BAYESIAN FORECASTING OF MULTIVARIATE LONGITUDINAL ZERO-INFLATED COUNTS: AN APPLICATION TO CIVIL CONFLICT

Beatrice Franzolini[1], Laura Bondi[2], Augusto Fasano[3], and Giovanni Rebaudo[4]

[1] Singapore Institute for Clinical Sciences (SICS), Agency for Science, Technology and Research (A*STAR), Singapore (e-mail: franzolini@pm.me)

[2] Medical Research Council (MRC) Biostatistics Unit (BSU), Cambridge University, United Kingdom (e-mail: laura.bondi@mrc-bsu.cam.ac.uk)

[3] Collegio Carlo Alberto, Italy (e-mail: augusto.fasano@carloalberto.org)

[4] Department of Economics, Social Studies, Applied Mathematics and Statistics (ESOMAS), University of Turin, Italy (e-mail: giovanni.rebaudo@unito.it)

ABSTRACT: Forecasting multiple dependent zero-inflated count processes is a problem encountered in many statistical applications. Standard parametric approaches typically rely on independence assumptions that fail to capture dependence structures. Here a Bayesian nonparametric approach is proposed to overcome this problem and showcased on a real dataset of civil conflicts in Asia. The forecasting model is obtained by generalizing the clustering methods proposed in Franzolini *et al.* (2023).

KEYWORDS: clustering, enriched Dirichlet, excess of zeros, mixtures of finite mixtures, rare events

1 Introduction

In statistical applications involving count data, it is common to encounter datasets showcasing a large number of zeros. Analyzing zero-inflated data requires statistical models that extend beyond standard count distributions, such as Binomial, Poisson, or Negative Binomial. Adding to the likelihood function a parameter specifically controlling the probability of observing a zero count is a popular strategy (Mullahy, 1986; Lambert, 1992), but this approach still relies on strong parametric assumptions regarding positive counts and is difficult to extend to multivariate count data: it requires a large number of parameters to avoid simplistic independence assumptions between multiple processes. Furthermore, when predicting future outcomes, the likelihood function is complicated by covariate values or autoregressive components, adding

to the complexity of the multivariate distribution of many zero-inflated processes. One flexible, yet parsimonious, solution has been recently proposed by Franzolini *et al.* (2023). They model joint probabilities of zero-inflation using a Bayesian enriched mixture of finite mixtures, obtained by combining the works of Wade *et al.* (2011) and Argiento & De Iorio (2022). The strength of the method relies on the fact that, within each mixture component, different processes are modeled with an independent kernel and the dependence across multiple count processes is captured by the underlying clustering structure. Thanks to the prior on the number of components of the mixture, the model automatically adjusts its complexity (measured by the number of parameters to be estimated) based on the data, ultimately requiring fewer parameters than traditional multivariate approaches when the data suggest so. Lastly, the method provides an additional interesting inferential outcome, i.e., a two-level clustering of subjects, based on the patterns of zero/non-zero counts (outer clustering) and values of positive counts (inner clustering). In Franzolini *et al.* (2023), the inferential goal is to detect groups of subjects with different count patterns and the data are cross-sectional. In this work, we extend their approach including in the model subject/time-specific covariates, autoregressive components, and random effects, aiming at predicting multiple longitudinal zero-inflated outcomes. We name the resulting model zero-inflated enriched mixture (ZIEM) regression.

2 ZIEM regression

The ZIEM regression is presented for a bivariate count process $(X_{i,t}, Y_{i,t})$, with multivariate predictors $Z_{i,t}$, where i and t denote subjects and time, respectively. The zero/non-zero components of the responses, i.e., $\tilde{X}_{i,t} = \mathbb{1}(X_{i,t} > 0)$ and $\tilde{Y}_{i,t} = \mathbb{1}(Y_{i,t} > 0)$, are modeled through a finite mixture model with bivariate kernel where mixture components are defined by the parameters of a logit regression with an autoregressive component, i.e.,

$$
\begin{aligned}
(\tilde{X}_{i,t}, \tilde{Y}_{i,t}) \mid p_{i,t}, q_{i,t} &\overset{ind}{\sim} \text{Bern}(p_{i,t}) \times \text{Bern}(q_{i,t}) \\
\text{logit}(p_{i,t}) &= \alpha_i^{(x)} + \beta_i^{(x)} X_{i,t-1} + (\eta^{(x)})^T Z_{i,t} \qquad (1) \\
\text{logit}(q_{i,t}) &= \alpha_i^{(y)} + \beta_i^{(y)} Y_{i,t-1} + (\eta^{(y)})^T Z_{i,t} \\
(\alpha_i^{(x)}, \alpha_i^{(y)}, \beta_i^{(x)}, \beta_i^{(y)}) \mid M_0, w, \theta &\overset{iid}{\sim} \sum_{m=1}^{M_0} w_m \delta_{\theta_m} \qquad (\eta^{(x)}, \eta^{(y)}) \sim \mathcal{N}(0, I) \\
\theta_m \mid M_0 &\overset{iid}{\sim} \mathcal{N}(0, I) \quad w \mid M_0 \sim \text{Dirichlet}_{M_0}(\gamma_0, \ldots, \gamma_0) \quad M_0 \sim \text{Poi}_0(\lambda_0)
\end{aligned}
$$

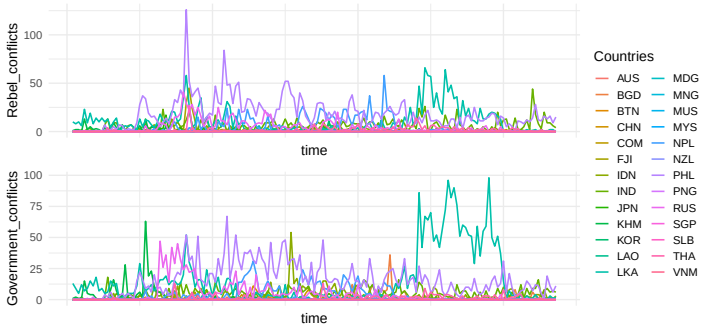

Figure 1. *Civil conflict data: data are part of a Defense Advanced Research Project Agency (DARPA) funded project which has created a dataset of over 2 million machine-coded daily events occurring between actors within the Asia-Pacific region.*

where $\theta = (\theta_1, \ldots, \theta_{M_0})$, with $\theta_m \in \mathbb{R}^4$, and Poi_0 denotes a shifted Poisson distribution on $\{1, 2, \ldots, \}$. The model (1) induces an outer clustering structure of the subjects denoted by i. Then, within each outer cluster m and independently across outer clusters, the positive component of the responses is modeled through a finite mixture model with bivariate Poisson kernel, i.e.,

$$\begin{aligned} (X_{i,t}, Y_{i,t}) \mid (X_{i,t} > 0), (Y_{i,t} > 0), \mu_i, \nu_i &\overset{ind}{\sim} \text{Poi}_0(\mu_i) \times \text{Poi}_0(\nu_i) \\ (\mu_i, \nu_i) \mid M_m, q_m, \xi_m &\overset{iid}{\sim} \sum_{m=1}^{M_m} q_{m,s} \delta_{\xi_{m,s}} \qquad (2) \\ q_m \mid M_m \sim \text{Dirichlet}_{M_m}(\gamma, \ldots, \gamma), \;\; \xi_{m,s} \mid M_m &\overset{iid}{\sim} Q_0, \;\; M_m \sim \text{Poi}_0(\lambda) \end{aligned}$$

where $\xi_m = (\xi_{m,1}, \ldots, \xi_{m,M_m})$, with $\xi_{m,s} \in (\mathbb{R}^+)^2$ and Q_0 is a bivariate Log-normal distribution with independent components. The extension to processes with dimension $d > 2$ is straightforward.

3 An application to civil conflict

We test the out-of-sample predictive performance of our model on a monthly bi-variate dataset concerning domestic civil conflicts from 1997 to 2010 in $n = 26$ countries in Asia. The observed responses are plotted in Figure 1. For a detailed description of the dataset, we refer to Bagozzi (2015). Monthly data from 1997 to 2009 are used to train our model (ZIEM regression), a zero-inflated Poisson (ZIP) regression, and a zero-inflated Negative Binomial re-

gression (ZINB) regression. ZIP and ZINB regressions are estimated with the `R` package `pscl` (Zeileis *et al.*, 2008). Data from the year 2010 are used to evaluate the prediction performance. All three models include an autoregressive component and three covariates (i.e., log-GDP per capita, GDP growth, log-population), which are used to predict the occurrence of a non-zero count. Table 1 summarizes the predictive performance of the three models, based on which we conclude that ZIEM regression outperforms the competitors.

Table 1. *Out-of-sample predictive performance: root mean squared error (RMSE), normalized root mean squared error (NRMSE), maximum squared error (*$\max\{\hat{e}^2\}$*), and squared error (*e^2*) distributions' quantiles. Bold values denote the best performance.*

				x s.t. $\hat{\Pr}(e^2 > x) = p$		
Model	RMSE	NRMSE	$\max\{\hat{e}^2\}$	p=0.15	p=0.25	p=0.50
ZIEM reg.	**6.72**	**0.1527**	**26.07**	**7.44**	**4.77**	**1.18**
ZIP reg.	7.52	0.1708	35.74	8.22	8.00	1.72
ZINB reg.	8.07	0.1834	36.90	8.11	7.10	5.26

References

ARGIENTO, R., & DE IORIO, M. 2022. Is infinity that far? A Bayesian nonparametric perspective of finite mixture models. *The Annals of Statistics*, **50**, 2641–2663.

BAGOZZI, B. E. 2015. Forecasting civil conflict with zero-inflated count models. *Civil Wars*, **17**, 1–24.

FRANZOLINI, B., CREMASCHI, A., VAN DEN BOOM, W., & DE IORIO, M. 2023. Bayesian clustering of multiple zero-inflated outcomes. *Philosophical Transactions of the Royal Society A*, **381**, 20220145.

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

MULLAHY, J. 1986. Specification and testing of some modified count data models. *Journal of Econometrics*, **33**, 341–365.

WADE, S., MONGELLUZZO, S., & PETRONE, S. 2011. An enriched conjugate prior for Bayesian nonparametric inference. *Bayesian Analysis*, **6**, 359–385.

ZEILEIS, A., KLEIBER, C., & JACKMAN, S. 2008. Regression models for count data in R. *Journal of Statistical Software*, **27**, 1–25.