

Article

Application of Mixture Models for Doubly Inflated Count Data

Monika Arora ¹  and N. Rao Chaganty ^{2,*}¹ Department of Mathematics, Indraprastha Institute of Information Technology, Delhi 110020, India² Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA 23529-0077, USA

* Correspondence: rchagant@odu.edu

Abstract: In health and social science and other fields where count data analysis is important, zero-inflated models have been employed when the frequency of zero count is high (inflated). Due to multiple reasons, there are scenarios in which an additional count value of $k > 0$ occurs with high frequency. The zero- and k -inflated Poisson distribution model (ZkIP) is more appropriate for such situations. The ZkIP model is a mixture distribution with three components: degenerate distributions at 0 and k count and a Poisson distribution. In this article, we propose an alternative and computationally fast expectation–maximization (EM) algorithm to obtain the parameter estimates for grouped zero and k -inflated count data. The asymptotic standard errors are derived using the complete data approach. We compare the zero- and k -inflated Poisson model with its zero-inflated and non-inflated counterparts. The best model is selected based on commonly used criteria. The theoretical results are supplemented with the analysis of two real-life datasets from health sciences.

Keywords: poisson regression; zero-inflated data; zero- and k -inflated data; EM algorithm; health science



Citation: Arora, M.; Chaganty, N.R. Application of Mixture Models for Doubly Inflated Count Data. *Analytics* **2023**, *2*, 265–283. <https://doi.org/10.3390/analytics2010014>

Received: 18 August 2022

Revised: 13 February 2023

Accepted: 6 March 2023

Published: 11 March 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

A categorical variable deals with a set of categories which could be based on a measurement scale. When there is natural ordering in the measurement scale, the categorical variable is ordinal; otherwise, it is known as a nominal categorical variable. Categorical variables arise not only in medical and social science but also in many other studies such as travel, agriculture, education, finance, ecology, and others. A categorical random variable with the number of counts as its categories is usually modeled by a Poisson distribution. The Poisson distribution has one unknown parameter, $\lambda > 0$. This parameter is also the mean and variance of the distribution. This property of Poisson distribution is known as equi-dispersion. In real-life applications, the count data are often not equi-dispersed; instead, they could be over- or under-dispersed. There could be different reasons for over-dispersion, and one such reason is an excess number of zeros in the data. An appropriate model for such count datasets is the zero-inflated Poisson (ZIP) distribution. In a seminal paper, Lambert [1] studied the ZIP regression model. ZIP models and their applications have been studied extensively in the literature. Ghosh et al. [2] and Agarwal et al. [3] studied ZIP models using a Bayesian approach. Random effects ZIP models were studied by Min and Agresti [4] and Yau and Lee [5]. Furthermore, Saffari and Adnan [6], Yang and Simpson [7], and Nguyen and Dupuy [8] have applied ZIP models for censored data. Recently, a review of various ZIP models was presented in [9,10]. Applications of the ZIP model and its variations can be found in health science [11,12], manufacturing [1,2], and transportation [13,14]. The models have made their mark in biology [15], ecology [16], psychology [17,18], education [19], economics [20–22], and social networks [23].

In count data, besides zero, there could be another count $k > 0$ that is inflated. The inflation could be due to various reasons such as the design of the study or types of responses. For example, the number of pap smear tests performed on women had zero and six inflated [24,25]. Similarly, the data on the number cigarettes smoked have zero

(non-smokers) and 20 (a pack of cigarettes) inflated [25], while in a survey on the number of divorces, counts zero and 1 are likely to be inflated. Arora and Chaganty [24] described such situations where, besides zero, another count $k > 0$ is also inflated, and applied zero- and k -inflated Poisson (ZkIP) models. The ZkIP model for count data is defined by a mixture of three distributions, which assumes that each count observation is a draw from a degenerate distribution at zero with probability π_1 , from a degenerate distribution at value $k > 0$ with probability π_2 , or from a Poisson distribution with probability $(1 - \pi_1 - \pi_2)$. The probabilities π_1 and π_2 can also be viewed as mixing weights. Lin and Tsai [25] studied ZkIP models using the maximum likelihood estimation method. Sheth et al. [26] presented two forms of the ZkIP model. A special case of ZkIP is when $k = 1$. The special case is known as zero- and one-inflated Poisson (ZOIP). The other special case is $\pi_1 = 0$; that is, only $k > 0$ is inflated and the corresponding model is a k -inflated Poisson (kIP) model, which also can be regarded as an extension of the ZIP model. Recently, Arora et al. [27] studied the kIP models using traditional and data science approaches. For doubly censored data, [28] studied zero and one inflation using power-normal distribution.

Most of these aforementioned articles deal with data that contain covariates besides the response variable and study regression models. However, at times the data have missing observations for the covariates. To build a regression model, a list-wise deletion is performed or missing observations are imputed. The deletion of observations could significantly reduce the sample size. On the other hand, the imputed observations could lead to misleading inferences. There is a need to develop inferential methods for data without covariates. These methods without covariates allow us to estimate the inflated proportion for count 0 and k categories. Furthermore, they could also be used as a preliminary step to detect the inflation before using the regression models. Models without covariates are easier to apply and more efficient for large datasets. The proposed model captures the double inflation and is parsimonious. The parameters are simple to interpret and the corresponding analysis is straightforward.

In this article, we deal with the situation where the covariate data are absent and develop an EM algorithm to obtain the estimates for the grouped count data with inflation at zero and $k > 0$. The EM algorithm provides maximum likelihood (ML) estimates when some data are missing or when latent variables are present. For the ZkIP data, the latent variables are the zero's and counts $k > 0$ coming from the degenerate distributions, as opposed to the Poisson distribution. The EM algorithm takes into account the missing information and allows us to obtain the ML estimates of the unknown parameters of the ZkIP model. The standard errors are obtained using the method described by Louis [29]. Our methods include the ZOIP and kIP as special cases. We compare the ZkIP model with the ZIP and Poisson models. We apply our methods on two real-life applications in health sciences. The outline of the article is as follows. Section 2 describes the distributions involved in detail. This includes the ZkIP distribution and its properties. For the grouped data, we present the likelihood function for the ZkIP model in Section 3. In Section 3.1, we present the mathematical details for the expectation–maximization (EM) method [30] to obtain the maximum likelihood estimates. In Section 3.2, we describe the method first described by Louis [29] on how to find the standard errors for the EM estimates for the ZkIP model. Section 4 describes the hypothesis tests for the unknown parameters. It also explains the methods used for model selection and measures to find a model that fits best. In Section 5, we perform two simulation studies. We compare the ZkIP model to ZIP and Poisson models using standardized bias and standardized mean squared error criteria. We also evaluate the coverage probabilities for various confidence levels. Finally, Section 6 contains the analysis of two real-life datasets.

2. Distributions

The Poisson distribution is normally used as a model for count data. The probability mass function of a random variable Y distributed as Poisson with mean $\lambda > 0$ is given by

$$P(Y = y) = \frac{e^{-\lambda} \lambda^y}{y!}, \quad y = 0, 1, 2, \dots$$

The probability mass function of a random variable Y following a zero-inflated Poisson (ZIP) distribution with parameters $\lambda > 0$ and $0 < \pi_1 < 1$ is given by

$$P(Y = y) = \begin{cases} \pi_1 + (1 - \pi_1) e^{-\lambda} & \text{when } y = 0 \\ (1 - \pi_1) \frac{\lambda^y e^{-\lambda}}{y!} & \text{when } y = 1, 2, \dots \end{cases} \quad (1)$$

A generalization of the ZIP model is the ZkIP model which accounts for inflated frequencies at zero and at $k > 0$. The ZkIP distribution is also a mixture model, similar to the ZIP. It is composed of mixing two degenerate distributions with a Poisson distribution. The probability mass function of Y distributed as ZkIP (λ, π_1, π_2) is given by

$$P(Y = y) = \begin{cases} \pi_1 + \pi_3 e^{-\lambda} & \text{when } y = 0 \\ \pi_2 + \pi_3 \frac{\lambda^k e^{-\lambda}}{k!} & \text{when } y = k \\ \pi_3 \frac{\lambda^y e^{-\lambda}}{y!} & \text{when } y = 1, 2, \dots, y \neq k, \end{cases} \quad (2)$$

where $\pi_3 = (1 - \pi_1 - \pi_2)$, $\lambda > 0$ and $0 < \pi_1 + \pi_2 < 1$. The corresponding cumulative distribution function (CDF) is given by

$$F_Y(y) = \begin{cases} 0 & \text{when } y < 0 \\ \pi_1 + \pi_3 \sum_{u=0}^{\lfloor y \rfloor} \frac{\lambda^u e^{-\lambda}}{u!} & \text{when } 0 \leq y < k \\ \pi_1 + \pi_2 + \pi_3 \sum_{u=0}^{\lfloor y \rfloor} \frac{\lambda^u e^{-\lambda}}{u!} & \text{when } y \geq k, \end{cases} \quad (3)$$

where $\lfloor y \rfloor$ is the floor function. Using (3), we can show that the probability generating function of Y is $G_Y(z) = E(z^Y) = \pi_1 + \pi_2 z^k + \pi_3 e^{\lambda(z-1)}$. The moment generating function is given by $M_Y(t) = E(e^{tY}) = \pi_1 + \pi_2 e^{tk} + \pi_3 e^{\lambda(e^t-1)}$. The mean $E(Y) = k\pi_2 + \pi_3\lambda$ and $Var(Y) = k^2\pi_2(1 - \pi_2) + \pi_3\lambda(1 + \pi_1\lambda + \pi_2\lambda - 2k\pi_2)$ can be obtained taking derivatives of $M_Y(t)$ with respect to t at $t = 0$.

The unknown parameters in a ZkIP distribution are λ , π_1 , and π_2 with $\lambda > 0$ and $0 < \pi_1 + \pi_2 < 1$. There are various methods for estimating the parameters and drawing inferences. In the next section, we develop the expectation–maximization (EM) algorithm to obtain the maximum likelihood estimates of the ZkIP model parameters and the corresponding standard errors.

3. Methodology

Suppose that we have a vector $\mathbf{y} = (y_1, y_2, \dots, y_n)$ consisting of a random sample of n observations potentially from a ZkIP distribution. The frequency distribution of the sample can be organized in a table as

j	0	1	...	k	...	K	Total
Observed frequency	n_0	n_1	...	n_k	...	n_K	n

Here, $n_j = \#$ of y_i 's that are equal to j and $K = \max\{y_i\}$. If the observations are truly from the ZkIP distribution, the values of n_0 and n_k will be large compared to the rest of the frequencies. The vector of observed frequencies (n_0, n_1, \dots, n_K) can be regarded as incomplete data in the sense that n_0 is actually $n_a + n_b$ and $n_k = n_c + n_d$, where the unknown n_a and n_c are frequencies from degenerate distributions at 0 and k , respectively. Using (3), we can write the likelihood function of the observed frequencies n_i 's as

$$\begin{aligned} L_{obs}(\pi_1, \pi_2, \lambda | \mathbf{y}) &\propto (\pi_1 + \pi_3 e^{-\lambda})^{n_0} \left(\pi_2 + \pi_3 \frac{\lambda^k e^{-\lambda}}{k!} \right)^{n_k} \prod_{j \neq 0, k}^K \left(\pi_3 \frac{\lambda^j e^{-\lambda}}{j!} \right)^{n_j} \\ &\propto (\pi_1 + \pi_3 p_0)^{n_0} (\pi_2 + \pi_3 p_k)^{n_k} \prod_{j \neq 0, k}^K (\pi_3 p_j)^{n_j}, \end{aligned} \quad (4)$$

where $p_j = (\lambda^j e^{-\lambda})/j!$ and $\pi_3 = (1 - \pi_1 - \pi_2)$. Note that when $\pi_2 = 0$, the ZkIP reduces to ZIP. Thus, the likelihood for the ZIP model is

$$L_{obs}(\pi_1, \lambda | \mathbf{y}) \propto (\pi_1 + (1 - \pi_1)e^{-\lambda})^{n_0} \prod_{j \neq 0}^K \left((1 - \pi_1) \frac{\lambda^j e^{-\lambda}}{j!} \right)^{n_j}.$$

If $\pi_1 = \pi_2 = 0$, the likelihood (4) becomes the likelihood function of the Poisson distribution given by

$$L_{obs}(\lambda | \mathbf{y}) = \prod_{j=0}^K \left(\frac{\lambda^j e^{-\lambda}}{j!} \right)^{n_j}.$$

3.1. EM Estimation

For the likelihood (4), the unknown parameter $\theta = (\pi_1, \pi_2, \lambda)$ can be estimated using the maximum likelihood (ML) approach. A computationally simple approach to get an ML estimate of θ is the expectation–maximum (EM) method, which was introduced by [30] in a seminal paper. The EM algorithm is a simple modification of the maximum likelihood and has become a popular alternative for ML estimation in cases where data are missing or incomplete. We describe the EM approach to study the ZkIP model for grouped data.

The frequency vector $(n_0, n_1, \dots, n_k, \dots, n_K)$ is the observed data. It can be viewed as partially incomplete data, in the sense that $n_0 = n_a + n_b$ and $n_k = n_c + n_d$, since the number, n_a , of zeros and the number, n_c , of k s are missing. Here n_a and n_c are the unknown number of observations from degenerate distributions at 0 and k , respectively. Thus, the complete data vector including the missing frequencies is $(n_a, n_b, n_1, \dots, n_c, n_d, \dots, n_K)$. The likelihood function of this complete data vector is

$$L_{comp}(\pi_1, \pi_2, \lambda | \mathbf{y}) \propto \pi_1^{n_a} \pi_2^{n_b} \pi_3^{(n - n_a - n_c)} p_0^{n_b} p_k^{n_d} \prod_{j \neq 0, k}^K p_j^{n_j} \quad (5)$$

where $p_j = (\lambda^j e^{-\lambda})/j!$ and $\pi_3 = (1 - \pi_1 - \pi_2)$. Our interest is to maximize the likelihood (5) or minimize the negative of the log-likelihood. The log-likelihood, $\ell_{comp} = \log L_{comp}$, can be written as

$$\begin{aligned}\ell_{comp}(\pi_1, \pi_2, \lambda | \mathbf{y}) &\propto n_a \log(\pi_1) + n_c \log(\pi_2) + (n - n_a - n_c) \log \pi_3 \\ &\quad + n_b \log p_0 + n_d \log p_k + \sum_{j \neq 0, k}^K n_j \log p_j \\ &\propto n_a \log(\pi_1) + n_c \log(\pi_2) + (n - n_a - n_c) \log \pi_3 \\ &\quad - n_b \lambda + n_d (-\lambda + k \log \lambda) + \sum_{j \neq 0, k}^K n_j (-\lambda + j \log \lambda),\end{aligned}\quad (6)$$

where the frequencies n_a and n_c are unknown. The expectation step in the EM algorithm replaces these frequencies with their expected values. To obtain the expected values, we assume there is a latent variable $\mathbf{z} = (z_1, z_2, z_3)$ distributed as a multinomial with parameter vector $(1, \pi_1, \pi_2, \pi_3)$, where the number of trials is one. Here, \mathbf{z} takes values $(1, 0, 0)$ with probability π_1 , $(0, 1, 0)$ with probability π_2 , and $(0, 0, 1)$ with probability π_3 . That is,

$$P(\mathbf{z} = (z_1, z_2, z_3)) = \begin{cases} \pi_1 & \text{if } z_1 = 1, z_2 = 0, z_3 = 0 \\ \pi_2 & \text{if } z_1 = 0, z_2 = 1, z_3 = 0 \\ \pi_3 & \text{if } z_1 = 0, z_2 = 0, z_3 = 1. \end{cases}\quad (7)$$

Furthermore, assume the conditional distribution of $Y | \mathbf{z}$ is

$$P(Y = y | \mathbf{z} = (z_1, z_2, z_3)) = \begin{cases} 1 & \text{for } z_1 = 1, y = 0 \\ 1 & \text{for } z_2 = 1, y = k \\ \frac{\lambda^y e^{-\lambda}}{y!} & \text{for } z_3 = 1, y = 0, 1, \dots \end{cases}\quad (8)$$

Now, the joint distribution of (Y, \mathbf{z}) obtained by multiplying (7) and (8) is

$$P(Y = y, \mathbf{z} = (z_1, z_2, z_3)) = \begin{cases} \pi_1 & \text{for } z_1 = 1, y = 0 \\ \pi_2 & \text{for } z_2 = 1, y = k \\ \pi_3 \frac{\lambda^y e^{-\lambda}}{y!} & \text{for } z_3 = 1, y = 0, 1, \dots \end{cases}\quad (9)$$

The marginal of Y can be obtained from (9) by summing over the three possible values of \mathbf{z} . Thus, we obtain

$$\begin{aligned}P(Y = 0) &= P(Y = 0, z_1 = 1) + P(Y = 0, z_2 = 1) + P(Y = 0, z_3 = 1) \\ &= \pi_1 + \pi_3 e^{-\lambda}, \\ P(Y = k) &= P(Y = k, z_1 = 1) + P(Y = k, z_2 = 1) + P(Y = k, z_3 = 1) \\ &= \pi_2 + \pi_3 \frac{\lambda^k e^{-\lambda}}{k!},\end{aligned}$$

and

$$\begin{aligned}P(Y = y) &= P(Y = y, z_1 = 1) + P(Y = y, z_2 = 1) + P(Y = y, z_3 = 1) \\ &= \pi_3 \frac{\lambda^y e^{-\lambda}}{y!}, \quad \text{for } y = 1, 2, \dots, y \neq k,\end{aligned}$$

which is equivalent to the ZkIP distribution defined by (3). Now, the conditional expected values can be computed by the posterior probabilities given in Table 1.

Table 1. $P(\mathbf{z} = (z_1, z_2, z_3) | Y = y)$ for ZkIP model.

$\mathbf{z} = (z_1, z_2, z_3)$	$y = 0$	$y = k$	$y \neq 0, k$
$(1, 0, 0)$	$\frac{\pi_1}{\pi_1 + \pi_3 p_0}$	0	0
$(0, 1, 0)$	0	$\frac{\pi_2}{\pi_2 + \pi_3 p_k}$	0
$(0, 0, 1)$	$\frac{\pi_3 p_0}{\pi_1 + \pi_3 p_0}$	$\frac{\pi_3 p_k}{\pi_2 + \pi_3 p_k}$	1

The three latent variables z_i s are the indicator variables for the three distributions in the ZkIP mixture model. More specifically,

$$\begin{aligned}\hat{n}_a &= n_0 E(z_1 | y = 0) = n_0 P(z_1 = 1 | y = 0) = n_0 \frac{\pi_1}{\pi_1 + \pi_3 p_0}, \\ \hat{n}_c &= n_k E(z_2 | y = k) = n_k P(z_2 = 1 | y = k) = n_k \frac{\pi_2}{\pi_2 + \pi_3 p_k}.\end{aligned}\quad (10)$$

The maximization step or the M-step in the EM algorithm involves maximizing the log-likelihood (6) after substituting these estimates for n_a and n_c . However, this maximization is easy since the score equations have closed-form solutions. Indeed, equating partial derivatives with respect to the three parameters of (6) to zero we obtain,

$$\frac{\partial \ell_{comp}(\pi_1, \pi_2, \lambda)}{\partial \pi_1} = 0 \iff \hat{\pi}_1 = \frac{n_a(1 - \pi_2)}{n - n_c}, \quad (11)$$

$$\frac{\partial \ell_{comp}(\pi_1, \pi_2, \lambda)}{\partial \pi_2} = 0 \iff \hat{\pi}_2 = \frac{n_c(1 - \pi_1)}{n - n_a}, \quad (12)$$

$$\frac{\partial \ell_{comp}(\pi_1, \pi_2, \lambda)}{\partial \lambda} = 0 \iff \hat{\lambda} = \frac{\sum_{j=0}^K j n_j}{n - n_a - n_c}. \quad (13)$$

We summarize the steps of the EM algorithm as follows:

1. Choose the initial values of π_1^0 , π_2^0 , and λ^0 for π_1 , π_2 and λ , respectively.
2. E-step: Calculate \hat{n}_a and \hat{n}_c using (10), and set $\hat{n}_b = n_0 - \hat{n}_a$ and $\hat{n}_d = n_1 - \hat{n}_c$.
3. M-step: Update the estimates of π_1 , π_2 , and λ using the formulas in (11), (12), and (13).
4. Iterate the E-step and M-step until the estimates $\hat{\pi}_1$, $\hat{\pi}_2$, and $\hat{\lambda}$ converge.

We have developed an R code for this algorithm and used it for the two data analysis examples in Section 6.

3.2. Standard Errors of EM Estimates

The optimization algorithms routinely output a numerically computed Hessian matrix for the functions that are being optimized. However, calculation of the standard errors will be more accurate if analytical expressions are available. To compute the standard errors of the estimates obtained by the EM algorithm, we follow the approach described by Louis [29]. The relation between the likelihood of complete, observed, and missing data is given as

$$L_{comp}(\boldsymbol{\theta} | \mathbf{y}, \mathbf{z}) = L_{obs}(\boldsymbol{\theta} | \mathbf{y}) L_{miss}(\boldsymbol{\theta} | (\mathbf{z} | \mathbf{y})), \quad (14)$$

where \mathbf{y} and \mathbf{z} stand for the observed and missing data, respectively. From (14), taking logs we obtain

$$\ell_{comp}(\boldsymbol{\theta} | \mathbf{y}, \mathbf{z}) = \ell_{obs}(\boldsymbol{\theta} | \mathbf{y}) + \ell_{miss}(\boldsymbol{\theta} | (\mathbf{z} | \mathbf{y})). \quad (15)$$

Taking second order partial derivatives, we can see that from Equation (15), the information matrices for the complete, observed, and missing data satisfy the following identity

$$\mathcal{I}_{comp} = \mathcal{I}_{obs} + \mathcal{I}_{miss}$$

or

$$\mathcal{I}_{obs} = \mathcal{I}_{comp} - \mathcal{I}_{miss}. \quad (16)$$

Since the right hand side of Equation (16) depends on the missing data, Louis [29] suggested to take the expected value of the missing data given the observed. This gives us the identity

$$\mathcal{I}_{obs} = E(\mathcal{I}_{obs} | \mathbf{y}) = E(\mathcal{I}_{comp} | \mathbf{y}) - E(\mathcal{I}_{miss} | \mathbf{y}). \quad (17)$$

In other words, an estimate of the observed information matrix is given by

$$\widehat{\mathcal{I}}_{obs} = E(\mathcal{I}_{comp} | \mathbf{y}) - E(\mathcal{I}_{miss} | \mathbf{y}). \quad (18)$$

Regularity conditions under which these information matrices are non-singular are given in [31]. Without going into technical details, we can say the salient regularity conditions are (1) the ranges of the random variables do not depend on the parameter; (2) the partial derivatives of the pdf with respect to the parameters exist; and (3) the integrals of the partial derivatives are same as the partial derivatives of the integrals with respect to the parameters. Under these conditions, the standard errors of the parameter estimates can be obtained taking the square root of the diagonal elements of inverse of the observed information matrix (18). Note that

$$\mathcal{I}_{comp} = \begin{bmatrix} -\frac{\partial^2 \ell_{comp}}{\partial \pi_1^2} & -\frac{\partial^2 \ell_{comp}}{\partial \pi_1 \partial \pi_2} & -\frac{\partial^2 \ell_{comp}}{\partial \pi_1 \partial \lambda} \\ -\frac{\partial^2 \ell_{comp}}{\partial \pi_2 \partial \pi_1} & -\frac{\partial^2 \ell_{comp}}{\partial \pi_2^2} & -\frac{\partial^2 \ell_{comp}}{\partial \pi_2 \partial \lambda} \\ -\frac{\partial^2 \ell_{comp}}{\partial \lambda \partial \pi_1} & -\frac{\partial^2 \ell_{comp}}{\partial \lambda \partial \pi_2} & -\frac{\partial^2 \ell_{comp}}{\partial \lambda^2} \end{bmatrix}. \quad (19)$$

From (6), the elements of the information matrix \mathcal{I}_{comp} are

$$\begin{aligned} -\frac{\partial^2 \ell_{comp}}{\partial \pi_1^2} &= \frac{n_a}{\pi_1^2} + \frac{n - n_a - n_c}{\pi_3^2} \\ -\frac{\partial^2 \ell_{comp}}{\partial \pi_1 \partial \pi_2} &= -\frac{\partial^2 \ell_{comp}}{\partial \pi_2 \partial \pi_1} = \frac{n - n_a - n_c}{\pi_3^2} \\ -\frac{\partial^2 \ell_{comp}}{\partial \pi_2^2} &= \frac{n_c}{\pi_2^2} + \frac{n - n_a - n_c}{\pi_3^2} \\ -\frac{\partial^2 \ell_{comp}}{\partial \lambda^2} &= \frac{n_d k}{\lambda^2} + \frac{\sum_{j \neq 0, k}^K j n_j}{\lambda^2}. \end{aligned}$$

The other elements $-\partial^2 \ell_{comp} / \partial \pi_1 \partial \lambda$ and $-\partial^2 \ell_{comp} / \partial \pi_2 \partial \lambda$ are equal to zero. Since n_a and n_c are missing, we replace them by their expected values

$$E(n_a | n_0) = \frac{n_0 \pi_1}{\pi_1 + \pi_3 p_0} \quad \text{and} \quad E(n_c | n_k) = \frac{n_k \pi_2}{\pi_2 + \pi_3 p_k}.$$

Thus, the nonzero elements of $E(\mathcal{I}_{comp} | \mathbf{y}) = E(\mathcal{I}_{comp} | n_0, n_k)$ are

$$E \left[-\frac{\partial^2 \ell_{comp}}{\partial \pi_1^2} \right] = \frac{n}{\pi_3^2} + \frac{n_0}{\pi_1(\pi_1 + \pi_3 p_0)} - \frac{n_0 \pi_1}{\pi_3^2(\pi_1 + \pi_3 p_0)} - \frac{n_k \pi_2}{\pi_3^2(\pi_2 + \pi_3 p_k)}$$

and

$$\begin{aligned} E \left[-\frac{\partial^2 \ell_{comp}}{\partial \pi_1 \partial \pi_2} \right] &= E \left[-\frac{\partial^2 \ell_{comp}}{\partial \pi_2 \partial \pi_1} \right] = \frac{n}{\pi_3^2} - \frac{n_0 \pi_1}{\pi_3^2(\pi_1 + \pi_3 p_0)} - \frac{n_k \pi_2}{\pi_3^2(\pi_2 + \pi_3 p_k)} \\ E \left[-\frac{\partial^2 \ell_{comp}}{\partial \pi_2^2} \right] &= \frac{n}{\pi_3^2} - \frac{n_0 \pi_1}{\pi_3^2(\pi_1 + \pi_3 p_0)} + \frac{n_k}{\pi_2(\pi_2 + \pi_3 p_k)} - \frac{n_k \pi_2}{\pi_3^2(\pi_2 + \pi_3 p_k)} \\ E \left[-\frac{\partial^2 \ell_{comp}}{\partial \lambda^2} \right] &= \frac{n_k k}{\lambda^2} - \frac{n_k k \pi_2}{\lambda^2(\pi_2 + \pi_3 p_k)} + \frac{1}{\lambda^2} \sum_{j \neq 0, k}^K j n_j. \end{aligned}$$

Next, to compute the second term, $E(\mathcal{I}_{miss} | \mathbf{y})$, in Equation (16), we proceed as follows. The likelihood of the observed and complete data are given in (4) and (5), respectively. Hence, the likelihood of the missing data is obtained taking the ratio of these likelihoods and it is given by

$$L_{miss}(\pi_1, \pi_2, \lambda | \mathbf{z}) \propto \pi_1^{n_a} \pi_2^{n_c} (p_0 \pi_3)^{n_b} (p_k \pi_3)^{n_d} \left(\frac{1}{\pi_1 + \pi_3 p_0} \right)^{n_0} \left(\frac{1}{\pi_2 + \pi_3 p_k} \right)^{n_k}.$$

Thus, the log-likelihood of the missing data is

$$\begin{aligned} \ell_{miss}(\pi_1, \pi_2, \lambda | \mathbf{y}) &\propto n_a \log(\pi_1) + n_c \log(\pi_2) - n_0 \log(\pi_1 + \pi_3 p_0) \\ &\quad - n_k \log(\pi_2 + \pi_3 p_k) + (n_b + n_d) \log(\pi_3) \\ &\quad - (n_b + n_d) \lambda + (n_d k) \log(\lambda). \end{aligned} \quad (20)$$

We can easily check that the first-order partial derivatives are

$$\begin{aligned} \frac{\partial \ell_{miss}}{\partial \pi_1} &= \frac{n_a}{\pi_1} - n_0 \left(\frac{1 - p_0}{\pi_1 + \pi_3 p_0} \right) - \frac{n_b + n_d}{\pi_3} + \frac{n_k p_k}{\pi_2 + \pi_3 p_k} \\ \frac{\partial \ell_{miss}}{\partial \pi_2} &= \frac{n_c}{\pi_2} + n_0 \left(\frac{p_0}{\pi_1 + \pi_3 p_0} \right) - \frac{n_b + n_d}{\pi_3} - \frac{n_k(1 - p_k)}{\pi_2 + \pi_3 p_k} \\ \frac{\partial \ell_{miss}}{\partial \lambda} &= \frac{n_d k}{\lambda} - (n_b + n_d) + n_0 \left(\frac{\pi_3 p_0}{\pi_1 + \pi_3 p_0} \right) \\ &\quad - \frac{n_k \pi_3 p_k}{\pi_2 + \pi_3 p_k} \left(\frac{k}{\lambda} - 1 \right). \end{aligned}$$

and the negative of the second-order partial derivatives are

$$\begin{aligned}
 -\frac{\partial^2 \ell_{miss}}{\partial \pi_1^2} &= \frac{n_a}{\pi_1^2} - \frac{n_0(1-p_0)^2}{(\pi_1 + \pi_3 p_0)^2} - \frac{n_k p_k^2}{(\pi_2 + \pi_3 p_k)^2} + \frac{(n_b + n_d)}{\pi_3^2} \\
 -\frac{\partial^2 \ell_{miss}}{\partial \pi_1 \partial \pi_2} &= \frac{n_0 p_0(1-p_0)}{(\pi_1 + \pi_3 p_0)^2} + \frac{n_k p_k(1-p_k)}{(\pi_2 + \pi_3 p_k)^2} + \frac{(n_b + n_d)}{\pi_3^2} \\
 -\frac{\partial^2 \ell_{miss}}{\partial \pi_1 \partial \lambda} &= \frac{n_0(1-\pi_2)p_0}{(\pi_1 + \pi_3 p_0)^2} - \frac{n_k \pi_2 p_k}{(\pi_2 + \pi_3 p_k)^2} \left(\frac{k}{\lambda} - 1 \right) \\
 -\frac{\partial^2 \ell_{miss}}{\partial \pi_2^2} &= \frac{n_c}{\pi_2^2} - \frac{n_0 p_0^2}{(\pi_1 + \pi_3 p_0)^2} - \frac{n_k(1-p_k)^2}{(\pi_2 + \pi_3 p_k)^2} + \frac{(n_b + n_d)}{\pi_3^2} \\
 -\frac{\partial^2 \ell_{miss}}{\partial \pi_2 \partial \lambda} &= \frac{n_0 \pi_1 p_0}{(\pi_1 + \pi_3 p_0)^2} - \frac{n_k(1-\pi_1)p_k}{(\pi_2 + \pi_3 p_k)^2} \left(\frac{k}{\lambda} - 1 \right) \\
 -\frac{\partial^2 \ell_{miss}}{\partial \lambda^2} &= \frac{n_0 \pi_1 \pi_3 p_0}{(\pi_1 + \pi_3 p_0)^2} + \frac{n_k \pi_2 \pi_3 p_k}{(\pi_2 + \pi_3 p_k)^2} \left(\frac{k}{\lambda} - 1 \right)^2 \\
 &\quad - \frac{k}{\lambda^2} \frac{n_k \pi_3 p_k}{\pi_2 + \pi_3 p_k} + \frac{k n_d}{\lambda^2}.
 \end{aligned}$$

Once again, using the expected values

$$\begin{aligned}
 E(n_a|n_0) &= \frac{n_0 \pi_1}{\pi_1 + \pi_3 p_0} \quad \text{and} \quad E(n_c|n_k) = \frac{n_k \pi_2}{\pi_2 + \pi_3 p_k}, \\
 E(n_b|n_0) &= \frac{n_0 \pi_3 p_0}{\pi_1 + \pi_3 p_0} \quad \text{and} \quad E(n_d|n_k) = \frac{n_k \pi_3 p_k}{\pi_2 + \pi_3 p_k},
 \end{aligned}$$

we obtain the elements of $E(\mathcal{I}_{miss}|\mathbf{y}) = E(\mathcal{I}_{miss}|n_0, n_k)$ as follows

$$\begin{aligned}
 E \left[-\frac{\partial^2 \ell_{miss}}{\partial \pi_1^2} \right] &= \frac{n_0}{\pi_1(\pi_1 + \pi_3 p_0)} - \frac{n_0(1-p_0)^2}{(\pi_1 + \pi_3 p_0)^2} - \frac{n_k p_k^2}{(\pi_2 + \pi_3 p_k)^2} \\
 &\quad + \frac{n_0 p_0}{\pi_3(\pi_1 + \pi_3 p_0)} + \frac{n_k p_k}{\pi_3(\pi_2 + \pi_3 p_k)} \\
 E \left[-\frac{\partial^2 \ell_{miss}}{\partial \pi_1 \partial \pi_2} \right] &= \frac{n_0 p_0(1-p_0)}{(\pi_1 + \pi_3 p_0)^2} + \frac{n_k p_k(1-p_k)}{(\pi_2 + \pi_3 p_k)^2} \\
 &\quad + \frac{n_0 p_0}{\pi_3(\pi_1 + \pi_3 p_0)} + \frac{n_k p_k}{\pi_3(\pi_2 + \pi_3 p_k)}
 \end{aligned}$$

and

$$\begin{aligned}
 E \left[-\frac{\partial^2 \ell_{miss}}{\partial \pi_1 \partial \lambda} \right] &= \frac{n_0(1-\pi_2)p_0}{(\pi_1 + \pi_3 p_0)^2} - \frac{n_k \pi_2 p_k}{(\pi_2 + \pi_3 p_k)^2} \left(\frac{k}{\lambda} - 1 \right) \\
 E \left[-\frac{\partial^2 \ell_{miss}}{\partial \pi_2^2} \right] &= \frac{n_k}{\pi_2(\pi_2 + \pi_3 p_k)} - \frac{n_0 p_0^2}{(\pi_1 + \pi_3 p_0)^2} - \frac{n_k(1-p_k)^2}{(\pi_2 + \pi_3 p_k)^2} \\
 &\quad + \frac{n_0 p_0}{\pi_3(\pi_1 + \pi_3 p_0)} + \frac{n_k p_k}{\pi_3(\pi_2 + \pi_3 p_k)} \\
 E \left[-\frac{\partial^2 \ell_{miss}}{\partial \pi_2 \partial \lambda} \right] &= \frac{n_0 \pi_1 p_0}{(\pi_1 + \pi_3 p_0)^2} - \frac{n_k(1-\pi_1)p_k}{(\pi_2 + \pi_3 p_k)^2} \left(\frac{k}{\lambda} - 1 \right) \\
 E \left[-\frac{\partial^2 \ell_{miss}}{\partial \lambda^2} \right] &= \frac{n_0 \pi_1 \pi_3 p_0}{(\pi_1 + \pi_3 p_0)^2} + \frac{n_k \pi_2 \pi_3 p_k}{(\pi_2 + \pi_3 p_k)^2} \left(\frac{k}{\lambda} - 1 \right)^2.
 \end{aligned}$$

The remaining elements follow by symmetry. Matrices $E(\mathcal{I}_{comp}|\mathbf{y})$ and $E(\mathcal{I}_{miss}|\mathbf{y})$ are positive definite and so they are non-singular [32,33].

4. Goodness of Fit and Model Selection

Hypothesis testing usually follows parameter estimation to check the significance of the parameters. In the presence of competing models, there is a need to compare and find the best model. There are various measures useful for model comparisons, the most popular being the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). It is also important to check the goodness of fit of the models. This is accomplished using the Pearson chi-square statistic and the sum of absolute error. In this section, we will discuss the aforementioned topics, namely hypothesis testing, model selection, and goodness of fit.

4.1. Hypothesis Testing and Measures of Model Selection

The unknown parameter in the ZkIP distribution is $\theta = (\pi_1, \pi_2, \lambda)$. The parameter λ is the rate parameter of the Poisson distribution, and thus $\lambda > 0$, while π_1 represents the proportion of zeros, π_2 is the proportion of $k > 0$ from the degenerate distributions, and $0 < \pi_1 + \pi_2 < 1$. To study the statistical significance of the unknown parameter θ , we can perform various hypothesis tests. Under standard regularity conditions, the EM estimate $\hat{\theta}$ of θ is asymptotically normal with mean θ^0 and a covariance matrix given by $(\widehat{\mathcal{I}}_{obs})^{-1}$, where θ^0 is its true value, which we assume lies in the interior of the parameter space. We can use this result to construct a Wald's test for the null hypothesis $H_0 : \lambda = \lambda_0$ versus the alternative $H_0 : \lambda \neq \lambda_0$. The test statistic $\hat{\lambda}/SE(\hat{\lambda})$ is asymptotically normal. Similarly, the Wald test could be used to test the hypotheses for a specified proportion $\pi_2 = \pi_2^0$ of observations coming from a degenerate distribution at k or a specified proportion $\pi_1 = \pi_1^0$ coming from the degenerate distribution at zero.

The FMM and Countreg procedures in SAS use the parameters $\gamma = \log(\pi_1/\pi_3)$ and $\delta = \log(\pi_2/\pi_3)$ and test for the hypothesis $H_0 : (\gamma, \delta) = (0, 0)$. This hypothesis is equivalent to testing $H_0 : (\pi_1, \pi_2) = (\pi_1^0 = 1/3, \pi_2^0 = 1/3)$, which we could do using Wald's test because $\pi_1^0 = 1/3$ and $\pi_2^0 = 1/3$ are values in the interior of the respective parameter spaces.

As mentioned in Section 3, when $\pi_2 = 0$, ZkIP reduces to ZIP, and additionally if $\pi_1 = 0$, the model simply is a Poisson distribution. Thus, we can say ZkIP, ZIP, and Poisson are nested models. We can perform likelihood ratio tests (LRT) for model reduction for these nested models. To test the significance of inflation at $k > 0$, the appropriate hypothesis would be $H_0 : \pi_2 = 0$ versus $H_1 : \pi_2 > 0$. Similarly, the significance of inflation at zero could be tested by $H_0 : \pi_1 = 0$ versus $H_1 : \pi_1 > 0$. The problem is that the null hypothesis is on the boundary of the parameter space in both scenarios, and therefore the regularity conditions are not satisfied. However, Chant [34] and Shapiro [35] have shown that the test statistic $-2 \log L(\hat{\theta})$ is asymptotically distributed under the null hypothesis as $0.5\chi_0^2 + 0.5\chi_1^2$, a mixture of chi-square distributions. We could use this result to test the hypotheses.

To find the best model we could use various criteria, the most popular being the Akaike information criterion (AIC). It selects the best model based on the expected difference between the hypothesized model and the observed data. The minimum difference, that is, the model with minimum AIC, is considered as the best among the analyzed models. The AIC is given by $-2 \log L(\hat{\theta}) + 2m$. Here, $\log L(\hat{\theta})$ is the log-likelihood of the model at the ML estimates, while m is the number of parameters in the model. Recall that for a Poisson model, there is only one parameter λ . In ZIP there are two parameters, λ and π_1 , and ZkIP has an additional parameter π_2 . Akaike [36] has suggested not just the minimum value of the AIC that is of relevance, but also the difference between the AICs of various models. A rule of thumb to select the best model from a set of competing models for data, can be based on the difference between the AICs as given in Table 2.

Table 2. Rules of thumb [36] for $\Delta_i = AIC_i - AIC_{min}$.

Δ_i	Level of Empirical Support of Model i
0–2	Substantial
4–7	Considerably less
>10	Essentially none

The other popularly used measure to select the best model is Bayesian information criterion (BIC). The BIC is given by $-2\log L(\hat{\theta}) + m \log n$, where n is the sample size. Similar to the AIC, the model with the minimum value of BIC among the competing models is the best. The AIC and BIC both penalize for adding more parameters to the model. The rules of thumb to study the difference between the BICs are given in Table 3. To choose the best model, we select the model with the minimum BIC. When the difference, $\Delta_i = BIC_i - BIC_{min}$, is high then there is sufficient evidence against the competing models and the model with minimum BIC is the best.

Table 3. Rules of thumb [37] for $\Delta_i = BIC_i - BIC_{min}$.

Δ_i	Evidence Against a Candidate Model to Be the Best Model
$0 \leq \Delta_i \leq 2$	Not worth more than a bare mention
$2 < \Delta_i \leq 6$	Positive
$6 < \Delta_i \leq 10$	Strong
$\Delta_i > 10$	Very strong

4.2. Model Checking

There is also a need to check how well the best model among completing models fits the data. The goodness of fit of a model is studied using various measures. A commonly used measure is the Pearson statistic $\chi^2 = \sum (O_i - E_i)^2 / E_i$, where O_i is the observed frequency and E_i is the expected frequency of the i -th count. This statistic, under the null hypothesis, asymptotically follows a chi-square distribution with $(\kappa - 1)$ degrees of freedom, where κ is the total number of categories. Large values of the test statistic lead to rejection of the null hypothesis. For inflated data, the χ^2 values are usually high and thus tend to reject the null hypothesis. In such scenarios, a better measure is the sum of absolute errors (ABE) given by

$$\text{sum ABE} = \sum |O_i - E_i|.$$

We employ these model checking criteria for the analysis of two real-life datasets in Section 6.

5. Simulations

To study the performance of the proposed EM algorithm, we have conducted some simulation studies. Data $Y = (Y_1, \dots, Y_n)$ of sample size n are generated from the ZkIP distribution with parameter vector $\theta = (\pi_1, \pi_2, \lambda)$. For varying values of θ and values of $n = 200, 500, 1000$ and 2000 , we simulated $N = 10,000$ datasets. Using these simulated data, we compare the performance of the ZkIP model to ZIP and ordinary Poisson using the standardized bias (SBias), standardized mean squared error (SMSE), and coverage probability criteria. The SBias and SMSE are more informative than Bias and MSE, respectively, and thus are preferable [38]. The standardized bias is given by

$$\begin{aligned} SBias(\theta) &= E(\hat{\theta} - \theta) / \theta \\ &\approx \left(\sum_{i=1}^N \frac{\hat{\theta}^i - \theta}{\theta} \right) / N \end{aligned}$$

The standardized mean squared error is given by

$$\begin{aligned} SMSE(\theta) &= E(\hat{\theta} - \theta)^2 / \theta^2 \\ &\approx \left(\sum_{i=1}^N \frac{(\hat{\theta}^i - \theta)^2}{\theta^2} \right) / N \end{aligned}$$

The coverage probability of the parameters θ is the proportion of times the confidence interval contains the true value of the parameter. We considered 90%, 95%, and 99% confidence intervals for all of the parameters and various sample sizes.

5.1. Simulation I

In our first simulation study, we generated data from ZkIP with $\lambda = 2$ and a probability at zero of $\pi_1 = 0.2$, and assumed that $k = 2$ is inflated with probability $\pi_2 = 0.4$. The data were independently generated $N = 10,000$ times for each value of $n = 200, 500, 1000$, and 2000 . For ZIP and Poisson models, the standardized bias was negative for each sample size. This indicates that the models underestimate the parameters. The SBias for ZkIP was close to zero for all the parameters and all the sample sizes. Similarly, the SMSE was the smallest for the parameters of the ZkIP model, irrespective of the sample size. As expected, the SMSE decreased as the sample size increased. In this simulation exercise, we observed that the mean estimated values of the ZkIP parameters are close to the true values, and the SBias and SMSE are very close to zero (see Tables 4 and 5). Thus, we conclude that the performance of the proposed EM algorithm is precise and accurate in this case.

To obtain the confidence intervals, we evaluated the EM estimates and SE of the parameters at each iteration using the methods proposed in Sections 3.1 and 3.2, respectively. Table 6 shows that for all confidence levels (90%, 95%, and 99%), the coverage probabilities are close to the nominal levels for all the parameters irrespective of the sample size.

Table 4. Comparison of standardized bias (SBias) of the simulated data. True values $\lambda = 2$, $\pi_1 = 0.2$, $\pi_2 = 0.4$, and $k = 2$.

<i>n</i>	Parameters	ZkIP	ZIP	Poisson
2000	$\hat{\lambda}$	−0.0002	−0.1076	−0.2007
	$\hat{\pi}_1$	0.0020	−0.4785	−
	$\hat{\pi}_2$	< 0.0001	−	−
1000	$\hat{\lambda}$	−0.0015	−0.1081	−0.2006
	$\hat{\pi}_1$	−0.0016	−0.4825	−
	$\hat{\pi}_2$	< 0.0001	−	−
500	$\hat{\lambda}$	0.0014	−0.1070	−0.2002
	$\hat{\pi}_1$	0.0009	−0.4798	−
	$\hat{\pi}_2$	0.0021	−	−
200	$\hat{\lambda}$	0.0025	−0.1063	−0.1987
	$\hat{\pi}_1$	−0.0085	−0.4865	−
	$\hat{\pi}_2$	0.0025	−	−

Table 5. Comparison of standardized MSE (SMSE) of the simulated data. True values $\lambda = 2$, $\pi_1 = 0.2$, $\pi_2 = 0.4$, and $k = 2$.

n	Parameters	ZkIP	ZIP	Poisson
2000	$\hat{\lambda}$	0.0009	0.0118	0.0405
	$\hat{\pi}_1$	0.0031	0.2329	–
	$\hat{\pi}_2$	0.0013	–	–
1000	$\hat{\lambda}$	0.0019	0.0121	0.0406
	$\hat{\pi}_1$	0.0063	0.2408	–
	$\hat{\pi}_2$	0.0025	–	–
500	$\hat{\lambda}$	0.0037	0.0123	0.0408
	$\hat{\pi}_1$	0.0122	0.2453	–
	$\hat{\pi}_2$	0.0053	–	–
200	$\hat{\lambda}$	0.0095	0.0135	0.0413
	$\hat{\pi}_1$	0.0328	0.2769	–
	$\hat{\pi}_2$	0.0125	–	–

Table 6. Comparison of coverage probabilities of the simulated data. True values $\lambda = 2$, $\pi_1 = 0.2$, $\pi_2 = 0.4$, and $k = 2$.

n	Parameters	90%	95%	99%
2000	$\hat{\lambda}$	0.8920	0.9440	0.9890
	$\hat{\pi}_1$	0.8970	0.9590	0.9930
	$\hat{\pi}_2$	0.8910	0.9530	0.9930
1000	$\hat{\lambda}$	0.8950	0.9430	0.9860
	$\hat{\pi}_1$	0.9070	0.9500	0.9890
	$\hat{\pi}_2$	0.9070	0.9550	0.9880
500	$\hat{\lambda}$	0.9030	0.9600	0.9900
	$\hat{\pi}_1$	0.9080	0.9630	0.9920
	$\hat{\pi}_2$	0.8930	0.9540	0.9880
200	$\hat{\lambda}$	0.9090	0.9520	0.9850
	$\hat{\pi}_1$	0.9110	0.9590	0.9940
	$\hat{\pi}_2$	0.8970	0.9550	0.9950

5.2. Simulation II

In our second simulation study, we generated data from ZkIP ($\lambda = 5$, $\pi_1 = 0.4$, and $\pi_2 = 0.1$), and the inflation points were zero and $k = 3$. For each sample size $n = 200, 500, 1000$, and 2000 we generated $N = 10,000$ datasets. The average estimated value of λ for the N replications using our method for the ZkIP model is $4.9950 \leq \hat{\lambda} \leq 5.0024$ for all the sample sizes. Similarly, the ranges of the average estimated values of π_1 and π_2 for N replications are $0.3995 \leq \hat{\pi}_1 \leq 0.4003$, and $0.0994 \leq \hat{\pi}_2 \leq 0.1004$, respectively. These results clearly demonstrate that our method of estimation is very precise and accurate. Table 7 contains the standard bias (SBias) calculated from the simulated data. The SBias is at a minimum and close to zero for all the parameters of the ZkIP model and for all the sample sizes. The SMSE values are also less for the ZkIP model compared to the ZIP and Poisson models for all the parameters and for all the sample sizes, as shown in Table 8. Thus, the proposed EM algorithm efficiently estimates the true parameters in this second simulation study as well. This conclusion is also supported by the coverage probabilities, which are close to nominal levels, especially for large sample sizes (see Table 9).

Table 7. Comparison of standardized bias (SBias) of the simulated data. True values $\lambda = 5$, $\pi_1 = 0.4$, $\pi_2 = 0.1$, and $k = 3$.

n	Parameters	ZkIP	ZIP	Poisson
2000	$\hat{\lambda}$	<0.0001	−0.0707	−0.4397
	$\hat{\pi}_1$	−0.0012	−0.0072	−
	$\hat{\pi}_2$	0.0044	−	−
1000	$\hat{\lambda}$	0.0005	−0.0698	−0.4399
	$\hat{\pi}_1$	0.0008	−0.0052	−
	$\hat{\pi}_2$	−0.0062	−	−
500	$\hat{\lambda}$	−0.0010	−0.0712	−0.4399
	$\hat{\pi}_1$	−0.0013	−0.0073	−
	$\hat{\pi}_2$	−0.0042	−	−
200	$\hat{\lambda}$	<0.0001	−0.0700	−0.4397
	$\hat{\pi}_1$	<0.0001	−0.0061	−
	$\hat{\pi}_2$	−0.0096	−	−

Table 8. Comparison of standardized MSE (SMSE) of the simulated data. True values $\lambda = 5$, $\pi_1 = 0.4$, $\pi_2 = 0.1$, and $k = 3$.

n	Parameters	ZkIP	ZIP	Poisson
2000	$\hat{\lambda}$	0.0002	0.0052	0.1935
	$\hat{\pi}_1$	0.0007	0.0008	−
	$\hat{\pi}_2$	0.0094	−	−
1000	$\hat{\lambda}$	0.0005	0.0052	0.1938
	$\hat{\pi}_1$	0.0016	0.0016	−
	$\hat{\pi}_2$	0.0188	−	−
500	$\hat{\lambda}$	0.0009	0.0057	0.1942
	$\hat{\pi}_1$	0.0031	0.0031	−
	$\hat{\pi}_2$	0.0352	−	−
200	$\hat{\lambda}$	0.0024	0.0066	0.1950
	$\hat{\pi}_1$	0.0073	0.0074	−
	$\hat{\pi}_2$	0.0937	−	−

Table 9. Comparison of coverage probabilities of the simulated data. True values $\lambda = 5$, $\pi_1 = 0.4$, $\pi_2 = 0.1$, and $k = 3$.

n	Parameters	90%	95%	99%
2000	$\hat{\lambda}$	0.9010	0.9560	0.9930
	$\hat{\pi}_1$	0.9110	0.9500	0.9920
	$\hat{\pi}_2$	0.9020	0.9510	0.9920
1000	$\hat{\lambda}$	0.9100	0.9500	0.9880
	$\hat{\pi}_1$	0.8990	0.9510	0.9870
	$\hat{\pi}_2$	0.9150	0.9610	0.9870
500	$\hat{\lambda}$	0.9080	0.9540	0.9900
	$\hat{\pi}_1$	0.8950	0.9540	0.9920
	$\hat{\pi}_2$	0.9260	0.9680	0.9900
200	$\hat{\lambda}$	0.9061	0.9566	0.9899
	$\hat{\pi}_1$	0.9162	0.9626	0.9949
	$\hat{\pi}_2$	0.9263	0.9636	0.9869

6. Applications

In this section, we illustrate the application of the zero- and k -inflated Poisson (ZkIP) model to analyze two real-life dataset examples. The first example (sunburn data) has

counts zero and one which are inflated, and the second example (off days data) has inflated frequencies for zero and count $k = 2$. Both datasets were extracted from the National Health Interview Survey (NHIS) conducted by the National Center for Health Sciences (NCHS) in 2010. The NHIS has questionnaires and sampling designs for collecting data from US residents. NHIS collects data annually on topics related to health such as immunizations, depression, hepatitis, cancer, use of tobacco, and other variables related to the health and demographics of the subjects.

6.1. Sunburn Data

In this example, we study the prevalence of sunburn in adults in the US. It has been established that sunburn is one of the leading causes of developing skin cancer. Here, the response variable is the number of times the sunburn has occurred in the last 12 months. The sample data were collected on 3917 subjects. The mean and variance of the sample are 0.69 and 1.60, respectively. There are 64.05% of zeros and 19.35% of ones. The zeros are more than 50%, which strongly indicates the existence of inflation at zero, while one is probably also inflated. We first fit the Poisson model to the data and its inflated extensions. The results are shown in Table 10. All the unknown parameters (π_1 , π_2 , λ) are significant in all of the models. The estimated proportion of inflation at zero for ZIP is 0.54 and for ZOIP it is 0.61. From the ZOIP, the inflation at $k = 1$ is 0.13. The LRT statistic between ZIP and Poisson is $-2 \log L = 977.16$, and the p -value is < 0.0001 . Thus, we reject the null hypothesis and conclude that the inflation at zero is significant or the ZIP model fits significantly better than the simple Poisson distribution. Similarly, comparing the ZOIP model to the ZIP model, the LRT statistic is $-2 \log L = 155.78$ and the p -value is < 0.0001 . The AIC is 8982.41 and the BIC is 9026.05 for the ZOIP model. The AIC difference between ZOIP and ZIP is $\Delta_{ZIPAIC} = 153.78$, while that between ZOIP and Poisson is $\Delta_{PoissonAIC} = 1128.94$. Similarly, the BIC difference between ZOIP and ZIP is $\Delta_{ZIPBIC} = 122.68$ and $\Delta_{PoissonBIC} = 1099.84$. According to the AIC and BIC rules of thumb mentioned in Tables 2 and 3, the ZOIP model gives the best fit when compared to the ZIP and Poisson models.

Table 10. Parameter estimates for sunburn data.

Parameters	ZOIP	ZIP	Poisson
$\hat{\lambda}$	2.1415 * (0.0739)	1.4868 * (0.0357)	0.6906 * (0.0133)
$\hat{\pi}_1$	0.6096 * (0.0093)	0.5355 * (0.0104)	–
$\hat{\pi}_2$	0.1273 * (0.0167)	–	–
$-2 \log L$	8976.41	9132.19	10,109.35
AIC	8982.41	9136.19	10,111.35
BIC	9026.05	9148.73	10,125.89

* These estimates are significant.

A comparison between observed and expected frequencies from the ZOIP, ZIP, and Poisson models is shown in Table 11. The expected frequencies from the Poisson model are not close to the observed frequencies and thus the sum of absolute error and χ^2 values are very high. The ZIP model shows an improvement. It perfectly captures the inflation at zero but it does not provide a good fit for counts 1 to 8. The ZOIP model captures the inflations both at zero and at count one. The sum of absolute error is equal to 274.92 and $\chi^2 = 309.22$; both these numbers are smaller when compared to the other two models. For these data, the ZOIP model seems to be the best based on LRT, AIC, and BIC criteria. It also fits the data best based on absolute error and chi-square goodness of fit measures. The estimated inflation at zero is about 61% and at one is about 13%, and clearly both are significant.

Table 11. Observed and expected frequencies of sunburn data.

Count	Observed Frequency	ZOIP	ZIP	Poisson
0	2509	2508.87	2508.94	1963.53
1	758	757.90	611.63	1355.98
2	374	277.61	454.67	468.20
3	127	198.17	225.33	107.78
4	40	106.09	83.75	18.61
5	47	45.44	24.90	2.57
6	27	16.22	6.17	0.30
7	19	4.96	1.31	0.03
8	16	1.33	0.24	0.003
ABE	–	274.92	445.56	1384.36
χ^2	–	309.22	1462.95	117,569.90

6.2. Off Days Data

Back pain is a chronic disease among adults, and occasionally it can be severe, forcing many to take days off from work. In these off days data, the count variable is the number of days off taken due to back pain. The number of people surveyed was 2548. The sample mean and variance are 0.37 and 0.88, respectively. In the data, the zeros are 83% and 10% are equal to 2. Both these proportions are indicators of inflation and suggest that ZkIP with $k = 2$ may be an appropriate model. We first fitted the simpler Poisson model for comparison purposes. Due to the high proportions of zeros, we then implemented the zero-inflated Poisson (ZIP) model. Furthermore, to test the significance of inflation at two, we embarked on the zero- and k -inflated Poisson (ZkIP) model. The estimates and standard errors of the parameters are in Table 12. The rate parameter, λ , is significant in all three models. The ZIP and ZkIP models have a significant π_1 . The ZkIP model also has a significant π_2 , indicating that along with the significant inflation at zero there is significant inflation at count 2. Table 12 also lists the negative log-likelihood, AIC, and BIC values for the models.

The comparison between ZIP and ZkIP models based on the LRT criterion gives a p -value less than 0.0001. Thus, the ZkIP model is significantly better than the ZIP model. The p -value for comparing Poisson and ZIP is also very small (< 0.0001). Thus, ZIP is significantly better than Poisson. Since the models are nested, we can conclude that ZkIP outperforms both the ZIP and the Poisson models. Now, using the AIC and BIC criteria, the best model turns out to be ZkIP. Furthermore, the Δ_{ZIPAIC} difference between the ZIP and ZkIP model is 166.29, while $\Delta_{PoissonAIC} = 1335.28$. This clearly indicates that empirically there is no significant support for the Poisson or ZIP models. Similarly, when comparing the models using the BIC criterion, $\Delta_{PoissonBIC} = 1339.28 \gg 10$ and $\Delta_{ZIPBIC} = 168.29 \gg 10$.

Table 12. The model description of off days data.

Parameters	ZkIP	ZIP	Poisson
$\hat{\lambda}$	2.0674 * (0.1075)	1.8569 * (0.0869)	0.3662 * (0.0120)
$\hat{\pi}_1$	0.8204 * (0.0075)	0.8028 * (0.0089)	–
$\hat{\pi}_2$	0.0755 * (0.0121)	–	–
$-2 \log L$	3321.44	3489.73	4660.72
AIC	3327.44	3493.73	4662.72
BIC	3337.13	3505.42	4676.41

* These estimates are significant.

The goodness of fit measures are shown in Table 13. The expected frequencies from the Poisson model are nowhere close to the observed frequencies, resulting in a high sum of absolute error and χ^2 statistic. The ZIP model is able to capture the inflation at zero and thus has a relatively low error when compared to the Poisson model. The ZkIP model, along with inflation at zero, also captures the inflation at count two, thus it gives a good fit to all the counts. The sum of absolute error and χ^2 statistic is at a minimum for the ZkIP model. Thus, the statistical significance of inflation parameters, π_1 and π_2 , in Table 12, and the minimum value for the sum of absolute errors indicates that the ZkIP model is a good model for this off days dataset.

Table 13. Observed and expected frequencies of the off days dataset.

Count	Observed Frequency	ZkIP	ZIP	Poisson
0	2124	2123.94	2123.99	1766.75
1	84	69.38	145.70	646.93
2	264	264.01	135.27	118.44
3	25	49.42	83.73	14.46
4	23	25.54	38.87	1.32
5	14	10.56	14.43	0.10
>5	14	3.64	4.47	0.06
ABE	–	55.54	274.99	1125.86
χ^2	–	46.02	216.65	36,207.74

7. Discussion

This article proposes a mixture model, ZkIP, for grouped count data with high frequencies for zero and another count of $k > 0$. The ZIP and kIP models are special cases of the ZkIP model. The ZkIP model has just one more parameter than ZIP and kIP, and it captures both the inflation at zero and k . Hence, the ZkIP model is a parsimonious model for studying doubly inflated count data. The model provides more accurate estimates of the probabilities when compared to the model for ungrouped data. The estimated probabilities at zero and k give the estimated count of zeros and ks in excess. The ZkIP model has applications in manufacturing, transportation, econometrics, ecology, and other disciplines. An algorithm is developed using the expectation–maximization (EM) approach to obtain the ML estimates for the ZkIP model. This is a computationally fast approach and extends the estimation method first proposed by Lambert [1] to study the ZIP model. To obtain the standard errors, instead of using the Hessian matrix, we implement the method given by Louis [29] that is based on complete data. We illustrate our algorithm and methodologies on two simulated and two real-life examples from health science. Using various criteria, we show that the ZkIP model is the most appropriate model for the sample data. We are currently extending our methods for zero- and k -inflated Conway–Maxwell–Poisson distributions for grouped and ungrouped data.

Author Contributions: Conceptualization, N.R.C.; data curation, M.A. and N.R.C.; writing—original draft, M.A. and N.R.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are openly available in “Data Files” at <https://www.cdc.gov/nchs/nhis/1997-2018.htm> (accessed on 27 May 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lambert, D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **1992**, *34*, 1–14. [\[CrossRef\]](#)
2. Ghosh, S.K.; Mukhopadhyay, P.; Lu, J.C. Bayesian analysis of zero-inflated regression models. *J. Stat. Plan. Inference* **2006**, *136*, 1360–1375. [\[CrossRef\]](#)
3. Agarwal, D.K.; Gelfand, A.E.; Citron-Pousty, S. Zero-inflated models with application to spatial count data. *Environ. Ecol. Stat.* **2002**, *9*, 341–355. [\[CrossRef\]](#)
4. Min, Y.; Agresti, A. Random effect models for repeated measures of zero-inflated count data. *Stat. Model.* **2005**, *5*, 1–19. [\[CrossRef\]](#)
5. Yau, K.; Lee, A. Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. *Stat. Med.* **2001**, *20*, 2907–2920. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Saffari, S.E.; Adnan, R. Zero-inflated Poisson regression models with right censored count data. *Matematika* **2011**, *27*, 21–29.
7. Yang, Y.; Simpson, D.G. Conditional decomposition diagnostics for regression analysis of zero-inflated and left-censored data. *Stat. Methods Med. Res.* **2012**, *21*, 393–408. [\[CrossRef\]](#)
8. Nguyen, V.T.; Dupuy, J.F. Asymptotic results in censored zero-inflated Poisson regression. *Commun. Stat. Theory Methods* **2021**, *50*, 2759–2779. [\[CrossRef\]](#)
9. Altun, E. A new zero-inflated regression model with application. *J. Stat. Stat. Actuar. Sci.* **2018**, *2*, 73–80.
10. Bakouch, H.; Chesneau, C.; Karakaya, K.; Kuş, C. The Cos-Poisson model with a novel count regression analysis. *Hacet. J. Math. Stat.* **2021**, *50*, 559–578. [\[CrossRef\]](#)
11. Gupta, P.L.; Gupta, R.C.; Tripathi, R.C. Analysis of zero-adjusted count data. *Comput. Stat. Data Anal.* **1996**, *23*, 207–218. [\[CrossRef\]](#)
12. Umbach, D. On inference for a mixture of a Poisson and a degenerate distribution. *Commun. Stat. Theory Methods* **1981**, *10*, 299–306. [\[CrossRef\]](#)
13. Lord, D.; Washington, S.P.; Ivan, J.N. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. *Accid. Anal. Prev.* **2005**, *37*, 35–46. [\[CrossRef\]](#)
14. Qin, X.; Ivan, J.N.; Ravishanker, N. Selecting exposure measures in crash rate prediction for two-lane highway segments. *Accid. Anal. Prev.* **2004**, *36*, 183–191. [\[CrossRef\]](#)
15. Ridout, M.; Demetrio, C.; Hinde, J. Models for count data with many zeros. In Proceedings of the International Biometric Conference, Cape Town, South Africa, 14–18 December 1998.
16. Welsh, A.; Cunningham, R.; Donnelly, C.; Lindenmayer, D. Modelling the abundance of rare species: Statistical models for counts with extra zeros. *Ecol. Model.* **1996**, *88*, 297–308. [10.1016/0304-3800\(95\)00113-1](#). [\[CrossRef\]](#)
17. Atkins, D.; Gallop, R. Rethinking how family researchers model infrequent outcomes: A tutorial on count regression and zero-inflated models. *J. Fam. Psychol.* **2007**, *21*, 726–735. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Loeys, T.; Moerkerke, B.; De Smet, O.; Buysse, A. The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression. *Br. J. Math. Stat. Psychol.* **2012**, *65*, 163–180. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Salehi, M.; Roudbari, M. Zero-inflated Poisson and negative binomial regression models: application in education. *Med. J. Islam. Repub. Iran* **2015**, *29*, 297.
20. Cameron, A.C.; Trivedi, P.K. *Regression Analysis of Count Data*; Cambridge Press: London, UK, 2013.
21. Greene, W. *Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models*; Working Papers; New York University: New York, NY, USA, 1994.
22. Gurmu, S.; Trivedi, P. Excess zeros in count models for recreational trips. *J. Bus. Econ. Stat.* **1996**, *14*, 469–477.
23. Motalebi, N.; Owlia, M.S.; Amiri, A.; Fallahnezhad, M.S. Monitoring social networks based on zero-inflated Poisson regression model. *Commun. Stat. Theory Methods* **2023**, *52*, 2099–2115. [\[CrossRef\]](#)
24. Arora, M.; Chaganty, N.R. EM estimation for zero- and k-inflated Poisson regression model. *Computation* **2021**, *9*, 94. [\[CrossRef\]](#)
25. Lin, T.H.; Tsai, M.H. Modeling health survey data with excessive zero and k responses. *Stat. Med.* **2012**, *32*, 1572–1583. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Sheth-Chandra, M.; Chaganty, N.R.; Sabo, R.T. *A Doubly Inflated Poisson Distribution and Regression Model*; Springer International Publishing: Berlin, Germany, 2019; pp. 131–145.
27. Arora, M.; Kalyani, Y.; Shanker, S. A comparative study on inflated and dispersed count data. In Proceedings of the 10th International Conference on Data Science, Technology and Applications (DATA 2021), Online, 6–8 July 2021; Volume 1, pp. 29–38.
28. Martínez-Flórez, G.; Bolfarine, H.; Gómez, H.W. Doubly censored power-normal regression models with inflation. *TEST* **2015**, *24*, 265–286. [\[CrossRef\]](#)
29. Louis, T.A. Finding the observed information matrix when using the EM algorithm. *J. R. Stat. Soc. Ser. (Methodol.)* **1982**, *44*, 226–233.
30. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. (Methodol.)* **1977**, *39*, 1–22.
31. Schervish, M.J. *Theory of Statistics*; Springer: New York, NY, USA, 1995.
32. Rao, C.R. *Linear Statistical Inference and Its Applications*; John Wiley and Sons Inc.: New York, NY, USA, 1965.
33. Wald, A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Am. Math. Soc.* **1943**, *54*, 426–482. [\[CrossRef\]](#)

34. Chant, D. On asymptotic tests of composite hypotheses in nonstandard conditions. *Biometrika* **1974**, *61*, 291–298. [[CrossRef](#)]
35. Shapiro, A. Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika* **1985**, *72*, 133–144. [[CrossRef](#)]
36. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [[CrossRef](#)]
37. Kass, R.E.; Raftery, A.E. Bayes Factors. *J. Am. Stat. Assoc.* **1995**, *90*, 773–795. [[CrossRef](#)]
38. Mallick, A.; Joshi, R. Parameter Estimation and Application of Generalized Inflated Geometric Distribution. *J. Stat. Theory Appl.* **2018**, *17*, 491. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.