The International Journal of Biostatistics

Volume 6, Issue 1

2010

Article 27

Bivariate Zero-Inflated Regression for Count Data: A Bayesian Approach with Application to Plant Counts

Anandamayee Majumdar, *Arizona State University*

Corinna Gries, *University of Wisconsin-Madison*

Recommended Citation:

Majumdar, Anandamayee and Gries, Corinna (2010) "Bivariate Zero-Inflated Regression for Count Data: A Bayesian Approach with Application to Plant Counts," *The International Journal of Biostatistics*: Vol. 6: Iss. 1, Article 27.

DOI: 10.2202/1557-4679.1229

Bivariate Zero-Inflated Regression for Count Data: A Bayesian Approach with Application to Plant Counts

Anandamayee Majumdar and Corinna Gries

Abstract

Lately, bivariate zero-inflated (BZI) regression models have been used in many instances in the medical sciences to model excess zeros. Examples include the BZI Poisson (BZIP), BZI negative binomial (BZINB) models, etc. Such formulations vary in the basic modeling aspect and use the EM algorithm (Dempster, Laird and Rubin, 1977) for parameter estimation. A different modeling formulation in the Bayesian context is given by Dagne (2004). We extend the modeling to a more general setting for multivariate ZIP models for count data with excess zeros as proposed by Li, Lu, Park, Kim, Brinkley and Peterson (1999), focusing on a particular bivariate regression formulation. For the basic formulation in the case of bivariate data, we assume that X_i are (latent) independent Poisson random variables with parameters λ_i , $i = 0, 1, 2$. A bi-variate count vector (Y_1, Y_2) response follows a mixture of four distributions; p_0 stands for the mixing probability of a point mass distribution at $(0, 0)$; p_1 , the mixing probability that $Y_2 = 0$, while $Y_1 = X_0 + X_1$; p_2 , the mixing probability that $Y_1 = 0$ while $Y_2 = X_0 + X_2$; and finally $(1 - p_0 - p_1 - p_2)$, the mixing probability that $Y_i = X_i + X_0$, $i = 1, 2$. The choice of the parameters $\{p_i, \lambda_i, i = 0, 1, 2\}$ ensures that the marginal distributions of Y_i are zero inflated Poisson ($\lambda_0 + \lambda_i$). All the parameters thus introduced are allowed to depend on co-variables through canonical link generalized linear models (McCullagh and Nelder, 1989). This flexibility allows for a range of real-life applications, especially in the medical and biological fields, where the counts are bivariate in nature (with strong association between the processes) and where there are excess of zeros in one or both processes. Our contribution in this paper is to employ a fully Bayesian approach consolidating the work of Dagne (2004) and Li et al. (1999) generalizing the modeling and sampling-based methods described by Ghosh, Mukhopadhyay and Lu (2006) to estimate the parameters and obtain posterior credible intervals both in the case where co-variables are not available as well as in the case where they are. In this context, we provide explicit data augmentation techniques that lend themselves to easier implementation of the Gibbs sampler by giving rise to well-known and closed-form posterior distributions in the bivariate ZIP case. We then use simulations to explore the effectiveness of this estimation using the Bayesian BZIP procedure, comparing the performance to the Bayesian and classical ZIP approaches. Finally, we demonstrate the methodology based on bivariate plant count data with excess zeros that was collected on plots in the Phoenix metropolitan area and compare the results with independent ZIP regression models fitted to both processes.

KEYWORDS: Bayesian inference, regression model, bivariate zero-inflated Poisson, data augmentation, generalized linear model, multivariate zero-inflated Poisson

Author Notes: The authors would like to thank Steven S. Carroll for sampling design; L. Dugan, D. Gonzales, E. Holmes, Q. Stewart, J. Walker, M. Feldner, S. Whitcomb, and X. Zhuo for field and lab assistance; Salt River Project for the donation of helicopter time; Cities of Phoenix, Scottsdale and Tempe, Maricopa County Parks, Tonto National Forest, Arizona State Lands Department, Sky Harbor Airport and all the private property owners involved for giving permission to access their land. This material is based upon work supported by the National Science Foundation under Grant No. DEB-0423704, Central Arizona-Phoenix Long-Term Ecological Research (CAP LTER). Any opinions, findings and conclusions or recommendation expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF). Majumdar's research was funded by NSF grant.

1. Introduction

Recently, there has been increased interest in statistical models to account for ‘excess’ zeros in data. These models are often called the zero inflated (ZI) models. A review of literature on such models can be obtained in Ridout, Demetrio and Hinde (1998) where examples from econometrics, manufacturing, medicine, patents applications, road safety, species abundance, use of recreational activities and sexual behavior have been cited. Li, Lu, Park, Kim, Brinkley and Peterson, 1999 discuss the use of multivariate zero inflated Poisson models in manufacturing processes. Dagne (2004) discusses a hierarchical regression model in the Bayesian context, where the formulation is based on prior distributions, but the likelihood is not based on bi-variate ZIP modeling. In the current paper, we consider the following example from life sciences where the plant counts of certain species have been collected; In a fast developing landscape as in Phoenix, Arizona, urban ecologists aim to find associations of plant counts with other factors like income per capita, land use, historic agricultural land use, and elevation. As part of the Central Arizona-Phoenix Longterm Ecological Research(CAP LTER) project scientists conducted in the year 2005 a large scale, intensive survey of 204 plots. One objective was to compare plant preferences in the urban area with the surrounding open desert. Another aim was to investigate whether observations on socio-economic factors such as income per capita, historical agricultural use of land, land use at the present time and geophysical factors like elevation could be used for prediction of various plant counts as well as make statistical inference about plant preferences at a specific desert/urban location. Even though data was collected on 38 plant species, we restrict our attention to two such species of interest, namely *Prosopis Glandulosa* and *Larrea Tridentata*. We use this data to motivate the use of multivariate zero-inflated Poisson models proposed by Li *l* (1999), focusing on the bi-variate case. We introduce bi-variate ZIP models, bi-variate ZIP regression (BZIPR) modeling and general multivariate ZIP models in 2. We demonstrate the Bayesian approach of handling the BZIP as well as BZIP regression models in 3, including novel data augmentation techniques. We then turn to several simulation results and a real-life application in section 4 where we apply the methodology on plant counts in Phoenix metropolitan area. Finally, we summarize our findings and end with some comments in section 5.

2. Bivariate ZIP distribution

We now turn to the discussion of bi-variate zero-inflated Poisson models.

A bi-variate ZIP can be constructed as a mixture of a bi-variate Poisson, two univariate Poisson, and a point mass at $(0, 0)$:

$$\begin{aligned}
 (Y_1, Y_2)^T &\sim (0, 0) \text{ with probability } p_0 \\
 &\sim (\text{Poisson}(\lambda_1), 0) \text{ with probability } p_1 \\
 &\sim (0, \text{Poisson}(\lambda_2)) \text{ with probability } p_2 \\
 &\sim \text{bi-variate Poisson}(\lambda_{10}, \lambda_{20}, \lambda_{00}) \\
 &\quad \text{with probability } (1 - p_0 - p_1 - p_2)
 \end{aligned} \tag{1}$$

A bi-variate Poisson distribution with parameters $(\lambda_{10}, \lambda_{20}, \lambda_{00})$ can be represented as follows (Marshall and Olkin, 1985):

$$Y_1 = W_1 + W_0 \text{ and } Y_2 = W_2 + W_0 \tag{2}$$

where W_1 , W_2 and W_0 are independent and have respective means λ_{10} , λ_{20} and λ_{00} .

Then the marginal distributions of the bi-variate Poisson distribution (2) are Poisson with means $\lambda_{10} + \lambda_{00}$ and $\lambda_{20} + \lambda_{00}$. The assumptions that $\lambda_1 = \lambda_{10} + \lambda_{00}$ and $\lambda_2 = \lambda_{20} + \lambda_{00}$ are necessary to guarantee that the marginal distributions of Y_1 and Y_2 are univariate ZIP's.

There is a natural extension of the bi-variate ZIP to an K -dimensional multivariate ZIP as in Li, Lu, Park, Kim, Brinkley and Peterson, 1999. First a multivariate Poisson is needed. There are several ways to construct multivariate Poisson. For instance, a general way to construct would be to use one common term (Johnson and Kotz, 1969):

$$Y_1 = W_1 + W_0, Y_2 = W_2 + W_0, \dots, Y_K = W_K + W_0. \tag{3}$$

where the W_i 's are independent Poisson variables. This constructive approach ensures a manageable number of parameters to handle. The same ideas of constructing the bi-variate ZIP used before in this section can be used here.

To extend the Bivariate ZIP to a bi-variate ZIP regression model, first, define the variables $\mathbf{Z}_j = (Z_{0j}, Z_{1j}, Z_{2j}, Z_{3j})^T$ such that $\mathbf{Z}_j = (1, 0, 0, 0)^T$ if and

only if $j = (Y_{1j}, Y_{2j})^T$ is degenerate at $(0, 0)$; $\mathbf{Z}_j = (0, 1, 0, 0)^T$ iff Y_{1j} is degenerate at 0 and $Y_{2j} \sim \text{Poisson}(\lambda_{0j} + \lambda_{2j})$; $\mathbf{Z}_j = (0, 0, 1, 0)^T$ iff Y_{2j} is degenerate at 0 and $Y_{1j} \sim \text{Poisson}(\lambda_{0j} + \lambda_{1j})$; otherwise, $\mathbf{Z}_j = (0, 0, 0, 1)^T$, in which case, $(Y_{1j}, Y_{2j}) \sim \text{Bivariate Poisson}(\lambda_{0j}, \lambda_{0j}, \lambda_{2j})$. Define the $4N \times 1$ vector $\mathbf{Z}$ as $\mathbf{Z} = (\mathbf{Z}_1^T, \dots, \mathbf{Z}_N^T)^T$. We next define the vector $\mathbf{p} = (p_{01}, \dots, p_{0N}, p_{11}, \dots, p_{1N}, p_{21}, \dots, p_{2N})^T$ where $\mathbf{p}_j = (p_{0j}, p_{1j}, p_{2j})^T$ are given as:

$$P\left((Y_{1j}, Y_{2j})^T \text{ is degenerate at } (0, 0)\right) = P\left(\mathbf{Z}_j = (1, 0, 0, 0)^T\right) = p_{0j}$$

$$P(Y_{1j} \text{ is degenerate at } 0, Y_{2j} \sim \text{Poisson}(\lambda_{0j} + \lambda_{2j})) = P\left(\mathbf{Z}_j = (0, 1, 0, 0)^T\right) = p_{1j}$$

$$P(Y_{2j} \text{ is degenerate at } 0, Y_{1j} \sim \text{Poisson}(\lambda_{0j} + \lambda_{1j})) = P\left(\mathbf{Z}_j = (0, 0, 1, 0)^T\right) = p_{2j}$$

$$P((Y_{1j}, Y_{2j}) \text{ is Bivariate Poisson } (\lambda_{0j}, \lambda_{1j}, \lambda_{2j})) = P\left(\mathbf{Z}_j = (0, 0, 0, 1)^T\right) = 1 - p_{0j} - p_{1j} - p_{2j} \quad (4)$$

The Poisson parameters $\boldsymbol{\lambda} = (\lambda_{01}, \dots, \lambda_{0N}, \lambda_{11}, \dots, \lambda_{1N}, \lambda_{21}, \dots, \lambda_{2N})^T$ are modeled via canonical link GLMs as $\log(\lambda_{ij}) = \mathbf{B}_j \boldsymbol{\beta}_i$ and $\log(p_{ij}/(1 - \sum_{i=0}^2 p_{ij})) = \mathbf{G}_j \boldsymbol{\gamma}_i$ for the j -th rows $\mathbf{B}_j$ and $\mathbf{G}_j$ of the design matrices $\mathbf{B}$ and $\mathbf{G}$ respectively, $j = 1, \dots, N$. In our life science example involving two plant species, the design matrix $\mathbf{G}$, corresponds to a matrix with four columns, correspond to four regressor variables, namely, intercept, ever in agriculture (a 0-1 variable signifying whether the spatial point had been ever used for agriculture or not), income, and elevation (in meters), respectively. The design matrix corresponds to a matrix with seven columns, where the first four columns are the same as the matrix $\mathbf{G}$, while the last three columns correspond to landuse (categorical variables). In our final analysis we omitted landuse column from $\mathbf{G}$ since the corresponding estimates were all non-significant, while their inclusion yielded unstable estimates. Income and elevation were standardized by subtracting mean and dividing mean to yield stable estimates. From equation (4), the full likelihood of the data $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_N^T)^T$ can be derived. Details of the bi-variate ZIP regression model described in this section has been fitted using maximum likelihood methods in Li, Lu, Park, Kim, Brinkley and Peterson (1999), and the likelihood derived. In our case, we shall proceed to use a fully Bayesian approach for dealing with this model as described in the following section.

3. Bayesian analysis

In this section we develop a new Bayesian approach for analyzing BZIP and BZIP regression models.

3.1 Prior choice for non-regression case

We now specify prior distribution of the parameters in the non-regression case described in section 2. Here we assume the prior distributions for $\mathbf{p} = (p_0, p_1, p_2, (1 - p_0 - p_1 - p_2))^T$ and $\lambda_0, \lambda_1, \lambda_2$ are mutually independent, and use the following conditionally conjugate priors:

$$\mathbf{p} \sim \text{Dirichlet}(b_1, b_2, b_3, b_4), \lambda_i \sim \text{Gamma}(a_{i1}, a_{i2}) \quad i = 0, 1, 2.$$

(While Ghosh, Mukhopadhyay and Lu (2006) choose a Beta distribution and a conjugate prior for the power series family for the conditional conjugate priors, the authors use Dirichlet and Gamma distributions.) The hyperparameters $a_{01}, \dots, a_{22}, b_1, \dots, b_4$ will be assumed known, and specified in a manner to ensure a large variance for each of the parameters in the model. In particular, $b_1 = \dots = b_4 = 1$ gives the uniform prior on the unit rectangle on $\mathcal{P}$. Small values of a_{i2} result in non-informative priors for λ_i . The joint distribution of $(\mathbf{p}, \lambda_0, \lambda_1, \lambda_2)$ in this case is not of closed form, and to tackle this aspect, a Bayesian paradigm is used. In particular, the Gibbs sampling method (Gelfand and Smith, 1990) is exploited to obtain a large random sample from the posterior distribution. We observe that under this set up, independence of priors will not imply independence in the posterior distributions of the random parameters. In addition to finding the posterior mean, median and variance, having the entire posterior distributions of parameters helps one to compute related intervals for univariate and multivariate parameters (e.g., 95% posterior credible interval), and predict future values of interest, which are not readily available via a frequentist approach.

3.2 Prior choice for regression case

We next proceed to choosing priors for the BZIP regression model discussed in section 2. As in the case of the non-regression case, we choose priors with large variance (non-informative priors) to rule out prior-sensitivity. When the sample size is relatively large such priors will yield estimators close to the maximum likelihood estimators. Assuming that the elements of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$ with $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)^T$ and $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_k)^T$ are independently distributed, the joint prior distribution of the parameters is given by

$$\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i=1}^k \pi(\beta_i) \pi(\gamma_i) \quad (5)$$

where the prior $\pi(\beta_i)$ and $\pi(\gamma_i)$ are multivariate Normal with $\mathbf{0}$ and variance covariance matrix $\sigma_i^2 (\mathbf{I})^{-1}$ and $\tau_i^2 (\mathbf{G}^T \mathbf{G})^{-1}$ respectively. The hyperparameters σ_i^2 and τ_i^2 are assigned Inverse Gamma(2, 1) priors which implies that the mean is 1 and variance is infinite.

3.3 The data augmentation method

The Gibbs sampling method based on full conditional distributions obtained from the BZIP regression makes it necessary to use non-closed expressions in the posterior distributions of the parameters involved. This situation is less attractive due to the computational burdens associated with using the Metropolis Hastings algorithm within the Gibbs sampler as outlined in Gelman and Rubin, 1992. To override this difficulty, a general procedure is proposed for the case without co-variables, with the idea of augmenting to the bi-variate data. A data augmentation technique was also used for the univariate case in Ghosh, Mukhopadhyay and Lu (2006). (Y_1, Y_2) , using the latent variables $(\mathbf{W}, \mathbf{Z})$, with the representation

$$Y_1 = (1 - Z_0)(1 - Z_2)(W_0 + W_1), \quad Y_2 = (1 - Z_0)(1 - Z_1)(W_0 + W_2)$$

where $\mathbf{W} = (W_0, W_1, W_2)^T$ and $\mathbf{p} = (p_0, p_1, p_2, (1 - p_0 - p_1 - p_2))^T$ such that $W_i \stackrel{\text{i.i.d}}{\sim} \text{Poisson}(\lambda_i)$ and $\mathbf{Z} = (Z_0, Z_1, Z_2, (1 - Z_0 - Z_1 - Z_2))^T \sim$

Multinomial(1, $\mathbf{p}$) are mutually independent. Instead of obtaining samples directly from the posterior of $(\mathbf{p}, \boldsymbol{\lambda})$ given the data, samples from the posterior of $(\mathbf{p}, \boldsymbol{\lambda}, \mathbf{W}, \mathbf{Z})$ are obtained given the data (Y_1, Y_2) where $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \lambda_2)^T$. To implement this procedure within a Gibbs sampling algorithm, the conditional distribution $(\mathbf{W}, \mathbf{Z})$ given $(Y_1, Y_2, \mathbf{p}, \boldsymbol{\lambda})$ is required. Sampling from this conditional distribution is sometimes referred to as the “data augmentation” step (Tanner, and Wong, 1987) or the “imputation step”. Ghosh, Mukhopadhyay and Lu (2006) investigate a similar procedure, but because our bivariate response increases the complexity, we present our procedure in Appendix B. A data augmentation method for the BZIP regression case can be similarly deduced.

3.4 Fitting BZIP regression

The joint posterior distribution of the unknown parameters and the hyperparameters, $(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\sigma}^2, \boldsymbol{\tau}^2)$ discussed in the previous section conditional on the data is obtained by combining (4) and (5) via Bayes theorem, as:

$$\pi(\mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma}^2, \boldsymbol{\tau}^2 | \mathbf{y}) = L(\mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma} | \mathbf{y}) \prod_{i=1}^k \{\pi(\beta_i) \pi(\gamma_i) \pi(\sigma_i^2) \pi(\tau_i^2)\} \quad (6)$$

where $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_k^2)^T$ and $\boldsymbol{\tau}^2 = (\tau_1^2, \dots, \tau_k^2)^T$.

Bayesian estimation requires that we determine the conditional posterior distribution of the parameters including the hyperparameters, given the data. Unfortunately, it is not possible to do this analytically in closed form, nor is it feasible to do it numerically because of the large number of parameters. However, we can use a Markov chain Monte Carlo (MCMC) algorithm, such as the Gibbs sampler, that has the posterior distribution as its long run distribution. The Gibbs sampler is a Markovian updating procedure that can be used to generate samples from a joint distribution by iteratively sampling from full conditional distributions (Gamerman, 1997; Gelman, Carlin, Stern, Rubin, 1995; Gelman and Rubin (1992); Gilks, Richardson, and Spiegelhalter, 1996).

Full conditional posterior distributions are required to implement the Gibbs sampler algorithm, and they are given in Appendix C of the paper.

4. Data application and simulation studies

4.1 Simulation studies

To establish validity of the proposed Bayesian method, this section presents four simulation studies based on the BZIP model without regressors to keep the model simpler and the study more focused. In simulation study I, true parameter values were fixed at $\mathbf{p}^T = (0.25, 0.25, 0.25, 0.25)$, $\lambda_0 = 0.5$, $\lambda_1 = 0.5$, $\lambda_2 = 0.5$, which results into a BZIP model for which the marginal distribution of Y_1 follows $ZIP(p = 0.5, \lambda = 1)$. We note that there are many choices of parameters, $\{\mathbf{p}, \boldsymbol{\lambda}\}^T$ which yield the same marginal distribution of Y_1 and we have chosen one such. We also have the probability of zero for the first process $Pr(Y_1 = 0) = (p_0 + p_2) + (1 - p_0 - p_2) \exp\{-\lambda_0 - \lambda_1\}$. This study thus is a specific study of BZIP for which the marginal distribution of Y_1 coincides with the study carried out in Table 3 (page 1368) of Ghosh, Mukhopadhyay and Lu (2006) (study I). Our inclusion of this study shall help us to look at the same model when analyzed by BZIP models from Bayesian approaches. In simulation study-II, the true parameter values were fixed at $\mathbf{p}^T = (0.45, 0.05, 0.45, 0.05)$, $\lambda_0 = 0.5$, $\lambda_1 = 0.5$, $\lambda_2 = 0.5$, which is a case of the BZIP model for which the marginal distribution of Y_1 follows $ZIP(p = 0.9, \lambda = 1)$. Again, there are many choices of parameters, $\{\mathbf{p}, \boldsymbol{\lambda}\}^T$ which can yield the

same marginal distribution of Y_1 and we have chosen one such case. Again, this marginal study coincides with the study carried out in Table 3 of Ghosh, Mukhopadhyay and Lu (2006) (study II). As in Ghosh, Mukhopadhyay and Lu (2006), for each of the approaches and methods used, we run 600 iterations and find the empirical coverage probabilities and average 95% confidence intervals (or credible intervals in the case of a Bayesian approach) for estimators of (i) $p = p_0 + p_2$ (ii) $\lambda = \lambda_0 + \lambda_1$ and (iii) $P(Y_1 = 0) = p + (1 - p)\exp(-\lambda)$ as well as (iv) $E(Y_1) = E(W_0 + W_1)E(1 - Z_0)(1 - Z_2) = (\lambda_0 + \lambda_1)(1 - p_0 - p_2)$, (v) $P(Y_1 = 0, Y_2 = 0) = p_0 + p_1e^{\{-\lambda_0 - \lambda_1\}} + p_2e^{\{-\lambda_0 - \lambda_2\}} + p_3e^{\{-\lambda_0 - \lambda_1 - \lambda_2\}}$ (vi) p_0 (vii) p_1 (viii) p_2 (ix) λ_0 (x) λ_1 (xi) λ_2 , along with their average biases, and s.e. of the biases. We note that the first four parameters (i) - (iv) correspond to the marginal distribution of the first process of which the first three parameters have been examined by Ghosh, Mukhopadhyay and Lu (2006) in the ZIP case. Tables 2 and 3 provide the exact numbers for comparison, but a quick summary is given in Table 1.

First we examine the coverage probabilities of the mean posterior credible intervals. For Study I, the Bayesian BZIP (BaBZIP) case yields mean intervals with higher coverage probabilities than either of the classical (CZIP) or the Bayesian ZIP (BaZIP) approaches for parameters p and λ . For $\Pr(Y_1 = 0)$, the coverage probability 94.6% of the BaBZIP model is larger than the BaZIP model, but smaller than the CZIP approach in study I – owing to the fact that the classical approach yields yet too large intervals. In Study II, the coverage probability 96.7%, corresponding to the BaBZIP estimate is slightly smaller when compared to both the BaZIP and the CZIP estimates. Since throughout the studies, BaBZIP gives coverage probabilities close to or more than 95%, we may conclude that the BaBZIP model seems to be working well in terms of coverage probability of the posterior intervals.

Next, we study the interval sizes. In study I, the mean confidence intervals of the CZIP and the BaZIP models corresponding to p are 141.56% larger and 84.42% smaller than the one corresponding to the BaBZIP model. But in study II, these numbers are 157.63% and 115.31%, respectively. This indicates the BaBZIP model works much better in estimating the p for the classical approach and works better than the BaZIP approach when the p is large. We note that the corresponding comparisons in Study I for the parameter λ for the CZIP and the BaZIP yields 118.59% and 105.72% larger confidence intervals, while in Study II these numbers become 223.74% and 101.86%, respectively. So the BaBZIP approach works better in all cases. Finally we focus on the parameter $\Pr(Y_1 = 0)$. In study I, the CZIP and BaZIP approaches yield 325.21% larger and 91.03% smaller intervals compared to the BaBZIP approach. On the other hand, in study II, the ZIP and BaZIP approaches yield 444.60% and

Table 1

Brief summary of estimation performance of the 3 models – BaBZIP or Bayesian BZIP, BaZIP or Bayesian ZIP, and CZIP or Classical ZIP. The parameters chosen for estimation are in the first column. The top half of the table compares coverage probabilities. The bottom half of the table compares the sizes of the confidence intervals.

Model	BaBZIP	BaZIP	CZIP
Mean coverage probability	Comparison with others		
<i>Study I</i>			
p	Highest		
λ	Highest		
$\Pr(Y_1 = 0)$	Medium highest		
<i>Study II</i>			
p	Lowest (slight margin)		
λ	Lowest (slight margin)		
$\Pr(Y_1 = 0)$	Lowest (slight margin)		
Sizes of intervals		Compare with BaBZIP	Compare with BaBZIP
<i>Study I</i>			
p		84.42% smaller	141.56% larger
λ		105.72% larger	118.59% larger
$\Pr(Y_1 = 0)$		91.03% smaller	325.21% larger
<i>Study II</i>			
p		115.31% larger	157.63% larger
λ		101.86% larger	223.74% larger
$\Pr(Y_1 = 0)$		237.41% smaller	444.60% larger

237.41% larger intervals compared to the BaBZIP approach. This indicates the BaBZIP model works much better than the classical approach in estimating $\Pr(Y_1 = 0)$ and works better than the BaZIP approach when $\Pr(Y_1 = 0)$ is large.

We next focus our attention to the bias and the s.e. of the bias. Ghosh, Mukhopadhyay and Lu (2006) note that in the ZIP case, Bayes estimates are not unbiased (under squared error loss). The absolute value of the average biases and corresponding s.e.'s are mostly smaller for the BaBZIP case as compared to the BaZIP case. In general we may note from these studies, that the strength of the BaBZIP procedure is in the bias reduction of the BaZIP, and producing tighter interval estimates with good coverage probability, based on the finite sample distribution.

To summarize, the BaBZIP model is satisfactory in performance with respect to prediction coverage for all kinds of parameter choices that we have examined, and overall yield tighter (thus more accurate) credible intervals for the parameters considered than both the BaZIP and the CZIP models, when. The slight increase of computing time might be compensated by the use of freely available software packages.

4.2 Data analysis: no co-variables

We return to the motivating example of Central Arizona-Phoenix Longterm Ecological Research (CAP LTER) project in analyzing two plant species *Larrea tridentata* and *Prosopis Glandulosa* using a BZIP model, without any co-variables with notations used as in section 2 of the paper. The posterior findings are summarized in the Table 4, where a five-number summary, consisting of the posterior mean, sd (standard deviation), 2.5 percentile, median and 97.5 percentile, has been presented. Note that the 2.5 and 97.5 percentiles provide an equal-tail 95% posterior interval estimate for the parameters. Alternatively, the highest posterior density (HPD) interval can also be constructed (especially if the distribution is bimodal).

The software **WinBUGS** was used to generate samples from the posterior distribution of the parameters. To check convergence of the MCMC method, the expected values of the two species *Prosopis Glandulosa* and *Larrea Tridentata*, $E(Y_1)$, $E(Y_2)$, and the probability of zeros, namely, $\Pr(Y_1 = 0, Y_2 = 0)$, $\Pr(Y_1 = 0)$ and $\Pr(Y_2 = 0)$ were monitored. The convergence of the Gibbs sampler was checked using a Splus code **CODA** (also available from the same website as **WinBUGS**), as outlined in , Best, Cowles and Vines (2006). In particular, from the initial runs with three dispersed starting values, the dependence factor of the Raftery-Lewis convergence diagnostic was found to be 1.1, 1.07,

Table 2

Simulation studies: coverage probability based on a CI with nominal coverage of 95% C.I.

	ZIP classical	ZIP Bayesian	BZIP Bayesian
Study-I			
$p = 0.5$	94.4 (0.153, 0.698) 0.0337 (0.0155)	94.5 (0.378, 0.703) -0.0666 (0.0723)	94.8 (0.354, 0.739) -0.0588 (0.0830)
$\lambda = 1$	95.9 (0.528, 1.689) 0.0040 (0.0302)	96.2 (0.687, 1.722) 0.1478 (0.2231)	99.0 (0.612, 1.591) -0.0328 (0.2062)
$Pr(Y_1 = 0)$ $= 0.684$	98.2 (0.241, 0.802) -0.0022 (0.0614)	92.8 (0.602, 0.815) -0.0331 (0.0502)	94.6 (0.587, 0.821) -0.0265 (0.0545)
Study-II			
$p_0 = 0.9$	93.0 (0.507, 0.972) 0.1571 (0.2928)	94.3 (0.634, 0.973) 0.0676 (0.1476)	98.3 (0.663, 0.957)) 0.0479 (0.0479)
$\lambda = 1$	91.8 (0.207, 2.610) 0.0926 (0.6773)	94.1 (0.867, 1.961) -0.16 (0.4301)	96.0 (0.555, 1.629) -0.009 (0.2633)
$Pr(Y_1 = 0)$ $= 0.937$	97.7 (0.366, 0.984) 0.0007 (0.0355)	97.1 (0.655, 0.985) 0.0067 (0.0338)	96.7 (0.834, 0.973) 0.0191 (0.0284)

Table 3*Simulation studies: coverage probability based on 95% C.I.*

	Mean 95% C. I	Percent coverage	Mean Bias	S.E (Bias)
Study-I				
$E(Y_1) = 0.5$	(0.281, 0.755)	96.8	0.0104	0.1066
$P(Y_1 = Y_2 = 0) = 0.489$	(0.374, 0.624)	95.7	-0.007	0.0590
$p_0 = 0.25$	(0.041, 0.454)	98.7	0.0129	0.0855
$p_1 = 0.25$	(0.114, 0.455)	95.1	-0.0169	0.0813
$p_2 = 0.25$	(0.163, 0.528)	92.3	-0.0717	0.0809
$\lambda_0 = 0.5$	(0.142, 1.108)	98.5	-0.1053	0.1667
$\lambda_1 = 0.5$	(0.129, 1.271)	98.6	-0.0601	0.2288
$\lambda_2 = 0.5$	(0.092, 1.044)	96.7	0.0724	0.1863
Study-II				
$E(Y_1) = 0.1$	(0.041, 0.316)	96.3	-0.0408	0.0555
$P(Y_1 = Y_2 = 0) = 0.645$	(0.504, 0.749)	96.8	0.0144	0.0580
$p_0 = 0.45$	(0.100, 0.6135)	95.3	0.0765	0.1076
$p_1 = 0.05$	(0.013, 0.258)	97.1	-0.0410	0.0444
$p_2 = 0.45$	(0.269, 0.739)	97.1	-0.0285	0.1039
$\lambda_0 = 0.5$	(0.051, 1.156)	100	-0.0099	0.1715
$\lambda_1 = 0.5$	(0.053, 1.833)	99.3	-0.0965	0.2942
$\lambda_2 = 0.5$	(0.053, 1.207)	99.6	0.0005	0.1871

1.07, 1.19 and 1.08 for $E(Y_1)$, $E(Y_2)$, $Pr(Y_1 = 0, Y_2 = 0)$, $Pr(Y_1 = 0)$ and $Pr(Y_2 = 0)$, respectively. Also the 95% shrink factors of the Gelman-Rubin diagnostics (based on three parallel chains) were 1, 1, 1, 1.05 and 1 respectively. Thus, these summary diagnostics indicate no potential problem with the convergence of the sampler. For more details on these numerical diagnostics, the reader is referred to the CODA manual. These diagnostics should not be taken as a proof of convergence of the chains, however if there were any problems, usually the diagnostic factors point to some potential problems.

Based on the above diagnostic results, a burn-in of 2000 samples was used with three parallel chains and then every 100-th sample was kept, until 9000 observations were obtained from each chain. Thus, a total of 27000 observations were generated from the joint posterior distribution of the parameters. The posterior summary of the parameters, based on the final 27000 observations, is presented in Table 4. WinBUGS took about 33-45s on an Intel core 2 quad 2.66 GHz processor, to perform the job for the BZIP models.

The posterior mean (and median) estimate of $Pr(Y_1 = 0, Y_2 = 0)$ from the BZIP distribution is close to the empirical percentage of zero-defects, which is 0.45 for our data, with corresponding 95% posterior credible interval [0.375, 0.528]. The posterior mean (and median) estimate of $Pr(Y_1 = 0)$ and $P(Y_2 = 0)$ are also close to the respective empirical percentages, which are 0.91 and 0.51 in our data, with corresponding 95% posterior credible intervals [0.860, 0.947] and [0.443, 0.597]. This indicates a good fit of the BZIP distribution. Again, the posterior mean (and median) of the expected counts of *Prosopis Glandulosa* and *Larrea Tridentata*, $E(Y_1)$ and $E(Y_2)$, are close to the empirical means of 0.13 and 13.02 respectively with corresponding 95% posterior credible intervals [0.073, 0.202] and [0.443, 0.597].

When two independent ZIP models were fitted to Y_1 and Y_2 the summary findings give a 95% credible interval for $E(Y_1)$ is [0.069, 0.220], with mean (sd) 0.132 (0.039) and median 0.127. Thus the posterior credible interval is 117% wider and the standard deviation 118% larger than the one in 4. The 95% credible interval of $P(Y_1 = 0)$ is [0.853, 0.949] while the mean (sd) is 0.908 (0.024) and median 0.910. Hence the posterior credible interval is 110% wider and the standard deviation 109% larger than the BZIP model results. The credible interval for $E(Y_2)$ is [10.80, 15.30], while the mean (sd) and median are 13.02 (1.165) and 13.01. This indicates that the posterior credible interval is 108.4% wider while the standard deviation is 108.7% larger than that in 4. Again the 95% credible interval of $P(Y_2 = 0)$ is [0.432, 0.595] with mean (sd) and median are 0.514 (0.042) and 0.514. So the credible interval is 105.84% wider and sd is 107.69% larger than the estimates of the BZIP model. Finally, the 95% credible interval of $P(Y_1 = 0, Y_2 = 0)$ is [0.389, 0.544]

with mean (sd) and median 0.467 (0.040) and median 0.467. Again, we find that these estimates yield a posterior credible interval 108.39% wider and sd 102.56% larger than that in 4. Hence the BZIP model performs slightly better in estimation for all the relevant parameters for this particular data set.

4.3 Data analysis: with co-variables

We continue to examine the example of the CAP LTER data of two plant species *Prosopis Glandulosa* and *Larrea Tridentata* using a BZIPR model, with four co-variables with notations used as in section 2 of the paper. From Table 4, we observe that the regressor variables do not affect most of the $\log(p_{ij}/(1-p_{0j}-p_{1j}-p_{2j}))$ components of the model significantly; the γ_{ij} 's are all insignificant but for the γ_{12} which corresponds to the income regression coefficient term of $\log(p_{12}/(1-p_{02}-p_{12}-p_{23}))$, the negative sign indicative of the fact that the chance of having no *Prosopis Glandulosa* decreases with higher income – in other words the chance of having *Prosopis Glandulosa* in the plot increases with higher income. On the other hand, some of the regressor coefficients corresponding to the Poisson component of the BZIP models, that is λ_i , are significant. In particular, we find that $\beta_{2,j}, j = 1, \dots, 4$ are all significant – which implies the significance of all the regressor terms in explaining $\log(\lambda_{2i})$. The signs of the posterior credible intervals corresponding to ever in agriculture, income and elevation are all negative, so this implies that frequency of *Larrea Tridentata* are negatively associated with these three regressor variables. We further have the regressor coefficients $b_{ij}, i = 0, 1, 2, j = 1, 2, 3$ in the model. These coefficients correspond to the land-use categories (we have four landuse categories: (i) Desert (ii) Urban residential (iii) Transportation or under construction (iv) Urban non residential). Hence “ b_{01} is significant” shall mean that urban residential landuse is significantly different than desert landuse in explaining $\log(\lambda_{0j})$, and hence is significantly different in explaining both the *Prosopis Glandulosa* and *Larrea Tridentata* species, and so on. We observe that b_{21}, b_{22}, b_{23} are significant with negative sign, and this implies that desert landuse is associated with significantly larger frequency of *Larrea Tridentata* – so we expect to see more *Larrea Tridentata* in the desert plots than each of the other plots. These results are consistent with the fact that *Prosopis Glandulosa* is considered a garden-plant while *Larrea Tridentata* is largely a desert-plant.

As before, WinBUGS was used to fit the BZIP regression models. The posterior summary values, presented in Table 5 is based on the final 27000 samples (after a burn-in of 2000 samples with three parallel chains and then every 100-th sample was kept, until 9000 observations were obtained from

each chain). WinBUGS took about 9h to perform the job for the BZIP regression model. To check convergence of the MCMC method, the average value of the expected values of the two species *Prosopis Glandulosa* and *Larrea Tridentata*, $E(\bar{Y}_1)$, $E(\bar{Y}_2)$, and the average of the probability of zeros, namely, $\sum_{i=1}^n Pr(Y_{1i} = 0, Y_{2i} = 0)/n$, $\sum_{i=1}^n Pr(Y_{1i} = 0)/n$ and $\sum_{i=1}^n Pr(Y_{2i} = 0)/n$ were monitored. The convergence of the Gibbs sampler was checked using a Splus code CODA as earlier. In particular, from the initial runs with three dispersed starting values, the dependence factor of the Raftery-Lewis convergence diagnostic was found to be 1.05, 0.97, 1.05, 1.05 and 0.97 for $E(\bar{Y}_1)$, $E(\bar{Y}_2)$, $\sum_{i=1}^n Pr(Y_{1i} = 0, Y_{2i} = 0)/n$, $\sum_{i=1}^n Pr(Y_{1i} = 0)/n$ and $\sum_{i=1}^n Pr(Y_{2i} = 0)/n$ respectively. Also the 95% shrink factors of the Gelman-Rubin diagnostics (based on three parallel chains) were 1, 1, 1, 1.05 and 1 respectively. Thus, these summary diagnostics indicate no potential problem with the convergence of the sampler.

The posterior mean (and median) estimate of $\frac{1}{n} \sum_{i=1}^n Pr(Y_{1i} = 0, Y_{2i} = 0)$ from the BZIP regression distribution is close to the empirical percentage of zero-defects, which is 0.45 for our data, with a 95% posterior credible interval [0.382, 0.516]. The posterior mean (and median) estimates of $\frac{1}{n} \sum_{i=1}^n Pr(Y_{1i} = 0)$ and $\frac{1}{n} \sum_{i=1}^n Pr(Y_{2i} = 0)$ with respective posterior credible intervals [0.861, 0.942] and [0.428, 0.568], are also close to the respective empirical percentages, which are 0.91 and 0.51 in our data. The posterior means (and medians) of the average expected counts of *Prosopis Glandulosa* and *Larrea Tridentata*, i.e. $E(\bar{Y}_1)$ and $E(\bar{Y}_2)$ are close to the empirical estimates of the average of these species, 0.13, and 13.02 with corresponding posterior intervals [0.078, 0.204] and [12.53, 13.6]. All the parameters have considerably shorter posterior credible intervals and smaller standard errors in this model, compared to the estimates in Table 4. Thus the model with co-variates seems to be performing better than the model without co-variates, in terms of the performance of estimation.

Again, we conduct a model comparison with two independent ZIP regression models fitted to Y_1 and Y_2 . Here, the summary findings give a 95% credible interval for $E(Y_1)$ is [0.081, 0.201], with mean (sd) 0.136 (0.031) and median 0.134. Thus the posterior credible interval is 98.36% shorter and the standard deviation 96.88% smaller than the respective estimates in Table 5. The 95% credible interval of $P(Y_1 = 0)$ is [0.854, 0.943] while the mean (sd) is 0.902 (0.022) and median 0.910. Hence the posterior credible interval is 109.88% wider and the standard deviation 110% larger than the BZIP regression model results. The credible interval for $E(Y_2)$ is [12.43, 13.63], while the mean (sd) and median are 13.03 (0.302) and 13.01. This indicates that the posterior credible interval is 112.15% wider while the standard deviation is

Table 4
Posterior summary of the BZIP model of CAP LTER data

Parameter value	mean	sd	2.50%	median	97.50%
γ_0	1.115	3.762	-13.02	2.006	3.484
γ_1	0.930	0.611	-0.152	0.907	2.142
γ_2	2.04	0.862	-0.469	2.082	3.547
β_0	-4.25	7.695	-27.36	-0.665	0.289
β_1	-17.29	18.05	-61.18	-11.88	0.151
β_2	3.266	0.027	3.202	3.269	3.317
p_1	0.351	0.114	3.441E-7	0.383	0.475
p_2	0.160	0.113	0.048	0.127	0.533
p_3	0.419	0.069	0.179	0.431	0.511
p_4	0.068	0.062	0.013	0.052	0.290
$E(Y_1)$	0.127	0.033	0.073	0.126	0.202
$E(Y_2)$	13.05	1.072	10.82	13.11	14.97
$P(Y_1 = 0, Y_2 = 0)$	0.447	0.039	0.375	0.447	0.528
$P(Y_1 = 0)$	0.908	0.022	0.860	0.909	0.947
$P(Y_2 = 0)$	0.511	0.039	0.443	0.509	0.597

105.23% larger than the corresponding estimates in 5. So the 95% credible interval of $P(Y_2 = 0)$ is $[0.435, 0.586]$ with mean (sd) and median are 0.511 (0.038) and 0.512. So the credible interval is 107.86% wider and sd is 108.57% larger than the estimates of the BZIP regression model. Finally, the 95% posterior credible interval of $P(Y_1 = 0, Y_2 = 0)$ is $[0.395, 0.539]$ with mean (sd) and median 0.465 (0.036) and median 0.463. These estimates yield a posterior credible interval 107.46% wider and sd 112.5% larger than the corresponding ones in 5. Hence the BZIP regression model performs overall slightly better in estimation for all the relevant parameters for this particular data set.

5. Summary

Bivariate time series of counts with excess zeros relative to the Poisson process are common in many bioscience applications, for example Wang, Lee, Yau, and Carrivick (2003) and Lee, Wang, Carrivick, Yau, and Stevenson (2005) discuss a class of bi-variate zero-inflated Poisson autoregression models to accommodate the zero-inflation and the inherent serial dependency between successive observations where parameter estimation is achieved via an EM algorithm (Dempster, Laird and Rubin, 1977).

Table 5

Posterior summary of the BZIP regression model of CAP LTER data

Parameter value	mean	sd	2.50%	median	97.50%
$\gamma_{0,1}$	-1.064	1.508	-5.402	-0.729	0.721
$\gamma_{0,2}$	-0.679	1.738	-4.943	-0.4679	1.912
$\gamma_{0,3}$	0.615	0.765	-0.616	0.53	2.745
$\gamma_{0,4}$	0.627	0.970	-1.283	0.574	2.812
$\gamma_{1,0}$	-0.104	0.557	-1.261	-0.094	1.118
$\gamma_{1,1}$	-0.495	1.021	-2.327	-0.565	1.615
$\gamma_{1,2}$	-1.895	0.791	-3.668	-1.783	-0.629
$\gamma_{1,3}$	1.098	0.6915	-0.1001	1.019	2.865
$\gamma_{2,0}$	-0.283	1.294	-3.416	-0.069	1.398
$\gamma_{2,1}$	0.223	1.538	-3.026	0.352	2.728
$\gamma_{2,2}$	0.723	0.681	-0.343	0.666	2.303
$\gamma_{2,3}$	0.029	1.015	-2.106	0.129	2.354
$b_{0,1}$	-1.942	4.315	-7.965	-3.5	4.849
$b_{0,2}$	-13.98	34.71	-109.2	-4.052	1.278
$b_{0,3}$	-16.44	29.41	-127.3	-6.262	2.707
$b_{1,1}$	2.346	4.207	-11.19	3.523	8.167
$b_{1,2}$	-13.74	27.36	-97.49	-4.72	7.194
$b_{1,3}$	0.551	8.34	-19.37	1.988	18.04
$b_{2,1}$	-2.423	0.275	-3.027	-2.41	-1.965
$b_{2,2}$	-9.124	9.38	-36.0	-6.165	-3.648
$b_{2,3}$	-21.66	29.55	-94.9	-11.92	-4.985
$\beta_{0,0}$	-3.183	1.171	-5.598	-3.101	-1.015
$\beta_{0,1}$	-1.401	2.26	-7.554	-0.766	1.071
$\beta_{0,2}$	1.851	0.736	0.818	1.839	3.277
$\beta_{0,3}$	-0.158	0.771	-1.638	-0.079	1.078
$\beta_{1,0}$	-7.368	7.16	-30.27	-5.247	-2.242
$\beta_{1,1}$	0.112	1.966	-4.093	0.328	2.548
$\beta_{1,2}$	1.142	1.168	-1.512	1.114	3.638
$\beta_{1,3}$	-0.173	1.093	-2.586	-0.153	2.167
$\beta_{2,0}$	3.686	0.037	3.616	3.683	3.755
$\beta_{2,1}$	-0.3319	0.055	-0.430	-0.334	-0.215
$\beta_{2,2}$	-0.215	0.073	-0.356	-0.217	-0.081
$\beta_{2,3}$	0.258	0.061	0.143	0.255	0.381

Table 6

*Posterior summary of the BZIP regression model of CAP LTER data
(continued)*

Parameter value	mean	sd	2.50%	median	97.50%
$E(\bar{Y}_1)$	0.135	0.032	0.078	0.133	0.204
$E(\bar{Y}_2)$	13.03	0.287	12.53	13.03	13.6
$\bar{p}_1$	0.135	0.065	0.035	0.130	0.265
$\bar{p}_2$	0.292	0.072	0.165	0.296	0.427
$\bar{p}_3$	0.289	0.117	0.052	0.307	0.485
$\bar{p}_4$	0.282	0.114	0.098	0.267	0.518
$\frac{1}{n} \sum_{i=1}^n P(Y_{1i} = 0, Y_{2i} = 0)$	0.457	0.032	0.382	0.460	0.516
$\frac{1}{n} \sum_{i=1}^n P(Y_{1i} = 0)$	0.902	0.020	0.861	0.900	0.942
$\frac{1}{n} \sum_{i=1}^n P(Y_{2i} = 0)$	0.506	0.035	0.428	0.509	0.568

Figure 1. Jittered plot of Larrea Tridentata and Prosopis Glandulosa occurrences across the plots

In this article Bayesian methodologies have been used to model bi-variate ecological data involving plant species, using sampling-based methods. In earlier data exploration, the simpler techniques had failed to yield any meaningful estimation or prediction from the data given. Also, the association between the occurrence of these two plant species was non-estimable. In this paper, the occurrence of the species given some other explanatory variables (landuse, income, elevation and the information whether a plot was ever used for agriculture) - a mixture of socio-economic and natural variables has enabled us a framework for predicting the number of occurrence (and non-occurrence) of the two plant species, together and separately.

Appendix A

```
model
{
  #Likelihood:
  for(i in 1:n)
  {
    Y1[i] ~ dpois(mu1[i])
    Y2[i] ~ dpois(mu2[i])
    mu1[i] <- (1-u[i, 1])*(1-u[i, 3])*
              (lambda0 + lambda1)
    mu2[i] <- (1-u[i, 1])*(1-u[i, 2])*
              (lambda0+lambda2)
    u[i, 1:4] ~ dmulti(p[], 1)
  }
  for(i in 1:3)
  {
    log(q[i]) <-alpha[i]
    p[i] <- q[i]*p[4]
  }
  p[4] <- 1/(q[1]+q[2]+q[3]+1)
  log(lambda0) <-beta[1]
  log(lambda1) <-beta[2]
  log(lambda2) <-beta[3]
  zdp <- p[1]+p[2]*(exp(-lambda0 -lambda1))
        +p[3]*(exp(-lambda0 -lambda2)) +
        p[4]*(exp(-lambda0 -lambda1 -lambda2))
        #P(Y1=Y2=0)
  zdp1 <- p[1]+p[3]+(p[2]+p[4])*(exp(-lambda0
```

```

        -lambda1)) #P(Y1=0)
zdp2 <- p[1]+p[2]+(p[3]+p[4])*(exp(-lambda0
        -lambda2))#P(Y2=0)
e1 <- (p[2] +p[4])*(lambda0 + lambda1) # E(Y1)
e2 <- (p[3] +p[4])*(lambda0 + lambda2) # E(Y2)
#Priors:
for(j in 1:3)
{
    beta[j]~ dnorm(0, 0.001)
    alpha[j]~dnorm(0, 0.001)
}
}

```

Appendix B: Data augmentation method

The following conditional probabilities can then be simplified as

$$P(\mathbf{W} = \mathbf{w}, \mathbf{Z} = (1, 0, 0, 0) | \mathbf{y}) = \begin{cases} \frac{p_0 P(\mathbf{W}=\mathbf{w})}{P(=\mathbf{y})} & \text{if } \mathbf{y} = \mathbf{0} \\ 0 & \text{o.w.} \end{cases}$$

$$P(\mathbf{W} = \mathbf{w}, \mathbf{Z} = (0, 1, 0, 0) | \mathbf{y}) = \begin{cases} \frac{p_1 P(\mathbf{W}=\mathbf{w})}{P(=\mathbf{y})} & \text{if } y_2 = w_1 = 0 \\ 1 & \text{if } y_2 = w_1 > 0 \\ 0 & \text{o.w.} \end{cases}$$

$$P(\mathbf{W} = \mathbf{w}, \mathbf{Z} = (0, 0, 1, 0) | \mathbf{y}) = \begin{cases} \frac{p_2 P(\mathbf{W}=\mathbf{w})}{P(=\mathbf{y})} & \text{if } y_1 = w_2 = 0 \\ 1 & \text{if } y_1 = w_2 > 0 \\ 0 & \text{o.w.} \end{cases}$$

$$P(\mathbf{W} = \mathbf{w}, \mathbf{Z} = (0, 0, 0, 1) | \mathbf{y}) = \begin{cases} \frac{(1-p_0-p_1-p_2)P(\mathbf{W}=\mathbf{w})}{P(=\mathbf{y})} & \text{if } y_1 = w_0 + w_1, y_2 = w_0 + w_2 \\ 0 & \text{o.w.} \end{cases}$$

That is, for $i = 1, \dots, n$,

1. If $i = \mathbf{0}$ and we have $\mathbf{W}_i^T = (W_{i1}, W_{i2}, W_{i3})^T = \mathbf{0}$, then generate a Multinomial random variable $\mathbf{Z}_i$ with number of trials = 1 and probability vector = $\mathbf{p}$.

- (a) If the first element of $\mathbf{Z}_i$ i.e. $Z_{i0} = 1$, generate $W_{ik} \stackrel{\text{indep.}}{\sim} \text{Poisson}(\lambda_k)$ $k = 0, 1, 2$.
 - (b) For $k = 1$ or $k = 2$, if the $(k + 1)$ -th element of $\mathbf{Z}_i$ i.e. $Z_{ik} = 1$ set $W_{ik} = Y_k = 0$. Also generate $W_{i0} \sim \text{Poisson}(\lambda_0)$ and $W_{ic} \sim \text{Poisson}(\lambda_c)$ independently, where $c = 2$ if $k = 2$ and $c = 1$ otherwise.
 - (c) If the last element of $\mathbf{Z}_i$ is equal to 1, then set $W_{ik} = 0; k = 0, 1, 2$.
2. If $i = \mathbf{0}$ and we have $\mathbf{W}_i^T = (0, W_{i1}, 0)^T$ with $W_{i1} > 0$, then set $\mathbf{Z}_i = (Z_{i0}, 0, 1 - Z_{i0}, 0)$ where $Z_{i0} \sim \text{Bernoulli}(\frac{p_{i0}}{(p_{i0} + p_{i2})})$. If $Z_{i0} = 1$ then generate $W_{ik} \sim \text{Poisson}(\lambda_k), k = 0, 1, 2$. On the other hand, if $Z_{i0} = 0$, then set $\mathbf{W}_i = (0, W_{i1}, 0)$ where $W_{i1} \sim \text{Poisson}(\lambda_1)$.
 3. If $i = \mathbf{0}$ and we have $\mathbf{W}_i^T = (0, 0, W_{i2})^T$ with $W_{i2} > 0$, then set $\mathbf{Z}_i = (Z_{i0}, 1 - Z_{i0}, 0, 0)$ where $Z_{i0} \sim \text{Bernoulli}(\frac{p_{i0}}{(p_{i0} + p_{i1})})$. If $Z_{i0} = 1$ then now set $W_{ik} \sim \text{Poisson}(\lambda_k), k = 0, 1, 2$. On the other hand, if $Z_{i0} = 0$, then set $\mathbf{W}_i = (0, 0, W_{i2})$ with $W_{i2} \sim \text{Poisson}(\lambda_2)$.
 4. If $i = \mathbf{0}$ and we have $\mathbf{W}_i^T = (W_{i0}, 0, 0)^T$ with $W_{i0} > 0$ or if and we have $\mathbf{W}_i^T = (0, W_{i1}, W_{i2})^T$ with $W_{i1}, W_{i2} > 0$, then set $W_{ik} \sim \text{Poisson}(\lambda_k), k = 0, 1, 2$, and $\mathbf{Z}_i = (1, 0, 0, 0)^T$.
 5. Next, if $Y_{i1} > 0, Y_{i2} = 0$ is observed,
 - (a) with $\mathbf{W}_i^T = (W_{i0}, 0, 0)^T, W_{i0} > 0$ or $\mathbf{W}_i^T = (0, 0, W_{i2})^T, W_{i2} > 0$ then set $\mathbf{Z}_i = (0, 1, 0, 0)^T$, and next set $W_{i0} \sim \text{Bin}(Y_{i1}, \frac{\lambda_0}{(\lambda_0 + \lambda_1)})$, $W_{i1} = Y_{i1} - W_{i0}$ and finally set $W_{i2} \sim \text{Poisson}(\lambda_2)$.
 - (b) with $\mathbf{W}_i^T = (0, W_{i1}, 0)^T, W_{i1} > 0$, then set $\mathbf{Z}_i = (0, Z_{i1}, 0, 1 - Z_{i1})$ where $Z_{i1} \sim \text{Bernoulli}(\frac{p_{i1}}{(1 - p_{i0} - p_{i2})})$. If $Z_{i1} = 1$, then set $W_{i0} \sim \text{Bin}(Y_{i1}, \frac{\lambda_0}{(\lambda_0 + \lambda_1)})$, $W_{i1} = Y_{i1} - W_{i0}$ and finally set $W_{i2} \sim \text{Poisson}(\lambda_2)$. On the other hand if $Z_{i1} = 0$, then set $\mathbf{W}_i^T = (0, Y_{i1}, 0)^T$.
 6. Moreover, if $Y_{i1} = 0, Y_{i2} > 0$ is observed,
 - (a) with $\mathbf{W}_i^T = (W_{i0}, 0, 0)^T, W_{i0} > 0$ or $\mathbf{W}_i^T = (0, 0, W_{i2})^T, W_{i2} > 0$ then set $\mathbf{Z}_i = (0, 0, 1, 0)^T$, and next set $W_{i0} \sim \text{Bin}(Y_{i2}, \frac{\lambda_0}{(\lambda_0 + \lambda_2)})$, $W_{i2} = Y_{i2} - W_{i0}$ and finally set $W_{i1} \sim \text{Poisson}(\lambda_1)$.
 - (b) with $\mathbf{W}_i^T = (0, 0, W_{i2})^T, W_{i2} > 0$, then set $\mathbf{Z}_i = (0, 0, Z_{i2}, 1 - Z_{i2})$ where $Z_{i2} \sim \text{Bernoulli}(\frac{p_{i2}}{(1 - p_{i0} - p_{i1})})$. If $Z_{i2} = 1$, then set $W_{i0} \sim$

$\text{Bin}(Y_{i2}, \frac{\lambda_0}{(\lambda_0 + \lambda_2)})$, $W_{i2} = Y_{i2} - W_{i0}$ and finally set $W_{i1} \sim \text{Poisson}(\lambda_1)$.

On the other hand if $Z_{i2} = 0$, then set $\mathbf{W}_i^T = (0, 0, Y_{i2})^T$.

7. Finally, if $Y_{i1} > 0, Y_{i2} > 0$ is observed, then set $Z_{i0} = Z_{i1} = Z_{i2} = 0$, $Z_{i3} = 1$. Now generate non-negative discrete random variable W_{i0} where the probability mass function is given by $P(W_0 = w_0) \propto \{\frac{\lambda_0}{\lambda_1 \lambda_2}\}^{w_0} / w_0! (y_1 - w_0)! (y_2 - w_0)!$ for $w_0 = 0, \dots, \min(y_1, y_2)$. Hence, W_{i0} would have to be generated by inverting its cumulative distribution function (the cumulative distribution of a random variable is known to be a Uniform(0, 1) random variable). Then set $W_{i1} = Y_{i1} - W_{i0}$ and $W_{i2} = Y_{i2} - W_{i0}$.

Next sample $\mathbf{p}$ from Dirichlet($\sum_{i=1}^n Z_{i0} + b_1, \sum_{i=1}^n Z_{i1} + b_2, \sum_{i=1}^n Z_{i2} + b_3, n - (\sum_{i=1}^n \sum_{k=0}^2 Z_{ik}) + b_4$) using the completed data consisting of Z_{ik} and W_{ik} , obtained in the data augmentation step. Similarly, sample λ_k from a distribution Gamma($a_{1k} + \sum_{i=1}^n W_{ik}, a_{2k} + n$). With these sampled values of $\mathbf{p}$ and $\boldsymbol{\lambda}$, return to the “data augmentation step” and iterate the whole process until the convergence of the sampling distribution to the posterior distribution of $(\mathbf{p}, \boldsymbol{\lambda}, \mathbf{Z}, \mathbf{W})$. For most inferential procedures, only the sequence $\{\mathbf{p}^{(l)}, \boldsymbol{\lambda}^{(l)}\}$, $l = 1, \dots, L$ for some large L is saved. This sample can be used to make inference about $(\mathbf{p}$ and $\boldsymbol{\lambda})$ and any function of $(\mathbf{p}$ and $\boldsymbol{\lambda})$ for the MZIP model. The above sampling procedure can be easily implemented using WINBUGS.

Appendix C: Full conditional distribution of parameters in BZIP regression model

- $\pi(\beta_i | \boldsymbol{\theta}, \beta_i, \mathbf{Z}, \boldsymbol{\sigma}^2, \boldsymbol{\tau}^2, \mathbf{y}) \propto L(\mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma} | \mathbf{y}) \pi(\beta_i)$
- $\pi(\gamma_i | \boldsymbol{\theta}, \gamma_i, \mathbf{Z}, \boldsymbol{\sigma}^2, \boldsymbol{\tau}^2, \mathbf{y}) \propto L(\mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma} | \mathbf{y}) \pi(\gamma_i)$
- Probability of membership: We next proceed to calculate the distribution of $\mathbf{Z}_j | \mathbf{y}, \boldsymbol{\theta}$. It follows that

1. If $\mathbf{y}_j = \mathbf{0}$, then $\mathbf{Z}_j | \mathbf{y}, \boldsymbol{\theta} \sim \text{Multinomial}(1; \mathbf{q}_j)$ where $\mathbf{q}_j \propto (p_{0j}, p_{1j}e^{-(\lambda_{0j} + \lambda_{2j})}, p_{2j}e^{-(\lambda_{0j} + \lambda_{1j})}, (1 - \sum_{i=0}^2 p_{ij})e^{-\sum_{i=0}^2 \lambda_{ij}})$.
2. If $\mathbf{y}_j = (0, a)$, where a is non-zero, then $\mathbf{Z}_j = (0, Z_{1j}, 0, 1 - Z_{1j})^T$ such that $Z_{1j} | \mathbf{y}, \boldsymbol{\theta} \sim \text{Bernoulli}(q_j)$ where $q_j : (1 - q_j) \propto (p_{1j}e^{-(\lambda_{0j} + \lambda_{2j})}(\lambda_{0j} + \lambda_{2j})^{y_{2j}} : (1 - \sum_{i=0}^2 p_{ij})e^{-\sum_{i=0}^2 \lambda_{ij}})$.

3. If $\mathbf{y}_j = (a, 0)$, where a is non-zero, then $\mathbf{Z}_j = (0, 0, Z_{2j}, 1 - Z_{2j})^T$ such that $Z_{2j}|\mathbf{y}, \boldsymbol{\theta} \sim \text{Bernoulli}(q_j)$ where $q_j : (1 - q_j) \propto \left(p_{2j} e^{-(\lambda_{0j} + \lambda_{1j})} (\lambda_{0j} + \lambda_{1j})^{y_{1j}} : (1 - \sum_{i=0}^2 p_{ij}) e^{-\sum_{i=0}^2 \lambda_{ij}} \right)$.
 4. If $\mathbf{y}_j = (a, a')$, where a, a' are both non-zero then clearly, $\mathbf{Z}_j = (0, 0, 0, 1)^T$.
- The posterior distribution of the hyperparameter $\sigma_i^2|\mathbf{y}, \boldsymbol{\theta} - \sigma_i^2 \sim \text{Inverse Gamma}(a, b)$ where the shape parameter a is given by $a = 2 + k_b/2$ and $b = 1 + \beta_i^T(T)\beta_i/2$. Further, the posterior distribution of the hyperparameter $\tau_i^2|\mathbf{y}, \boldsymbol{\theta} - \tau_i^2 \sim \text{Inverse Gamma}(a, b)$ where the shape parameter a is given by $a = 2 + k_g/2$ and $b = 1 + \gamma_i^T(\mathbf{G}^T\mathbf{G})\gamma_i/2$. Here ‘-’ denotes the sign of removal of a parameter and k_b and k_g denote the length of vector β_i and γ_i respectively; $i=0, 1, 2$.

We use programming language to run the MCMC algorithm using our own codes. The posterior distribution of the regression coefficient vectors β_i and γ_i do not have any closed form and hence we employ the Metropolis Hastings algorithm within the Gibbs sampler with a Normal density proposal.

References

- Dagne. G. A. (2004). Hierarchical Bayesian Analysis of Correlated Zero-inflated Count Data. *Biometrical Journal* **46**, 653 – 663.
- Dempster, A. P. , Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood Estimation From Incomplete Data via the EM algorithm. *Journal of the Royal Statistical Society, Ser. B*, **39**, 1-39.
- Gamerman, D. (1997). Markov Chain Monte Carlo, *Chapman and Hall, London*.
- Gelfand, A. E. and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Stat. Assqc.* **85**, 398 – 409.
- Gelman, A., Carlin, J., Stern, H. and Rubin, D. B. (1995). Bayesian Data Analysis. *Chapman and Hall, New York*.
- Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat. Science* **7**, 457 – 511.
- Ghosh, S. K., Mukhopadhyay, P. and Lu, J. C. (2006). Bayesian analysis of zero-inflated regression models. *J. Stat. Planning and Inference*, **136**, 1360 – 1375.
- Gilks, W. R., Richardson, S. and Spiegelhalter, D. J.(1996). Markov Chain Monte Carlo in Practice. *Chapman and Hall, London*.

- Johnson, N. L. and Kotz, S. (1969). Distributions in Statistics: Discrete Distributions. *Houghton Mifflin, Boston*
- Lee, A., Wang, K., Carrivick, P., Yau, K. K.W. and Stevenson, M. R. (2005). Modelling bivariate count series with excess zeros. *Mathematical Biosciences*, **196**, 226 – 237.
- Li, C., Lu, J., Park, J., Kim, K., Brinkley, P. A. and Peterson, J. P. (1999). Multivariate zero-inflated Poisson models and their applications. *Technometrics*, **41**, 29–38.
- Marshall, A. W. and Olkin, I. (1985). A family of bivariate distributions generated by the bivariate Bernoulli distribution. *J. Amer. Stat. Assoc.*, **80**, 332 – 338.
- McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models, 2nd edition. *Chapman and Hall, London*.
- Plummer, M., Best, N., Cowles, K. and Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, **6**, 1, 7–11.
- Ridout, M., Demetrio, C. G. B. and Hinde, J. (1998). Models for count data with many zeros. *Proceedings of the XIXth International Biometric Conference*.
- Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Stat. Assoc.*, **82**, 528 – 550.
- Wang, K., Lee, A. H., Yau, K. K. W. and Carrivick, P. J. W. (2003). A bivariate zero-inflated Poisson regression model to analyze occupational injuries. *Accident Analysis and Prevention*, **35**, 625 – 629.