Thus, we conclude that the genetic data introduced by Skelan (1948) is better fitted by extra-Binomial variability models.

5.2 – An example with covariates

In table 3, we have a data set introduced by Crowder (1978). A batch of tiny seeds is brushed onto a plate covered with a certain extract at a given dilution. The numbers of germinated and ungerminated seeds are subsequently counted. This data set is given as a 2 x 2 factorial layout. There are two types of seed, O. aegyptiaca 75 and O. aegyptiaca 73, and two root extracts, been and cucumber.

From inspection of the data of table 3, we observe that there is heterogeneity of proportions between replicates.

(I) O. aegyptiaca 75				(II) O. aegyptiaca 73			
Bean		Cucumber		Bean		Cucumber	
y_i	n_i	y_i	n_i	y_i	n_i	y_i	n_i
10	39	5	6	8	16	3	12
23	62	53	74	10	30	22	41
23	81	55	72	8	28	15	30
26	51	32	51	23	45	32	51
17	39	46	79	0	4	3	7
		10	13				

Table 3. Data for seeds O. aegyptiaca 75 and 73, been and cucumber root extracts

To analyse the data set of table 3, we first assume the logistic regression model (28) where,

$$p_i = \frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\beta_3 x_{1i}x_{2i}}}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\beta_3 x_{1i}x_{2i}}} \tag{33}$$

where X_{1i}= -1 for seed O. aegyptiaca 75, X_{1i}= 1 for seed O. aegyptiaca 73 and $X_{2i}=-1$ for root extract bean and $X_{2i}=1$ for root extract cucumber, $i=1,2,...,21$.

Considering the prior distirbutions (30) for β_0, β_1, β_2 and β_3, in the logistic regression model (33) with μ_0=0.1, $\sigma_0^2=0.3$, $\mu_1=0.1$, $\sigma_1^2=0.3$, $\mu_2=0.3$, $\sigma_2^2=0.3$, $\mu_3=-0.2$ and $\sigma_3^2=0.2$, we generated 5 separate Gibbs chains, each of which ran for 2000 iterations. We also monitored the convergence of the Gibbs samples using the Gelman and Rubin (1992) method. For each parameter, we discarded the 1000 first iterations (" burn-in-samples") and we considered the 10^{th}, 20^{th}, ... iterations. In table 4, we have the posterior summaries of interest. We also have in table 4, the estimated potential scale reductions $\hat{R}$ (see Gelman and Rubin, 1992) for all parameters.

Parameter	Mean	95% Credible Interval	$\hat{R}$
β_0	-0.0060	(-0.1553 ; 0.1379)	1.0011
β_1	-0.1010	(-0.2328 ; 0.0465)	1.0055
β_2	0.4512	(0.3054 ; 0.5936)	1.0007
β_3	-0.2043	(-0.3480 ; -0.0771)	0.9998

Table 4. Posterior summaries (logistic regression model)

From the results of table 4, we observe that β_2 and β_3 are different of zero (credible intervals do not contain zero). That is, the covariate X_2 (root extract) and the interation X_1X_2 (seed x root extract) have signifiant effects on the counts of germinated seeds.

Considering the $S = 500$ generated Gibbs samples, we find $c_1 = \prod_{i=1}^{N} \hat{f}\left(y_i / n_i, \mu_0, \mu_1, \mu_2, \mu_3, \sigma_0^2, \sigma_1^2, \sigma_2^2, \sigma_3^2\right) = 2.5194 \text{ x } 10^{-24}$ where $\hat{f}\left(y_i / n_i, \mu_0, \mu_1, \mu_2, \mu_3, \sigma_0^2, \sigma_1^2, \sigma_2^2, \sigma_3^2\right)$ is the Monte Carlo estimate for the predictive density of Y_i, $i = 1,2,..,N$.

Assuming the correlated-Binomial model (29) in presence of covariates with p_i given in (33), we consider the prior distributions (30) for β_0, β_1, β_2, β_3 and δ with μ_0=0.1 , $\sigma_0^2 = 0.2$, $\mu_1 = -0.2$, $\sigma_1^2 = 0.2$, $\mu_2 = 0.5$, $\sigma_2^2 = 0.3$, $\mu_3 = -0.3$, $\sigma_1^2 = 0.3$, $a = 0$ and $b = 0.04$. In this case, we also generated 5 separate Gibbs chains of size 2000 for each parameter using the Metropolis-Hastings algorithm. The selection of samples and verification of convergence of the Gibbs samples was similar as used for logistic regression model.

In table 5, we have the posterior summaries of interest.

Parameter	Mean	95% Credible Interval	$\hat{R}$
β_0	-0.0318	(-0.1776 ; 0.1316)	1.0010
β_1	-0.1522	(-0.2985 ; 0.0074)	1.0006
β_2	0.4319	(0.2826 ; 0.6026)	1.0021
β_3	-0.2074	(-0.354 ; -0.0377)	0.9999
δ	0.0217	(0.0057 ; 0.0392)	1.0000

Table 5. Posterior summaries (correlated-Binomial model in presence of covariates)

From the 95% credible intervals given in table 5, we observe that β_2, β_3 and δ are different of zero.

Considering the $S = 500$ generated Gibbs samples, we obtained Monte Carlo estimates for the predictive density of Y_i, $i = 1,...,N$ given $n_i, \mu_0, \mu_1, \mu_2, \mu_3, \sigma_0^2, \sigma_1^2, \sigma_2^2, \sigma_3^2, a$ and b, and $c_2 = \prod_{i=1}^{N} \hat{f}\left(y_i / n_i, \mu_0, \mu_1, \mu_2, \mu_3, \sigma_0^2, \sigma_1^2, \sigma_2^2, \sigma_3^2, a, b\right) =$ 8.2023 x 10^{-24}. This value for c_2 is larger than $c_1 =$ 2.5194 x 10^{-24}, considering the usual logistic regression model. That is, the correlated-Binomial model in presence of covariates gives better fit for the data set of table 3.

If we assume different intra-class correlations among the binary variables in the correlated-Binomial model in presence of covariates and the same prior distributions for β_0, β_1, β_2 and β_3 as considered for the case of δ homogeneous across $Y_1,...,Y_N$ and uniform prior distirbutions $U(0;0.04)$ for δ_i, $i = 1,....,N$, we have in table 6 the posterior summaries of interest based on $S = 500$ generated Gibbs samples.

Parameter	Mean	95% Credible Interval	$\hat{R}$
β_0	-0.0203	(-0.1665 ; 0.1561)	1.0037
β_1	-0.1557	(-0.3072 ; 0.0305)	0.9995
β_2	0.4306	(0.2739 ; 0.6054)	0.9958
β_3	-0.2077	(-0.3627 ; -0.0417)	1.0008
δ_1	0.0210	(0.0041 ; 0.0384)	0.9975
δ_2	0.0205	(0.0031 ; 0.0387)	1.0066
δ_3	0.0208	(0.0043 ; 0.0386)	1.0004
δ_4	0.0212	(0.0048 ; 0.0385)	0.9977
δ_5	0.0204	(0.0036 ; 0.0386)	0.9996
δ_6	0.0212	(0.0038 ; 0.0385)	0.9963
δ_7	0.0214	(0.0052 ; 0.0391)	0.9999
δ_8	0.0213	(0.0050 ; 0.0391)	1.0002
δ_9	0.0209	(0.0041 ; 0.0387)	1.0063
δ_{10}	0.0215	(0.0033 ; 0.0389)	0.9980
δ_{11}	0.0211	(0.0028 ; 0.0389)	0.9962
δ_{12}	0.0210	(0.0040 ; 0.0386)	1.0074
δ_{13}	0.0210	(0.0043 ; 0.0387)	1.0108
δ_{14}	0.0209	(0.0035 ; 0.0384)	1.0012
δ_{15}	0.0209	(0.0042 ; 0.0392)	0.9996
δ_{16}	0.0208	(0.0047 ; 0.0389)	0.9999
δ_{17}	0.0206	(0.0039 ; 0.0386)	0.9962
δ_{18}	0.0210	(0.0043 ; 0.0389)	0.9954
δ_{19}	0.0205	(0.0030 ; 0.0389)	0.9971
δ_{20}	0.0215	(0.0044 ; 0.0389)	1.0038
δ_{21}	0.0216	(0.0049 ; 0.0394)	0.9986

Table 6. Posterior summaries (correlated-Binomial model with different inta-class correlations).

Considering the $S = 500$ generated Gibbs samples, we get Monte Carlo estimates for the predictive density for Y_i, $i = 1,...,N$ and

$c_3 = \prod_{i=1}^{N} \hat{f}\left(y_i / n_i, \mu_0, \mu_1, \mu_2, \mu_3, \sigma_0^2, \sigma_1^2, \sigma_2^2, \sigma_3^2, a, b\right) =$ 7.1162 x 10^{-24}. This value is close to c_2 =8,2023 x 10^{-24}. In this case, we can assume the correlated-Binomial model with same intra-class correlation δ as the best model for the data set introduced by Crowder (1978).

6. Concluding remarks

The use of extra-Binomial variation models is needed for many applications, since the observed variation could be greater than expected under the ordinary Binomial assumption. Also, in many applications, we can have the presence of covariates.

The use of MCMC methods is a suitable way to get the posterior summaries of interest for these models.

References

ALTHAM, P.M.E (1978). Two Generalizations of the Binomial Distribution. *Applied Statistics*, n.27, p.162-167.

CROWDER, M.J. (1978). Beta-Binomial Anova for Proportions. *Applied Statistics*, n.27, p.34-37.

GELFAND, A.E.; SMITH, A.F.M. (1990). Sampling Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, n.85, p.398 – 409.

GELMAN, A.E.; RUBIN, D. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Sciences*, n.7, p.457-472.

GRIFFITHS, D.A. (1973). Maximum Likelihood Estimation for the Beta-Binomial Distribuition and an Application to the Household Distribution of the Total Number of Cases of a Diease. *Biometrics*, n.29, p.637-648.

HSIAO, C.K. (1994). Bayesian Tests of Extra-Binomial Variability with Emphasis on the Boundary Case, *PhD thesis*, Carnegie-Mellon University, U.S.A.

KUPPER, L.L.; HASEMAN, J.K. (1978). The Use of a Correlated Binomial Model for the Analysis of Certain Toxicological Experiments. *Biometrics*, n.34, p.69–76.

RUDOLFER, S.,M. (1990). A Markov Chain Model of Extra-Binomial Variation. *Biometrika*, n.77, p.255-264.

SKELLAM, J.G. (1948). A Probability Distribution Derived from the Binomial Distribution by Regarding the Probability of Sucess as Variable between the Sets of Trials. *Journal of the Royal Statistical Society*, B, p.257-261.

SMITH, A.F.M.; ROBERTS, G.O. (1993). Bayesian Methods via the Gibbs Sampler and related Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society*, B, n.55, p.3-23.

TANNER, M.A.; WONG, W.H. (1995). The Calculation of Posterior Distributions by Data Augmentation (with discussion). *Journal of the American Statistical Association*, n.82, p.528-550.

TARONE, R.E. (1979). Testing the Goodness-of-Fit of the Binomial Distribution. *Biometrika*, n.66, p.585-590.

TIERNEY, L.; KADANE, J.B. (1986). Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association*, n.81, p.82-86.

TITTERINGTON, D.M.; SMITH, A.F.M.; MARKOV, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions.* New York: Wiley.

WILLIAMS, D.A. (1982). Extra-Binomial Variation in Logistic Linear Models. *Applied Statistics*, p.31, n.144-148.

Resumo

Neste artigo, apresentamos uma análise Bayesiana para modelos de variabilidade extra-Binomial introduzidos na literatura. Também consideramos o uso de misturas de distribuições para modelar a variabilidade extra-Binomial e a introdução de covariáveis. Considerando os algoritmos Gibbs Sampling e Metropolis-Hastings, obtemos estimadores de Monte Carlo para as quantidades a posteriori de interesse. A metodologia é ilustrada com dois conjuntos de dados reais.

NOTAS DO ICMC

SÉRIE ESTATÍSTICA

062/2000 WRUCK, E.; ACHCAR, J.A.; MAZUCHELI, J. – Classification and discrimination for populations with mixture of multivariate normal distributions.

061/2000 ANDRADE, M.G.; MEIRA, S.A.; FRAGOSO, M.D.; CARNEIRO, A.A.F.M. – A bayesian approach to the stochastic flood control problem.

060/2000 ACHCAR, J.A.; JANEIRO, V. – A bayesian analysis for corralated binary data in the presence of covariates.

059/2000 MAZUCHELI, J.; ACHCAR, J.A.; KASS, R.E. – Regression models for lifetime data with mixture of normal distributions.

058/99 ACHCAR, J.A.; FORTULAN, V.C. – Meta analysis: a bayesian approach.

057/99 OLIVEIRA, S.C.; ACHCAR, J.A. - Confiabilidade de redes: um enfoque bayesiano.

056/98 RODRIGUES, J.; CHAVES, J.S. – A note on bayesian exponential regression model with censored data.

055/98 RODRIGUES, J.; SILVEIRA, V.D.R. – Bayesian computation for dichotomous variables with classification errors.

054/98 ACHCAR, J.A.; FORTULAN, V.C. – Relação entre o uso de hormônio e câncer em mulheres: um aplicação de meta-análise sob um enfoque bayesiano.

053/98 CID, J.E.R.; ACHCAR, J..A. – Bayesian inference for nonhomogeneous poisson processes in software reliability models assuming nonmonotonic intensity functions.