Here we review additional literature related to zero-altered models. This literature emphasizes the more common zero-inflation case, while zero-deflation tends only to be considered in general formulations which are meant to encompass a broad class of models.

The use of mixture models and conditional models are two common approaches to handling zero-inflation within the context of ecological and health data. The well known ZIP model (Lambert 1992) is a mixture of a degenerate zero mass and a Poisson distribution. On the other hand, Welsh, Cunningham, Donnelly and Lindenmayer (1996) formulate a two-component conditional model where the presence/absence data is modelled with a binomial distribution and the abundance at active sites is modelled using a truncated Poisson or truncated negative binomial distribution.

Conditional models, also known as two-part models, hurdle models, and compatible models are discussed by Mullahy (1986) and Heilbron (1994), for example. Conditional models consist of a zero mass, the so-called ‘hurdle’, and a truncated form of a standard discrete distribution such as the binomial, Poisson or negative binomial. Most applications of hurdle models assume independence between the linear predictor of the probability of overcoming the hurdle and the conditional mean of the counts, given that the hurdle is overcome. This orthogonality simplifies computation and provides a straightforward interpretation of the covariate effects. In the ecology setting, the hurdle may be habitat suitability, and the conditional mean may represent mean abundance given suitable habitat. Thus, the covariate effects on habitat suitability can
be interpreted independently of the effects on abundance within suitable habitat. Under the
conditional model, the likelihood is the sum of two independent likelihoods with no terms in
common. Therefore, it is fully efficient to fit the two components separately.

Lambert’s (1992) mixture model approach to zero-inflated Poisson regression also allows
the design matrices associated with the Poisson mean, $\mu$, and the probability of arising from
the zero mass, $\theta$, to contain different sets of experimental factors and covariates. Through the
use of a latent variable representing membership in the zero component, a complete likelihood
is formed and an EM algorithm (Dempster, Laird and Rubin 1977) used to obtain maximum
likelihood estimates (MLE). The complete log-likelihood conveniently splits into the sum of
two exponential family log likelihoods so that weighted logistic and Poisson regressions can be
used to obtain parameter estimates. Interval estimates rely on likelihood asymptotics and are
based on normal approximations which require the log-likelihood surface to be approximately
quadratic near the MLE.

Thus, we have two specifications of the zero-inflated model: 1) a mixture of a degenerate
distribution with mass at zero and a non-degenerate distribution such as the binomial or Poisson
distribution, and 2) a conditional specification where the the zero mass and the truncated
distribution of the non-zero counts are modelled independently.

Consider $\theta$, the probability of a true zero, and $\mu$, the mean parameter for the probability
mass function $f$ associated with the random variable $Y$. Then the mixture model specification
of the zero-altered model is:

$$ Y \sim \theta I_\theta + (1 - \theta) f(Y | \mu) $$

where $I_\theta$ is the degenerate distribution taking the value zero with probability one. This formu-
lation encompasses distributions such as the binomial, Poisson, and generalized Poisson with
\[ E(Y|\theta, \mu) = (1 - \theta)E_f(Y|\mu) \] and
\[ Var(Y|\theta, \mu) = \theta(1 - \theta)E_f(Y|\mu)^2 + (1 - \theta)Var_f(Y|\mu) \]
where \( E_f(Y|\mu) \) and \( Var_f(Y|\mu) \) denote the expectation and variance of a random variable with probability mass function \( f \). Under this formulation, it is possible for \( \theta \) to be negative, in which case we have a model which provides for zero-deflation, though the interpretation of the distribution as arising from a mixture would be lost. Additional flexibility may be added to such models by incorporating random effects into the distribution of \( f \), or \( \theta \), or both.

Under the conditional model we specify \( Z = 1 \) if \( Y = 0 \) and \( Z = 0 \) if \( Y > 0 \). Then,
\[ Z \sim \theta_c Z c (1 - \theta_c) (1 - Z) \]
\[ Y|Z = 0 \sim f_{trunc}(Y|\mu) \]

where \( f_{trunc} \) is a truncated distribution such as a truncated binomial or truncated Poisson distribution, and \( \theta_c \) is the probability of a zero count.

Ridout, Demetrio and Hinde (1998), Martin et al. (2005) and Kuhnert, Martin, Mengersen and Possingham (2005) make the distinction between different types of zeros in the ecological setting. True zeros may be structural. This would be the case for immune individuals or when an ecological effect creates unsuitable species habitat. On the other hand, sampling zeros occur simply by chance and false zeros may arise due to a failure to detect an occurrence because of observing too small an area or because of limited ability of a species to disperse to all parts of the region. Martin et al. (2005) explain that the type of zero represented by a particular observation depends on the study objective. If the goal is to quantify the instantaneous location of a species, and the species is temporarily absent from the study site, then the recording would not be considered a false zero. On the other hand, if the goal is to determine which sites are
inhabited by a species, then an absence would constitute a false zero.

Thus, the probability of a zero count under the conditional specification, $\theta_c$, corresponds to the probability of either a true or false zero, whereas the probability $\theta$ under the mixture specification corresponds to the probability of a true zero. That is, structural zeros and random zeros are not distinguished under the conditional specification, whereas under the mixture model, we can examine the different sources of error (Kuhnert et al., 2005). Martin et al. (2005) suggest the use of Bayesian methods in order that information on the contribution of false zeros may be incorporated through an informative prior on the detection probability.


A multivariate version of the zero-inflated Poisson distribution has been developed by Li et al. (1999). These authors assume that most of the data arises from the perfect state. However, outside of the manufacturing context in which they work, this is not always a reasonable assumption. Without simplifying assumptions, such extensions to the multivariate situation require a large number of parameters.

Another approaches to modelling zero-inflated data include the use of a mixture of Poisson distributions leading to the Neyman Type A distribution (Dobbie and Welsh 2001). This distribution is particularly helpful for multi-modal data. These authors consider three parameterizations of the Neyman type A distribution and extend each to incorporate covariates. Unfortunately, choice of parameterization is important and model fitting is complicated by infinite sums. Further, this distribution is not a member of the exponential family so that the
associated advantages are not available. Finally, good initial parameter estimates are needed but these are often difficult to obtain.

Birth processes and threshold models are also used to model zero-inflated data. See Ridout et al. (1998) for a review of these models. Other approaches to zero-inflation include the spatial probit model for zero-inflated data (Rathbun and Fei 2006) and the zero-inflated modified power series distribution used by Gupta, Gupta, and Tripathi (1996). References related to the development of zero-inflated models for continuous data can be found in Martin et al. (2005).

Overdispersion and zero-inflated models

In practice, it has been found that even after modelling excess zeros, some overdispersion related to the counts may remain. In the standard Poisson regression setting, ignoring overdispersion relative to the Poisson model generally leads to consistent parameter estimates. However, with a truncated distribution, as used in hurdle models, ignoring overdispersion leads to inconsistent parameter estimates (Grogger and Carson 1991). In order to avoid such problems, Gurmu (1997) developed a semi-parametric hurdle model.

Ridout, Hinde and Demetrio (2001) also discuss the bias in parameter estimates when non-zero counts are overdispersed in relation to the Poisson distribution. In the context of mixture models, one solution is the use of a negative binomial, generalized Poisson or beta binomial distribution for $f$.

Zero-inflated negative binomial models have found a wide range of applications in the literature. Martin et al. (2005) compare ZIP and ZINB models of bird counts; Welsh et al. (1996) compare the use of a truncated Poisson and a truncated negative binomial distribution for modelling abundance of Leadbeater’s Possum in south-eastern Australia; and Nobtvedt et
al. (2002) use the ZINB model in a longitudinal study of gastrointestinal parasite burdens in Canadian dairy cows; while Simons, Neal and Gaher (2006) used the ZINB model to study problems associated with drug use among college students.

The generalized Poisson distribution can accommodate underdispersion as well as overdispersion in count data through the addition of an extra parameter. For \( \lambda_1 > 0, |\lambda_2| < 1 \), the generalized Poisson distribution (Consul and Jain, 1973) is defined by

\[
p_x(\lambda_1, \lambda_2) = \lambda_1 (\lambda_1 + x\lambda_2)^{x-1}e^{-(\lambda_1 + x\lambda_2)}/x!, \quad x = 0, 1, 2, \ldots
\]  

such that

\[
p_x(\lambda_1, \lambda_2) = 0 \quad \text{for} \quad x \geq m \quad \text{if} \quad \lambda_1 + m\lambda_2 \leq 0
\]  

zero-inflated versions of the generalized Poisson distribution (ZIGP) have been utilized in the frequentist setting by, for example, Gupta, Gupta and Tripathi (1996) and Famoye and Singh (2006), while Angers and Biswas (2003) approach the ZIGP from a Bayesian perspective. Czado, Erhardt and Min (2006) discuss an extension of zero-inflated generalized Poisson (ZIGP) regression models to allow for regression on the overdispersion and zero-inflation parameters.

Famoye and Singh (2006) note that, in practice, inadequacies of the ZIP model and computational concerns related to the zero-inflated negative binomial distribution are common. As noted by other authors, iterative techniques used to estimate the parameters of the ZINB model often fail to converge. Thus, ZIGP models provide a nice alternative to the ZINB model in these cases.

A.2 Computing
Bohning (1998) discusses using C.A.MAN software as a diagnostic device for ZIP models. Considering the ZIP model as a special case of a wide class of mixture models with an unknown number of components, one can use C.A.MAN to obtain maximum likelihood estimates for a non-parametric mixing distribution. On the other hand, Dietz and Boehning (2000) present an EM algorithm for maximum likelihood estimation of the zero-truncated Poisson model using standard software. This algorithm can also be used to estimate the parameters of a zero-deflated Poisson model and to calculate a likelihood ratio test of zero modification. Finally, the ‘zicounts’ package for R software can be used to fit regression models for zero-inflated count data. The package can fit a Poisson, negative binomial, zero-inflated Poisson or zero-inflated negative binomial model. Covariates can be used with both the zero and non-zero components.

Spatial models can also be formulated in a Bayesian framework. Freely available WinBUGS software (Speigelhalter, Best and Lunn 2003) with the associated GeoBUGS package, can be used to fit such models. Any Bayesian analysis must be carefully scrutinized. Careful consideration must be paid to parameter convergence. As well, WinBUGS routinely reports the MC errors which indicate technical errors in computation. It is suggested that the MC error divided by the standard deviation of the posterior distribution ought to be less than 5%.

A.3 Score Tests

Finally, there is a large body of literature on score tests. Dean (1992) develops a score test for overdispersion in Poisson and binomial models. Many authors have considered score tests in the ZIP context: van den Broek (1995), Deng and Paul (2005), Xiang, Lee, Yau, McLachlan (2006). Ridout et al. (2001) discuss score tests for comparing ZIP models and ZINB models,
References


