ZIP Exponential families allow **easier interpretation** than standard ZIP

Check out our **izipr** package!

Zero-inflated Poisson exponential families, with applications to time-series modelling of counts

Thomas Fung[1,]
@thomas.fung.dr
thomas.fung@mq.edu.au
Alan Huang[2,]
[1] *Department of Mathematics and Statistics, Macquarie University*
[2] *School of Mathematics and Physics, University of Queensland*

Introduction

In many count data processes, zero observations occur more frequently than expected from a nominal distribution. Perhaps the most well-known model for such scenarios is the zero-inflated Poisson (ZIP) of **Lambert (1992)**. ZIP can be constructed via two independent latent variables, namely, $B \sim$ Bernoulli with some probability π of being zero and $P \sim$ Poisson with some rate λ. One desirable feature of ZIP is that the latent Bernoulli construction offers an explicit explanation of the excess zeros. However, the mean of the observed response can only be identified to the product $(1-\pi)\lambda$, using classical ZIP. The goodness-of-fit of ZIP models depends crucially on the individual models for π and λ, but this can be not easy to check as neither process is fully observed.

ZIP exponential families

Let $f(y|\pi, \lambda)$ be the mass function of a classical ZIP family with parameter π & λ). Construct a family $\{f_\theta(y); \theta \in \mathbb{R}\}$ of distributions, indexed by θ, via exponential tilting:

$$f_\theta(y) \propto \exp(\theta y) f(y|\pi, \lambda)\ , \quad \theta \in \mathbb{R}.$$

Each $f_\theta(y)$ remains a ZIP distribution with new parameters π_θ and λ_θ given respectively by

$$\pi_\theta = \frac{\pi}{\pi + (1-\pi)e^{\lambda(e^\theta - 1)}} \quad \text{and} \quad \lambda_\theta = \lambda e^\theta \lambda_\theta = \lambda e^\theta.$$

For mathematical convenience, we set

$$\lambda \equiv 1, \quad \nu = \pi/(1-\pi) \quad \text{and} \quad \mu = E(Y_\theta) = (1-\pi_\theta)\lambda_\theta.$$

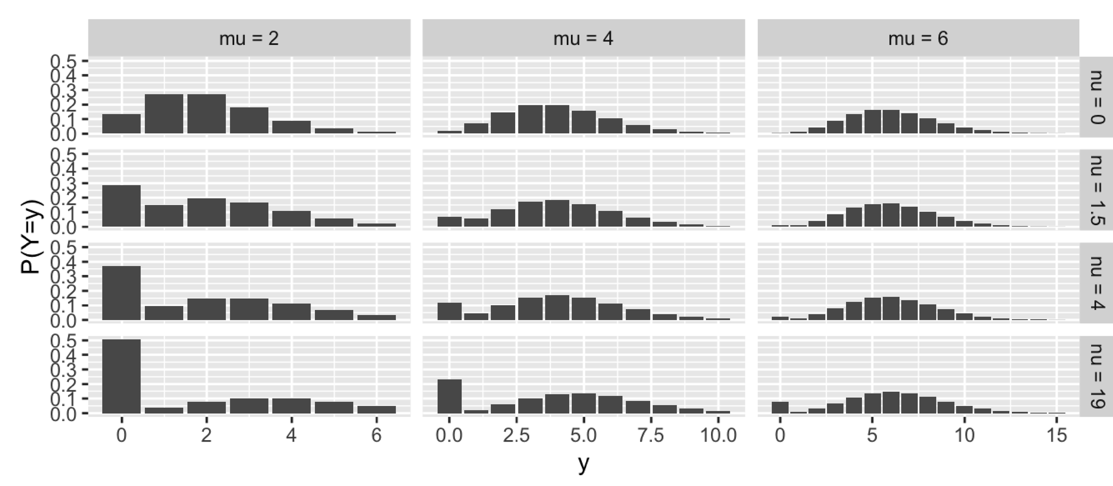
We then write the distribution as $\mathrm{ZIP}_\nu(\mu)$.

$\mathrm{ZIP}_\nu(\mu)$ allows us to construct simple, interpretable regression models via

$$Y|X \sim \mathrm{ZIP}_\nu(\mu(X^\top \beta))\ , \text{ for some } \nu \in [0, \infty)$$

where $E(Y|X) = \mu(X^\top \beta)$ for some mean function $\mu(\cdot)$. This is implemented, with model diagnostic tools, in `izipr` package of **Huang and Fung (2021)**.

Some pmf of $\mathrm{ZIP}_\nu(\mu)$



`bioChemist` dataset

The dataset contains the number of articles produced by 915 graduate students in biochemistry during the last three years of their PhD, along with some information on the graduates, such as gender, marital status, the number of kids under 5, how prestigious the department is and mentor's publication record over the same period.

We fitted the classical ZIP using the `zeroinfl()` in the `pscl` package as well as our own `glm.izip()` in the `izipr` package.

Table 1: Estimated coefficients, standard errors, AIC and BIC values for the 'bioChemist' dataset using the classical ZIP and ZIP$_\nu$ regression models

	ZIP				ZIP$_\nu$	
	Poisson component		Bernoulli component			
coefficients	**est.**	**se**	**est.**	**se**	**est.**	**se**
(Intercept)	0.641	0.121	-0.577	0.509	0.325	0.118
femWomen	-0.209	0.063	0.11	0.28	-0.229	0.063
marMarried	0.104	0.071	-0.354	0.318	0.159	0.071
kid5	-0.143	0.047	0.217	0.196	-0.19	0.046
phd	-0.006	0.031	0.001	0.145	0.01	0.03
ment	0.018	0.002	-0.134	0.045	0.025	0.002
ν	–	–	–	–	0.572	0.088
AIC		3230.0			3238.2	
BIC		3278.2			3267.1	

Interpretating the models

Suppose we want to interpret the effect of `kid5`, the number of kids under 5.

For the classical ZIP, interpretation requires two-steps. Each additional kid under 5 is associated with an increase in the log-odds of being in the subpopulation that *did not have the opportunity to produce a paper* of `0.217`, which translates to $\exp(0.217) = 1.242 \approx$ `24%` increase in odds.

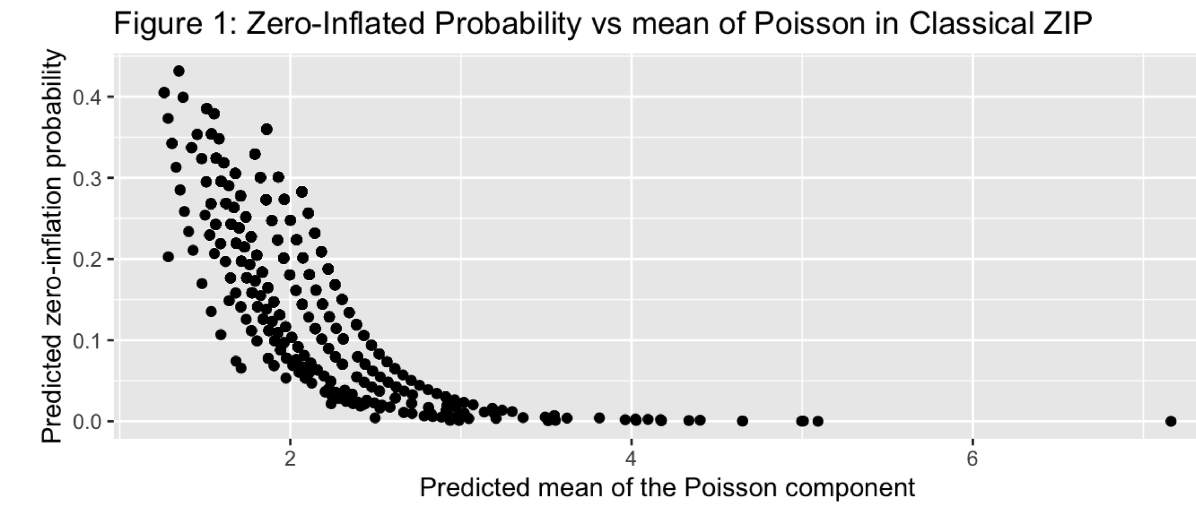
Given a graduate is in the other subpopulation that *have the opportunity to produce paper(s)* then each additional kid under 5 is associated with a decrease in the expected number of papers by a factor of exp(`-0.143`) = 0.87, i.e. `13%` decrease.

For the ZIP$_\nu$ model, model interpretation is similar to a log-linaer model. The effect of each additional kid under 5 is a multiplicative factor of exp(`-0.190`) = 0.82, i.e. `18%` decrease to the expected number of papers produced. This value has already been adjusted for zero-inflation.

Predicting the zero-inflation

Notice that all variables with a positive effect on the Poisson component of the classical ZIP model had a negative effect on the Bernoulli component. In other words, as the expected number of papers produced increases, the probability of being in the do not have opportunity to write a paper'' (i.e., zero-inflation) subpopulation tends to decrease, and vice versa.

Figure 1: Zero-Inflated Probability vs mean of Poisson in Classical ZIP



The strong negative relationship here provides a clear example of how constant zero-inflation can be unrealistic in practice. But this is the assumption used in some time series model for counts in the literature as in **Yau, Lee, and Carrivick (2004)**, **Zhu (2012)**, **Yang, Cavanaugh, and Zamba (2015)**,

ZIP$_\nu$ for count time-series

ZIP$_\nu$ distributions prove even more useful for modelling count time-series, as we only need to construct a single ARMA-type recursion for the conditional mean of the process, rather than two latent processes which are only partially observed.

An integer-valued generalized autoregressive conditionally heteroskedastic (INGARCH) time-series model of order (s, q) based on ZIP$_\nu$ distributions can be specified via

$$Y_t|\mathcal{F}_{t-1} \sim \mathrm{ZIP}_\nu(\mu_t)$$
$$\mu_t = \delta + \alpha_1\mu_{t-1} + \ldots \alpha_s\mu_{t-s} + \beta_1 Y_{t-1} + \ldots \beta_q Y_{t-q}$$

where $\delta, \alpha_1, \ldots, \alpha_s, \beta_1, \ldots, \beta_q > 0$. We call such processes ZIP$_\nu$-INGARCH(s, q). This is also implemented in the `izipr` package.

If you are interested in what the `izipr` package can do, please scan the QR-code below.

References

Huang, Alan, and Thomas Fung. 2021. *Izipr: Interpretable Zero-Inflated Poisson (and Related) Models in r.* URL: https://CRAN.R-project.org/package=izipr.
Lambert, Diane. 1992. "Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing." *Technometrics* 34 (1): 1. https://doi.org/10.2307/1269547.
Yang, Ming, Joseph E Cavanaugh, and Gideon KD Zamba. 2015. "State-Space Models for Count Time Series with Excess Zeros." *Statistical Modelling* 15 (1): 70–90. https://doi.org/10.1177/1471082X14535530.
Yau, Kelvin K. W., Andy H. Lee, and Philip J. W. Carrivick. 2004. "Modeling Zero-Inflated Count Series with Application to Occupational Health." *Computer Methods and Programs in Biomedicine* 74 (1): 47–52. https://doi.org/10.1016/S0169-2607(03)00070-1.
Zhu, Fukang. 2012. "Zero-Inflated Poisson and Negative Binomial Integer-Valued GARCH Models." *Journal of Statistical Planning and Inference* 142 (4): 826–39. https://doi.org/10.1016/j.jspi.2011.10.002.

MACQUARIE University