A GENERAL FRAMEWORK FOR MODELLING ZERO INFLATION

JOHN HASLETT[1], ANDREW PARNELL[2], JAMES SWEENEY[3]

[1] *School of Computer Science, Trinity College Dublin*

[2] *Hamilton Institute, Maynooth University*

[3] *School of Business, University College Dublin*

Abstract

We propose a new framework for the modelling of count data exhibiting zero inflation (ZI). The main part of this framework includes a new and more general parameterisation for ZI models which naturally includes both over- and under-inflation. It further sheds new theoretical light on modelling and inference and permits a simpler alternative, which we term as multiplicative, in contrast to the dominant mixture and hurdle models. Our approach gives the statistician access to new types of ZI of which mixture and hurdle are special cases. We outline a simple parameterised modelling approach which can help to infer both ZI type and degree and provide an underlying treatment that shows that current ZI models are themselves typically within the exponential family, thus permitting much simpler theory, computation and classical inference. We outline some possibilities for a natural Bayesian framework for inference; and a rich basis for work on correlated ZI counts.

The present paper is an incomplete report on the underlying theory. A later version will include computational issues and provide further examples.

1. Introduction

We consider regression modelling of observed count data $\mathbf{y} = \{y_1, \ldots, y_n\}$ on covariates $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ where many values of y_i are 0; termed zero-inflation (ZI). We treat the count data as realisations of a random variable $\tilde{Y}$, having pmf $\tilde{\pi}_y = \tilde{\pi}_y(\theta, \kappa), y \in \mathbb{Z}_0^+$, with the parameters θ and κ controlling the location and ZI processes respectively, with e.g. $\theta_i = \theta(\mathbf{x}_i \beta)$. The pmf $\tilde{\pi}_y$ is related to a simpler pmf $\pi_y = \pi_y(\theta)$ which characterises the standard count process, absent ZI; typically $\tilde{\pi}_0 > \pi_0$. Observations on $\tilde{Y}$ thus imperfectly reflect $\pi_y(\theta)$. We propose a new approach for the two main issues associated with such models: (a) performing inference on θ and κ given a specific $\tilde{\pi}_y$ and (b) constructing families of $\tilde{\pi}_y$ from families of π_y. Our main contribution is the modelling of the function $\tilde{\pi}_0 = \tilde{\pi}_0\big(\pi_0(\theta), \kappa\big)$ which allows us to diagnose a wide class of zero inflation types, some of which have not been identified in the literature to-date. Figure presents examples of such types. However we also identify a universal system of equations for (a).

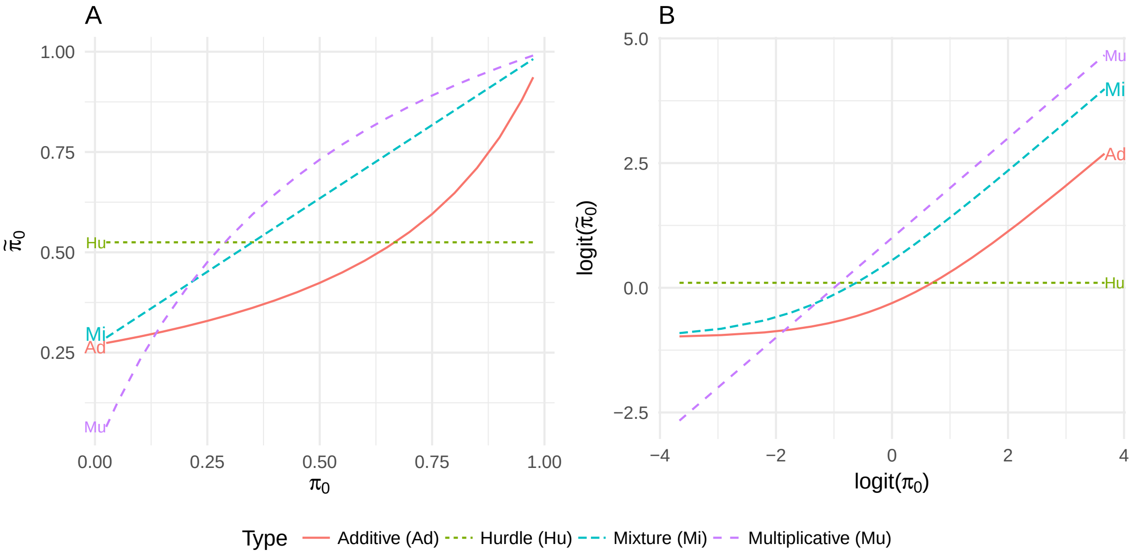

FIGURE 1. Multiplicative, mixture, hurdle and additive models. A: probability metric. B: logit metric

The seminal types of ZI are those of the so-called hurdle model (Mullahy 1986, 1997), and the mixture model of Lambert (1992). In the hurdle model, all instances of $\tilde{Y} = 0$ are generated by a Bernoulli process with $\tilde{\pi}_0$, such that $\tilde{\pi}_0$ is not a function of π_0; hence observed zeroes contain no information on θ. Conversely, all instances of $\tilde{Y} > 0$ are generated by a distribution π_y, defined solely on $y > 0$; typically a truncated distribution. In the mixture model, there are two latent variables: Y as defined above together with binary J such that $P(J = 1) = q$, where $q = q(\kappa)$. The observable is $\tilde{Y} = YJ$. Now $\tilde{\pi}_0 = (1 - q) + q\pi_0$ hence $\tilde{\pi}_0$ is a linear function of π_0 and a function of θ. In the mixture model only some instances of $\tilde{Y} = 0$ are relevant to inference on θ; the challenge is that we do not know which.

Since their publication, the literature has generated more than 1000 papers[1] with very many applications of these two models; excess zeroes arise in many contexts. There have been technical extensions, such as algorithms for mixed models for which the seminal paper is Hall (2000); the use of distributions other than the simple Poisson and binomial used in the early papers, and in particular of distributions that facilitate the separation of over-dispersion and zero-inflation, including Ridout et al. (2001), Xiang et al. (2007), and Kassahun et al. (2014). Several authors have pursued Hall's lead on random effects, including (as well as some of the above) Long et al. (2015), Min & Agresti (2005), Martin & Hall (2017), and Molas & Lesaffre (2010). These of course have much in common with other examples of correlated, zero- inflated data, such as arise in studies with a focus on longitudinal, time-series and spatial data, including Chebon et al. (2017), Yang et al. (2016), Agarwal et al. (2002), Ancelet et al. (2010). Multivariate data exhibiting ZI have been examined in papers such as Li et al. (1999) and Liu & Tian (2015). Several authors have pursued Bayesian

[1] Google Scholar search for "zero inflation" retrieved on 25/4/18

inference in ZI, including Angers & Biswas (2003), Rodrigues (2003), Dagne (2004), Ghosh et al. (2006), Klein et al. (2015), Lee et al. (2017), Neelon & Chung (2017) and some of those cited earlier. Thus is a very active area of research.

There are several useful reviews. Ridout et al. (1998) point out, inter-alia, that: (i) the hurdle and mixture models can in fact be seen as re-parameterisations of each other; and (ii) that the parameterisation $\tilde{\pi}_0 = (1-q) + q\pi_0$, although suggested by the mixture model, in fact allows $\kappa > 1$ (subject to $\tilde{\pi}_0 \in [0, 1]$); this can characterise zero deflation. Ghosh et al. (2012) draw attention to early precursors, such as Cohen (1960, although there does seems to be a citation error). Both Ridout et al. (1998) and Ghosh et al. (2012) emphasise that these ideas suggest different parameterisations; for in the simplest *iid* case all ZI models are equivalent by re-parameterisation. But, remarkably, there seem to be few if any attempts to set models of zero inflation in a wider modelling framework. In this paper we distinguish between (i) regression modelling of counts on covariates in the presence of ZI; and (ii) modelling of zero-inflation, per se. The latter focusses on both the type and degree of ZI, possibly also involving covariates. This modelling of ZI seems, surprisingly, to be completely unexplored.

The hurdle and mixture models are well defined in an abstract sense, or for Monte Carlo simulations. But are they really rich enough to complement and stimulate the more process based thinking of many subject matter specialists? Might they feel, at least sometimes, that they are being shoehorned by the statistical community into an unnatural framework? Chebon et al. (2017) raise the query in their title: "Models for zero-inflated, correlated count data with extra heterogeneity: when is it too complex?" Todem et al. (2016) remark "much of the literature on real applications of these models has centered around the latent class formulation where the mean response of the so-called at-risk or susceptible population and the susceptibility probability are both related to covariates. While this formulation in some instances provides an interesting representation of the data, it often fails to produce easily interpretable covariate effects on the overall mean response."

Several applied authors register similar anxieties. Miller & Miller (2008) report "The results suggest that the best-fitting zero-inflated model sometimes depends on the proportion of zeros and the distribution for the non-zeros. In fact, there are situations where the zero- inflated models are not necessary. Garay et al. (2011) says: "In order to study departures from the error assumption as well as the presence of outliers, we perform residual analysis .....illustrated with a real data set, where it is shown that, by removing the most influential observations, the decision about which model is best as the data changes. Similar is Fisher et al. (2017): "Although these models are often appropriate on statistical grounds, their interpretation may prove substantively difficult." This all suggests an un-met need for a wider range of ZI models easy to interpret and to criticise.

The first thought in such modelling is surely the question of whether, and if so *why*, there is a need to model zeroes differently. Is this need always well served by the two main models? We will argue that they are not even the simplest. At a more technical level, we observe that in the Mixture model, the null model corresponds to $q = 1$, an extremum in the parameter space. Are there natural model variations which are well defined in a region *around* the null?

The hurdle model is often described as permitting both zero inflation and zero deflation. But it is not strictly necessary that the underlying distribution π_y, truncated to have support only on $y > 0$, have any interpretation at $y = 0$. The common use of truncated distributions is most often a convenient way to marry extant statistical machinery to the necessity of a model defined only on $y > 0$. Its use does not always derive from a need to question, in a natural way, the evidence that zeroes are different. The null model is not - in any clear sense - nested within the hurdle model unless a truncated π_y is used; then, for $y > 0, \tilde{\pi}_y = \frac{1-\tilde{\pi}_0}{1-\pi_0}\pi_y$; but $\tilde{\pi}_0$ does not depend on θ. In this paper, we take the hurdle model to be defined with respect to a truncated distribution.

Lambert's 1992 formulation led to an EM algorithm, which dominates the implementation of the mixture model: the variable J is taken as providing the complete data likelihood. The algorithm leads to (a) down-weighting zeroes to estimate regression coefficients; and (b) a clever, but highly technical, use of binary regression of latent J on covariates, in order to estimate κ. We will see (a) does not need to appeal to EM for motivation, as it flows directly from maximum likelihood; on (b) we propose alternatives which do allow natural criticism of this choice of ZI model.

Lambert enters a caveat; she remarks that the the EM algorithm cannot be used if θ and κ are related, including via a dependence on common covariates. She is clear that the caveat is directed at the EM algorithm, rather than at the use of maximum likelihood *per se* to estimate the parameters of the Mixture model. She also remarks that in the event that (modulo notation) "If the same covariates affect κ and θ, it is natural to reduce the number of parameters by thinking of q as a function of κ". This saves computation, as she says. But it is of course perfectly sensible advice for seeking to simplify the statistical model, the rewards for which go well beyond computation. One parsimonious example is Salter-Townshend & Haslett (2012), where q is modelled as $\left(\frac{\mu_y(\theta)}{1+\mu_y(\theta)}\right)^{\gamma}$, where γ is a scalar characterising the degree of ZI, and $\mu_y(\theta)$ is the expected value of the pmf π_y. One conclusion of the current paper is that the mixture model is simple only in a rather specific sense.

It has been noted that ZI can be difficult to distinguish from over-dispersion (Perumean-Chaney et al. (2013)). This has led several researchers to build ZI models $\tilde{\pi}_y$ on richer examples of π_y. Researchers have considered the Negative Binomial, which can be considered as mixture of Poissons; e.g. Ridout et al. (2001), Moghimbeigi (2011), and Garay et al. (2011). Others have used the Conway-Maxwell-Poisson distribution (Sellers & Raim 2016) and Generalised Poisson distributions (Xie et al. 2009). But these are all specific to one example of π_y, the Poisson. Several ZI researchers have found it fruitful to use distributions from the power series family (e.g. Bhattacharya et al. 2008, Patil & Shirke 2011). Bizarrely there seems to be little interest in building $\tilde{\pi}_y$ from more classic families of π_y, such as the classic exponential-dispersion family of Jorgensen (1987), or the over-dispersed exponential family (Gelfand & Dalal 1990, Dey et al. 1997). We shall see below that our simplest ZI model fits very naturally with the exponential family.

In this paper we provide an apparently new and much wider framework for ZI modelling, which we distinguish from regression in the presence of ZI. Here, because its widespread acceptance, we use the ZI label to include various versions of zero-modification (e.g. Min & Czado 2010), including both over- and under-inflation (deflation). We focus on the univariate case, and work primarily within the exponential family, extending to multivariate in later sections. We concentrate on inference through the likelihood, primarily because it provides interesting insights on the options for ZI modelling, but the general ideas extend naturally to other frameworks. The key contribution lies in the next two sections. We hope that the framework will open new avenues for others to pursue.

2. A general model for ZI

We adopt the notation

$$\tilde{\pi}_y(\kappa, \theta) = \begin{cases} (1+\kappa)\rho\pi_0(\theta) & \text{if } y = 0 \\ \rho\pi_y(\theta) & \text{otherwise.} \end{cases}$$

where the function $\rho = \rho(\pi_0, \kappa)$ renormalises and κ controls ZI as before. Typically $\tilde{\pi}_0 > \pi_0$ corresponding to $\kappa > 0$ and $0 < \rho < 1$. However also permissible is $-1 < \kappa < 0$ which implies $\tilde{\pi}_0 < \pi_0$ and $\rho > 1$. The simplest expression for ρ is $\rho = \frac{1-\tilde{\pi}_0}{1-\pi_0}$, from which it follows that

$$\frac{\tilde{\pi}_0}{1-\tilde{\pi}_0} = (1+\kappa)\frac{\pi_0}{1-\pi_0} \qquad \text{or equivalently,} \qquad \text{logit}(\tilde{\pi}_0) = \omega + \text{logit}(\pi_0)$$

with $\omega = \log(1+\kappa)$. We refer to these as the odds ratio and the log-odds forms of ZI respectively. It is a simple matter to show that $\rho^{-1} = (1+\kappa\pi_0) = (1-\pi_0)+e^{\omega}\pi_0$. Note that central to the notation is an explicit functional relationship between $\tilde{\pi}_0$ and π_0, characterised by ω. Below we shall treat ω as a function of π_0, that is $\omega = \omega(\gamma, \pi_0)$ where γ is a scalar parameter controlling the degree of ZI. In regression (where θ, and thus $\pi_0(\theta)$, and also γ may vary with covariates) the choice of function is at the heart of the modelling of the type of ZI. A wide class of ZI models can be obtained this way.

The relationship between $\tilde{\pi}_y$ and π_y may be more compactly written as

$$\tilde{\pi}_y = (1+\kappa)^{\mathbb{I}\{y=0\}}\rho\pi_y; \text{ or equivalently } \log(\tilde{\pi}_y) = \omega\mathbb{I}\{y=0\} + \log(\rho) + \log(\pi_y)$$

Note the duality between ω and ρ. Each defines the other via normalisation; but a simple parametrisation for one may require a difficult parameterisation for the other. The simplest model in this notation is that in which ω is constant wrt π_0; but ρ is thus dependent on π_0. In contrast, as shown later, the mixture model has ρ independent of π_0; this leads to ω being dependent on π_0.

A natural and generic simulation mechanism for $\tilde{Y}$ is as below, with U denoting a realisation from $U(0,1)$:

(1) If $U \leq \tilde{\pi}_0$ generate $\tilde{Y} = 0$
(2) Otherwise generate $\tilde{Y}$ from π_y, rejecting all instances of 0

Thus in the latter case, with probability $1-\tilde{\pi}_0$ we sample from the truncated pmf $\frac{1}{1-\pi_0}\pi_y$. Observe that there are no latent variables involved in the data generation. This constructive formulation makes it apparent that the central parameters, for simulation and inference, are $(\tilde{\pi}_0, \theta)$.

The formulation we adopt includes, as a special case, the hurdle model, where $\omega = \gamma - \text{logit}(\pi_0)$, and thus $\text{logit}(\tilde{\pi}_0) = \gamma$ is independent of π_0. A further special case is the classic mixture model; here $\tilde{\pi}_0 = (1-q) + q\pi_0$, and $\tilde{\pi}_y = q\pi_y, \forall y > 0$ with q the mixture weight. We may identify q with ρ, noting in particular that, in this ZI model, the normalising function $\rho(\cdot)$ is typically modelled as independent of $\pi_0(\theta)$ *per* Lambert's caution on EM. From this identification, with $e^{-\gamma} = \frac{1-q}{q}$ we find $\kappa = \frac{e^{-\gamma}}{\pi_0}$; hence $\omega = \log(\pi_0 + e^{-\gamma}) - \log(\pi_0)$ is a function of $\pi_0(\theta)$; and in this sense it is not as simple as constant ω. However, $\tilde{\pi}_0$ is a linear function of π_0. These, and other, types of ZI are discussed below.

Further, the latent variable definition of the mixture model admits the very simple interpretation of positive κ; for $P(J = 1|\tilde{Y} = 0) = \frac{q\pi_0}{\tilde{\pi}_0} = \frac{1}{1+\kappa} = e^{-\omega}$. This probability is central to Lambert's EM algorithm, as we elaborate below. But we will see below that EM is natural only in the very specific context of the mixture model of ZI. Of some interest below will be the interpretation, as an expected value, rather than a probability, of $\frac{1}{1+\kappa} = e^{-\omega}$. It is thus simply interpretable for negative ω also, corresponding to $-1 < \kappa < 0$.

Under-inflation of zeroes can be thought of via probabilistic censoring. Consider the following data generating model: (i) generate Y from $\pi_y(\theta)$; (ii) return $\tilde{Y} = y$ if $y > 0$; but, (iii) when $y = 0$, return a value of *missing*, with probability r. The pmf of the *non-missing* $\tilde{Y}$ is $\tilde{\pi}_y$, and $P(\tilde{Y} = 0|\textit{not missing}) = \tilde{\pi}_0 = \frac{\pi_0(1-r)}{1-\pi_0 r}$. Thus, with $\kappa \in (-1, 0)$, we may identify κ with $-r$; note that the limiting case of $\kappa \to -1$ (or equivalently, $\omega \to -\infty$) corresponds to truncation at $y = 0$.

Equivalently, we may associate with each observed $\tilde{Y} = 0$ an unobserved random (integer) number M of instances of $Y = 0$, of which the observed zero is the sole survivor. Then (with $\kappa < 0$) $P(M = m) = (-\kappa)(1+\kappa)^{m-1}; m \in \mathbb{Z}_0^+$ with $E[M] = \frac{1}{1+\kappa}$; but M is only defined when $\tilde{Y} = 0$. We note that the same interpretation can apply to over-inflation, but now M is binary; $M = 0$ here corresponds to $J = 0$, an 'inflated' zero. Now $E[M] = \frac{1}{1+\kappa}$, a value shared with $E[J|\tilde{Y} = 0]$ in the mixture model. Note that in both cases $Var[M] = \frac{\kappa}{1+\kappa}$. Thus in one sense the variable M is a generalisation of the variable J which defines the mixture model.

Despite M being undefined if $\tilde{Y} \neq 0$, it is natural to extend its definition, for we can write $M = 1$ when $\tilde{Y} \neq 0$. Then $E[M|\tilde{Y}] = \left(\frac{1}{1+\kappa}\right)^{\mathbb{I}\{\tilde{Y}=0\}}$ covering all cases. The unconditional distribution of M is $\pi_M(m) = \sum_y P(M = m|\tilde{Y} = y)\tilde{\pi}_y; m \in \mathbb{Z}_0^+$. We consider separately the cases of positive and negative κ. With $\kappa > 0$, $\pi_M(m) = 0$ for $m > 1$. Then $\pi_M(0) = \frac{\kappa}{1+\kappa}\tilde{\pi}_0 = \frac{\kappa}{1+\kappa}(1+\kappa)\rho\pi_0 = 1 - \rho$, and $\pi_M(1) = \rho$; when $\rho = q$, as in the mixture model, this coincides with the distribution of J. For negative κ, we note that $m > 1$, $P(M = m|\tilde{Y} = y) = 0$ unless $y = 0$; thus $\pi_M(m) = \tilde{\pi}_0(-\kappa)(1+\kappa)^{m-1} = (\rho - 1)(1+\kappa)^m$, when $m > 1$. It follows that $\pi_M(1) = -\kappa^{-1}(\rho - 1)(1+\kappa)^2$.

The latent variable M is best understood as the number of instances of (unobserved) $Y = 0$ associated with every instance of observed $\tilde{Y} = 0$. It will be noted therefore that, given n_0 *iid* instances of $\tilde{Y} = 0$, then, on average, $\frac{n_0}{1+\kappa}$ are relevant to inference on θ. In a sample of n *iid* observations, the expected number of zeroes is $n\tilde{\pi}_0$, so the expected total number in the sample, relevant to such inference, is $n\left(1 - \frac{\kappa\tilde{\pi}_0}{1+\kappa}\right) = n\rho$. This may be thought of as an effective sample size, being greater or less than n for under-and over-inflation, respectively; this echoes Cohen (1960). We formalise this below in the context of $\pi_y(\theta)$ in the exponential family. But note that M is not part of the definition of our ZI model, in contrast to the role of J in the classic mixture model.

2.1. **Types of ZI.** From the parameterisation above, we see that $\tilde{\pi}_0$, as a function of π_0, is most simply expressed via the log-odds form $\text{logit}(\tilde{\pi}_0) = \omega + \text{logit}(\pi_0)$. The implicit function $\omega\Big(\gamma(\mathbf{x}\alpha), \pi_0(\theta(\mathbf{x}\beta))\Big)$ characterises the type of ZI, with any parameters α defining the degree. The functions $\theta(\cdot)$ and $\gamma(\cdot)$ are functions relating the covariates to $\tilde{\mu} = E[\tilde{Y}]$; $\gamma(\cdot)$ (and hence α) characterises degree of the ZI, within the type specified by the function $\omega(\gamma, \pi_0)$. In our simplest models we will typically consider these to be identity functions. Indeed the function $\omega(\cdot)$ only becomes visible in the presence of covariates x. For then, within a data set, observations y_i are associated with different $\theta_i = \theta(\mathbf{x}_i\beta)$ and $\gamma_i = \gamma(\mathbf{x}_i\beta)$, and thus with pairs $(\tilde{\pi}_{i0}, \pi_{i0})$, these being defined as $\pi_{i0} = \pi_0(\theta_i)$ and $\tilde{\pi}_{i0} = \pi_0(\theta_i, \omega(\gamma_i, \pi_0))$. Types of ZI are characterised by functions such as $\tilde{\pi}_0(\gamma, \pi_0)$ that are defined by a given function ω; the scalar parameter γ controls the degree of ZI.

The simplest model has $\text{logit}(\tilde{\pi}_0) - \text{logit}(\pi_0) = \log(1+\kappa) = \omega = \gamma$; here the function ω is constant with respect to π_0. It appears to be new. We refer to it as multiplicative ZI. It is strictly multiplicative only in the odds ratio sense; that is, ratios $\frac{\pi_0}{1-\pi_0}$ are increased (or decreased) to $\frac{\tilde{\pi}_0}{1-\tilde{\pi}_0}$ by a multiplicative factor. But we note that when π_0 is small (corresponding to parts of the covariate space where $\pi_0(\theta(\mathbf{x}\beta))$ is small) so necessarily must be $\tilde{\pi}_0$; we may argue similarly for large π_0 and $\tilde{\pi}_0$. In these circumstances, the nature of ZI is to accentuate the variation in π_{i0} that is induced by varying x_i in covariate space; it is in the sense of 'accentuate' that we use the term 'multiply'. In the context of Figure 1, when π_{i0} is very small, or very large, so also will be $\tilde{\pi}_{i0}$; but when $\pi_{i0} \approx 0.5$, $\tilde{\pi}_{i0}$ will be larger than 0.5, for positive ω, and smaller than 0.5 for negative ω. It is also uniformly inflationary in the sense that, for $\omega > 0$, $\tilde{\pi}_0(\omega, \theta) > \pi_y(\theta)$ for all θ, and conversely for $\omega < 0$.

As seen above, the classic mixture model above is another type of ZI, characterised here by the rather more awkward $\omega = \log(\pi_0 + e^{-\gamma}) - \log(\pi_0)$, where the awkward additive terms in the first expression reflect the typical difficulties of a mixture model. It is these difficulties that lead to some awkward computations, for which EM here can supply solutions. We may describe this model as having additive ZI, in the sense that even in parts of covariate space where π_0 is small, $\tilde{\pi}_0$ can be such that many zeroes are observed. As remarked, with $q > 1$ (subject to $0 \leq \tilde{\pi}_y \leq 1$), this type of ZI can be under-inflationary. This constraint is difficult to parameterise, however, and $q > 1$ is rarely used. This model is in fact also uniformly inflationary in the sense of the previous paragrpah.

Another well studied type is the hurdle model. This need not be cast as either over- or under-inflation, for $\pi_y(\theta)$ is only defined on $y > 0$. However, here we only use truncated versions of distributions such as the Poisson. In these circumstances we can say that π_0 is defined, at least implicitly, and the distinguishing feature of the truncated model is thus that, although $\tilde{\pi}_0$ is well defined, it is constant *wrt* $\pi_0(\theta)$. In our notation, the (truncated) hurdle type of ZI corresponds to $\omega = e^{\gamma} - \text{logit}(\pi_0)$. This ZI type is, unsurprisingly, not uniformly inflationary.

An apparently new type is available through $\omega = \mathbf{x}\alpha - \log(\pi_0)$. Now logit $(\tilde{\pi}_0) = \mathbf{x}\alpha - \log(1-\pi_0)$, an increasing function of π_0. This has the property that, for small π_0, $\tilde{\pi}_0 \approx \frac{e^{\mathbf{x}\alpha}}{1+e^{\mathbf{x}\alpha}}$. In this sense this parameterisation also has the additive property. We refer to it as additive, recognising however that that mixture model also has this property. As seen in Figure 1, it is not uniformly inflationary.

More generally, the multiplicative, additive, and hurdle are all special cases of $\omega = \gamma + \tau_1 \log(\pi_0) + \tau_2 \log(1 - \pi_0)$; they correspond to $(\tau_1, \tau_2) = (0, 0), (0, -1), (-1, -1)$, respectively. In this paper, apart from the classic mixture model, we shall restrict ourselves to ZI types of this form. However, the reader will note that there is no general restriction in ZI type; the function $\omega(\gamma, \pi_0)$ may be modelled very flexibly. We may, for example, extend our notation γ to refer both to coefficients of $\mathbf{x}$ in the link function $\gamma(\alpha\mathbf{x})$ and to coefficients of the chosen functions of π_0, such as the log functions used above. We shall take the τ coefficients as known, unless otherwise stated; but they too can be estimated from data.

Other functions $\tilde{\pi}_0(\gamma, \pi_0)$, equivalently other ZI types, can of course be defined as convenient. But their key properties are characterised by the function $\omega = \text{logit}(\tilde{\pi}_0) - \text{logit}(\pi_0)$. For example, the recursive use of ZI by Ghosh et al. (2012) corresponds to a new distribution $\tilde{\tilde{\pi}}_y$ being related to $\tilde{\pi}_y$ via the odds ratios:

$$OR(\tilde{\tilde{\pi}}_0) = (1 + \tilde{\kappa})OR(\tilde{\pi}_0) = (1 + \tilde{\kappa})(1 + \kappa)OR(\pi_0) = (1 + \tilde{\tilde{\kappa}})OR(\pi_0)$$

or equivalently via $\tilde{\tilde{\omega}} = \tilde{\omega} + \omega$, concatenating the two inflations. As the authors remark, these functions cannot be identified separately; but using different parameterisations for each is one way to motivate a new parameterisation for a single function ω.

Critically the type of ZI depends on the functional relationship between $\tilde{\pi}_0$ and π_0, moderated by the covariates $\mathbf{x}$ and captured by $\omega(\cdot)$. This is where lies the essence of ZI modelling, which we may contrast with regression of y on x in the presence of ZI. This relationship can be arbitrarily rich and need not be linear; but simplicity usually brings more insight.

We shall write

$$\omega = \text{logit}(\tilde{\pi}_0) - \text{logit}(\pi_0) = \gamma(\alpha\mathbf{x}) + \sum_k \tau_k f_k(\pi_0)$$

where the functions $f_k(\pi_0)$ typically include the log functions $\log(\pi_0), \log(1 - \pi_0)$ and the τ_k are coefficients independent of $\mathbf{x}$. When these are known, the ZI model is specified. We shall see that linear logistic regression of $\mathbb{I}\{y = 0\}$ on $\mathbf{x}$ and (if the ZI type is not known) on a small number of functions $f_k(\pi_0)$ suffice to estimate the coefficients α and τ. Compactly we can write this as $\omega = \gamma(\alpha\mathbf{x}) + \tau\mathbf{f}(\pi_0(\theta))$. Note that the additive form of ω allows us to separate degree γ from type τ. We shall show that the classic mixture version of ZI can be seen as involving parametric but non-linear logistic regression of $\mathbb{I}\{y = 0\}$ on $\mathbf{x}$.

We enter one caveat, illustrated by another single parameter ZI model with somewhat pathological behaviour that is only apparent with close inspection: $\omega = \gamma \log(\pi_0)$. When $\gamma \approx 1$, and except for small π_0, plots of this function show much in common with the mixture model. It too exhibits the additive property; and it is uniformly inflationary. But this model implies that $\tilde{\pi}_0 = (\gamma - 1)\log(\pi_0) - \log(1 - \pi_0)$. But when $\pi_0 \to 0$ we find: for $\gamma < 1$ that $\tilde{\pi}_0 \to 0$ - it is not strictly additive; and for $\gamma > 1$ we find $\tilde{\pi}_0 \to 1$; that is, $\tilde{\pi}_0(\gamma, \pi_0)$ is not monotone. Modelling care may be needed with choosing functions $f(\pi_0)$ and data-defined values of τ; there may be algorithmic issues as well, as discussed later.

2.2. **Inference and Modelling.** Typically observations are available as $(y_i, \mathbf{x}_i); i = 1, \ldots n$, which we write as $(\mathbf{y}, X)$. The parameters (θ, ω) themselves become parameterised as $\theta_i = \theta(\mathbf{x}_i\beta)$ and $\omega(\gamma(\alpha\mathbf{x}_i)) + \tau\mathbf{f}(\pi_0(\theta_i))$. Inference focusses on the vector parameters (α, β) and (if the type of ZI is not pre-specified) on τ. We proceed via the likelihood in classical inference, and via the likelihood and priors in Bayesian inference. This involves both computation and critical evaluation, the latter involving both model selection and criticism of the data, ideally in a collaboration between statistician and subject matter specialist.

Computation itself is no longer the main challenge for the size of data sets we consider; but there is natural interest in efficiency, and we turn to this below. The criticism will come most easily from computation that builds on the familiar tools of the statistical trade, which in regression often include graphics and residuals. As we shall see, this typically involves iterating until convergence between: (i) regression of y on $\mathbf{x}$, for given α and τ, yielding estimates of β and hence of $\pi_{0i} = \pi_0(\theta_i)$; and (ii) binary regression of $\mathbb{I}\{y = 0\}$ on both $\mathbf{x}$ and (given β) suitable functions of $\pi_{0i} = \pi_0(\theta_i)$, chosen to embody the modelling of the ZI itself, yielding new estimates of α and, if necessary, τ. In particular plots of (estimates of) pairs $(\tilde{\pi}_{0i}, \pi_{0i})$ seem likely to be informative. Subject to the (important) qualification that the variance estimates that emerge from these two separate regressions are conditional, not joint, they provide a conventional framework for criticism. We discuss below the details of these regressions and of theory for the appropriate joint inference.

Initially we explore the theory for *iid* observations. We subsequently consider the regression case where all parameters can be considered as additive functions of covariates $\mathbf{x}$ via coefficients upon which inference becomes focussed.

3. Maximum likelihood theory for a simple ZI model

Our interest here lies in likelihood theory for observations y regarded as independent realisations of $\tilde{Y}_i \sim \tilde{\pi}_y(\theta, \gamma)$ where $\log \tilde{\pi}_y(\theta, \gamma) = \omega(\gamma, \pi_0)\mathbb{I}\{y = 0\} + \log(\rho) + \pi_y(\theta)$, where (θ, γ) are scalar parameters. We restrict ourselves throughout to the usual case where the mle for $\pi_y(\theta)$ may be studied via examination of the stationary points of the log likelihood. We first consider general $\pi_y(\theta)$; but subsequently we focus the case of $\pi_y(\theta)$ in the exponential family. Subsequently we outline the many possibilities beyond this family.

In this section we work with *iid* observations. Our focus will be on the special treatment of instances of $y = 0$. Of course, with *iid* data, all two parameter models are equivalent; we note that the essential parameters here are $(\theta, \tilde{\pi}_0)$ and that, here, $\tilde{\pi}_0$ has an obvious estimator (n_0/n). Focus therefore rests on the estimation of θ. But the objective is to lay the groundwork for regression, where the choice of ZI model has implications for θ. We concentrate initially on three ZI models: multiplicative, mixture and hurdle. But our interest is wider, for $\omega = \omega(\gamma, \pi_0(\theta))$ provides much flexibility. We subsequently consider the case of regression where $\theta_i = \theta(\mathbf{x}_i\beta)$ and $\gamma_i = \gamma(\mathbf{x}_i\alpha)$.

3.1. General Properties.

We note first some simple properties of $\tilde{\pi}_y(\theta, \omega)$ and $\pi_y(\theta)$. It is simple to show that $\tilde{\mu} = E[\tilde{Y}] = \rho E[Y] = \rho\mu$. More generally, for any function $h(\cdot)$,

$$E[h(\tilde{Y})] = h(0)\tilde{\pi}_0 + \rho \sum_{y \neq 0} h(y)\pi_y = h(0)(\tilde{\pi}_0 - \rho\pi_0) + \rho E[h(Y)] = (1-\rho)h(0) + \rho E[h(Y)].$$

Further it is easy to show that, for any functions $h_1(\cdot), h_2(\cdot)$

$$Cov[h_1(\tilde{Y})(h_2(\tilde{Y}))] = \rho Cov[h_1(Y)(h_2(Y))] + \rho(1-\rho)(E[h_1(Y)] - h_1(0))(E[h_2(Y)] - h_2(0))$$

Hence $Var[\tilde{Y}] = \rho Var[Y] + \rho(1-\rho)\mu^2$.

We note also the useful identities below:

$$\begin{aligned}\frac{\partial \log(\rho)}{\partial\theta} &= -\frac{\partial}{\partial\theta}\log(1+\kappa\pi_0) = -\rho\left(\kappa\frac{\partial\pi_0}{\partial\theta} + \pi_0\frac{\partial(1+\kappa)}{\partial\theta}\right)\\ &= (\rho-1)\frac{\partial}{\partial\theta}\log(\pi_0) - \tilde{\pi}_0\frac{\partial\omega}{\partial\theta};\ \text{and}\\ \frac{\partial\log(\rho)}{\partial\omega} &= -\tilde{\pi}_0\end{aligned}$$

Further we write $\frac{\partial\omega}{\partial\theta} = \frac{\partial\omega}{\partial\pi_0}\frac{\partial\pi_0}{\partial\theta} = \left(\pi_0\frac{\partial\omega}{\partial\pi_0}\right)\frac{\partial}{\partial\theta}\log(\pi_0)$, which we write more simply as $u\frac{\partial}{\partial\theta}\log(\pi_0)$. We also write $v = \frac{\partial\omega}{\partial\gamma}$ and note that $\frac{\partial}{\partial\gamma}\log(\rho) = -v\tilde{\pi}_0$.

The important term $u = \pi_0\frac{\partial\omega}{\partial\pi_0}$ characterises, for mle, the crucial aspect of the relationship between $\tilde{\pi}_0$ and π_0, itself the essence of ZI modelling. One useful re-expression involves writing $g = \text{logit}(\pi_0)$ and $\tilde{g} = \text{logit}(\tilde{\pi}_0)$; whence $\omega = \tilde{g} - g$. Then we have an alternative formulation for u:

$$u = \pi_0\frac{\partial\omega}{\partial\pi_0} = \pi_0\frac{\partial\omega}{\partial g}\frac{\partial g}{\partial\pi_0} = \pi_0\left(\frac{\partial\tilde{g}}{\partial g} - 1\right)\frac{1}{\pi_0(1-\pi_0)} = \left(\frac{\partial\tilde{g}}{\partial g} - 1\right)\frac{1}{1-\pi_0} \tag{1}$$

The term $\frac{\partial\tilde{g}}{\partial g}$ is the slope of the curve in Figure 1B.

For the multiplicative model $\omega = \gamma$; thus $u = 0$ and $v = 1$. For the mixture model $\omega = \log(\pi_0 + e^{-\gamma}) - \log(\pi_0)$; thus $u = -\frac{e^{-\gamma}}{\pi_0+e^{-\gamma}}$, and $v = e^{-\gamma}u$. Recalling that $\tilde{\pi}_0 = e^{\omega}\rho\pi_0 = \rho(\pi_0 + e^{-\gamma})$ for this ZI model, we can re-express this as $u = -\frac{1-\rho}{\tilde{\pi}_0} = -\frac{\kappa}{1+\kappa}$. One implication is that here $\frac{\partial}{\partial\theta}\log(\rho) = 0$. This, of course, follows more directly from $\rho = q$, independently of π_0, which is the distinguishing characteristic of the mixture model. In this it contrasts with the distinguishing characteristic of the multiplicative model, which is that ω is independent of π_0. For the hurdle, $\omega = \gamma - \text{logit}(\pi_0)$; thus $u = -\frac{1}{1-\pi_0}$ and $v = 1$. For the additive $\omega = \gamma - \log(\pi_0)$; thus $u = -1, v = 1$.

3.2. **Score functions.** We recall that $\log(\tilde{\pi}_y(\theta,\omega)) = \mathbb{I}\{y=0\}\omega(\gamma,\pi_0) + \log(\rho) + \log(\pi_y(\theta))$. It is useful to denote the score functions *wrt* θ, for $\tilde{\pi}_y$ and π_y, as $\tilde{S}_\theta(y) = \frac{\partial}{\partial\theta}\log(\tilde{\pi}_y(\theta,\omega))$ and $S_\theta(y) = \frac{\partial}{\partial\theta}\log(\pi_y(\theta))$. For γ, we write $\tilde{S}_\gamma(y) = \frac{\partial}{\partial\gamma}\log(\tilde{\pi}_y(\theta,\gamma))$.

Then the score functions are:

$$\begin{aligned}\tilde{S}_\theta(y) &= \frac{\partial}{\partial\theta}\left(\omega\mathbb{I}\{y=0\} + \log(\rho) + \log(\pi_y)\right)\\ &= (\mathbb{I}\{y=0\} - \tilde{\pi}_0)uS_\theta(0) + (\rho-1)S_\theta(0) + S_\theta(y)\,, \qquad (2)\\ \tilde{S}_\gamma(y) &= (\mathbb{I}\{y=0\} - \tilde{\pi}_0)v \qquad (3)\end{aligned}$$

For the ZI models above, $\tilde{S}_\theta(y)$ simplifies. For the mixture model, with $u = -\frac{\kappa}{1+\kappa}$, we have

$$\tilde{S}_\theta(y) = (1+\kappa)^{-\mathbb{I}\{y=0\}}S_\theta(y)$$

The implication of this simplification is that $S_\theta(y)$ carries a weight of $(1+\kappa)^{-1} = e^{-\omega}$ when $y = 0$. When $\kappa > 0$ this weight is $P(Y=0|\tilde{Y}=0)$, in the conventional notation of this ZI model, and one important component of Lambert's EM algorithm. But note that the multiplier has value $(1+\kappa)^{-1}$ for *all* κ, and can always can be interpreted as the expected number of instances of $Y = 0$ associated with each $\tilde{Y} = 0$, which exceeds one for $\omega < 0$, that is, for under-inflation. Recall that under-inflation in this model must be constrained to $\tilde{\pi}_0 > 0$, and is thus subject to subject to $\gamma < -\log(\pi_0)$.

For the hurdle, recalling that $\rho = \frac{1-\tilde{\pi}_0}{1-\pi_0}$, we find that

$$\begin{aligned}\tilde{S}_\theta(y) &= -\frac{1-\tilde{\pi}_0}{1-\pi_0} S_\theta\left(0\right) + (\rho-1)S_\theta\left(0\right) + S_\theta\left(0\right) = 0, \text{ for } y=0 \text{ and} \\ &= \frac{\pi_0}{1-\pi_0} S_\theta\left(0\right) + S_\theta\left(y\right), \text{ for } y > 0\end{aligned}$$

The implication of $\tilde{S}_\theta(0) = 0$ is that here instances of $\tilde{Y} = 0$ do not contribute to inference on θ, as is standard in the hurdle. It is easy to establish that $\tilde{S}_\theta(y) = \frac{\partial}{\partial\theta}\log(\pi_y^+(\theta))$, where $\pi_y^+(\theta) = \frac{\pi_y}{1-\pi_0}$ is the truncated version of $\pi_y(\theta)$.

For the multiplicative, with $u = 0$, we have $\tilde{S}_\theta(y) = (\rho-1)S_\theta\left(0\right) + S_\theta\left(y\right)$. We will see below that this is particularly simple for π_y in the exponential family.

Compactly we may write, with $m_Y = S_\theta\left(0\right)$,

$$\tilde{S}(y) = \left[\begin{array}{c}\tilde{S}_\theta(y)\\ \tilde{S}_\gamma(y)\end{array}\right] = G\left[\begin{array}{c}S_\theta\left(y\right)\\ \mathbb{I}\{y=0\}-\tilde{\pi}_0\end{array}\right] + \left[\begin{array}{c}(\rho-1)m_Y\\ 0\end{array}\right] \tag{4}$$

where $G = \left(\begin{array}{cc}1 & uS_\theta\left(0\right)\\ 0 & v\end{array}\right)$. We shall see that, when $\pi_y(\theta)$ is in the exponential family, $m_Y = E[Y]$.

3.3. **Information Matrices.** Fisher's expected information, $\tilde{\mathcal{I}}_{\theta,\gamma} = Var\left[\begin{array}{c}\tilde{S}_\theta(\tilde{Y})\\ \tilde{S}_\gamma(\tilde{Y})\end{array}\right]$ follows directly from (4) via the consideration of $Var\left[\begin{array}{c}S_\theta\left(\tilde{Y}\right)\\ \mathbb{I}\{\tilde{Y}=0\}-\tilde{\pi}_0\end{array}\right]$. From Section 3.1 we have

$$Var\left[\tilde{S}_\theta(\tilde{Y})\right] = \rho Var\left[S_\theta\left(Y\right)\right] + \rho(1-\rho)m_Y^2.$$

Clearly $Cov\left[S_\theta\left(Y\right), \mathbb{I}\{Y=0\}\right] = m_Y$ and $Var\left[\mathbb{I}\{\tilde{Y}=0\}\right] = \tilde{\pi}_0(1-\tilde{\pi}_0)$. Then $\tilde{\mathcal{I}}_{\theta,\gamma} = GHG^T$, where

$$H = \left(\begin{array}{cc}\rho Var\left[S_\theta\left(Y\right)\right] + \rho(1-\rho)m_Y^2 & m_Y\\ m_Y & \tilde{\pi}_0(1-\tilde{\pi}_0)\end{array}\right)$$

.

The observed information matrix is available, via additional identities for $\frac{\partial\rho}{\partial\theta}$ and $\frac{\partial\tilde{\pi}_0}{\partial\theta}$ which follow from $\frac{\partial\rho}{\partial\theta} = \rho\frac{\partial}{\partial\theta}\log(\rho)$ and $\frac{\partial\tilde{\pi}_0}{\partial\theta} = \tilde{\pi}_0\frac{\partial}{\partial\theta}\log(\tilde{\pi}_0)$ and from the second derivatives of $\omega(\gamma,\pi_0)$. For the mixture model these second derivatives are not insightful. But for simpler cases where $\omega = \gamma + \tau\mathbf{f}(\pi_0)$ only the second derivative *wrt* θ are non-zero, leading to simplifications.

The information matrix provides access to an approximation to the *joint* variance of the estimators $(\hat{\theta}, \hat{\gamma})$.

3.4. **The case of *iid* data for general $\pi_y(\theta)$.** Our focus on the *iid* case comes via the inferential role of the instances of $y_i = 0$. Of course here all two parameter models are equivalent, for all may be characterised by $(\theta, \tilde{\pi}_0)$. However, the expected information is not the same for all ZI models, reflecting the fact that the parameter γ has different interpretations in different models.

Given n *iid* observations $\mathbf{y}$, of which n_0 have value zero the mle's $(\hat{\theta}, \hat{\gamma})$ satisfy: $\tilde{S}_\theta\big(\mathbf{y}\big) = \sum \tilde{S}_\theta\big(y_i\big) = 0$ and $\tilde{S}_\gamma\big(\mathbf{y}\big) = \sum \tilde{S}_\gamma\big(y_i\big) = 0$ where

$$\tilde{S}_\theta\big(\mathbf{y}\big) = \Big(\sum(\mathbb{I}\{y_i = 0\} - \tilde{\pi}_0)\Big) uS_\theta\,(0) + n(\rho - 1)S_\theta\,(0) + \sum_i S_\theta\,(y_i)$$

$$\tilde{S}_\gamma\big(\mathbf{y}\big) = \left(\sum_i \mathbb{I}\{y_i = 0\} - n\tilde{\pi}_0\right) v = (n_0 - n\tilde{\pi}_0)v$$

From the latter, $\hat{\tilde{\pi}}_0 = \tilde{\pi}_0(\hat{\theta}, \hat{\gamma}) = \frac{n_0}{n}$, a trivial result in the *iid* case. The former then simplifies to estimators, including $\hat{\theta}, \hat{\pi}_0 = \pi_0(\hat{\theta}), \hat{\omega} = \omega(\hat{\gamma}, \hat{\pi}_0) = \log(1 + \hat{\kappa})$ and similarly $\hat{\rho} = \frac{1-\hat{\tilde{\pi}}_0}{1-\hat{\pi}_0}$, being such that they satisfy

$$\tilde{S}_\theta\big(\mathbf{y}\big) = \left(n(\hat{\rho} - 1) + n_0\right) S_\theta\,(0) + \sum_{y_i > 0} S_\theta\,(\theta; y_i)$$

The terms (u, v) have cancelled in this *iid* case. Thus confirms, of course, that here all two-parameter ZI models have the same mle's $(\hat{\theta}, \hat{\tilde{\pi}}_0)$.

3.5. **π_y in the exponential family.** Alternative forms of the score equations arise for $\pi_y(\theta)$ from within the exponential family; that is $\log(\pi_y(\theta)) = \theta y - A(\theta)$ to within an additive function of y. For then we have:

$$\log(\tilde{\pi}_y(\theta, \omega)) = \mathbb{I}\{y = 0\}\omega + \theta y + \log(\rho) - A(\theta) = \mathbb{I}\{y = 0\}\omega + \theta y - \tilde{A}(\omega, \theta)$$

with $\tilde{A}(\omega, \theta) = A(\theta) - \log(\rho)$.

Now, when $\omega(\gamma, \pi_0) = \gamma$ is independent of π_0, then in this context $\tilde{\pi}(\theta, \gamma)$ is a member of the two parameter exponential family with natural parameters (θ, γ). This confirms, in a precise sense, the earlier statement that multiplicative ZI is the simplest case. Now $\tilde{S}_\theta\big(y\big) = y - \rho\mu$ (with $\mu = A'(\theta)$) may be interpreted as $\tilde{S}_\theta\big(y\big) = y - E[\tilde{Y}]$; similarly $\tilde{S}_\gamma\big(y\big) = \mathbb{I}\{y = 0\} - E[\mathbb{I}\{\tilde{Y} = 0\}]$. In the *iid* case the sufficient statistics are $n_0 = \sum \mathbb{I}\{y_i = 0\}$ and $\sum y_i$. Note that the sample mean $\bar{y}$ is the mle for $E[\tilde{Y}] = \rho\mu(\theta)$. Given ρ, the rescaled sample mean is unbiased for μ. But ρ is not known.

For a general ZI model Equations (2) and (3) are now written more simply as

$$\tilde{S}_\theta\big(y\big) = -(\mathbb{I}\{y = 0\} - \tilde{\pi}_0)u\mu + (y - \rho\mu) \tag{5}$$

$$\tilde{S}_\gamma\big(y\big) = (\mathbb{I}\{y = 0\} - \tilde{\pi}_0)v \tag{6}$$

The expected information is

$$\begin{aligned} \tilde{\mathcal{I}}_{\theta,\gamma} &= \begin{pmatrix} 1 & -u \\ 0 & v \end{pmatrix} Var\begin{bmatrix} \tilde{Y} \\ \mathbb{I}\{\tilde{Y} = 0\} \end{bmatrix} \begin{pmatrix} 1 & 0 \\ -u & v \end{pmatrix} \\ &= \begin{pmatrix} 1 & -u \\ 0 & v \end{pmatrix} \begin{pmatrix} \rho Var[Y] + \rho(1-\rho)\mu^2 & -\rho\mu\tilde{\pi}_0 \\ -\rho\mu\tilde{\pi}_0 & \tilde{\pi}_0(1 - \tilde{\pi}_0) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -u & v \end{pmatrix} \end{aligned} \tag{7}$$

since $Cov[\tilde{Y}, \mathbb{I}\{\tilde{Y}=0\}] = E[\tilde{Y}\ \mathbb{I}\{\tilde{Y}=0\}] - E[\tilde{Y}]E[\mathbb{I}\{\tilde{Y}=0\}] = -\rho\mu\tilde{\pi}_0$ as $\tilde{Y}\ \mathbb{I}\{\tilde{Y}=0\} = 0$. This is particularly simple when $u = 0$ for the multiplicative ZI.

3.6. **The case of *iid* data for $\pi_y(\theta)$ in the exponential family.** The score equations are $\sum_i \tilde{S}_\theta(y_i) = 0$ and $\sum_i \tilde{S}_\gamma(y_i) = 0$. As previously the terms (u, v) cancel, and $\hat{\tilde{\pi}}_0 = \frac{n_0}{n}$. But now for π_y in the exponential family, $\hat{\theta}$ is the solution to $\sum(y_i - \rho\mu) = 0$.

For multiplicative ZI, $u = 0$, whence $\phi = \rho, \psi = 1$: thus $\tilde{S}_\theta(y) = y - \rho\mu$. Thus for multiplicative ZI, $(\hat{\theta}, \hat{\gamma})$, or equivalently for $(\hat{\theta}, \hat{\rho})$, are the solution to the equations $\sum_i(y_i - \rho\mu(\theta)) = 0$ and $\sum_i(\mathbb{I}\{y_i = 0\} - \tilde{\pi}_0) = 0$. A natural iterative algorithm begins with $\hat{\theta}^{(1)} = \bar{y}$ from an initial $\hat{\rho}^{(0)} = 1$; then with $\hat{\tilde{\pi}}_0 = \frac{n_0}{n}$ and $\hat{\rho}^{(1)} = \frac{1-\hat{\tilde{\pi}}_0}{1-\tilde{\pi}_0}$, we have $\hat{\theta}^{(2)} = \left(\hat{\rho}^{(1)}\right)^{-1}\bar{y}$; repeating until convergence. We refer to this as an iterative re-scaling algorithm.

Observe that there is no special treatment of zeroes, as arose with the mixture and hurdle formulations of ZI; yet for the *iid* case these solutions must lead to the same estimators. It is insightful, both here and for the later purpose of studying regression, to re-express (5) as

$$\tilde{S}_\theta(y) = -\mathbb{I}\{y=0\}u\mu + y - (\rho - u\tilde{\pi}_0)\mu = \psi^{\mathbb{I}\{y=0\}}(y - \phi\mu) \tag{8}$$

where $\phi = \rho - u\tilde{\pi}_0$ and $\psi = \frac{u+\phi}{\phi} = 1 + \frac{u}{\phi}$. The term ψ is a special weight on instances of $y_i = 0$.

Thus we can write the score equation for θ in a general form, which will be important in regression.

$$\tilde{S}_\theta(\mathbf{y}) = \sum_i \psi^{\mathbb{I}\{y_i=0\}}(y_i - \phi\mu) = 0 \tag{9}$$

Then the solution must satisfy the weighted average $\phi\mu = \frac{\sum \psi^{\mathbb{I}\{y_i=0\}}y_i}{\sum \psi^{\mathbb{I}\{y_i=0\}}} = \frac{\sum y_i}{\sum \psi^{\mathbb{I}\{y_i=0\}}}$. Here the term $n^* = \sum \psi^{\mathbb{I}\{y_i=0\}}$ may be thought of as an effective sample size (ESS); this formulation thus generalises Cohen (1960). All ZI models, and thus all (ϕ, ψ) terms, lead here to the same ESS as we outline.

It is useful to consider the mixture version of Equation (9). This solves:

$$\tilde{S}_\theta(y) = \sum (1+\kappa)^{-\mathbb{I}\{y_i=0\}}(y_i - \mu) = 0 \tag{10}$$

The natural algorithm model now (in the case of $\pi_y(\theta)$ in the exponential family) leads to $\hat{\theta}$ being the solution obtained by iterative re-weighting of instances of $y = 0$. That is, from an initial $\hat{\kappa}^{(0)} = 0$ we obtain $\hat{\theta}^{(1)}$ satisfying $\mu(\theta) = \bar{y}$; the equation $\tilde{\pi}_0 = \frac{n_0}{n}$ leads to $\hat{\kappa}^{(1)}$; the iteration proceeds to convergence. As we shall see, in the wider context of regression where the ZI models lead to different solutions, the natural algorithm solves equations based on (10) by iteratively re-scaling and re-weighting. In the *iid* case, the exponential family formulation leads to all versions having the same ESS.

However, there is a subtle theoretical point for n_0 that can be seen as a realisation of $N_0 = \sum \mathbb{I}\{y_i = 0\}$. Then the expected value of the (random) effective sample size N^* is $E[N^*] = E\left[\frac{N_0}{1+\kappa} + (n - N_0)\right] = n\rho$. But $E\left[\hat{\theta}\right] = E\left[\frac{\sum_i \tilde{Y}_i}{N^*}\right] \neq \frac{\sum_i E[\tilde{Y}_i]}{E[N^*]} = \frac{n\rho\mu}{n\rho} = \mu$. Thus, unlike the re-scaling estimate with known ρ, the weighted estimate with known κ is not unbiased for μ. It is of course asymptotically unbiased. In the special case of *iid* data and unknown parameters the estimators are identical, but both are biased.

3.6.1. *ϕ and ψ are non-negative.* It is useful to know, for the purposes of algorithms, that for all $\pi_y(\theta)$ in the exponential family the both of the weights ϕ and ψ are non-negative, the latter subject to the important constraint that $\tilde{\pi}_0$ is a non-decreasing function of π_0. We demonstrate below by exploring the implications of $\phi = 0$ and $\psi = 0$.

Starting from (1):

$$\phi = \rho - \tilde{\pi}_0 u = \frac{1-\tilde{\pi}_0}{1-\pi_0} + \frac{\tilde{\pi}_0}{1-\pi_0}\left(1 - \frac{\partial \tilde{g}}{\partial g}\right) = \frac{1}{1-\pi_0}\left(1 - \tilde{\pi}_0 \frac{\partial \tilde{g}}{\partial g}\right)$$

Thus $\phi \geq 0$ when $\frac{\partial \tilde{g}}{\partial g} \leq \frac{1}{\tilde{\pi}_0}$ or, equivalently, when $\frac{\partial g}{\partial \tilde{g}} \geq \tilde{\pi}_0 = \frac{e^{\tilde{g}}}{1+e^{\tilde{g}}}$. But $\phi = 0$ leads to $\frac{\partial g}{\partial \tilde{g}} = \frac{e^{\tilde{g}}}{1+e^{\tilde{g}}}$, which requires that $g = \gamma + \log(1+e^{\tilde{g}}) = \gamma - \log(1-\tilde{\pi}_0)$ for some constant γ; or, equivalently, when $\tilde{\pi}_0 = 1 - e^{\gamma}\frac{1-\pi_0}{\pi_0} = (1+e^{\gamma}) - \frac{e^{\gamma}}{\pi_0}$, providing, of course, that $0 \leq \tilde{\pi}_0 \leq 1$. But this in turn requires that $\frac{e^{\gamma}}{1+e^{\gamma}} \leq \pi_0 \leq 1$; that is, π_0 is bounded below. However, this is contradicted by π_y being in the exponential family unless $e^{\gamma} = 0$, and this in turn requires that the degenerate case of $\tilde{\pi}_0 = 1$ is the only model for which $\phi = 0$. We thus assert that, apart from this degenerate case, $\phi > 0$ for π_y in the exponential family.

As regards ψ, we note that $\psi = \frac{u+\phi}{\phi}$ and focus is on the sign of

$$u + \phi = \rho + (1-\tilde{\pi}_0)u = \rho + \frac{1-\tilde{\pi}_0}{1-\pi_0}\left(\frac{\partial \tilde{g}}{\partial g} - 1\right) = \rho\frac{\partial \tilde{g}}{\partial g}.$$

Thus if $\frac{\partial \tilde{g}}{\partial g} \geq 0$, that is if $\tilde{\pi}_0$ is an increasing function of π_0, then $\psi \geq 0$, since $\phi \geq 0$. Note that $\frac{\partial \tilde{g}}{\partial g} = 0$ characterises the hurdle model.

We note that the attractive but possibly pathological case of $\omega = \gamma\log(\pi_0)$ has $u = \gamma$ and $v = \log(\pi_0)$. But recall that when $\gamma > 1$, $\tilde{\pi}_0$ is a *decreasing* function of π_0, for small π_0. It follows that the special weight ψ, on instances of $y_i = 0$ can be negative.

We now turn to the wider context of ZI regression. Here ZI is modelled by arbitrary functions $\omega(\gamma_i, \pi_{0i})$, where $\gamma_i = \gamma(\mathbf{x}_i\alpha)$ and $\pi_0(\theta(\mathbf{x}_i\alpha))$, the key new issue is that we no longer have the simple $\hat{\tilde{\pi}}_0 = \frac{n_0}{n}$. The estimation of α is the essential new issue; this is ZI modelling. The estimation of β is that of regression in the presence of ZI.

4. Regression models for θ and γ in the exponential family

We now consider maximum likelihood for the case where the elements y_i of $\mathbf{y}$ are taken as independent realisations of $\tilde{Y}_i \sim \tilde{\pi}_y(\theta_i, \gamma_i)$ where $\theta_i = \mathbf{x}_i\beta$ and $\gamma_i = \mathbf{x}_i\alpha$ in the context of $\omega(\gamma, \pi_0) = \omega(\gamma(\alpha\mathbf{x})) + \tau\mathbf{f}(\pi_0(\theta))$. We initially consider the ZI type, and thus assume that the coefficients τ and the functions $\mathbf{f}(\cdot)$, to be known. We focus exclusively on the case of the exponential family. We consider separately the estimation of β for given α and of α for given β, envisaging an algorithm that iterates between these from an initial α corresponding to the absence of ZI. The former builds on the *iid* estimating equation at (10); the latter generalises the trivial result for the *iid* case,

$\hat{\tilde{\pi}}_0 = \frac{n_0}{n}$, and is thus the main focus of this section. Initially we focus on the case where the ZI model is itself known; that is, the coefficients τ and the functions $\mathbf{f}(\pi_0)$ are known. Subsequently we consider the modelling of ZI itself.

4.1. β **parameters, for given** α**.** The score equations for each β_j follow directly from (5). It follows that, for case $(y_i, \mathbf{x}_i)$,

$$\tilde{S}_\beta\left(y_i\right) = \nabla_\beta \log(\tilde{\pi}_{y_i}(\theta_i, \omega_i)) = \tilde{S}_\theta\big(y_i\big)\nabla_\beta(\theta_i) = \psi_i^{\mathbb{I}\{y_i=0\}}(y_i - \phi_i\mu(\theta_i))\mathbf{x}_i \tag{11}$$

where $\omega_i = \omega(\gamma_i, \pi_0(\theta_i))$ and $\gamma_i = \alpha\mathbf{x}_i$, and (ψ_i, ϕ_i) are defined by these. Thus

$$\tilde{S}_\beta\left(\mathbf{y}\right) = \sum_i \psi_i^{\mathbb{I}\{y_i=0\}}(y_i - \phi_i\mu(\theta_i))\mathbf{x}_i = \sum_i \left(\psi_i^{\mathbb{I}\{y_i=0\}}\phi_i\right)(\dot{y}_i - \mu(\theta_i))\mathbf{x}_i \tag{12}$$

where $\dot{y}_i = \phi_i^{-1}y_i$. In the second form we recognise, for given (ψ_i, ϕ_i), the estimating equations for a weighted quasi-(log)likelihood, the weights being $\psi_i^{\mathbb{I}\{y_i=0\}}\phi_i$, the likelihood being based on that of $\pi_y(\theta)$; see for example Pawitan (2001) Ch 14. Note that $\dot{Y} = \phi_i^{-1}Y$ is no longer integer; $\pi_y(\theta)$ is not its pmf. One solution is thus via iterative, weighted, quasi-loglikelihood, the weights depending on γ_i and thus on α. Ultimately, such an algorithm is iteratively re-weighted least squares. But typically the coefficients α are themselves not known.

4.2. α **parameters, for given** θ_i**.** The estimation of the α parameters follow from score equations built on (6); but here the terms $v_i = \frac{\partial\omega_i}{\partial\gamma}$ no longer cancel. Formally the mle for α satisfies

$$\tilde{S}_\alpha\left(\mathbf{y}\right) = \sum_i(\mathbb{I}\{y_i = 0\} - \tilde{\pi}_0(\theta_i))\mathbf{x}_i \tag{13}$$

It is useful to interpret this in the light of Figure 1. We consider first the simplest case, that of Multiplicative ZI with $\omega = \gamma$ being a scalar independent of $\mathbf{x}_i$ and thus common to all cases. Then an estimate of γ is returned, as the intercept, by regular linear logistic regression of $\mathbb{I}\{y_i = 0\}$ on (known) $\text{logit}(\pi_0(\theta_i))$, in which the coefficient of $\text{logit}(\pi_0(\theta_i))$ is forced to unit value. The fitted values are estimates of $\tilde{\pi}_0(\gamma, \theta_i)$ with here the common γ being estimated by the intercept. This is most simply achieved by using $\text{logit}(\pi_0(\theta_i))$ as an offset term in a linear logistic regression of $\mathbb{I}\{y_i = 0\}$ on the (common) unit vector. For this ZI model there is a linear relationship, with unit slope, between fitted $\text{logit}(\tilde{\pi}_0(\theta_i))$ and $\text{logit}(\pi_0(\theta_i))$ as in Figure 1. The algorithmic solution of (13) is thus straightforward, for given fitted β. The joint estimation of (α, β) thus involves iterating between (12) and (13) until convergence.

The extension to $\omega_i = \alpha\mathbf{x}_i$ is similarly achieved by logistic regression of the $\mathbb{I}\{y_i = 0\}$ on covariates $\mathbf{x}_i$, again with an offset. Of course, if there is significant variation on γ_i the plot of $\text{logit}(\tilde{\pi}_{0i})$ against $\text{logit}(\pi_{0i})$ will not be as simple as in Figure 1. Nevertheless, the routine diagnostics of linear logistic regression will be of some assistance in criticising this choice of ZI model, albeit that without modification, these are strictly interpretable on in the sense that it is conditional on θ_i being known. We outline below the route to the joint inference, and thus to the modification.

The case of the hurdle model of ZI corresponds, in the simplest case, to the logistic regression of $\mathbb{I}\{y_i = 0\}$ on the unit value. The fitted values of $\text{logit}(\tilde{\pi}_0(\theta_i))$ will thus be independent of $\text{logit}(\pi_0(\theta_i))$. The extension to $\omega_i = \alpha\mathbf{x}_i$ is as in the simplest case. The additive model of ZI involves a regression on $\mathbf{x}_i$ and $\log(1 - \pi_0(\theta_i))$, with the coefficient of $\log(1 - \pi_0(\theta_i))$ forced to value

-1, again achieved by offset. More generally, when the ZI model is specified by known values τ on known functions $f(\pi_0(\theta_i))$, then these form the offset in a regression of $\mathbb{I}\{y_i = 0\}$ on the covariates, subject to the caveat illustrated above by the case of $\omega = \gamma \log(\pi_0)$.

The mixture model is more challenging algorithmically. But it too may be seen as logistic regression of $\mathbb{I}\{y_i = 0\}$ on the non-linear function $\text{logit}((1-q) + q\pi_0(\theta_i))$ of $\pi_0(\theta_i)$, which is most naturally parameterised as $\text{logit}\left(\frac{1+e^{\gamma_i}\pi_0(\theta_i)}{1+e^{\gamma_i}}\right)$ where $\gamma_i = \alpha\mathbf{x}_i$. The real imagination in Lambert's EM is the realisation that this can be achieved by conventional, albeit weighted, linear logistic regression. For, given an estimate of β and thus values of $\pi_0(\theta_i)$, it is possible to compute the equivalent of $\frac{1}{1+\kappa_i} = e^{-\omega_i}$ for each case. When $\kappa_i > 0$, this can be interpreted as the case specific value of $Pr(J_i = 1|\tilde{Y}_i = 0)$ for each of the n_0 instances of $y_i = 0$ in the data set. Of course, $Pr(J_i = 1|\tilde{Y}_i = y_i) = 1$ for each of the $(n - n_0)$ instances of $y_i > 0$.

Lambert's EM algorithm logistically regresses (latent) J against covariates $\mathbf{x}$ by the device of logistically regressing, with weights, $n_0 + n$ values of J on corresponding $\mathbf{x}$ as follows: for each of the $n - n_0$ cases with $y_i > 0$, create a case with $J_i = 1$, associating it with its covariate vector $\mathbf{x}_i$ and unit weight; for each of the n_0 cases with $y_i = 0$ create two sets of cases: n_0 cases with $J_i = 1$, corresponding $\mathbf{x}_i$ and weight $\frac{1}{1+\kappa_i} = e^{-\omega_i}$; and a further set of n_0 cases with $J_i = 0$, the same corresponding $\mathbf{x}_i$ but now weight $1 - e^{-\omega_i}$; we refer to this artificial data set as (J^+, X^+). The weighted logistic regression of these $n + n_0$ values of J^+ against their corresponding $\mathbf{x}^+$ yields the mle for coefficients α. Although this is efficient, it is somewhat opaque to critical evaluation of the mixture model for ZI. Irrespective of which algorithm is used to fit the unknown α, plots of n pairs $\left(\hat{\tilde{\pi}}_{0i}, \hat{\pi}_{0i}\right)$ in either the linear or logistic metrics (as in Figure 1) may be helpful, here as with all ZI models.

4.3. **Expected information.** The Expected information provides access to joint inference. In the case of the multiplicative model it is exact, for here $\tilde{\pi}_y$ is in the two-parameter exponential family. It comes directly from (7). Denoting as row-vectors $\mathbf{x}_i^\beta$ and $\mathbf{x}_i^\alpha$ the subsets of covariates involved modelling θ_i and γ_i we find

$$\tilde{\mathcal{I}}_{\beta,\alpha} = \sum_i \begin{pmatrix} \mathbf{x}_i^\beta \\ \mathbf{x}_i^\alpha \end{pmatrix}^T \tilde{\mathcal{I}}_{\theta_i,\gamma_i} \begin{pmatrix} \mathbf{x}_i^\beta \\ \mathbf{x}_i^\alpha \end{pmatrix} \tag{14}$$

Note that the subsets $\mathbf{x}_i^\beta$ and $\mathbf{x}_i^\alpha$ may in principle overlap, contrary to a restriction imposed by the EM algorithm when used to fit the Mixture model. It is to be anticipated that, as always but especially here, multi-collinearity will be encountered if the model is over-parameterised. But the options for avoiding this are increased by the availability of multiple different models for the ZI aspect of the model.

4.4. **Choice of ZI model.** If the choice of ZI model can be restricted to that which can be written as $\omega = \gamma + \sum_k \tau_k f_k(\pi_0)$, with $\gamma = \alpha\mathbf{x}$, then this is a simple extension of the above linear logistic regression. For now we may regress $\mathbb{I}\{y_i = 0\}$ against both $\mathbf{x}_i$ and the $f_k(\pi_{0i})$, estimating the coefficients τ_k, and thus the ZI model.

Arguably, this may be the simplest way to choose the ZI model, especially if flexibility in the τ coefficients permits a parsimonious model for γ. Conventional diagnostics may provide access to some criticism of the chosen model; for then the resultant plot of (the fitted values of) $\tilde{\pi}_{0i}$ vs π_{0i} may provide the simplest way to present the ZI model, which is no longer required to take one of a pre-chosen suite of models. Recall that current practice at best offers a comparison of the two extant ZI models (mixture and hurdle), typically in terms of a portmanteau measure such as BIC.

Indeed this may be enriched further to cater for the possibility that the ZI model is itself a function of the covariates $\mathbf{x}$; for this corresponds to allowing for interaction between the covariates and the functions $f_k(\pi_0)$. It is to be anticipated however that only data sets with a very strong signal to noise ratio will have the power to allow discrimination between alternative ZI models; and even fewer will allow meaningful identification of interactions. The caveat on non-decreasing ZI models may also constrain the search.

In closing we draw attention to other possibilities. It is not strictly necessary that the binary regression, with offset of logit(π_0), be *logistic* regression; for example probit($\cdot$) or other functions may be used to map from $\Re$ to the interval (0,1). For example, the model used by Ghosh et al. (2006), in addressing problems with very large numbers of observed zeroes, may be thought of as using the cdf of the beta-binomial distribution. There is a need howerer for careful consideration of the behaviour as $\pi_0 \to 0$ and $\pi_0 \to 1$.

5. Examples

We illustrate the power of this new approach with a simple example illustrated through Bayesian Inference via the Hamiltonian Monte Carlo package `rstan` (Stan Development Team 2018). We use a simple zero-inflated fishing data set to illustrate the use of model comparison techniques to determine the type of zero inflation amongst distinct sub-groups according to a covariate. The example is further illuminated by the plot of $(\pi_0, \tilde{\pi}_0)$ which displays the type of ZI behaviour. Code to run the examples is available at www.github.com/andrewcparnell/ZIpaper.

We use a data set from the UCLA Academic Technology Services website (http:// www.ats.ucla.edu) and previously analysed in Saffari et al. (2013). The data concern an environmental survey by a state wildlife biologist of groups that went to a park and attempted to catch fish. The number of fish they caught was recorded (variable `count`), as was further data on each group. We use the additional variables `persons` (number of people attending) and `camper` (whether or not they brought a camper van). Two counts had excessively high values of 65 and 149 and are removed from this exploratory analysis. The count response we write as $y_i, i = 1, \ldots, n$ and covariates $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ where $n = 248$. A histogram of the response and boxplots of the main covariates are shown in Figure 2.

We define the new distribution as $ZIPo$ with pmf:

$$\tilde{\pi}_y(\kappa, \theta) = \begin{cases} (1+\kappa)\rho\pi_0(\theta) & \text{if } y = 0 \\ \rho\pi_y(\theta) & \text{otherwise.} \end{cases}$$

with $\omega = \log(1+\kappa)$ and $\rho = (1+\kappa\pi_0)^{-1}$.

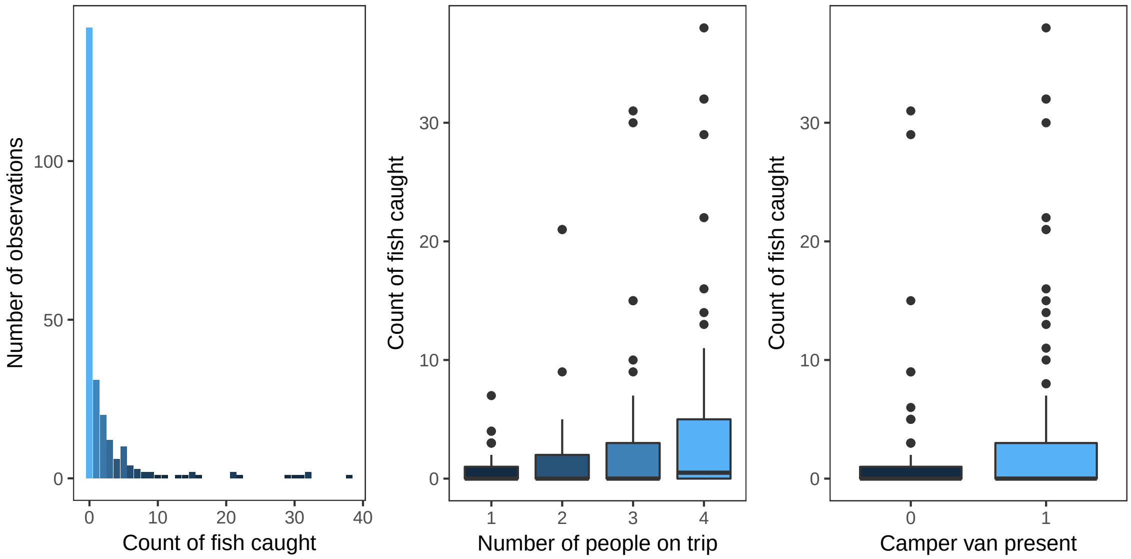

FIGURE 2. A plot of the fish data set. The left panel shows a histogram of the response, the middle panel a boxplot of the response against the covariate `persons`, and the right panel a boxplot of the response against the covariate `camper`.

For simplicity, we use the Poisson as the base distribution. The model can be written out hierarchically as:

$$\begin{aligned} y_i &\sim ZIPo(\omega_i, \theta_i) \\ \omega_i &= \alpha - \tau_1\theta_i + \tau_2 \log(1 - e^{-\theta_i}) \\ \log(\theta_i) &= \beta \mathbf{x}_i \\ \alpha &\sim N(0, 10^2) \\ \beta_k &\sim N(0, 10^2) \end{aligned}$$

where k is the covariate number and vague priors are given to the hyper-parameters. The different ZI models can be found through the values of τ. They are $(\tau_1, \tau_2) = (0, 0), (0, -1), (-1, -1)$ for the multiplicative, additive, and hurdle respectively. We identify these by fitting the model once for each type, and subsequently calculating the WAIC values (Watanabe 2013, Gelman et al. 2013).

We run the models for the default 2000 iterations across 4 chains with 1000 as burn-in and check convergence using the standard R-hat diagnostic (Brooks et al. 2011). We subsequently compute WAIC values using the `loo` package (Vehtari et al. 2016). Table 1 shows the results. The hurdle model seems to be strongly selected over the others. However we can extend this model to include situations where different ZI approaches may be preferred by different covariate values. The plot of (π_0 vs $\tilde{\pi}_0$ is shown in Figure 3 which, most interestingly, shows some elements of non-Hurdle like behaviour.

ZI type	WAIC	se(WAIC)
Hurdle	1,277	143
Multiplicative	1,359	159
Additive	1,365	159

Table 1. WAIC values with standard errors for a simple new ZI model applied to the fish data

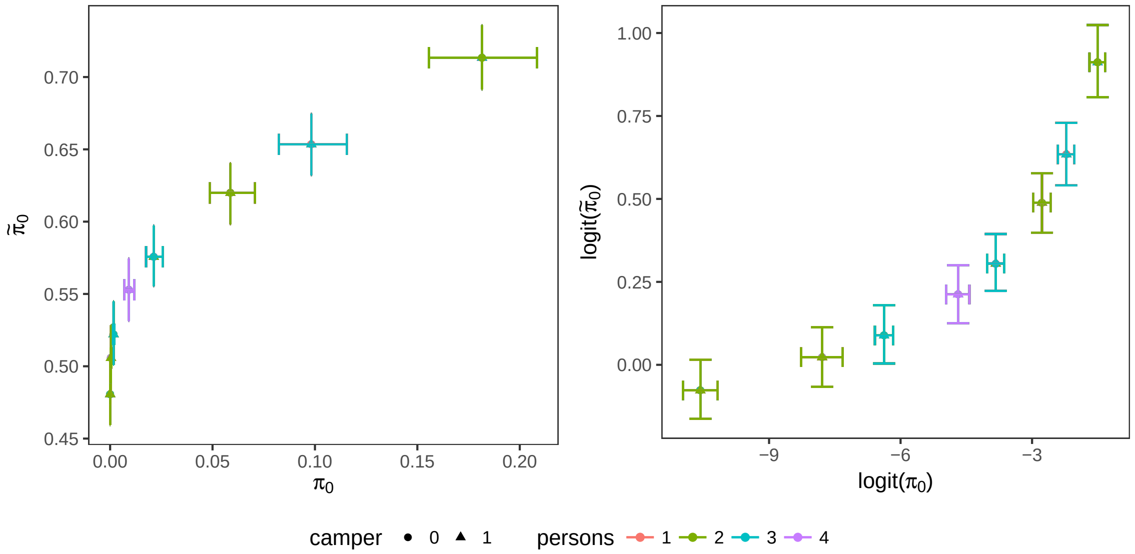

Figure 3. A plot of the posterior medians of π_0 vs $\tilde{\pi}_0$ (left panel) and the same on the logit scale on the right panel. This plot should be contrasted with Figure 1 to identify behaviour. The different covariate values used for the plot are shown in the legend.

6. Extensions

We see multiple extensions. These include: Bayesian inference with latent processes; the use of other families of univariate count distributions; multiple inflations of count data; extensions to continuous distributions; and multivariate zero-inflations of count data. We outline some options.

6.1. **Bayesian inference with latent processes.** With $(\theta, \gamma) = (\theta(\mathbf{x}\beta), \gamma(\mathbf{x}\alpha))$ we seek the posterior conditional distribution $[\beta, \alpha|\tilde{\mathbf{Y}} = \mathbf{y}]$ which for simplicity of notation, we write in this section as $[\beta, \alpha|\tilde{\mathbf{Y}}]$. In is natural to approach this by sampling from $[\beta, \alpha|\tilde{\mathbf{Y}}, \mathbb{I}\{\tilde{\mathbf{Y}} = 0\}, X]$, Gibbs fashion, by successively sampling from the full conditionals:

$$[\beta|\tilde{\mathbf{Y}}, \mathbb{I}\{\tilde{\mathbf{Y}} = 0\}, \alpha, X] = [\beta|\alpha, \tilde{\mathbf{Y}}, X] \qquad \text{and } [\alpha|\tilde{\mathbf{Y}}, \mathbb{I}\{\tilde{\mathbf{Y}} = 0\}, \pi_0(\theta), X] = [\alpha|\mathbb{I}\{\tilde{\mathbf{Y}} = 0\}, \pi_0(\theta), X].$$

The latter indicates the usual sampling in classical Bayesian inference for logistic (or more general binary) regression of observed instances of $\mathbb{I}\{\tilde{Y}=0\}$ on suitable functions of $\pi_0(\theta)$. We thus presume that $[\alpha|\mathbb{I}\{\tilde{\mathbf{Y}}=0\},\pi_0(\theta),X]$ is available and focus therefore on the former.

We shall presume the existence of an algorithm to sample from $[\beta|\mathbf{Y}=\mathbf{y},X]$, that is, Bayesian inference in the regression of count data, absent ZI, on covariates $\mathbf{x}$. Here we outline how to modify $[\beta|\mathbf{Y},X]$, in the presence of a known type of ZI, to build an algorithm to sample from $[\beta|\tilde{\mathbf{Y}},\alpha,X]$. We then propose the use of this algorithm to sample from $[\beta,\alpha|\tilde{\mathbf{Y}},X]$ above.

We recall that we can associate, with every $\tilde{Y}=y$, a latent integer M, denoting the implicit number of observations of $Y=y$. Clearly when $y\neq 0$ then $M=1$. But implicit in our model of ZI, is that when $y=0$, $M=1$ is only one possibility. In particular, with over-inflation, M is binary, with $P(M=1|\tilde{Y}=0)=E[M|\tilde{Y}=0]=\frac{1}{1+\kappa}=e^{-\omega}$; with under-inflation, however, M has a geometric distribution on $\mathbb{Z}_0^+$ with $E[M|\tilde{Y}=0]=\frac{1}{1+\kappa}=e^{-\omega}$. We note also that

$$[\beta|\tilde{\mathbf{Y}},\alpha,X]=E_{\mathbf{M}|\tilde{\mathbf{Y}},\alpha}\left[\beta|\alpha,\mathbf{M},\tilde{\mathbf{Y}},X\right]=E_{\mathbf{M}|\mathbb{I}\{\tilde{\mathbf{Y}}=0\},\alpha}\left[\beta|\alpha,\mathbf{M},\tilde{\mathbf{Y}},X\right].$$

But the joint knowledge that $\tilde{Y}_i=y_i$ and $M_i=m_i$ tells us that we can treat y_i, for inferential purposes, as m_i copies of $Y_i=y_i$ and of its covariates $\mathbf{x}_i$; we write this as $\left(y_i^M,\mathbf{x}_i^M\right)=\left(rep(y_i,m_i),rep(\mathbf{x}_i,m_i)\right)$ and use $\left(y^M,X^M\right)$ to refer to the artificial data set so constricted; that is, $\left[\beta|\alpha,M,\tilde{Y},X\right]=\left[\beta|Y=y^M,X^M\right]$

Then the vector of observations is $y_i;i=1,\dots n$, and each is accompanied by $m_i;i=1,\dots n$. This formulation is equivalent, for inference on β, to the regression of the vector of y^M on X^M. Formally $[\beta|\tilde{\mathbf{Y}},\mathbf{M},X]=[\beta|\mathbf{Y}^{\mathbf{M}},X^M]$. Integrating over $\mathbf{M}$ leads to

$$[\beta|\tilde{\mathbf{Y}},\alpha,X]=E_{\mathbf{M}|\tilde{\mathbf{Y}},\alpha}\left[\beta|\alpha,\tilde{\mathbf{Y}},\mathbf{M},X^{\mathbf{M}}\right]=E_{\mathbf{M}|\tilde{\mathbf{Y}},\alpha}[\beta|\mathbf{Y}^{\mathbf{M}},X^{\mathbf{M}}]$$

Clearly this can be more simply achieved by weighting each case $(y_i,\mathbf{x}_i)$ by random M_i, if that is supported. The natural integration method is by Monte Carlo.

Another possibility exploits the fact that we can write

$$[\beta|\mathbf{Y}^{\mathbf{M}}]\propto[\mathbf{Y}^{\mathbf{M}}|\beta][\beta]=\prod_i\left[Y_i^{M_i}|\beta\right][\beta]=\prod_i[Y_i,\beta]^{M_i}[\beta]$$

But $E_{M_i|\mathbb{I}\{\tilde{\mathbf{Y}}=0\},\alpha}[Y_i|\beta]^{M_i}=E_{M_i|\mathbb{I}\{y_i=0\},\alpha}[Y_i|\beta]^{M_i}$ is analytically available from the pgf of $[M_i|\mathbb{I}\{y_i=0\},\alpha]$. Further this distribution is simple, being concentrated on $M_i=1$ when $y_i>0$; binary when $y_i=0$ and $\omega>0$, and geometric when $y_i=0$ and $\omega<0$.

6.2. **Other distribution families.** There is a considerable interest in disentangling the effects of over-dispersion and zero-inflation. As mentioned, several authors have used distributions such as the negative-binomial, Conway-Poisson-Maxwell and power series distributions for this purpose, for such distributions already have a parameter for over-inflation. In practice, it seems likely that only in data sets with a very strong signal will such disentangling be feasible.

Two general families are are particularly attractive, but seem not to have been much pursued: the classic exponential-dispersion family, and the exponential over dispersion model of Dey et al. (1997). These seem to complement the notation of this paper, and specifically, the multiplicative model of ZI. In the former, with $\log(\pi_y(\theta,\eta)) = \frac{\theta y - A(\theta)}{\eta}$ to within an additive function of y, we may similarly define $\tilde{\pi}_y(\theta,\eta,\gamma) = \omega\mathbb{I}\{y=0\} + \frac{\theta y - A(\theta)}{\eta}$. In the latter, following Dey et al. (1997): $\log(\pi_y(\theta,\zeta)) = \theta y + \zeta Z(y) + \chi(\theta,\zeta)$ is a member of the 2-dimensional exponential family, and thus $\log(\tilde{\pi}_y(\theta,\zeta,\omega)) = \mathbb{I}\{y=0\}\omega + \theta y + \zeta Z(y) + \chi(\theta,\zeta)$ is a member of the 3-dimensional exponential family if ω is independent of (θ,ζ). Of course, we may go further for this ZI model, for if $\pi_y(\theta)$ for k-dimensional θ has the form of a k-dimensional exponential family distribution, then $\log(\pi_y(\theta,\omega)) = \omega\mathbb{I}\{y=0\} + \log(\pi_y(\theta))$ is $(k+1)$-dimensional exponential family.

6.3. **Multiple and continuous inflations.** We have focussed on zero inflation, that is the inflation of π_0 to $\tilde{\pi}_0$. But some authors (e.g. Sweeney 2012, Deng & Zhang 2015, Tian et al. 2015) in dealing with ZI binomial distributions $Bin(n,p)$, have remarked that the probabilities of both $(y=0)$ and $(y=n)$ can be inflated. Trivially the notation above may be adapted to the inflation of the probability of any other values of Y, such as $y=k\in K$ where interest will instead focus on multiple k and specifically on π_k and $\tilde{\pi}_k$ for $k\in K$ via

$$\log(\tilde{\pi}_y(\theta,\omega)) = \omega_k(K)\mathbb{I}\{Y=k\in K\} + \rho + \log(\pi_y(\theta)). \tag{15}$$

For example, in the binomial case, with $K=\{0,n\}$, $\omega_0(\{0,n\})$ and $\omega_n(\{0,n\})$ control, respectively, the inflations of π_0 and π_n. Clearly each can be parameterised via γ_k and hence involve covariates.

The zero-inflation of a continuous distribution has often been remarked on (e.g. Lambert 1992). The classic example is rainfall; more generally values less than some (possibly small) threshold are returned as zero. If $\pi(y,\theta)$ denotes the pdf of a continuous distribution then with K now denoting a continuous interval

$$\log(\tilde{\pi}(y,\theta,\omega)) = \omega\mathbb{I}\{y\in K\} + \rho + \log(\pi(y,\theta))$$

models a single parameter inflation of the pdf $\pi(y,\theta)$ for $y\in K$. This will have the form of a continuous pdf with a probability mass at $y=0$.

However this itself can be extended by $\omega=\omega(y,K)$ taken to be a continuous function. Then the resultant $\tilde{\pi}(y,\theta,\omega)$ will be a conventional continuous pdf, inflated continuously over $y\in K$.

6.4. **Multivariate ZI.** Several authors (e.g. Li et al. 1999, Dong et al. 2014, Liu & Tian 2015, Lee et al. 2017) have reported on the ZI of multivariate counts. The simplest discrete multivariate distribution is the Multinomial. But here the issue is multivariate ZI modelling. For illustrative purposes we consider bivariate counts $y=(y_a,y_b)$, and bivariate $\pi_y(\theta)$. There are three versions of zero: (i) $(0,\cdot),(\cdot,0),(0,0)$; here for example, $(0,\cdot)\equiv(y_a=0,y_b>0)$. Extending the notation of (15) we have three versions of ω : being $\omega_{(0,\cdot)},\omega_{(\cdot,0)},\omega_{(0,0)}$. It is often the case that (0,0) is never observed in a multinomial setting.

The only new issue is parsimony; for with k-dimensional y there are 2^k-1 such parameters. A simpler version of the above is however available: $\omega_{(0,0)} = \omega_{(0,\cdot)} + \omega_{(\cdot,0)}$; we have dropped the interaction terms. For higher dimensional cases, we can similarly drop all interactions, or all interactions higher than two-way, etc. The real challenge is the paucity of models $\pi_y(\theta)$ for discrete multivariate data (ie Absent ZI).

References

Agarwal, D. K., Gelfand, A. E. & Citron-Pousty, S. (2002), 'Zero-inflated models with application to spatial count data', *Environmental and Ecological statistics* **9**(4), 341–355.

Ancelet, S., Etienne, M.-P., Benoît, H. & Parent, E. (2010), 'Modelling spatial zero-inflated continuous data with an exponentially compound poisson process', *Environmental and Ecological Statistics* **17**(3), 347–376.

Angers, J.-F. & Biswas, A. (2003), 'A bayesian analysis of zero-inflated generalized poisson model', *Computational statistics & data analysis* **42**(1-2), 37–46.

Bhattacharya, A., Clarke, B. S., Datta, G. S. et al. (2008), A bayesian test for excess zeros in a zero-inflated power series distribution, *in* 'Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen', Institute of Mathematical Statistics, pp. 89–104.

Brooks, S., Gelman, A., Jones, G. & Meng, X.-L. (2011), *Handbook of markov chain monte carlo*, CRC press.

Chebon, S., Faes, C., Cools, F. & Geys, H. (2017), 'Models for zero-inflated, correlated count data with extra heterogeneity: when is it too complex?', *Statistics in medicine* **36**(2), 345–361.

Cohen, A. C. (1960), 'An extension of a truncated poisson distribution', *Biometrics* **16**(3), 446–450.

Dagne, G. A. (2004), 'Hierarchical bayesian analysis of correlated zero-inflated count data', *Biometrical Journal* **46**(6), 653–663.

Deng, D. & Zhang, Y. (2015), 'Score tests for both extra zeros and extra ones in binomial mixed regression models', *Communications in Statistics-Theory and Methods* **44**(14), 2881–2897.

Dey, D. K., Gelfand, A. E. & Peng, F. (1997), 'Overdispersed generalized linear models', *Journal of Statistical Planning and Inference* **64**(1), 93–107.

Dong, C., Clarke, D. B., Yan, X., Khattak, A. & Huang, B. (2014), 'Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections', *Accident Analysis & Prevention* **70**, 320–329.

Fisher, W. H., Hartwell, S. W. & Deng, X. (2017), 'Managing inflation: On the use and potential misuse of zero-inflated count regression models', *Crime & Delinquency* **63**(1), 77–87.

Garay, A. M., Hashimoto, E. M., Ortega, E. M. & Lachos, V. H. (2011), 'On estimation and influence diagnostics for zero-inflated negative binomial regression models', *Computational Statistics & Data Analysis* **55**(3), 1304–1318.

Gelfand, A. E. & Dalal, S. R. (1990), 'A note on overdispersed exponential families', *Biometrika* **77**(1), 55–64.

Gelman, A., Hwang, J. & Vehtari, A. (2013), 'Understanding predictive information criteria for Bayesian models'.

Ghosh, S., Gelfand, A. E., Zhu, K. & Clark, J. S. (2012), 'The k-ZIG: Flexible modeling for zero-inflated counts', *Biometrics* **68**(3), 878–885.

Ghosh, S. K., Mukhopadhyay, P. & Lu, J.-C. J. (2006), 'Bayesian analysis of zero-inflated regression models', *Journal of Statistical planning and Inference* **136**(4), 1360–1375.

Hall, D. B. (2000), 'Zero-inflated poisson and binomial regression with random effects: a case study', *Biometrics* **56**(4), 1030–1039.

Jorgensen, B. (1987), 'Exponential dispersion models', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 127–162.

Kassahun, W., Neyens, T., Faes, C., Molenberghs, G. & Verbeke, G. (2014), 'A zero-inflated overdispersed hierarchical poisson model', *Statistical Modelling* **14**(5), 439–456.

Klein, N., Kneib, T. & Lang, S. (2015), 'Bayesian generalized additive models for location, scale, and shape for zero-inflated and overdispersed count data', *Journal of the American Statistical*

Association **110**(509), 405–419.

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

Lee, K. H., Coull, B. A., Moscicki, A.-B., Paster, B. J. & Starr, J. R. (2017), 'Bayesian variable selection for multivariate zero-inflated models: Application to microbiome count data', *arXiv preprint arXiv:1711.00157* .

Li, C.-S., Lu, J.-C., Park, J., Kim, K., Brinkley, P. A. & Peterson, J. P. (1999), 'Multivariate zero-inflated poisson models and their applications', *Technometrics* **41**(1), 29–38.

Liu, Y. & Tian, G.-L. (2015), 'Type i multivariate zero-inflated poisson distribution with applications', *Computational Statistics & Data Analysis* **83**, 200–222.

Long, D., Preisser, J. S., Herring, A. H. & Golin, C. E. (2015), 'A marginalized zero-inflated poisson regression model with random effects', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **64**(5), 815–830.

Martin, J. & Hall, D. B. (2017), 'Marginal zero-inflated regression models for count data', *Journal of Applied Statistics* **44**(10), 1807–1826.

Miller, J. M. & Miller, D. (2008), 'No zero left behind: comparing the fit for zero-inflation models as a function of skew and proportion of zeros', *Interstat Statistics on the Internet.(Virginia Polytechnic Institute and State University: Blacksburg, VA)* .

Min, A. & Czado, C. (2010), 'Testing for zero-modification in count regression models', *Statistica Sinica* pp. 323–341.

Min, Y. & Agresti, A. (2005), 'Random effect models for repeated measures of zero-inflated count data', *Statistical modelling* **5**(1), 1–19.

Moghimbeigi, A. (2011), 'A score test for extra zeros in negative binomial mixed models', *Journal of Statistical Computation and Simulation* **81**(5), 635–644.

Molas, M. & Lesaffre, E. (2010), 'Hurdle models for multilevel zero-inflated data via h-likelihood', *Statistics in medicine* **29**(30), 3294–3310.

Mullahy, J. (1986), 'Specification and testing of some modified count data models', *Journal of econometrics* **33**(3), 341–365.

Mullahy, J. (1997), 'Heterogeneity, excess zeros, and the structure of count data models', *Journal of Applied Econometrics* pp. 337–350.

Neelon, B. & Chung, D. (2017), 'The LZIP: A bayesian latent factor model for correlated zero-inflated counts', *Biometrics* **73**(1), 185–196.

Patil, M. & Shirke, D. (2011), 'Tests for equality of inflation parameters of two zero-inflated power series distributions', *Communications in Statistics-Theory and Methods* **40**(14), 2539–2553.

Pawitan, Y. (2001), *In all likelihood: statistical modelling and inference using likelihood*, Oxford University Press.

Perumean-Chaney, S. E., Morgan, C., McDowall, D. & Aban, I. (2013), 'Zero-inflated and overdispersed: what's one to do?', *Journal of Statistical Computation and Simulation* **83**(9), 1671–1683.

Ridout, M., Demétrio, C. G. & Hinde, J. (1998), Models for count data with many zeros, *in* 'Proceedings of the XIXth international biometric conference', Vol. 19, pp. 179–192.

Ridout, M., Hinde, J. & DeméAtrio, C. G. (2001), 'A score test for testing a zero-inflated poisson regression model against zero-inflated negative binomial alternatives', *Biometrics* **57**(1), 219–223.

Rodrigues, J. (2003), 'Bayesian analysis of zero-inflated distributions', *Communications in Statistics-Theory and Methods* **32**(2), 281–289.

Saffari, S. E., Adnan, R. & Greene, W. (2013), 'Investigating the impact of excess zeros on hurdle-generalized poisson regression model with right censored count data', *Statistica Neerlandica*

67(1), 67–80.

Salter-Townshend, M. & Haslett, J. (2012), 'Fast inversion of a flexible regression model for multivariate pollen counts data', *Environmetrics* **23**(7), 595–605.

Sellers, K. F. & Raim, A. (2016), 'A flexible zero-inflated model to address data dispersion', *Computational Statistics & Data Analysis* **99**, 68–80.

Stan Development Team (2018), 'RStan: the R interface to Stan'. R package version 2.17.3.
URL: *http://mc-stan.org/*

Sweeney, J. (2012), Advances in Bayesian Model Development and Inversion in Multivariate Inverse Inference Problems: With Application to Palaeoclimate Reconstruction, PhD thesis, Trinity College Dublin.

Tian, G.-L., Ma, H., Zhou, Y. & Deng, D. (2015), 'Generalized endpoint-inflated binomial model', *Computational Statistics & Data Analysis* **89**, 97–114.

Todem, D., Kim, K. & Hsu, W.-W. (2016), 'Marginal mean models for zero-inflated count data', *Biometrics* **72**(3), 986–994.

Vehtari, A., Gelman, A. & Gabry, J. (2016), 'Practical bayesian model evaluation using leave-one-out cross-validation and waic', *Statistics and Computing* .

Watanabe, S. (2013), 'A Widely Applicable Bayesian Information Criterion', *Machine Learning Research* **14**(1), 867–897.

Xiang, L., Lee, A. H., Yau, K. K. & McLachlan, G. J. (2007), 'A score test for overdispersion in zero-inflated poisson mixed regression model', *Statistics in medicine* **26**(7), 1608–1622.

Xie, F.-C., Wei, B.-C. & Lin, J.-G. (2009), 'Score tests for zero-inflated generalized poisson mixed regression models', *Computational Statistics & Data Analysis* **53**(9), 3478–3489.

Yang, H., Li, R., Zucker, R. A. & Buu, A. (2016), 'Two-stage model for time varying effects of zero-inflated count longitudinal covariates with applications in health behaviour research', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **65**(3), 431–444.