 DOI: 10.1111/j.1541-0420.2011.01729.x

The k-ZIG: Flexible Modeling for Zero-Inflated Counts

Souparno Ghosh,[1,*] Alan E. Gelfand,[1,] Kai Zhu[2,***] and James S. Clark[2,****]**

[1]Department of Statistical Science, Duke University, Durham, North Carolina 27708-0251, U.S.A.
[2]Nicholas School of the Environment, Duke University, Durham, North Carolina 27708-0251, U.S.A.
[*]*email*: sg147@stat.duke.edu
[**]*email*: alan@stat.duke.edu
[***]*email*: kai.zhu@duke.edu
[****]*email*: jimclark@duke.edu

Summary. Many applications involve count data from a process that yields an excess number of zeros. Zero-inflated count models, in particular, zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models, along with Poisson hurdle models, are commonly used to address this problem. However, these models struggle to explain extreme incidence of zeros (say more than 80%), especially to find important covariates. In fact, the ZIP may struggle even when the proportion is not extreme. To redress this problem we propose the class of k-ZIG models. These models allow more flexible modeling of both the zero-inflation and the nonzero counts, allowing interplay between these two components. We develop the properties of this new class of models, including reparameterization to a *natural* link function. The models are straightforwardly fitted within a Bayesian framework. The methodology is illustrated with simulated data examples as well as a forest seedling dataset obtained from the USDA Forest Service's Forest Inventory and Analysis program.

Key words: Abundance; Bayesian modeling; Link function; Log score loss; Poisson-Gamma; Presence/absence.

1. Introduction

Count data with an excess of zeros arise in many contexts. In an influential paper, Lambert (1992) describes an industrial application where a reliable manufacturing process moves back and forth between a perfect state in which defects are extremely rare and an imperfect state in which the number of defects follow a Poisson distribution. A natural way to model such data is to put an additional point mass p at 0. That is, with probability p, we sample a degenerate distribution at 0 and with probability $(1-p)$ we sample a nondegenerate distribution $G(.|\Theta)$. When $G(.|\Theta)$ is Poisson, $Po(\lambda)$, we obtain the well-known zero-inflated Poisson (ZIP) models, which we denote as ZIP(p, λ). This model has been studied extensively. Cohen (1963) and Johnson and Kotz (1969) discuss the ZIP model without covariates, Lambert (1992) employs the ZIP model in a linear regression setup, using canonical links for both p and λ. She obtains the maximum likelihood estimates of the parameters using the E-M algorithm. Ghosh, Mukhopadhyay, and Lu (1998) discuss fully Bayesian methods for fitting ZIP models. Agarwal, Gelfand, and Citron-Pousty (2002) introduce spatial random effects to model spatial count data that exhibit excess zeros. They fit the model in a Bayesian framework and discuss the issue of posterior propriety in detail. Rathbun and Fei (2006) describe a spatial ZIP model in which the excess zeros are generated by a spatial probit model. Under this model, an excess zero is generated at a given site if the realization of a Gaussian random field falls below a threshold.

The contribution here is to address two limitations of the ZIP. First, customarily, either a logit or probit model is used for p as a function of covariate information to explain the occurrence of the excess zeros. As a result of considerable simulation and real data analysis (some reported in Section 3), we find that these links are limited in their ability to explain an extreme number of zeros. That is, both of these links will struggle to find covariates providing significant explanation. This may be due, in part, to a lack of variablity in the data but, as we show in our simulation examples, even when the data have enough variablity, these links are quite sensitive to model misspecification. Their symmetry (i.e., $\log\frac{(1-p)}{p} = -\log\frac{p}{(1-p)}$ and $\Phi(1-p) = -\Phi(p)$) is a further restriction. Wang and Dey (2010) and references therein provide a summary of flexible parametric link functions developed to model binary response data, for example, the complementary log–log link and the generalized extreme value link. However, these links are not purpose-built for handling an excessive amount of zeros in the context of modeling zero-contaminated count data.

Second, the choice of G to be Poisson implies limited shape and tail behavior. Alternative possibilities include finite Poisson mixtures and continuous Poisson mixtures, e.g., the Poisson Gamma (equivalently, negative binomial). The negative binomial model is more flexible than its Poisson counterpart in accommodating overdispersion (Lawless, 1987). Cameron and Trivedi (1986) study a variety of stochastic models, including the zero-inflated negative binomial model, for count data. Gurmu, Rilstone, and Stern (1999) develop a continuous mixture model for count data based on series expansion for the unknown density of the unobserved heterogeneity. They also discuss how to modify the mixture model to ac-

commodate excess zeros. Cui and Wang (2009), on the other hand develop a discrete mixture model for zero-inflated count data.

Often, random effects are introduced if excess heterogeneity is observed even after adjusting for covariates (Agarwal et al., 2002; Neelon, O'Malley, and Normand, 2010). These random effects can be incorporated in the occurrence part and/or in the abundance part. However, in our experience, with such high occurrence of zeros (potentially more than 80%), including random effects, correlated or otherwise, in the occurrence model leads to highly unstable model fitting and, in turn, computational intractibility. This was also observed by Agarwal et al. (2002).

In this article, we present a new parametric class of zero-inflated count models, again obtained by adding a point mass at 0 in the presence of covariate information. This class arises through an intuitive motivation: let G itself be a zero inflated model with G_0 handling the nondegenerate part. Then, an additional point mass for the probability of observing a zero results. Formally, we introduce an additional parameter, k, to accomplish this. We can also rewrite the resulting model as a zero inflated model with a novel link function. We need not choose a link function from a catalog. Rather, the link emerges under natural modeling to enable increased flexibility in explaining the zeros. A further feature of this model specification is that the choice of G_0 which is part of the definition of G can still be flexible but now the overall model allows for interplay between k and this G_0. We refer to such models as k-ZIG models, where the parameter k, in conjunction with the foregoing p, controls the amount of mass needed at 0. The standard zero inflated models in the literature are special cases of our k-ZIG model.

An important distinction in the literature is between the ZIP model and the Poisson hurdle model (Welsh et al., 1996; Cameron and Trivedi, 1998). The difference is simple to describe but its implications are primarily at the process level. That is, practically, it will be difficult to distinguish them in model fitting since the latter can be viewed as a special case of the former. In particular, the ZIP is a mixture model which allows zeros to come from two sources. One source is the implicit Bernoulli trial, the other is through the Poisson law. A natural setting is modeling abundance for a rare species. We would expect zeros because some sampled habitats are unsuitable for the species. However, even within a suitable habitat, the species might not be present due to its rarity. The Poisson hurdle model allows only one source for zeros, a Bernoulli model, and then models the nonzeros as a Poisson truncated greater than 0. Once a "hurdle" is passed, presences will be observed. The natural setting here is a habitat suitability model which, for a given species, distinguishes the processes of colonization (presence/absence) from growth (abundance). These models are easier to fit because they enable separation of the likelihood for these two processes. In a subsequent manuscript we will compare Poisson hurdle models to our k-ZIGs in ecological settings. Here, we confine ourselves to discussion of the latter.

We present the formal details of the proposed model and its properties. We propose model fitting and inference within a Bayesian framework. We illustrate the benefits of using this class of models with both simulated and real data, the latter in the context of studying presence and absence along with abundance of seedlings. In particular, we employ data from the USDA Forest Service's Forest Inventory and Analysis (FIA) program and present an analysis of the seedling counts observed in the FIA plots covering the entire eastern United States. We demonstrate that our class of models can handle the excess heterogeneity prevalent in the data, after adjusting for covariates, without resorting to the introduction of random effects. Finally, we offer a useful predictive cross-validation criterion to enable out-of-sample model comparison. With this criterion, we demonstrate substantial benefit in out-of-sample prediction using these new models.

The format of the article is as follows. In Section 2 we develop the modeling details, properties, model fitting and inference, and a model comparison criterion. In Section 3 we outline a simulation experiment and furnish the results as a proof of concept. In Section 4 we present the FIA data analysis. We conclude with Section 5, giving a brief summary and discussion of future work.

2. The k-ZIG Model

Let Y be a count random variable. Let $G(y|\Theta),\ \ y = 0, 1, 2, \ldots$ denote a probability mass function associated with these counts. The zero-inflated distribution associated with $G(y|\Theta)$ is then defined as

$$\begin{aligned} P(Y = 0|p, \Theta) &= p + (1 - p)G(0|\Theta) \\ P(Y = y|p, \Theta) &= (1 - p)G(y|\Theta),\ y > 0. \end{aligned} \tag{1}$$

Equivalently, we can write

$$\pi(y|p, \Theta) = p\delta_{\{0\}} + (1 - p)G(y|\Theta). \tag{2}$$

Usual choices for $G(y|\Theta)$ include the Poisson and negative binomial distributions. Suppose, instead, that we assume $G(y|\Theta)$ to be a zero-inflated distribution itself with a point mass q at zero and $G_0(y|\Theta_0)$ is the probability mass function corresponding to the nondegenerate part. Denote this zero-inflated distribution as $\text{ZIG}(q, \Theta_0)$. Inserting this G into (1) implies that a further proportion $1 - p$ is taken from the $\text{ZIG}(q, \Theta_0)$ and assigned to the event $\{y = 0\}$ thereby further enhancing the probability of this event. Writing out (1) explicitly with $G(y|\Theta)$ replaced by $\text{ZIG}(q, \Theta_0)$, we get

$$\begin{aligned} \pi(y|p, q, \Theta_0) &= p\delta_{\{0\}} + (1 - p)(q\delta_{\{0\}} + (1 - q)G_0(y|\Theta_0)) \\ &= (p + (1 - p)q)\delta_{\{0\}} + (1 - p)(1 - q)G_0(y|\Theta_0). \end{aligned} \tag{3}$$

It can be easily shown that p and q are not identifiable in (3). So we reparameterize (3) by assuming $(1 - p) = (1 - \theta)^{k-1}$ and $(1 - q) = (1 - \theta)$, yielding

$$\pi(y|\theta, \Theta_0, k) = (1 - (1 - \theta)^k)\delta_{\{0\}} + (1 - \theta)^k G_0(y|\Theta_0). \tag{4}$$

We refer to the distribution in (4) as a k-$\text{ZIG}(\theta, \Theta_0)$. Now, the expected number of zeros is controlled by the parameters k and θ. When $k = 1$, and $G_0(.)$ is Poisson(λ), (4) yields the familiar ZIP model. With the goal of modeling excess zeros, we would imagine $k > 1$. However, in principle, k need not be integer valued and can be less than 1; we can use the data to learn about k. In subsequent sections and through simulations we show that, under a suitable prior, the full conditional distribution for k has a closed form and the data provides consequential information about k. From (4) one can immediately

notice that defining $\tilde{\theta} = 1-(1-\theta)^k$, we obtain the standard zero-inflated form as in (2)

$$\pi(y|\tilde{\theta}, \Theta_0) = \tilde{\theta}\delta_{\{0\}} + (1-\tilde{\theta})G_0(y|\Theta_0). \tag{5}$$

Thus the k-ZIG(θ, Θ_0) can be interpreted as a ZIG$(\tilde{\theta}, \Theta_0)$. However, this does not mean that fitting (5) is equivalent to fitting (2), as we clarify in Section 2.2.

2.1 *Properties of the k-ZIG model*

From (4) it is clear that the overall probability of zeros under the k-ZIG model is given by

$$\begin{aligned} P(Y=0|\theta, \Theta_0, k) &= (1-[1-\theta]^k) + (1-\theta)^k G_0(0|\Theta_0) \\ &= 1-(1-\theta)^k(1-G_0[0|\Theta_0]). \end{aligned} \tag{6}$$

Since $0 \le (1-G_0[0|\Theta_0]) \le 1$, it is clear that $P(Y=0|\theta, \Theta_0, k) \to 1$ as $k \to \infty$, regardless of $\theta < 1$. From ZIG$(\tilde{\theta}, \Theta_0)$ in (5), we can see that unconditional moments are given by

$$\begin{aligned} E(Y|\tilde{\theta}, \Theta_0) &= (1-\tilde{\theta})E_{G_0}(Y|\Theta_0), \\ \text{var}(Y|\tilde{\theta}, \Theta_0) &= \tilde{\theta}(1-\tilde{\theta})[E_{G_0}(Y|\Theta_0)]^2 + (1-\tilde{\theta})\text{var}_{G_0}(Y|\Theta_0), \end{aligned}$$

where $E_{G_0}(Y|\Theta_0)$ and $\text{var}_{G_0}(Y|\Theta_0)$ denote the expectation and variance of the random variable Y with probability mass function given by $G_0(.|\Theta_0)$. If we assume that $G_0(y|\Theta_0)$ is Po(λ), then the moments turn out to be

$$\begin{aligned} E(Y|\theta, \lambda, k) &= (1-\theta)^k\lambda. \\ \text{var}(Y|\theta, \lambda, k) &= (1-\theta)^k\lambda(1+\left[1-\{1-\theta\}^k\right]\lambda). \end{aligned}$$

Note that the expectation under the k-ZIP model is a decreasing function of k. Hence for $k \ge 1$, the expectation under the k-ZIP model is less than that under the standard ZIP model. The variance, however, does not show this monotonicity. According to the value of θ, we can induce greater (smaller) variability in the k-ZIP as compared to the standard ZIP.

2.2 *Incorporating the covariates*

As a result of the connection between the k-ZIG(θ, Θ_0) model and the standard ZIG$(\tilde{\theta}, \Theta_0)$ model, do we want to specify a regression model on θ or on $\tilde{\theta}$? Suppose we adopt a logit link on θ. Denoting the set of covariates that influence the occurrences by $\mathbf{X}_1$ and the corresponding regression parameters by $\boldsymbol{\beta}_1$, we get $\log\frac{\theta}{1-\theta} = \mathbf{X}_1\boldsymbol{\beta}_1$ yielding

$$\tilde{\theta} = 1-\left(\frac{1}{1+\exp[\mathbf{X}_1\boldsymbol{\beta}_1]}\right)^k. \tag{7}$$

So, with a logit link on θ, we can fit the k-ZIG model described in (4) as a ZIG$(\tilde{\theta}, \Theta_0)$. In either case, the likelihood is a function of k and it is possible to draw inference about it.

If, however, we specify a logit link on $\tilde{\theta}$, we get $\log\frac{\tilde{\theta}}{1-\tilde{\theta}} = \mathbf{X}_1\boldsymbol{\beta}_1$ so

$$\theta = 1-\left(\frac{1}{1+\exp[\mathbf{X}_1\boldsymbol{\beta}_1]}\right)^{1/k}. \tag{8}$$

Now, regardless of whether we fit the ZIG model with a logit link on $\tilde{\theta}$ or k-ZIG model with link (8) on θ, the likelihood is independent of k. This parameterization makes it impossible to learn about k from the data. The above discussion clarifies the remark below (5); evidently, we will work with (7).

Figure 1 plots the quantiles of η $(= \mathbf{X}_1\boldsymbol{\beta}_1)$ against the given values of $\tilde{\theta}$ for five different values of k. The plot illustrates how the link function (7) behaves for different values of k as compared to the standard logit link on $\tilde{\theta}$. We can see the introduction of asymmetry as k moves away from 1. However, when using (7), care is needed in interpreting the regression parameters $\boldsymbol{\beta}$. The exponentiation resulting from k implies that the regression parameters mean different things for different values of k.

2.3 *Model fitting and inference*

Suppose a sample such that the Y_i are independently distributed as k-ZIG$(\theta_i, \Theta_{0,i})$. Then for $(Y_i, \mathbf{X}_{1i}, \mathbf{X}_{2i})$, $i = 1, 2, \ldots, n$, the conditional likelihood is given by

$$\begin{aligned} [\mathbf{Y}|\theta, \Theta_0, k] &= \prod_{i=1}^{n}\left([1-\{1-\theta_i\}^k]\delta_{\{0\}} + [1-\theta_i]^k G_0[y_i|\Theta_{0,i}]\right), \\ \text{with } \theta_i &= \frac{\exp(\mathbf{X}_{1i}\boldsymbol{\beta}_1)}{1+\exp(\mathbf{X}_{1i}\boldsymbol{\beta}_1)}, \end{aligned} \tag{9}$$

where $\mathbf{X}_1$ denotes the covariates that are used to explain the Bernoulli zeros. Corresponding to a ZIP model in (2), we would assume $G_0(y|\Theta_0)$ to be Po(λ) and define a parametric form for λ,

$$\lambda = \exp(\mathbf{X}_2\boldsymbol{\beta}_2), \tag{10}$$

where $\mathbf{X}_2$ denotes the covariates that influence abundance and the $\boldsymbol{\beta}_2$ are corresponding regression parameters. A more flexible model is obtained when, instead of defining a parametric form for λ, we assume that the λs are driven by a mixture model. In particular, suppose we assume $\lambda \sim$ Gamma(a, b) (i.e., $E(\lambda) = ab$) and incorporate the covariates by setting

$$b = \exp(\mathbf{X}_2\boldsymbol{\beta}_2). \tag{11}$$

The marginal model obtained after integrating over λ is, therefore, the zero-inflated negative binomial. The Poisson-gamma parameterization in convenient to work with since it allows us to specify the mean function through $\log(b)$.

So, returning to (9), we consider $G_0(y_i|\Theta_{0,i})$ as either a Po(λ_i) with $\lambda_i = \exp(\mathbf{X}_{2i}\boldsymbol{\beta}_2)$ or a Poisson-Gamma(a, b_i) with $b_i = \exp(\mathbf{X}_{2i}\boldsymbol{\beta}_2)$. The former results in the full Bayesian model

$$\Pi_i[Y_i|\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, k, X_{1i}, X_{2i}][\boldsymbol{\beta}_1][\boldsymbol{\beta}_2][k]. \tag{12}$$

The latter yields

$$\Pi_i[Y_i|\boldsymbol{\beta}_1, \lambda, k, X_{1i}]\Pi_i[\lambda_i|a, \boldsymbol{\beta}_2, X_{2i}][\boldsymbol{\beta}_1][\boldsymbol{\beta}_2][a][k]. \tag{13}$$

Priors: We impose proper priors on $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$, and k. The priors for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are Normal$(0, \mathbf{D}_{\boldsymbol{\beta}_1})$ and Normal$(0, \mathbf{D}_{\boldsymbol{\beta}_2})$, respectively. Specification of the covariance matrices $\mathbf{D}_{\boldsymbol{\beta}_1}$ and $\mathbf{D}_{\boldsymbol{\beta}_2}$ is described in Appendix A.1. For the parameter k, we adopt a Gamma(k_a, k_b) prior. Under this specification, it is possible to obtain a closed form for its full conditional, as derived in the Appendix A.2.

2.4 *A Model Comparison Criterion*

We judge models by out-of-sample performance. Because the models are nonstandard with support $\{0, 1, 2, \ldots\}$, we propose to use a predictive log-score loss criterion. That is, for

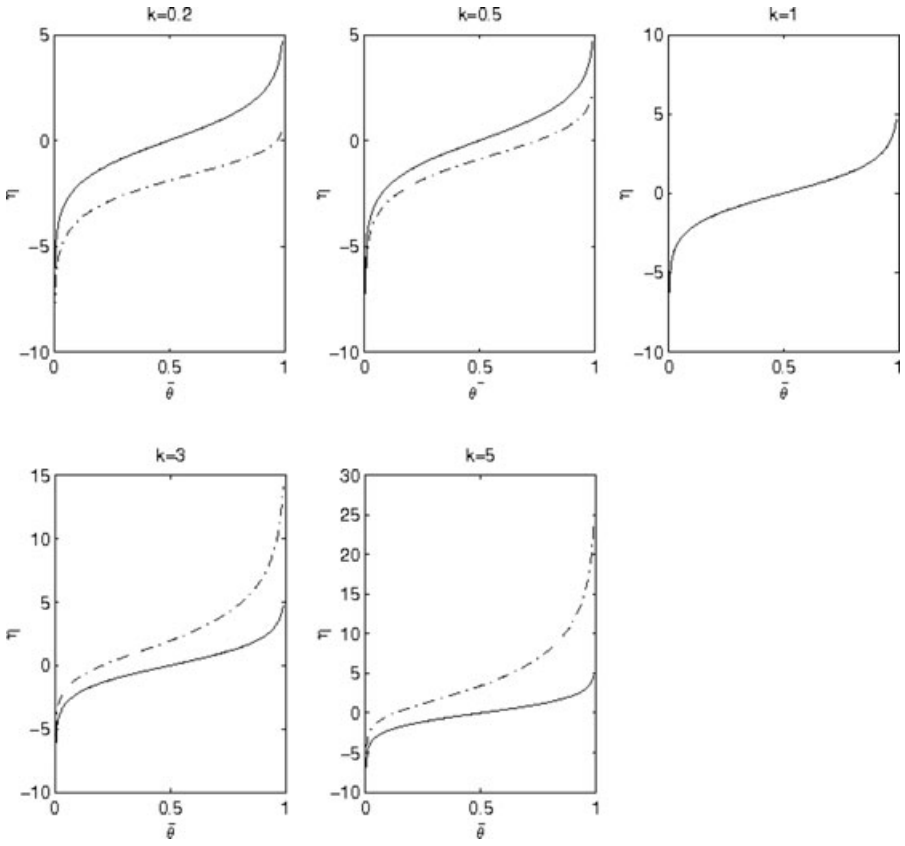

Figure 1. Quantile plot of η ($= \mathbf{X}_1\boldsymbol{\beta}_1$) for the link (7) shown in ($-.-.$) and the logit link (solid).

the discrete density, $f(y|\eta, x)$, we take -2log$f(y_{\text{obs}}|\eta, x_{\text{obs}})$ as the loss associated with a holdout y_{obs} at covariate level x_{obs} given η. The more *likely* y_{obs} is, the less we lose. We split the dataset randomly into a training set and validation set. We fit the competing models on the training set and obtain the posterior samples of the parameters. For validation data $\{(y_{\text{obs},l}, x_{\text{obs},l}), l = 1, 2, \ldots, L\}$, we obtain -$2\sum_l \log f(y_{\text{obs},l}|\eta, x_{\text{obs},l})$. In particular, we use the k-ZIP(θ, λ) as the score function and compute the log validation likelihood for each posterior sample of the parameters. This yields a posterior sample of predictive losses which we could display for each model to compare posteriors. Of course, we can obtain a scalar summary, say the posterior mean of the predictive loss, as a Monte Carlo integration, using the posterior sample of the parameters. The latter, which we define as $\bar{L}(\mathbf{Y}_{\text{obs}}|\hat{\Theta})$ takes the form

$$-\frac{2}{B}\sum_{j=1}^{B}\sum_{l=1}^{L}\Big[\log\Big\{(1-(1-\hat{\theta}_{lj})^{\hat{k}_j})\delta_{\{0\}} + (1-\hat{\theta}_{lj})^{\hat{k}_j}\, G(y_{\text{obs},l}|\hat{\lambda}_{lj})\Big\}\Big], \tag{14}$$

where $\hat{\theta}_j, \hat{\lambda}_j, \hat{k}_j, j = 1, 2, \ldots, B$ are the posterior samples and $G(y_{\text{obs},l}|\hat{\lambda}_{lj})$ is the mass function of a Poisson distribution with mean $\hat{\lambda}_{lj}$ evaluated at $y_{\text{obs},l}$. We prefer models with smaller values of the loss function (14). For a log score loss such as in (14), model comparison would typically compare specifications for θ_i and λ_i. Here we employ this loss to make comparison between the Poisson and Poisson-Gamma. We justify this by noting that, for a given k, both the Poisson and Poisson-Gamma mixture model have the same "conditional" likelihood.

Thus, for the former, once we have posterior draws of $\boldsymbol{\beta}_2$, we can create posterior samples of λ using the relation (10), for the latter, once we have posterior samples of a and $\boldsymbol{\beta}_2$, we can draw posterior samples of λ using composition sampling. With the λs, we can compute the "conditional" validation likelihood under both models and obtain the predictive log-score loss values to compare performance.

3. Simulation Examples

We perform three simulation studies. In the first case we simulate data from a k-zero-inflated Poisson (k-ZIP) model for various values of k (detailed results are provided in the supplementary web material) and in the second example we simulate data from a k-zero-inflated Poisson-Gamma (k-ZIPG) model for various values of k. In both simulations we examine whether it is possible to retrieve the true model from the simulated data and obtain out-of-sample model comparison results using the predictive log-score loss criterion. A third example illustrates the performance of the model when we use a bimodal $G_0(.)$. We only study whether it is possible to retrieve the true model from the simulated data, omitting model comparison. Detailed results pertaining to this simulation are provided in the supplementary web material. All computations, simulations and model fittings, are performed on a Matlab 7.8 platform in a dual core 2500 MHz processor

Table 1
The out-of-sample model comparison using the log-score loss criterion, with the k-ZIP score function, under k-ZIPG sampling model.

k	n	ZIP	ZIPG	k-ZIP	k-ZIPG
1	500	427.85	404.76	488.12	451.70
	10,000	2290.26	1870.62	2177.79	1874.02
3	500	283.51	245.03	222.04	239.43
	10,000	1324.37	1269.52	1273.11	1068.93
5	500	173.76	160.94	140.07	159.05
	10,000	1476.51	917.16	671.28	498.62

Table 2
The posterior mean and 95% CI (in parenthesis) of the parameters under k-ZIPG sampling model for two different sample sizes.

Parameters	True value	Posterior summary $n = 500$	Posterior summary $n = 10,000$
k	1	1.17 (0.36, 1.38)	0.96 (0.81, 1.12)
Percentage of zeros	–	45	42
β_{11}	−0.3	−0.06 (−2.06, 2.10)	−0.23 (−2.07, −0.04)
β_{12}	1	1.13 (0.08, 4.12)	1.05 (−0.29, 1.40)
β_{21}	1	0.92 (−1.27, 3.64)	1.03 (0.29, 3.12)
β_{22}	3	1.96 (0.74, 3.88)	2.86 (0.80, 4.07)
k	3	2.89 (2.71, 4.04)	3.03 (2.80, 3.24)
Percentage of zeros	–	75	76
β_{11}	−0.3	−0.82 (−3.12, 1.44)	−0.40 (−1.10, 0.49)
β_{12}	1	0.38 (−1.98, 2.47)	0.95 (0.21, 2.45)
β_{21}	1	1.24 (−0.87, 3.51)	1.09 (0.78, 2.48)
β_{22}	3	3.25 (0.76, 5.28)	3.08 (1.01, 4.06)
k	5	5.15 (4.68, 5.59)	4.95 (4.78, 5.08)
Percentage of zeros	–	94	93
β_{11}	−0.3	−0.87 (−3.38, 1.16)	−0.33 (−2.07, 0.46)
β_{12}	1	1.21 (−2.26, 1.27)	1.03 (−1.22, 1.97)
β_{21}	1	1.14 (−1.51, 3.39)	0.93 (−0.50, 2.26)
β_{22}	3	2.84 (0.53, 5.78)	2.95 (0.74, 4.76)

work station. Approximate computation time is around 90 minutes.

Simulation 1–Sampling from a k-ZIP: We generate a single covariate ranging from –1 to 1. We employ both an intercept and slope, fixing $\boldsymbol{\beta}_1 = [-0.3, 1]$ and $\boldsymbol{\beta}_2 = [1, 2]$. We then generate $\mathbf{Y}$ from the k-ZIP model (4) using the regression models described in (9) and (10). We need fairly large sample sizes to learn about the model parameters, especially k. We generate samples of size 500 and 10,000 using this model. We randomly select 50 and 1000 data points, respectively, to serve as the validation set and train the model on the remaining data points. We fit k-ZIP, k-ZIPG, the standard ZIP and ZIPG model (with $k = 1$) on the simulated data and compare them using the loss function described in Section 2.4. Web Table 1 shows the out-of-sample model comparison results for the competing models. The results suggest that we prefer the "reduced" model, the k-ZIP, when it is true. We also ascertain the performance of the fitted k-ZIP model for three k's and two sample sizes. The results presented in the Web Table 2 suggest that the posteriors capture the true values, that inference improves with the larger n. Clearly, this demonstrates that it is possible to learn about k and that the precision of the estimates increase with increase in sample size. Finally, though the priors on the regression parameters have relatively large variance with $c = 10$, they are elicited from the data. So we check that the empirical coverage probabilities of the credible intervals associated with the parameters are close to the nominal value. Web Table 3 shows the empirical coverage probabilities obtained from 30 simulated datasets each with 10,000 data points. They are all comparable to the nominal level of 95%.

Table 3
Assessing coverage probabilities based on 30 simulated datasets of 10,000 data points each under k-ZIPG sampling model.

Parameters	True value	Nominal level	Coverage probability $c_j = 10,\ j = 1, 2$
k	1	95%	93%
β_{11}	−0.3	95%	90%
β_{12}	1	95%	93%
β_{21}	1	95%	96%
β_{22}	3	95%	96%
k	3	95%	93%
β_{11}	−0.3	95%	90%
β_{12}	1	95%	93%
β_{21}	1	95%	90%
β_{22}	3	95%	93%
k	5	95%	93%
β_{11}	−0.3	95%	93%
β_{12}	1	95%	90%
β_{21}	1	95%	90%
β_{22}	3	95%	96%

Simulation 2—Sampling from the k-ZIPG model: As in the first simulation, we generate a single covariate ranging from –1 to 1. We fix $\boldsymbol{\beta}_1 = [-0.3, 1]$ and $\boldsymbol{\beta}_2 = [1, 3]$. We then generate $\mathbf{Y}$ from the model (4) using the regression model described in (9) and (11). That is, we assume the λs are driven by Gamma mixing and incorporate the covariate information using (11). We generate 500 (10,000) observations from the k-ZIPG model. Again, we randomly select 50 (1000) data points to serve as the validation set and train the model on the remaining data points. We fit both k-ZIP and k-ZIPG model on the simulated data and compare them using the loss function described in Section 2.4. Table 1 shows the out-of-sample model comparison results for the k-ZIPG model. Again, we include the results for the ZIP and the ZIPG. Note that, when $k = 3$ or 5, we need large samples to select the k-ZIPG when it is true. Table 2 shows the performance of the fitted k-ZIPG model for three k's and two sample sizes. Table 3 shows the empirical coverage probabilities of the model parameters based on 30 simulated datasets each of 10,000 data points. Once again we see that they agree with the nominal level.

Simulation 3—Sampling from the k-ZIG model: We illustrate the performance of the k-ZIG model when $G_0(.)$ is bimodal. For simplicity, the simulation is performed without covariates. We assume $G_0(.)$ to be a two component Poisson mixture, given by $G_0(y|\Theta_0) = \pi_1 \text{Poisson}(y|\lambda_1) + (1 - \pi_1)\text{Poisson}(y|\lambda_2)$. We fix $\lambda_1 = 5,\ \lambda_2 = 50$ to represent very distinct subpopulations and $\pi_1 = 0.7$ and $\theta = 0.4$. We

Table 4
The posterior mean and 95% CI (in the parenthesis) of the parameters obtained from fitting k-ZIP and k-ZIPG model to the ACRU seedling data.

Species	Parameters	Posterior summary under k-ZIP	Posterior summary under k-ZIPG
ACRU	k	1.23 (0.94, 1.69)	1.17 (0.83, 1.68)
	$\beta_1^{(0)}$	−0.16 (−0.38, 0.07)	0.09 (−0.35, 0.51)
	$\beta_1^{(1)}$	−0.19 (−0.27, −0.08)	−0.20 (−0.31, −0.07)
	$\beta_1^{(2)}$	0.05 (−0.03, 0.12)	0.06 (−0.01, 0.14)
	$\beta_1^{(3)}$	0.10 (−0.02, 0.19)	0.12 (−0.02, 0.23)
	$\beta_1^{(4)}$	0.08 (−0.06, 0.24)	0.09 (−0.05, 0.26)
	$\beta_1^{(5)}$	0.03 (−0.02, 0.09)	0.04 (−0.01, 0.11)
	β_2^{int}	−1.95 (−2.01, −1.89)	−2.56 (−2.64, −2.47)
	β_2^{ba}	0.03 (−0.04, 0.12)	0.06 (−0.04, 0.16)
	$\beta_2^{ba_{all}}$	−0.14 (−0.18, −0.09)	−0.15 (−0.21, −0.06)

Table 5
The posterior mean and 95% CI (in the parenthesis) of the parameters obtained from fitting k-ZIP and k-ZIPG model to the LITU seedling data.

Species	Parameters	Posterior summary under k-ZIP	Posterior summary under k-ZIPG
LITU	k	2.77 (2.21, 3.55)	2.40 (1.78, 2.98)
	$\beta_1^{(0)}$	0.18 (−0.17, 0.51)	0.92 (0.31, 1.65)
	$\beta_1^{(1)}$	−0.29 (−0.37, −0.18)	−0.34 (−0.46, −0.21)
	$\beta_1^{(2)}$	0.04 (−0.06, 0.13)	0.05 (−0.04, 0.14)
	$\beta_1^{(3)}$	−1.62 (−1.81, −1.41)	−2.05 (−2.60, −1.67)
	$\beta_1^{(4)}$	1.66 (1.44, 1.87)	2.11 (1.69, 2.64)
	$\beta_1^{(5)}$	0.21 (0.12, 0.28)	0.26 (0.15, 0.39)
	β_2^{int}	−2.57 (−2.69, −2.48)	−3.14 (−3.27, −3.02)
	β_2^{ba}	0.06 (−0.07, 0.18)	0.05 (−0.04, 0.16)
	$\beta_2^{ba_{all}}$	−0.47 (−0.53, −0.39)	−0.45 (−0.57, −0.31)

consider three different k's. For a sampling model with a given true k, we generate 10,000 observations and fit the k-ZIG model on the entire dataset. Web Table 4 shows that it is possible to retrieve the sampling model from the data under this moderately large sample size.

4. FIA Data

The USDA Forest Service's Forest Inventory and Analysis (FIA) program is the primary source for information about the extent, condition, status, and trends of forest resources in the United States (Smith et al., 2009). FIA applies a nationally consistent sampling protocol using a quasi-systematic design with the sample intensity of one plot per 2428 hectare. FIA inventory plots consist of four, 7.2 m fixed-radius subplots spaced 36.6m apart in a triangular arrangement with one subplot in the center (Bechtold and Patterson, 2005). All trees (standing live and dead) with a diameter at breast height (dbh) of at least 12.7 cm are inventoried on forested subplots. Within each subplot, a 2.07 m microplot offset 3.66 m from subplot center is established where only live trees with a dbh between 2.5 and 12.7 cm are inventoried. Within each microplot, all live tree seedlings are tallied according to species.

Here we analyze two species, *Acer rubrum* (ACRU) and *Liriodendron tulipifera* (LITU). ACRU is a generalist species with its range all over and even beyond the eastern U.S. It is regarded to be adaptable to a very wide range of site conditions. LITU is a specialist species with its range mostly restricted to the southern part of the eastern U.S. It is considered to be shade-intolerant and commonly associated with the early stage of forest succession. The data consists of the number of seedlings observed in 43,396 plots each of area $53.98\,\mathrm{m}^2$ spread over the entire eastern United States. Although the FIA plots are spatially referenced and numerous, with regard to the eastern U.S., they are not geographically close to each other. So, spatial modeling through spatial random effects is not introduced.

The dataset is randomly split into a training set containing 42,396 data points while the remaining 1000 data points are used for validation. We scale the expected counts by the plot area so that λ is interpreted per unit area. That is, we set

$$G(y_i|\lambda) \propto \exp(-A\lambda)(A\lambda)^{y_i},$$

where A is the area of the FIA plot. Canham and Thomas (2010) suggest that climate variables predict occurrence but not the abundance. We use temperature (temp) and precipitation (precip) as covariates in the Bernoulli part of the model. We attempt to capture competition and facilitation by using basal area of the trees of the corresponding species in each plot(ba) and basal area of all the trees present in the plot ($\mathrm{ba_{all}}$) as covariates in the Poisson/Poisson-Gamma part. As an exploratory tool to get an idea of the functional relationship between occurrence of seedlings and the climatological covariates, we divide temperature and precipitation into 20 bins and obtain the relative frequency of the FIA plots with zero counts in each bin. Web Figure 1 and Web Figure 2 show these plots for ACRU and LITU, respectively. Recalling (7), these plots suggest that we model θ as

$$\begin{aligned}\mathrm{logit}(\theta) = \beta_1^{(0)} + \beta_1^{(1)}\mathrm{Precip} + \beta_1^{(2)}(\mathrm{Precip})^2 + \beta_1^{(3)}\mathrm{Temp} \\ + \beta_1^{(4)}(\mathrm{Temp})^2 + \beta_1^{(5)}\mathrm{Temp}\times\mathrm{Precip}\end{aligned}$$

We fit both the k-ZIP and k-ZIPG model on both species and ascertain their performance using the predictive log-score loss function described in Section 2.4. For the ACRU species, the predictive log-score loss under the k-ZIP model (6374.28) is considerably lower than the competing models. Even the loss under the k-ZIPG model (6396.43) is lower than that of the standard ZIP (6422.12) and standard ZIPG (6597.29) models. For the LITU species, the predictive log-score loss obtained by fitting the k-ZIPG model is 1006.05, considerably less than that of the k-ZIP model (1066.91). The loss incurred in fitting standard ZIP and ZIPG models are 1116.88 and 1397.02, respectively. Evidently, the "k" models are preferred, the k-ZIP for ACRU, the k-ZIPG for LITU. Since the k-ZIPG adds only one parameter relative to the k-ZIP, we need not worry about a penalty adjustment for model dimension. Table 4 shows the estimates of the model parameters under the k-ZIP and k-ZIPG model for ACRU and Table 5 shows the same for LITU. We see, the data suggests that $k = 1$ is acceptable for ACRU but $k \neq 1$ for LITU, in accord with the model comparison. We also see different explanatory models for the two species. In particular, despite 92% of zeros for LITU, we can still find a significant climate story for presence/absence.

5. Discussion

We have developed and employed a novel class of zero-inflated models for count data. These models introduce a purpose-built link to allow increased flexibility to capture extremes in the number of zeros as well as the distribution of the nonzero observations. We have further shown that this class of link functions has symmetric links, like the logit link, embedded into it and thereby provides a guard against link misspecification. We have observed that the k-ZIG model works well with large sample sizes. However, the posited model includes simpler models, like the ZIP, and hence can be fitted regardless of the sample size.

We further note that our approach can handle the case of excesses at other count values. For example, Melkersson and Roth (2000) find excess of zeros and twos when looking at the number of offsprings in a family. We could provide a probability model with inflated zeros and twos analogous to (2). We could then further inflate the incidence of zeros and twos analogous to (3). We omit further details.

Future work will see us adding a spatial aspect to the modeling in a setting where spatial dependence between the counts at locations is expected. Further, we can add a dynamic aspect to the modeling, for instance, with daily or weekly disease counts for a collection of locations, e.g., hospitals. We might also consider joint modeling of the counts, e.g., modeling ACRU and LITU jointly in the FIA dataset.

6. Supplementary Materials

Web Tables 1–4 referenced in Section 3; Web Figures 1 and 2 referenced in Section 4; the MATLAB codes for the k-ZIP and k-ZIPG models and the seedling count data used in this article are available with this article at the *Biometrics* website on Wiley Online Library.

Acknowledgements

The authors thank Pankaj Agarwal and Thomas Mølhave for valuable conversations. The work of the authors was supported in part by NSF-CDI 0940671.

References

Agarwal, D. K., Gelfand, A. E., and Citron-Pousty, S. (2002). Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics* **9,** 341–355.

Bechtold, W. A. and Patterson, P. L. (2005). The enhanced forest inventory and analysis program: National sampling design and estimation procedures. *General Technical Report* (SRS-80 edition). Asheville, NC: USDA Forest Service, Southern Research Station.

Cameron, A. C. and Trivedi, P. K. (1986). Econometric models based on count data: comparisons and applications of some estimators and tests. *Journal of Applied Econometrics* **1,** 29–53.

Cameron, A. C. and Trivedi, P. K. (1998). *Regression Analysis of Count Data*. New York: Cambridge University Press.

Canham, C. D. and Thomas, R. Q. (2010). Frequency, not relative abundance, of temperate tree species varies along climate gradients in eastern North America. *Ecology* **91,** 3433–3440.

Cohen, A. C. (1991). *Truncated and Censored Samples*. New York: Marcel Dekker.

Cui, Y. and Yang, W. (2009). Zero-inflated generalized Poisson regression mixture model for mapping quantitative trait loci underlying count trait with many zeros. *Journal of Theoretical Biology* **256,** 276–285.

Ghosh, S. K., Mukhopadhyay, P., and Lu, J. C. (1998). Bayesian analysis of zero-inflated regression models. *Technical Report*, Department of Statistics, North Carolina State University.

Gurmu, S., Rilstone, P., and Stern, S. (1999). Semiparametric estimation of count regression models, *Journal of Econometrics* **88,** 123–150

Johnson, N. L. and Kotz, S. (1969). *Distributions in Statistics: Discrete Distributions*. Boston, MA: Houghton Mifflin.

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

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics* **15,** 209–225.

Melkersson, M. and Rooth, D-O. (2000). Modeling female fertility using inflated count data models. *Journal of Population Economics* **13,** 189–203.

Neelon, B. H., O'Malley, A. J., and Normand, S. L. (2010). A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Statistical Modelling* **10,** 421–439.

Rathbun, S. L. and Fei, S. (2006). A spatial zero-inflated poisson regression model for oak regeneration. *Environmental and Ecological Statistics* **13,** 409–426.

Smith, W. B., Miles, P. D., Perry, C. H., and Pugh, S. A. (2009). Forest resources of the United States, 2007. General Technical Report (WO-78 edition). Washington, DC: USDA Forest Service, Washington Office.

Wang, X. and Dey, D. K. (2010). Generalized extreme value regression for binary response data: An application to B2B electronic payments system adoption. *Annals of Applied Statistics* **4,** 2000–2023.

Welsh, A. H., Cunningham, R. B., Donnellly, C. F., and Lindenmayer, D. B. (1996). Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling* **88,** 297–308.

Received March 2011. *Revised November* 2011. *Accepted November* 2011.

Appendix

A.1. Prior Elicitation for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$

We first fit a Poisson regression using only the nonzero responses to obtain $\mathbf{D}^{(1)}_{\boldsymbol{\beta}_2}$, the asymptotic covariace matrix associated with $\widetilde{\boldsymbol{\beta}}_2$, the estimated $\boldsymbol{\beta}_2$. We perform the same operation on all the responses to get $\mathbf{D}^{(2)}_{\boldsymbol{\beta}_2}$. Then we set $\mathbf{D}_{\boldsymbol{\beta}_2}$ to be a diagonal matrix whose entries are the diagonal elements of $c_2(\mathbf{D}^{(1)}_{\boldsymbol{\beta}_2} + \mathbf{D}^{(2)}_{\boldsymbol{\beta}_2})/2$. Following some sensitivity analysis, we set $c_2 = 10$ in accord with Agarwal et al. (2002).

To obtain $\mathbf{D}_{\boldsymbol{\beta}_1}$, we first need to simulate a set of binary responses. To perform this simulation we set all $\lambda_i = \lambda$ and assume all $\theta_i = \theta$. Let N_0 denote the number of zeros observed in a sample of size n. Then using (6), we set

$$\frac{N_0}{n} = 1 - (1-\theta)^k(1-e^{-\lambda}). \tag{A.1}$$

Let $\overline{\mathbf{Y}}$ and $s^2_{\mathbf{Y}}$ denote the observed mean and variance of the data. Then we set

$$\overline{\mathbf{Y}} = (1-\theta)^k\lambda \tag{A.2}$$

$$s^2_{\mathbf{Y}} = (1-\theta)^k \lambda(1+[1-\{1-\theta\}^k]\lambda). \tag{A.3}$$

Then the equations (A.1.), (A.2.), and (A.3.) provide a "method of moments" estimate of θ. Using the estimates $\widetilde{\boldsymbol{\beta}}_2$ obtained via fitting the Poisson regression and the MoM estimates of θ and k we can simulate a set of binary variables with probability $\frac{\tilde{\theta}}{\tilde{\theta}+(1-\tilde{\theta})Po(0|\widetilde{\boldsymbol{\lambda}})}$, where $\widetilde{\boldsymbol{\lambda}} = \exp(\mathbf{X}_2\widetilde{\boldsymbol{\beta}}_2)$. We then run a standard binary regression on the simulated binary responses and obtain the asymptotic covariance matrix, $\widetilde{\mathbf{D}}_{\boldsymbol{\beta}_1}$, associated with the estimated $\boldsymbol{\beta}_1$. $\mathbf{D}_{\boldsymbol{\beta}_1}$ is set as a diagonal matrix whose entries are the diagonal elements of $\widetilde{\mathbf{D}}_{\boldsymbol{\beta}_1}$. We weaken the prior assumption on $\boldsymbol{\beta}_1$ by multiplying its prior covariance matrix with a constant $c_1 = 10$, following the earlier argument, and after experimenting with several values of c_1 to assess prior sensitivity. This method for prior specification is discussed in greater detail in Agarwal et al. (2002).

A.2. Posterior Distribution of k

Without loss of generality, we assume that the first r observations are 0 and the remaining $n-r$ observation are positive counts. To obtain the posterior distribution of k given all the other parameters, we introduce the following notations:

$$\begin{aligned}
p_i &\triangleq (1-\theta_i)\\
C_i &\triangleq G_0(y_i|\Theta_{0,i})\\
\bar{C}_i &\triangleq 1-G_0(y_i|\Theta_{0,i})\\
p^* &\triangleq \left(\prod_{l=r+1}^{n} p_l\right)\\
C^* &\triangleq \prod_{l=r+1}^{n} C_l\\
\bar{C}^* &\triangleq \prod_{i=1}^{r} \bar{C}_i.
\end{aligned}$$

Using this notations, the likelihood in (9) becomes

$$[\mathbf{Y}|\theta,\Theta_0,k] = \prod_{i=1}^{n}\left\{(1-p_i^k)\delta_{\{0\}} + p_i^k C_i\right\}. \tag{A.4}$$

Then (A.4.) can be expanded as

$$\begin{aligned}
[\mathbf{Y}|\cdots] &\propto \left(1-p_1^k\bar{C}_1\right)\left(1-p_2^k\bar{C}_2\right)\cdots\left(1-p_r^k\bar{C}_r\right)p^{*k}C^*\\
&= \left[1-\sum_{i=1}^{r}p_i^k\bar{C}_i + \sum_{i<j=1}^{r}(p_ip_j)^k\bar{C}_i\bar{C}_j\right.\\
&\quad \left. -\cdots(-1)^r\left(\prod_{i=1}^{r}p_i\right)^k\bar{C}^*\right]p^{*k}C^*\\
&= p^{*k}C^* - \sum_{i=1}^{r}(p_ip^*)^k\bar{C}_iC^* + \sum_{i<j=1}^{r}(p_ip_jp^*)^k\bar{C}_i\bar{C}_jC^*\\
&\quad -\cdots(-1)^r\left(\prod_{i=1}^{n}p_i\right)^k\bar{C}^*C^*\\
&= w_0\alpha_0^k - \sum_{i=1}^{r}w_{1i}\alpha_{1i}^k + \sum_{i<j=1}^{r}w_{2i}w_{2j}(\alpha_{2i}\alpha_{2j})^k\\
&\quad -\cdots(-1)^r\prod_{i=1}^{r}w_{ri}\left(\prod_{i=1}^{r}(\alpha_{ri})\right)^k,
\end{aligned} \tag{A.5}$$

where $\alpha_0 = p^*, w_0 = C^*, \quad \alpha_{1i} = p_ip^*, w_{1i} = \bar{C}_iC^*, \quad \alpha_{2i} = p_i(p^*)^{1/2}, w_{2i} = \bar{C}_iC^{*1/2}$, and $\alpha_{ri} = p_ip^{*1/r}, w_{ri} = \bar{C}_iC^{*1/r}$

Using (A.5.) and assuming a Gamma (k_a, k_b) prior on k, its posterior distribution conditional on all the other parameters $\Theta_0^{(-k)}$ is given by

$$\begin{aligned}
[k|\Theta_0^{(-k)},\mathbf{Y}] &\propto \left[w_0\alpha_0^k - \sum_{i=1}^{r}w_{1i}\alpha_{1i}^k + \sum_{i<j=1}^{r}w_{2i}w_{2j}(\alpha_{2i}\alpha_{2j})^k\right.\\
&\quad \left. -\cdots(-1)^r\prod_{i=1}^{r}w_{ri}\left(\prod_{i=1}^{r}(\alpha_{ri})\right)^k\right]\times k^{k_a-1}e^{-k/k_b}\\
&= \left[w_0e^{k\log\alpha_0} - \sum_{i=1}^{r}w_{1i}e^{k\log\alpha_{1i}}\right.\\
&\quad + \sum_{i<j=1}^{r}w_{2i}w_{2j}e^{k\log(\alpha_{2i}\alpha_{2j})}\\
&\quad \left. -\cdots(-1)^r\prod_{i=1}^{r}w_{ri}e^{k\log\left(\prod_{i=1}^{r}\alpha_{ri}\right)}\right]\times k^{k_a-1}e^{-k/k_b}\\
&= w_0k^{k_a-1}e^{-k(1/k_b-\log(\alpha_0))}\\
&\quad - \sum_{i=1}^{r}w_{1i}k^{k_a-1}e^{-k(1/k_b-\log(\alpha_{1i}))}\\
&\quad + \sum_{i<j=1}^{r}w_{2i}w_{2j}k^{k_a-1}e^{-k(1/k_b-\log(\alpha_{2i}\alpha_{2j}))}\cdots\\
&\quad (-1)^r\prod_{i=1}^{r}w_{ri}k^{k_a-1}e^{-k\left(1/k_b-\log\left(\prod_{i=1}^{r}\alpha_{ri}\right)\right)}\\
&= w_0\pi_{\text{Gamma}}(k_a,(1/k_b-\log(\alpha_0)^{-1})\\
&\quad - \sum_{i=1}^{r}w_{1i}\pi_{\text{Gamma}}(k_a,(1/k_b-\log(\alpha_{1i}))^{-1})\\
&\quad + \sum_{i<j=1}^{r}w_{2i}w_{2j}\pi_{\text{Gamma}}(k_a,(1/k_b-\log(\alpha_{2i}\alpha_{2j}))^{-1})\\
&\quad -\cdots(-1)^r\prod_{i=1}^{r}w_{ri}\pi_{\text{Gamma}}\\
&\quad \times\left(k_a,\left(1/k_b-\log\left(\prod_{i=1}^{r}\alpha_{ri}\right)\right)^{-1}\right)
\end{aligned}$$

Note that, since $\alpha_{li} \in (0,1)$, $\forall\ (l,i) = 1,2,\ldots,r$, the scale parameters of the Gamma distributions are necessarily positive.