DOI: 10.1111/biom.12558

The LZIP: A Bayesian Latent Factor Model for Correlated Zero-Inflated Counts

Brian Neelon* **and Dongjun Chung**

Department of Public Health Sciences, Medical University of South Carolina, Charleston, South Carolina, U.S.A.
**email:* brian.neelon@musc.edu

Summary. Motivated by a study of molecular differences among breast cancer patients, we develop a Bayesian latent factor zero-inflated Poisson (LZIP) model for the analysis of correlated zero-inflated counts. The responses are modeled as independent zero-inflated Poisson distributions conditional on a set of subject-specific latent factors. For each outcome, we express the LZIP model as a function of two discrete random variables: the first captures the propensity to be in an underlying "at-risk" state, while the second represents the count response conditional on being at risk. The latent factors and loadings are assigned conditionally conjugate gamma priors that accommodate overdispersion and dependence among the outcomes. For posterior computation, we propose an efficient data-augmentation algorithm that relies primarily on easily sampled Gibbs steps. We conduct simulation studies to investigate both the inferential properties of the model and the computational capabilities of the proposed sampling algorithm. We apply the method to an analysis of breast cancer genomics data from The Cancer Genome Atlas.

Key words: Bayesian analysis; Cancer genomics; Data augmentation; Latent factor model; Negative multinomial distribution; Zero-inflated Poisson model

1. Introduction

Count data with an abundance of zeros arise commonly in many fields. In infectious disease epidemiology, for example, patients with no infections will have a response value of zero. Likewise, in cancer genomic studies investigating copy number variations (CNVs), cancer-relevant pathways with no CNVs will have a response of zero. When the number of zeros is greater than expected under a standard count model, the data are said to be "zero inflated" relative to the standard model. Zero-inflated count data often require flexible two-part mixture models to accommodate both the excess zeros and the heterogeneous distribution of nonzero counts. A common choice is the zero-inflated model (Lambert, 1992), which is a mixture of a point mass that accounts for the excess zeros and a count distribution for the remaining values. Because the point mass and count components each accommodate zeros, zero-inflated models implicitly partition the zeros into two types. The first type, often termed the "structural" zero, corresponds to individuals who are not at risk for an event, and therefore have no opportunity for a positive count. For example, in infectious disease studies, the structural class might comprise patients who are in a pre-infectious state. Similarly, in cancer genomic studies, the structural class might represent patients with dormant cancer-relevant pathways that produce no genes with CNVs.

The second type of zero, termed the "at-risk" zero, represents a latent class of individuals who are at risk for an event and, therefore, have the potential for a nonzero count. In cancer genomic studies, for instance, this latent class might include patients with "activated" cancer-related pathways that are at risk for producing CNVs. Note that zeros may still be observed for the at-risk class. In principle, however, these individuals have the potential for a nonzero response. In contrast, subjects in the structural class have no such opportunity.

Often interest lies in predicting whether individuals belong to the latent at-risk class, and if so, modeling the count response given that they are at risk. In our motivating cancer genomics study, for example, we are interested in first predicting whether certain cancer-related pathways are activated and thus at risk for producing genes with CNVs. Next, our goal is to model the number of genes with CNVs given pathway activation. The observed counts, however, provide only partial information about pathway activation. Pathways with CNVs necessarily belong to the at-risk class; otherwise, there is no opportunity for a positive count. However, when no CNVs are observed, there is no immediate information as to whether the observed zero is structural or at risk. A chief aim of the article, then, is to develop a novel Bayesian approach to predict pathway activation status even when a zero is observed, and to subsequently model the count response among pathways that are activated.

A second aim of the article is to develop a computationally efficient multivariate model for correlated zero-inflated outcomes. While there is a rich literature on univariate zero-inflated models (see, e.g., Neelon et al., 2010), less attention has been devoted to multivariate zero-inflated outcomes, particularly within a computationally tractable Bayesian framework. Majumdar and Gries (2010) introduced a Bayesian bivariate ZIP model implemented using the software package WinBUGS (Lunn et al., 2000). More recently, Fox (2013) proposed a hybrid Bayesian zero-inflated model comprising a univariate model for the binary at-risk probability and a joint random effects model for multiple count responses conditional on being at risk. Metropolis steps were used for posterior computation.

Here, we develop a computationally convenient Bayesian latent factor model for the analysis of multivariate zero-inflated counts. We start by recognizing that zero-inflated models can be expressed as mixtures of two discrete random variables: a binary at-risk indicator variable, and a count variable that describes the response conditional on being at risk. In many cases, the binary and count variables are correlated. For example, biological pathways with higher activation probabilities may also have higher CNV counts given that they are activated. It is therefore desirable to build this association into the model. To do so, we make use of the latent Poisson representation proposed by Dunson and Herring (2005) for mixed discrete outcomes. Next, we introduce a set of subject-specific latent factors to account for marginal dependence between 1) the binary and count components for each zero-inflated outcome, and 2) the different zero-inflated responses for a given subject. The former can be viewed as a "within-outcome" association, and the latter as a "between-outcome" association. The latent factors and loadings are assigned conditionally conjugate gamma priors that admit closed-form full conditionals while accounting for overdispersion. For categorical covariates (e.g., cancer stage), the regression parameters also have conjugate full conditionals, resulting in computationally efficient posterior sampling. We conduct simulation studies to evaluate the properties of the model, and illustrate the approach using data from The Cancer Genome Atlas (TCGA) for breast-invasive carcinoma (Cancer Genome Atlas Network, 2012).

2. Latent Variable Representation of the Univariate ZIP Model

We begin by considering the univariate ZIP model given by

$$\Pr(Y=0) = (1-\phi) + \phi e^{-\mu}, \quad 0<\phi<1$$

$$\Pr(Y=y) = \phi\frac{\mu^y e^{-\mu}}{y!}, \quad \mu>0;\ y=1,2,\ldots;\ \text{or, more compactly,}$$

$$Y \sim (1-\phi)\mathbb{1}_{(W=0)} + \phi \text{Poi}(\mu)\mathbb{1}_{(W=1)}, \tag{1}$$

where Y is a count outcome, $\mathbb{1}_{(\cdot)}$ is the indicator function, W is a latent "at-risk" indicator such that with probability $1-\phi$, $Y=0$ and with probability ϕ, Y is drawn from a Poisson distribution with mean μ. Thus, ϕ denotes the probability of being in the at-risk class, while μ denotes the mean count among the at-risk population.

We can extend model (1) to the regression setting by modeling ϕ and μ as a function of covariates. Let Y_i denote the count response for the i-th subject ($i=1,\ldots,n$). Choosing a complementary log-log link for ϕ and a log link for μ yields the following ZIP model:

$$\begin{aligned} Y_i &\sim (1-\phi_i)\mathbb{1}_{(W_i=0)} + \phi_i \text{Poi}(\mu_i)\mathbb{1}_{(W_i=1)} \\ \text{cloglog}(\phi_i) &= \text{cloglog}[\Pr(W_i=1)] = \boldsymbol{x}_i'\boldsymbol{\beta}_1 \\ \log(\mu_i) &= \log[\text{E}(Y_i|W_i=1)] = \boldsymbol{x}_i'\boldsymbol{\beta}_2, \end{aligned} \tag{2}$$

where $\boldsymbol{x}_i$ is a $p\times 1$ vector of subject-specific covariates and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are corresponding regression coefficients for the binary and count components of the model. The complementary log-log link has been previously used to model zero-inflated data (Heilbron, 1994) and has an appealing latent Poisson representation. Specifically, setting $\text{cloglog}(\phi_i) = \boldsymbol{x}_i'\boldsymbol{\beta}_1$ implies that $\phi_i = 1-\exp(-\mu_{i1})$, where $\mu_{i1} = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_1)$. This is equivalent to the probability that a Poisson variable with mean μ_{i1} exceeds 0. If we let Z_{i1} denote this latent Poisson variable, it follows that $W_i=1$ (and hence subject i is at risk) if and only if $Z_{i1}>0$. Thus, Z_{i1} can be viewed as a latent count variable reflecting the propensity to be at risk. Likewise, we can introduce a second latent Poisson variable Z_{i2} with mean $\mu_{i2} = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_2)$ such that $Z_{i2} = (Y_i|Z_{i1}>0)$. This represents the count conditional on being at risk. In this way, we recast the classic ZIP model intuitively in terms of two latent count processes: one governing the propensity to be at risk and the other governing the count outcome given risk. Expressed in terms of Z_{i1} and Z_{i2}, our model becomes:

$$\begin{aligned} Y_i &\sim (1-\phi_i)\mathbb{1}_{(Z_{i1}=0)} + \phi_i \text{Poi}(z_{i2};\mu_{i2})\mathbb{1}_{(Z_{i1}>0)} \\ \phi_i &= \Pr(W_i=1) = \Pr(Z_{i1}>0) = 1-\exp(-\mu_{i1}) \\ \mu_{i1} &= \text{E}(Z_{i1}) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_1) \\ \mu_{i2} &= \text{E}(Z_{i2}) = \text{E}(Y_i|Z_{i1}>0) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_2). \end{aligned} \tag{3}$$

This is analogous to the latent Poisson model introduced by Dunson and Herring (2005) for mixed discrete outcomes. Because $Z_{i2} = (Y_i|Z_{i1}>0)$, Z_{i2} and Y_i are distinct random variables. When $Z_{i1}>0$ (subject i is at risk), $Z_{i2}=Y_i$ and hence Z_{i2} is observed. However, when $Z_{i1}=0$ (subject i is not at risk), $Y_i=0$, whereas Z_{i2} can take any non-negative integer value. Here, Z_{i2} can be viewed as the "potential" count that would have been observed had subject i been at risk. In this respect, model (3) has the flavor of a Heckman selection model (Heckman, 1979). By conditioning on Z_{i1} and Z_{i2} as part of a data-augmented Markov chain Monte Carlo (MCMC) algorithm, we can gather information about a subject's class allocation, which can in turn be used to estimate the regression parameters $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$.

In many cases, it is reasonable to assume that Z_{i1} and Z_{i2} are positively correlated. For example, in our cancer genomics application, we might expect patients with increased risk of pathway activation to also have more genes with CNVs given activation. We can accommodate this association by modeling μ_{i1} and μ_{i2} as a multiplicative function of subject-specific latent factors, $\boldsymbol{\xi}_i$, resulting in a latent factor ZIP (LZIP) model:

$$\begin{aligned} Y_i|\boldsymbol{\xi}_i &\sim (1-\phi_i)\mathbb{1}_{(Z_{i1}=0)} + \phi_i \text{Poi}(z_{i2};\mu_{i2})\mathbb{1}_{(Z_{i1}>0)} \\ \phi_i &= \Pr(W_i=1|\boldsymbol{\xi}_i) = \Pr(Z_{i1}>0|\boldsymbol{\xi}_i) = 1-\exp(-\mu_{i1}) \\ \mu_{i1} &= \text{E}(Z_{i1}|\boldsymbol{\xi}_i) = \boldsymbol{\lambda}_1'\boldsymbol{\xi}_i\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_1) \\ \mu_{i2} &= \text{E}(Z_{i2}|\boldsymbol{\xi}_i) = \text{E}(Y_i|Z_{i1}>0,\boldsymbol{\xi}_i) = \boldsymbol{\lambda}_2'\boldsymbol{\xi}_i\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_2), \end{aligned} \tag{4}$$

where $\boldsymbol{\xi}_i = (\xi_{i1},\ldots,\xi_{iL})'$ is an $L\times 1$ vector of subject-specific latent factors, with $\xi_{il}>0$ for all l to ensure $\mu_{ik}>0$ $(k=1,2)$, and $\boldsymbol{\lambda}_k = (\lambda_{k1},\ldots,\lambda_{kL})'$ is an $L\times 1$ vector of loadings for the k-th component of model (4), again with $\lambda_{kl} > 0$ for all l. For now we assume the number of latent factors, L, is fixed; below, we discuss practical strategies for accommodating unknown L. Further, for identifiability purposes, we assume that $\xi_{il} \overset{\text{ind}}{\sim} \text{Ga}(\alpha, \alpha)$, where $\text{Ga}(a, b)$ denotes a gamma distribution with shape parameter a and rate parameter b. In principle, any value $\alpha > 0$ is permitted. However, as noted by Dunson and Herring (2005), placing diffuse gamma priors on both the latent factors and the loadings can lead to poor posterior convergence. Following their recommendation, we anchor ξ_{il} by setting $\alpha = 1$, thus allowing the factor loadings to be unconstrained.

The latent factors, $\boldsymbol{\xi}_i$, account for between-subject heterogeneity potentially due to unmeasured subject-level confounding. Because $\text{E}(\xi_{il}) = 1$, it follows that the marginal (or population-average) mean of Z_{ik} is

$$\text{E}(Z_{ik}) = \left(\sum_{l=1}^{L} \lambda_{kl}\right) \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_k) = \exp(\nu_k + \boldsymbol{x}_i'\boldsymbol{\beta}_k), \;\; k = 1, 2, \quad (5)$$

where $\nu_k = \log\left(\sum_l \lambda_{kl}\right)$. Thus, ν_k functions as a population-average intercept for component k; for this reason, we exclude fixed intercept terms from $\boldsymbol{x}_i$ and $\boldsymbol{\beta}_k$ to ensure identifiability. An appealing feature of the model is that the regression coefficients, β_{kh} ($h = 1, \ldots, p$), have both subject-specific and population-average interpretations as the incremental change in the log mean of Z_{ik} per unit change in covariate x_{ih}. Consequently, the regression coefficients in the LZIP model have same interpretation as in the conventional ZIP model.

Assuming $\alpha = 1$, the marginal probability that subject i is at risk is

$$\Pr(W_i = 1) = \Pr(Z_{i1} > 0) = 1 - \prod_{l=1}^{L} \left[\frac{1}{1 + \lambda_{1l} \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_1)}\right]. \quad (6)$$

The derivation is provided in Web Appendix B. Furthermore, when $\alpha = 1$, $\text{Var}(\xi_{il}) = 1$, and hence the marginal variance of Z_{ik} is

$$\text{Var}(Z_{ik}) = \text{E}(Z_{ik}) + \left(\sum_{l=1}^{L} \lambda_{kl}^2\right) \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_k)^2. \quad (7)$$

As a result, Z_{ik} is overdispersed relative to the Poisson distribution. Finally, with $\alpha = 1$, the marginal covariance of Z_{i1} and Z_{i2} is

$$\text{Cov}(Z_{i1}, Z_{i2}) = \left(\sum_{l=1}^{L} \lambda_{1l}\lambda_{2l}\right) \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_1 + \boldsymbol{x}_i'\boldsymbol{\beta}_2). \quad (8)$$

The derivation is provided in Web Appendix A. Thus, while Z_{i1} and Z_{i2} are conditionally independent, marginally they are positively correlated, which in turn induces a positive association between the binary and count components of the LZIP model.

3. Extension to Multiple Outcomes

It is straightforward to extend the model to multiple zero-inflated outcomes. Assuming J outcomes per subject, the multivariate LZIP model is given by

$$\begin{aligned}
Y_{ij}|\boldsymbol{\xi}_i &\sim (1 - \phi_{ij})\mathbb{1}_{(Z_{ij1}=0)} + \phi_{ij}\text{Poi}(z_{ij2}; \mu_{ij2})\mathbb{1}_{(Z_{ij1}>0)} \\
\phi_{ij} &= \Pr(W_{ij} = 1|\boldsymbol{\xi}_i) = \Pr(Z_{ij1} > 0|\boldsymbol{\xi}_i) = 1 - \exp(-\mu_{ij1}) \\
\mu_{ij1} &= \text{E}(Z_{ij1}|\boldsymbol{\xi}_i) = \boldsymbol{\lambda}_{j1}'\boldsymbol{\xi}_i \exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta}_{j1}) \\
\mu_{ij2} &= \text{E}(Z_{ij2}|\boldsymbol{\xi}_i) = \text{E}(Y_{ij}|Z_{ij1} > 0, \boldsymbol{\xi}_i) = \boldsymbol{\lambda}_{j2}'\boldsymbol{\xi}_i \exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta}_{j2}), \quad (9)
\end{aligned}$$

where Y_{ij} denotes the j-th zero-inflated count for subject i, Z_{ij1} is the latent Poisson associated with the j-th at-risk indicator W_{ij}, Z_{ij2} is the latent count for subject i and outcome j conditional on being at risk for outcome j, and $\boldsymbol{\lambda}_{jk} = (\lambda_{jk1}, \ldots, \lambda_{jkL})'$ for outcome j and component $k = 1, 2$. Because Z_{ij1} varies across j, subjects can be at risk for some outcomes and not others. In our case study, for example, some cancer pathways may be activated while others are not. As in models (3) and (4), if we condition on $Z_{ij1} > 0$, so that subject i is at risk for outcome j, then $Z_{ij2} = Y_{ij}$ and hence Z_{ij2} is observed. Conversely, when $Z_{ij1} = 0$, subject i is not at risk for outcome j, and hence Z_{ij2} is not observed. In this case, Z_{ij2} represents a "potential" count that would have been observed had the subject been at risk for outcome j.

From model (9), we can establish the following useful proposition:

Proposition 1. *For $i = 1, \ldots, n$ and $l = 1, \ldots, L$, let $\xi_{il} \overset{ind}{\sim} \text{Ga}(\alpha, \alpha)$. Then, for outcomes $j = 1, \ldots, J$ and model components $k = 1, 2$, the joint marginal distribution of the $2J$ random variables, $(Z_{i11}, \ldots, Z_{iJ2})$, is product negative multinomial (Guo, 1996). That is,*

$$\text{p}(z_{i11}, \ldots, z_{iJ2}) \sim \prod_{l=1}^{L} \text{NegMult}(\alpha, \pi_{i11l}, \ldots, \pi_{iJ2l}), \quad (10)$$

where $\pi_{ijkl} = \eta_{ijkl}/(\eta_{il} + \alpha)$, $\eta_{ijkl} = \lambda_{jkl} \exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta}_{jk})$, and $\eta_{il} = \sum_{j,k} \eta_{ijkl}$.

Proposition 1 allows us to derive several population-average quantities of interest. For example, we can derive the population-average probability that subject i is at risk for at least one of the J outcomes, denoted ψ_i:

$$\begin{aligned}
\psi_i &= 1 - \Pr(W_{i1} = 0 \cap \cdots \cap W_{iJ} = 0) \\
&= 1 - \Pr(Z_{i11} = 0 \cap \cdots \cap Z_{iJ1} = 0) \\
&= 1 - \prod_{l=1}^{L} \left[\frac{\alpha}{\alpha + \sum_{j=1}^{J} \lambda_{j1l} \exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta}_{j1})}\right]^{\alpha}, \quad (11)
\end{aligned}$$

which reduces to expression (6) for $J = \alpha = 1$. In our case study, for example, we can use expression (11) to predict the population-average probability that a subject has at least one activated cancer-related pathway. Moreover, the population-average mean count among those at risk for outcome j is given by

$$\mathrm{E}(Z_{ij2}) = \sum_{l=1}^{L} \eta_{ij2l} = \left(\sum_{l=1}^{L} \lambda_{j2l}\right) \exp(\boldsymbol{x}_{ij}' \boldsymbol{\beta}_{j2}), \tag{12}$$

which is analogous to expression (5) for a single outcome $J = 1$. The proofs of (10)–(12) are provided in Web Appendix B.

4. Bayesian Inference

4.1. *Prior Distributions*

For inference, we adopt a Bayesian approach and assign priors to the remaining LZIP parameters—namely, $\boldsymbol{\lambda}_{jk} = (\lambda_{jk1}, \ldots, \lambda_{jkL})'$ and $\boldsymbol{\beta}_{jk} = (\beta_{jk1}, \ldots, \beta_{jkp})'$ for each outcome $j = 1, \ldots, J$ and component $k = 1, 2$ in model (9). For each λ_{jkl} $(l = 1, \ldots, L)$, we assign an independent, conditionally conjugate $\mathrm{Ga}(a, b)$ prior, where the choice of a and b may vary by outcome and component depending on the availability of prior information. Our default choice is $a = b = 0.001$ for all j and k. Our experience suggests that values less than or equal to 1 yield similar factor loading estimates, although this will become more sensitive as the sample size decreases and the percentage of zeros increases. For categorical predictors, we assign independent $\mathrm{Ga}(c, d)$ priors to the exponentiated regression coefficients, $\exp(\beta_{jkh})$ $(h = 1, \ldots, p)$, as these are conditionally conjugate for our model. Our default choice for c and d is again 0.001. As part of the simulation study presented in Section 5, we explore sensitivity to the choice of hyperparameters. For continuous predictors, we assign independent normal priors to the regression coefficients and use Metropolis–Hastings steps to sample from the posterior, as described in the following sub-section.

4.2. *MCMC via Data Augmentation*

Posterior computation proceeds via MCMC with data augmentation steps to facilitate sampling. Below, we summarize the algorithm; derivations are given in Web Appendix C.

(i) *Data Augmentation Step 1.* For all (i, j), update the latent Poisson random variables, Z_{ij1} and Z_{ij2}, from their closed-form full conditionals. As described in Web Appendix A, the update for Z_{ij1} depends on the observed response y_{ij} and the current value of Z_{ij2}, whereas the update for Z_{ij2} depends on the current value of Z_{ij1}.

(ii) *Data Augmentation Step 2.* Following Dunson and Herring (2005), we can write Z_{ijk} as a sum of L independent Poisson random variables:

$$\begin{aligned} Z_{ijk} &= \sum_{l=1}^{L} Z_{ijkl}, \text{ where} \\ Z_{ijkl} &\overset{ind}{\sim} \mathrm{Poi}(\mu_{ijkl}) \\ \mu_{ijkl} &= \lambda_{jkl}\xi_{il} \exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta}_{jk}), \;\; j = 1, \ldots, J; \\ k &= 1, 2; \; l = 1, \ldots, L. \end{aligned} \tag{13}$$

Next, for all (i, j, k), update $\{Z_{ijk1}, \ldots, Z_{ijkL}\}$ jointly from a multinomial distribution.

(iii) For all i and l, update ξ_{il}, from its gamma full conditional.

(iv) For all (j, k, l), update λ_{jkl} from its gamma full conditional.

(v) The update for β_{jkh} $(j = 1, \ldots, J;\ k{=}1, 2;\ h{=}1, \ldots, p)$ depends on whether the corresponding covariate, x_{ijh}, is discrete or continuous. For categorical predictors, a $\mathrm{Ga}(c, d)$ prior on $\exp(\beta_{jkh})$ is conditionally conjugate, allowing for straightforward Gibbs sampling from a gamma full conditional. If x_{ijh} is ordinal or continuous, a random-walk Metropolis–Hastings step is used to update β_{jkh}.

Convergence is assessed using trace plots and suitable MCMC diagnostics, such as Geweke's z-score (Geweke, 1992). Because the sampler relies almost entirely on easily sampled Gibbs steps, it tends to converge rapidly to a stationary joint posterior distribution.

A computational challenge for Bayesian latent variable models is "label switching," in which the latent factors and loadings are assigned different labels, $l = 1, \ldots, L$, during the course of the MCMC run. In the presence of label switching, the posterior summaries of the loadings are uninterpretable. To reconcile this issue, we apply the equivalence classes representatives (ECR) approach proposed by Papastamoulis and Iliopoulos (2010). Note that label switching does not pose a problem for estimating the vector of regression coefficients, $\boldsymbol{\beta}_j$, since its full conditional depends only on the linear combination, $\boldsymbol{\lambda}_j'\boldsymbol{\xi}_i$, and is therefore marginalized with respect to the factor labels. Likewise, label switching is not a concern for models with a single latent factor (i.e., $L = 1$).

Our primary focus in this article is to develop a flexible LZIP model assuming a fixed number of latent factors, L. However, we can allow for unknown L by fitting models with varying values (e.g., $L = 0, 1, 2$) and applying a suitable model comparison measure to select the optimal model. A popular Bayesian measure is the "widely applicable information criterion" or WAIC (Watanabe, 2010), which combines a measure of the predictive accuracy of the fitted model with a penalty for model complexity. In our cancer application, we also conduct posterior predictive checks (Gelman et al., 1996) to further assess model fit. All models described below can be easily programmed using R software (R Core Team, 2015).

5. Illustrative Examples

5.1. *Model 1: Univariate ZIP with No Latent Factors*

To evaluate the validity of the proposed sampling approach, we simulated data from three models of increasing complexity. First, we generated data from a conventional, univariate ZIP model with no latent factors and compared the results to maximum likelihood estimates (MLEs) obtained using the R package `pscl` (Jackman, 2015). The aim was to ensure that the proposed Bayesian approach with weakly informative priors yielded similar regression estimates and uncertainty intervals to those obtained under a classical, frequentist approach. To do so, we generated data according to the

Table 1

Parameter estimates and 95% intervals for simulation study 1: conventional univariate ZIP model with no latent factors

					Estimate (95% Interval)[b]	
n	Percent zeros	Gamma hyperparameters[a]	Parameter	Simulated value	Proposed model	MLE
500	40	$(0.001, 0.001)$	β_{10}	-0.25	$-0.15\,(-0.34, -0.05)$	$-0.15\,(-0.34, -0.03)$
			β_{11}	0.75	$0.67\,(0.34, 1.03)$	$0.64\,(0.34, 1.02)$
			β_{20}	1.00	$1.01\,(0.90, 1.12)$	$1.01\,(0.90, 1.12)$
			β_{21}	-0.50	$-0.57\,(-0.75, -0.39)$	$-0.56\,(-0.74, -0.38)$
	40	$(1, 1)$	β_{10}	-0.25	$-0.13\,(-0.33, -0.05)$	—[c]
			β_{11}	0.75	$0.63\,(0.32, 0.96)$	—
			β_{20}	1.00	$1.00\,(0.89, 1.11)$	—
			β_{21}	-0.50	$-0.55\,(-0.73, -0.38)$	—
	70	$(0.001, 0.001)$	β_{10}	-0.50	$-0.56\,(-0.77, -0.36)$	$-0.56\,(-0.78, -0.36)$
			β_{11}	-0.50	$-0.58\,(-1.02, -0.13)$	$-0.61\,(-1.07, -0.13)$
			β_{20}	1.00	$0.90\,(0.76, 1.03)$	$0.90\,(0.76, 1.04)$
			β_{21}	-1.00	$-0.81\,(-1.20, -0.44)$	$-0.78\,(-1.14, -0.42)$
	70	$(1, 1)$	β_{10}	-0.50	$-0.56\,(-0.78, -0.35)$	—
			β_{11}	-0.50	$-0.60\,(-1.04, -0.16)$	—
			β_{20}	1.00	$0.88\,(0.75, 1.02)$	—
			β_{21}	-1.00	$-0.77\,(-1.14, -0.41)$	—
5000	40	$(0.001, 0.001)$	β_{10}	-0.25	$-0.25\,(-0.31, -0.19)$	$-0.25\,(-0.31, -0.19)$
			β_{11}	0.75	$0.80\,(0.69, 0.90)$	$0.79\,(0.69, 0.91)$
			β_{20}	1.00	$1.02\,(0.99, 1.06)$	$1.02\,(0.99, 1.06)$
			β_{21}	-0.50	$-0.53\,(-0.59, -0.47)$	$-0.53\,(-0.59, -0.48)$
	40	$(1, 1)$	β_{10}	-0.25	$-0.25\,(-0.31, -0.19)$	—
			β_{11}	0.75	$0.79\,(0.69, 0.91)$	—
			β_{20}	1.00	$1.02\,(0.99, 1.06)$	—
			β_{21}	-0.50	$-0.53\,(-0.59, -0.48)$	—
	70	$(0.001, 0.001)$	β_{10}	-0.50	$-0.54\,(-0.60, -0.48)$	$-0.54\,(-0.60, -0.48)$
			β_{11}	-0.50	$-0.39\,(-0.54, -0.25)$	$-0.40\,(-0.54, -0.26)$
			β_{20}	1.00	$1.03\,(0.99, 1.06)$	$1.03\,(0.99, 1.06)$
			β_{21}	-1.00	$-1.07\,(-1.19, -0.96)$	$-1.06\,(-1.18, -0.94)$
	70	$(1, 1)$	β_{10}	-0.50	$-0.54\,(-0.61, -0.48)$	—
			β_{11}	-0.50	$-0.39\,(-0.53, -0.24)$	—
			β_{20}	1.00	$1.02\,(0.99, 1.06)$	—
			β_{21}	-1.00	$-1.07\,(-1.19, -0.95)$	—

[a] Gamma hyperparameters for exponentiated regression coefficients.

[b] Posterior means and 95% credible intervals for the Bayesian model; MLEs and 95% confidence intervals for the frequentist model fit using the `pscl` package.

[c] Results unchanged from previous four lines.

following ZIP model:

$$
\begin{aligned}
Y_i &\sim (1-\phi_i)\mathbb{1}_{(W_i=0)} + \phi_i \mathrm{Poi}(\mu_i)\mathbb{1}_{(W_i=1)} \\
\mathrm{cloglog}(\phi_i) &= \boldsymbol{x}_i'\boldsymbol{\beta}_1 = \beta_{10} + x_i\beta_{11} \\
\log(\mu_i) &= \boldsymbol{x}_i'\boldsymbol{\beta}_2 = \beta_{20} + x_i\beta_{21}, \qquad (14)
\end{aligned}
$$

where x_i is a Bernoulli(0.5) covariate. We considered two sample sizes, $n = 500$ and $n = 5000$, and two sets of values for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$: $\boldsymbol{\beta}_1 = (-0.25, 0.75)'$, $\boldsymbol{\beta}_2 = (1, -0.50)'$; and $\boldsymbol{\beta}_1 = (-0.5, -0.5)'$, $\boldsymbol{\beta}_2 = (1, -1)'$. These corresponded to approximately 40% and 70% zeros, respectively, a range of values commonly observed in practice. For the Bayesian approach, we assigned a conjugate $\mathrm{Ga}(0.001, 0.001)$ for the exponentiated regression coefficients. As a sensitivity analysis, we also considered $\mathrm{Ga}(1, 1)$ priors. We ran the Gibbs sampler for 10,000 iterations, discarding the first 5000 as burn-in. Web Figure 1 presents trace plots for the regression parameters corresponding to $n = 5000$, 70% zeros, and gamma hyperparameter values of 0.001. The trace plots indicate almost immediate convergence.

Table 1 presents the regression coefficients and 95% intervals under the two estimation methods. For all settings, the results were comparable under the two approaches, with 95% intervals encompassing the simulated values. The Bayesian

Table 2

Posterior means and 95% credible intervals (CrIs) for simulation study 2: univariate LZIP model with a single latent factor

n	Percent zeros	Gamma hyperparameters[a]	Parameter	Simulated value	Posterior mean (95% CrI)
500	40%	(0.001, 0.001)	λ_1	1.25	1.09 (0.71, 1.64)
			λ_2	1.50	1.40 (1.13, 1.71)
			β_1	1.50	1.47 (0.83, 2.07)
			β_2	1.00	1.05 (0.82, 1.29)
		(1, 1)	λ_1	1.25	1.15 (0.77, 1.63)
			λ_2	1.50	1.42 (1.16, 1.72)
			β_1	1.50	1.26 (0.74, 1.66)
			β_2	1.00	1.03 (0.78, 1.26)
	70%	(0.001, 0.001)	λ_1	0.50	0.51 (0.37, 0.71)
			λ_2	1.50	1.38 (1.06, 1.75)
			β_1	−0.50	−0.63 (−1.11, −0.19)
			β_2	0.75	0.61 (0.29, 0.93)
		(1, 1)	λ_1	0.50	0.51 (0.36, 0.73)
			λ_2	1.50	1.41 (1.10, 1.78)
			β_1	−0.50	−0.61 (−1.10, −0.15)
			β_2	0.75	0.57 (0.22, 0.89)
5000	40%	(0.001, 0.001)	λ_1	1.25	1.26 (1.13, 1.41)
			λ_2	1.50	1.52 (1.43, 1.62)
			β_1	1.50	1.46 (1.25, 1.74)
			β_2	1.00	1.02 (0.95, 1.10)
		(1, 1)	λ_1	1.25	1.42 (1.26, 1.62)
			λ_2	1.50	1.47 (1.38, 1.56)
			β_1	1.50	1.35 (1.17, 1.55)
			β_2	1.00	0.97 (0.90, 1.04)
	70%	(0.001, 0.001)	λ_1	0.50	0.52 (0.47, 0.58)
			λ_2	1.50	1.43 (1.33, 1.55)
			β_1	−0.50	−0.58 (−0.72, −0.44)
			β_2	0.75	0.82 (0.71, 0.93)
		(1, 1)	λ_1	0.50	0.52 (0.47, 0.58)
			λ_2	1.50	1.43 (1.34, 1.53)
			β_1	−0.50	−0.58 (−0.73, −0.44)
			β_2	0.75	0.82 (0.73, 0.91)

[a] Gamma hyperparameters for factor loadings and exponentiated regression coefficients.

estimates were also robust to choice of hyperparameters. These results suggest that even for relatively small sample sizes with a large percentage of zeros, the proposed method provides a suitable alternative to frequentist estimation for ZIP models with a complementary log-log link. The Bayesian approach might prove especially attractive in cases where prior information can be incorporated to improve inference, as the frequentist approach does not allow for this possibility.

5.2. *Model 2: Univariate LZIP with One Latent Factor*

For the second simulation, we generated from a univariate LZIP with a single latent factor:

$$\begin{aligned} Y_i|\xi_i &\sim (1-\phi_i)\mathbb{1}_{(Z_{i1}=0)} + \phi_i \mathrm{Poi}(z_{i2};\mu_{i2})\mathbb{1}_{(Z_{i1}>0)} \\ \phi_i &= \Pr(Z_{i1} > 0|\xi_i) = 1 - \exp(-\mu_{i1}) \\ \mu_{i1} &= \mathrm{E}(Z_{i1}|\xi_i) = \lambda_1 \xi_i \exp(x_i\beta_1) \\ \mu_{i2} &= \mathrm{E}(Z_{i2}|\xi_i) = \lambda_2 \xi_i \exp(x_i\beta_2), \end{aligned} \qquad (15)$$

where x_i was again simulated from a Bernoulli(0.5) distribution. Recall, for identifiability, model (15) excludes a fixed intercept, as $\log(\lambda_k)$ $(k = 1, 2)$ functions in this role.

As in simulation study 1, we considered two sample sizes, $n = 500$ and $n = 5000$, as well as two sets of values for β_1 and β_2 corresponding to approximately 40% and 70% zeros (see Table 2 for simulated values). We generated the latent factors according to a Ga(1, 1) density. To assess sensitivity to choice of hyperparameters, we assigned conditionally conjugate Ga(0.001, 0.001) and Ga(1, 1) priors to both the factor loadings and to the exponentiated regression coefficients. We ran the Gibbs sampler for 10,000 iterations, discarding the first 5000 as burn-in. Trace plots indicated rapid convergence (Web Figure 2).

The posterior estimates are shown in Table 2. The posterior means closely approximated the simulated values, with 95% credible intervals (CrIs) encompassing the true values. The results were generally robust to choice of prior distribution.

Table 3

Posterior means and 95% *credible intervals (CrIs) for simulation study 3: bivariate LZIP model with two latent factors. Results are for simulation with* 40% *zeros and* $Ga(0.001, 0.001)$ *priors for both the factor loadings and the exponentiated regression coefficients for the binary predictor,* x_{ij1}.

n	Outcome	Model component	Parameter	Simulated value	Posterior mean (95% CrI)
500	Y_1	Binary	λ_{111}	2.50	2.56 (1.39, 4.81)
			λ_{112}	0.00	0.00 (0.00, 0.00)[a]
			β_{111}^{b}	1.00	0.89 (0.17, 1.60)
			β_{112}^{c}	0.50	0.48 (0.31, 0.64)
		Count	λ_{121}	0.00	0.00 (0.00, 0.00)
			λ_{122}	2.50	2.81 (2.16, 3.59)
			β_{121}	0.25	0.12 (−0.16, 0.41)
			β_{122}	−0.25	−0.20 (−0.26, −0.14)
	Y_2	Binary	λ_{211}	2.50	2.22 (1.22, 3.86)
			λ_{212}	0.00	0.00 (0.00, 0.00)
			β_{211}	0.75	0.76 (0.08, 1.43)
			β_{212}	0.25	0.10 (−0.08, 0.30)
		Count	λ_{221}	0.00	0.00 (0.00, 0.00)
			λ_{222}	2.50	2.93 (2.24, 3.79)
			β_{221}	0.50	0.34 (0.05, 0.63)
			β_{222}	−0.50	−0.46 (−0.53, −0.40)
5000	Y_1	Binary	λ_{111}	2.50	2.42 (1.98, 3.00)
			λ_{112}	0.00	0.00 (0.00, 0.00)
			β_{111}	1.00	1.13 (0.89, 1.35)
			β_{112}	0.50	0.56 (0.51, 0.62)
		Count	λ_{121}	0.00	0.00 (0.00, 0.00)
			λ_{122}	2.50	2.67 (2.47, 2.91)
			β_{121}	0.25	0.17 (0.07, 0.26)
			β_{122}	−0.25	−0.26 (−0.28, −0.24)
	Y_2	Binary	λ_{211}	2.50	2.31 (1.89, 2.79)
			λ_{212}	0.00	0.00 (0.00, 0.00)
			β_{211}	0.75	0.76 (0.53, 0.97)
			β_{212}	0.25	0.26 (0.21, 0.32)
		Count	λ_{221}	0.00	0.00 (0.00, 0.00)
			λ_{222}	2.50	2.56 (2.36, 2.78)
			β_{221}	0.50	0.48 (0.38, 0.57)
			β_{222}	−0.50	−0.51 (−0.53, −0.49)

[a] Estimates rounded to two decimal places.

[b] Regression coefficients for binary predictor, x_{ij1}, updated using conjugate Gibbs steps.

[c] Regression coefficients for continuous predictor, x_{ij2}, updated using random-walk Metropolis–Hastings steps.

5.3. *Model 3: Bivariate LZIP with Two Latent Factors*

For the final simulation, we generated data from the following bivariate LZIP model:

$$
\begin{aligned}
Y_{ij}|\boldsymbol{\xi}_i &\sim (1-\phi_{ij})\mathbb{1}_{(Z_{ij1}=0)} + \phi_{ij}\mathrm{Poi}(z_{ij2};\mu_{ij2})\mathbb{1}_{(Z_{ij1}>0)} \\
\phi_{ij} &= \Pr(Z_{ij1}>0|\boldsymbol{\xi}_i) = 1-\exp(-\mu_{ij1}) \\
\mu_{ij1} &= \mathrm{E}(Z_{ij1}|\boldsymbol{\xi}_i) = \boldsymbol{\lambda}'_{j1}\boldsymbol{\xi}_i\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{j1}) \\
\mu_{ij2} &= \mathrm{E}(Z_{ij2}|\boldsymbol{\xi}_i) = \boldsymbol{\lambda}'_{j2}\boldsymbol{\xi}_i\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{j2}), \\
i &= 1,\ldots,n;\ j=1,2, \qquad (16)
\end{aligned}
$$

where $\boldsymbol{\xi}_i = (\xi_{i1}, \xi_{i2})'$ is a 2×1 vector of latent factors, $\boldsymbol{\lambda}_{jk} = (\lambda_{jk1}, \lambda_{jk2})'$ are the loadings for outcome j ($j = 1, 2$) and component k ($k = 1, 2$); $\boldsymbol{x}_{ij} = (x_{ij1}, x_{ij2})'$ is a 2×1 vector consisting of one binary (x_{ij1}) and one continuous (x_{ij2}) predictor; and $\boldsymbol{\beta}_{jk} = (\beta_{jk1}, \beta_{jk2})'$ denotes the corresponding vector of regression coefficients for outcome j and component k.

As in the previous two simulations, we considered two sample sizes, $n = 500$ and $n = 5000$, as well as two sets of parameter values corresponding to approximately 40% and 70% zeros (Table 3). We generated the binary predictor, x_{ij1}, from a Bernoulli(0.5) distribution and the continuous predictor, x_{ij2}, from a N(0, 4) density. As before, we simulated the latent factors from a Ga(1, 1) density and assigned conditionally conjugate Ga(0.001, 0.001) and Ga(1, 1) priors to both the factor loadings and to $\exp(\beta_{jk1})$, the exponentiated regression coefficients for the binary predictor x_{ij1}. For the continuous predictor, we assigned a N(0, 1000) prior to each

β_{jk2} and updated the parameters using separate random-walk Metropolis–Hastings steps with a t_3 proposal density. We ran the MCMC sampler for 125,000 iterations, discarding the first 25,000 as burn-in. We retained every 25th iteration for a total of 4000 stored samples.

To investigate whether the model could detect null associations, we set half of the factor loadings to the limiting lower bound of 0, and half to nonzero values. As a result, the simulated model assumed a positive "between-outcome" association for both the binary components and continuous components, but no "within-outcome" association between the binary and count components of the same zero-inflated model. Web Figure 3 provides the factor loading matrix corresponding to this dependence structure. We also simulated data with more complex within- and between-outcome associations, but larger sample sizes (e.g., $n > 5000$) were generally needed to accurately estimate these more complex dependence structures. This may be due to the fact that we simulated factor loadings with smaller values than those estimated in our case study below. In general, we expect that smaller sample sizes will be sufficient for detecting many dependence structures observed in practice.

Web Figures 4 and 5 present the trace plots for the 40% model. The plots indicate adequate mixing for all model parameters, including those updated via Metropolis steps. Table 3 presents the posterior means and 95% CrIs for the model with 40% zeros and $\text{Ga}(0.001, 0.001)$ hyperparameters. Web Table 1 presents the corresponding results for the model with $\text{Ga}(1, 1)$ priors, and Web Tables 2 and 3 present the results for 70% zeros. In general, the parameter estimates were quite accurate, with 95% CrIs encompassing the true values. Of note, the model was able to reliably estimate the null factor loadings, which should prove useful when variable selection is of primary interest.

Not surprisingly, we observed some prior sensitivity in the factor loadings for smaller sample sizes—for example, $n = 500$ and 40% zeros (Web Table 1). However, when the sample size increased to $n = 5000$, similar estimates were observed for both the $\text{Ga}(0.001, 0.001)$ and $\text{Ga}(1, 1)$ priors (Table 3 and Web Table 1, respectively). Similar results were seen for 70% zeros (Web Tables 2 and 3), although some prior sensitivity was observed in the factor loadings for $n = 500$. However, for all scenarios, the estimates of the regression coefficients (β's) were robust to these two choices of prior distribution.

To assess the validity of the WAIC-based model comparison procedure, for each simulation we fit models with zero, one, and two latent factors (one of which corresponded to the simulated data). In each case, we computed the WAIC score for the three fitted models. The results are provided in Web Table 4. The rows correspond to the simulated (i.e., "true") model, while the columns denote the fitted model. In each case, the WAIC value was the smallest for the fitted model corresponding to the simulated data. Note that for simulation study 1 (no latent factors), the models with one and two factors failed to converge based on standard diagnostics. This is understandable, since the true model assumes no latent between-subject heterogeneity; hence, higher-order models accounting for this variation are difficult to identify. Nevertheless, the results suggest that the WAIC measure, coupled with standard MCMC diagnostics, provides a reliable approach to assessing the number of latent factors.

6. Application to Breast Cancer Genomics

We applied the proposed approach to a multi-level molecular study of breast cancer. Specifically, we were interested in predicting whether cancer-relevant pathways were activated during the progression of breast cancer (pathway-level analysis) and if so, modeling the number of genes with CNVs within each pathway (gene-level analysis). Using the Molecular Signatures Database (`http://www.broadinstitute.org/gsea/downloads.jsp`), we first downloaded the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway data. Next, using the cBioPortal for Cancer Genomics database (`http://www.cbioportal.org`), we downloaded the provisional version of The Cancer Genome Atlas (TCGA) data for breast invasive carcinoma (Cancer Genome Atlas Network, 2012). Specifically, we downloaded the putative CNV data for genes annotated in the KEGG pathways.

For this analysis, we focused on three KEGG pathways: MAPK signaling, cytokine–cytokine receptor (CCR) interaction, and endocytosis. Genes in the MAPK pathway are related to cell proliferation, differentiation, and migration, while the genes in the CCR interaction pathway are associated with inflammatory host defenses, cell growth, differentiation and death, and the restoration of homeostasis. The genes in the endocytosis pathway are related to mechanisms by which cells transport ligands, nutrients, proteins, and lipids from the cell surface to the cell interior. For each pathway, we recorded the number of genes with significant CNVs (valued as -2 or $+2$). We included cancer stage as a covariate in the model, excluding stage-IV patients due to small sample sizes ($n = 15$). We coded stages IA and IB as stage I, IIA and IIB as stage II, and IIIA, IIIB, and IIIC as stage III. We included only patients with molecular and clinical data, resulting in a total of 908 patients: 160 in stage I, 536 in stage II, and 212 in stage III. Among the 908 patients, 237 (26%), 276 (31%), and 262 (29%) had zero counts for pathways MAPK, CCR interaction, and endocytosis, respectively. Vuong's procedure (Vuong, 1989) indicated significant zero inflation for all three pathways.

Next, we fit LZIP models with 0, 1, and 2 latent factors. The zero-factor model corresponded to three independent ZIP regressions—one for each pathway. We assigned a $\text{Ga}(1, 1)$ prior the latent factors, and $\text{Ga}(0.001, 0.001)$ priors to the factor loadings and exponentiated regression coefficients for cancer stage. We ran the MCMC sampler for 110,000 iterations, discarding the first 10,000 as burn in. We retained every 25th iteration for a total of 4000 stored iterations. Trace plots and Geweke diagnostics indicated rapid convergence and efficient mixing for all models (Web Figures 6 and 7). Based on the WAIC scores, the one- and two-factor models vastly outperformed the zero-factor model ($\text{WAIC} = 21{,}934$ for the zero-factor model, 10,886 for the one-factor model, and 10,885 for two-factor model). Because the WAIC values for the one- and two-factor models were essentially identical, we selected the more parsimonious one-factor model as our final model. Parameter estimates and 95% credible intervals (CrIs) from this model can be found in Web Table 5. The

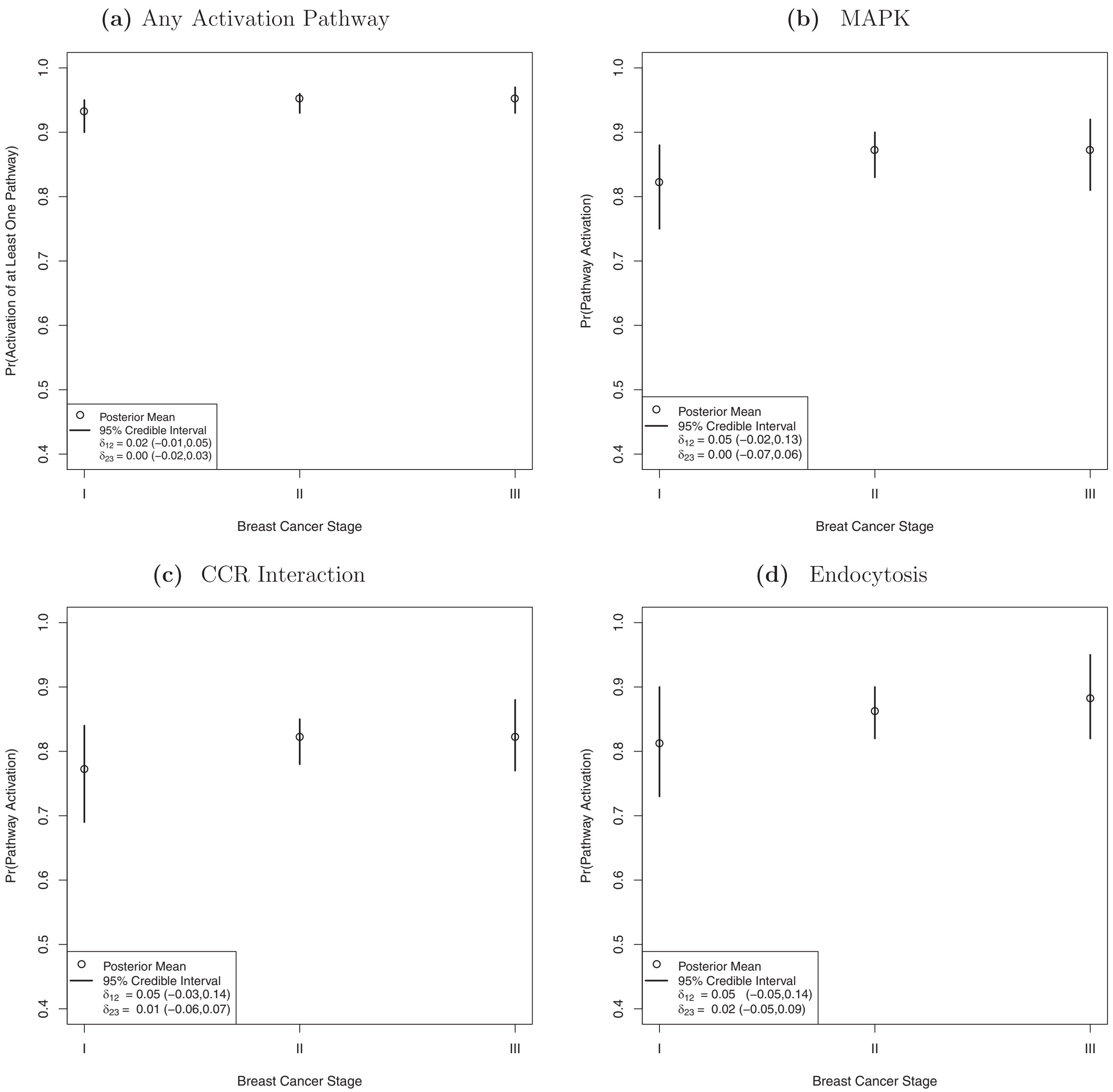


Figure 1. Population-average pathway activation probabilities. Panel (a): The probability of at least one activated pathway by cancer stage. Panels (b)–(d): probability of pathway activation for MAPK, CCR interaction, and endocytosis. Circles denote posterior mean estimates; solid lines are 95% credible intervals; and δ_{12} and δ_{23} are the differences between stages 1 and 2 and stages 2 and 3, respectively.

results suggest a steady increase in CNV output as cancer progresses, particular from stage I to stage II. This may represent an important transition period in which pathways "open up," producing increasing numbers of genes with CNVs. As a sensitivity check, we assigned Ga(1,1) and Ga(2,2) priors to the exponentiated regression coefficients and obtained similar results.

As discussed in Section 3, the LZIP yields several population-average summary measures that should prove useful in practice. For example, the population-average probabilities of at least one activated pathway, defined as ψ_i in equation (11), were 0.93, and 0.95, and 0.95 for stages I, II, and III, respectively (Figure 1, panel (a)). These findings are not surprising given that these are established cancer-relevant pathways, and therefore we would expect to observe at least one activated pathway at each stage. Nevertheless, there was a modest increase in the probability of some activation from stage I to stage II ($\delta = 0.02$, 95% CrI $= [-0.01, 0.05]$). This

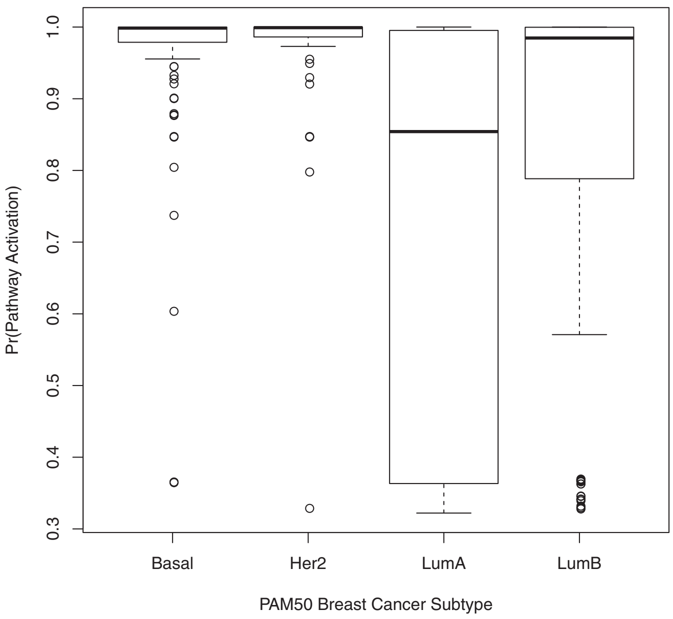


Figure 2. Boxplots showing the subject-level probabilities of MAPK pathway activation for patients of each PAM50 breast cancer subtype: Basal-like (Basal), HER2+ (Her2), Luminal-A (LumA), and Luminal-B (LumB).

is consistent with the fact that stage I has a more localized impact (e.g., cancers are confined to the breast), whereas stages II and III correspond to more regional effects in which tumors spread to surrounding tissue (American Cancer Society, 2013).

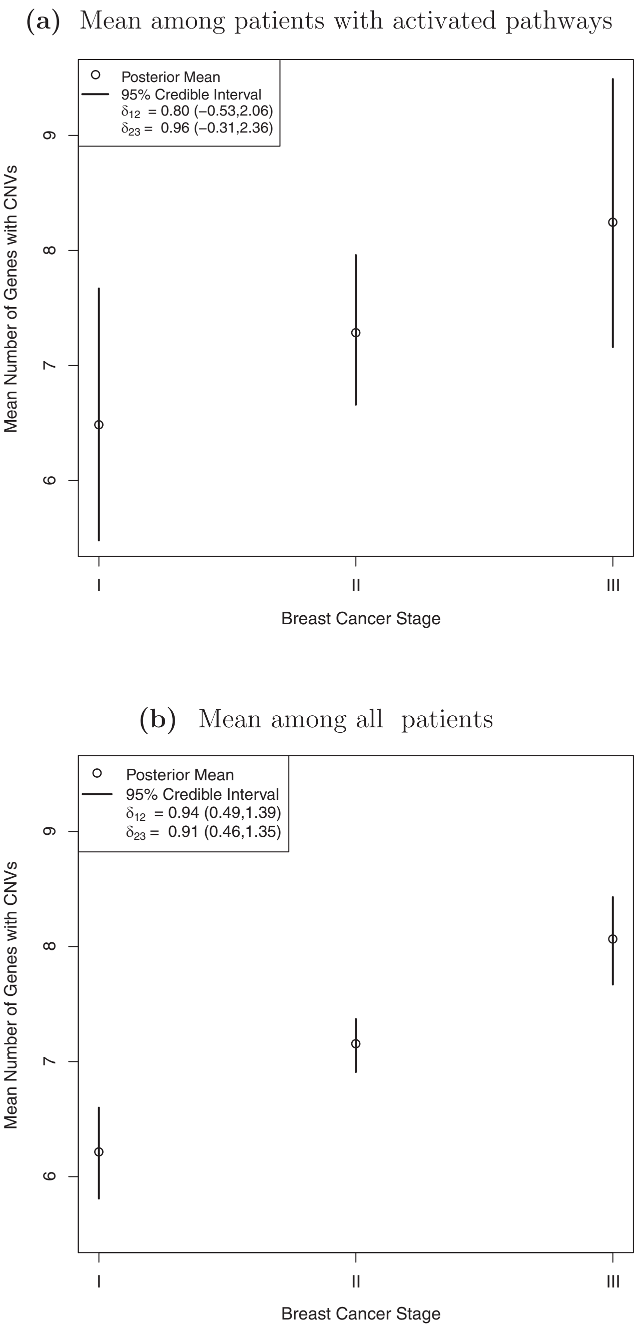
Next, for each pathway, we investigated the population-average probability of activation, which is represented by equation (6) for a single outcome. The results are shown in Figure 1, panels (b)–(d). The activation probabilities were nearly identical for MAPK and endocytosis (e.g., 0.82 and 0.81, respectively, for stage I), while CCR interaction had lower activation probabilities at all stages. In particular, the differences in activation probabilities between MAPK and CCR interaction were 0.05 (95% CrI = $[-0.03, 0.14]$) for stage I, 0.05 (95% CrI = $[0.01, 0.10]$) for stage II, and 0.04 (95% CrI = $[-0.02, 0.11]$) for stage III, respectively. The higher activation probabilities for MAPK and endocytosis are expected, since MAPK is among the most established biological pathways for breast cancer, while endocytosis has been recently considered an important molecular feature for multiple cancers (Mosesson et al., 2008).

To further refine our analysis, we investigated the distribution of activation probabilities across patients for four PAM50 breast cancer subtypes. PAM50 is a well-characterized qRT-PCR intrinsic subtyping classifier that measures expression of 50 genes selected as characteristic of breast cancer intrinsic subtypes (Parker et al., 2009). Using data from a recent TCGA Consortium study (Cancer Genome Atlas Network, 2012), we obtained PAM50 subtype information for 459 patients in our study. These patients comprised 84 Basal-like, 54 HER2+, 206 Luminal-A, and 115 Luminal-B subtypes. Figure 2 presents box-plots of the subject-specific MAPK activation probabilities for each subtype. The activation probabilities were extremely high for Basal-like and HER2+ subtypes (median = 0.99 for both subtypes), a result consistent with previous work suggesting that MAPK is a driver pathway for both subtypes (Liu and Hu, 2014). Moreover, the large variation in activation probabilities for Luminal-A supports earlier findings that Luminal-A is genomically and clinically the most heterogeneous subtype of breast cancer (Cancer Genome Atlas Network, 2012).

Next, we turned to a gene-level analysis. First, we estimated the population-average mean number of CNVs among patients with activated pathways, as expressed in equation (12). The results presented in Figure 3(a) for MAPK and in Web Figures 8(a) and 9(a) for the other two pathways. For all pathways, there was a modest increase in the mean number of CNVs as breast cancer progressed. MAPK had the most CNVs on average, ranging from 6.47 for stage I to 8.23 for stage III. For CCR interaction, the range was 5.70–7.29, and for endocytosis, the range was 4.42–5.40. Endocytosis had substantially fewer CNVs than MAPK, with a mean difference of 2.05 (95% CrI = $[1.46, 2.72]$) for stage I, 2.37 (95% CrI = $[2.02, 2.75]$) for stage II, and 2.87 (95% CrI = $[2.23, 3.52]$) for stage III. The high CNV count for MAPK is well supported in the literature, as multiple genes in this pathway have been linked to breast cancer (Cancer Genome Atlas Network, 2012).

Finally, we examined the population-average mean number of genes with CNVs among *all* patients (i.e., those with and without pathway activation), defined as $\mathrm{E}_{\xi_i}[\mathrm{E}(Y_{ij}|\xi_i)] = \mathrm{E}_{\xi_i}[\phi_{ij}\mu_{ij2}]$ in equation (9), where $L = 1$ and Monte Carlo integration was used to compute the expectation over ξ_i. The results for MAPK are presented in Figure 3(b). The results indicate that the mean number of CNVs increased steadily as the cancer progressed. For example, the mean number of CNVs among all patients increased by 0.94 (95% CrI = $[0.49, 1.39]$) from stage I to II, and by 0.91 (95% CrI = $[0.46, 1.35]$) from stage II to III. Similar results are seen for the other pathways (Web Figures 8(b) and 9(b)). There were also significant differences across pathways, with MAPK again having the most CNVs and endocytosis having the fewest.

To further evaluate model fit, we conducted a series of posterior predictive assessments, whereby the observed data are compared to data replicated from the posterior predictive distribution (Gelman et al., 1996). If the model fits well, the replicated data should resemble the observed data. To quantify the degree of similarity, one typically chooses a “discrepancy statistic,” such as a sample moment or quantile, that captures some important aspect of the data. A 95% predictive interval that includes the observed sample value suggests adequate model fit. Web Figures 10–12 present posterior predictive checks, by pathway and cancer stage, based on three discrepancy measures: 1) the mean number of genes with CNVs among *all* patients (with and without pathway activation); 2) the proportion of zeros; and 3) the sample skewness. In all cases, the 95% credible intervals overlapped the sample values, supporting the appropriateness of the one-factor LZIP in this analysis.



(a) Mean among patients with activated pathways

(b) Mean among all patients

Figure 3. Gene activation results for MAPK pathway. Panel (a): Population-average mean number of genes with CNVs, conditional on MAPK pathway activation. Panel (b): Population-average mean number of genes with CNVs among *all* patients (with and without pathway activation). Circles denote posterior mean estimates; solid lines are 95% credible intervals; and δ_{12} and δ_{23} are the differences between stages 1 and 2 and stages 2 and 3, respectively.

7. Discussion

We have developed a Bayesian latent factor model for the analysis of multivariate zero-inflated counts. The model has many appealing features: it accounts for zero inflation, within-subject association, dependence between model components, and overdispersion. Furthermore, the latent Poisson representation described in Section 2 allows us to learn about the "structural" and "at-risk" class allocations during the course of MCMC run. And finally, the model yields population-average quantities of practical interest. For posterior computation, we developed an efficient data-augmentation sampler based primarily (or in some cases entirely) on closed-form full conditionals.

Our simulation studies indicated that the model offers a suitable alternative to conventional ZIP models, but accommodates many more general settings, including multivariate zero-inflated data with complex dependence structures. Our simulations also demonstrated that the model can detect null factor loadings, reducing the risk of overfitting. A potential limitation, however, is that the model permits only positive associations between outcomes. Future work could extend the model to allow for negative associations. Future analyses might also include additional covariates to explore stage-by-covariate interactions.

Our approach should also prove useful in other applied settings, such as genome-wide association studies (GWAS) with rare variants. Here, the LZIP could provide information about genetic associations at both gene and single nucleotide polymorphism (SNP) levels, resulting in a unified analysis of both multi- and single-marker associations. More generally, the model should have broad applicability in settings where interest lies in modeling multivariate zero-inflated data within a computationally tractable Bayesian framework.

8. Supplementary Materials

The derivation of equation (8) referenced in Section 2, the proof of Proposition 1 referenced in Section 3, the MCMC algorithm referenced in Section 4.2, Web Figures 1–5 and Web Tables 1–4 referenced in Section 5, Web Figures 6–12 and Web Table 5 referenced in Section 6, and R code for fitting the models in Sections 5 and 6 are available with this article at the *Biometrics* website on Wiley Online Library.

References

American Cancer Society (2013). *Breast Cancer Facts and Figures 2013-2014*. Atlanta: American Cancer Society, Inc.

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70.

Dunson, D. B. and Herring, A. H. (2005). Bayesian latent variable models for mixed discrete outcomes. *Biostatistics* **6**, 11–25.

Fox, J.-P. (2013). Multivariate zero-inflated modeling with latent predictors: Modeling feedback behavior. *Computational Statistics & Data Analysis* **68**, 361–374.

Gelman, A., Meng, X., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **6**, 733–807.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds), *Bayesian Statistics 4*, pages 169–193, Oxford: Clarendon Press.

Guo, G. (1996). Negative multinomial regression models for clustered event counts. *Sociological Methodology* **37**, 59–163.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica* **47**, 153–161.

Heilbron, D. C. (1994). Zero-altered and other regression models for count data with added zeros. *Biometrical Journal* **36**, 531–547.

Jackman, S. (2015). *pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University*. Department of Political Science, Stanford University, Stanford, California. R package version 1.04.4.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1–14.

Liu, Y. and Hu, Z. (2014). Identification of collaborative driver pathways in breast cancer. *BMC Genomics* **15**, 605.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* **10**, 325–337.

Majumdar, A. and Gries, C. (2010). Bivariate zero-inflated regression for count data: A Bayesian approach with application to plant counts. *The International Journal of Biostatistics* **6**, 27.

Mosesson, Y., Mills, G. B., and Yarden, Y. (2008). Derailed endocytosis: An emerging feature of cancer. *Nature Reviews Cancer* **8**, 835–850.

Neelon, B. H., O'Malley, A. J., and Normand, S.-L. T. (2010). A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Statistical Modelling* **10**, 421–439.

Papastamoulis, P. and Iliopoulos, G. (2010). An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions. *Journal of Computational and Graphical Statistics* **19**, 313–331.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology* **27**, 1160–1167.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**, pp. 307–333.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* pages 3571–3594.

Received October 2015. *Revised May* 2016. *Accepted May* 2016.