Seminar in Statistics
Synopsis 1: *Markov regression models for count time series with excess zeros: A partial likelihood approach*

Anders Gantzhorn Kristensen - tzk942

13th October 2023

Paper Setup

In the paper the zero-inflated poisson (ZIP) regression model in the context of time series is introduced as a means of modelling count time series data that exhibit an excess of zeros. Initially, the authors: Ming Yang, Gideon K.D. Zamba, and Joseph E. Cavanaugh motivate the model in a small introduction. Hereafter, they introduce and extend the regular ZIP regression model to a Markov regression model. They write up a partial likelihood for the ZIP Markov regression model and prove by means of the martingale central limit theorem consistency and asymptotic normality of the Maximum Partial Likelihood Estimator (MPLE). The paper continues by mentioning a couple of known issues with maximizing the partial likelihood directly with for instance Newton-Raphson. Instead the authors advocate for the use of the EM-algorithm in order to obtain estimates for statistical inference. They show how one can fit the problem of maximizing the partial likelihood into the framework of the EM-algorithm while extensively explaining both the E- and the M-step. Afterwards, they return to some asymptotic results and derive the actual covariance matrix of the asymptotic distribution of our estimator. Then, they briefly discuss how one can do model selection with Akaike's information criterion (AIC) and Takeuchi's information criterion (TIC) and compare the two means of model selection, when the model is misspecified. That is they mention that TIC reduces to AIC, when a model is correctly specified, but is more robust to misspecifications. This as well as the finite sample behaviour of the MLPE is illustrated in a simulation study. Finally, they showcase the methodology in real data analysis by using the ZIP markov regression model for forecasting syphilis in Maryland, USA and conclude the paper in a brief summary.

Main Results

Theoretical results

The authors successfully introduce an auto-regressive model for count time series following the ZIP distribution, which can be viewed as an extension of the Poisson auto regressive model. That is each observation in some discrete count data $\{Y_t\}_{t=1}^N$ has the p.m.f.

$$f_{Y_t}\left(y_t|\mathcal{F}_{t-1};\theta\right) = \omega_t I_{(y_t=0)} + (1-\omega_t)\frac{\exp\left(-\lambda_t\right)\lambda_t^{y_t}}{y_t!}. \tag{1}$$

As the time series is assumed Markov, the filtration $\mathcal{F}_{t-1}$ only contains the information at time $t-1$. More specifically, the paper proposes the auto-regressive model where the dependence is modelled through the parameters λ_t and ω_t called the Poisson-intensity and zero-inflation parameter respectively in the following way

$$\log(\lambda_t) = \eta_t = x_{t-1}^{\top}\beta \tag{2}$$

$$\log\left(\frac{\omega_t}{1-\omega_t}\right) = \xi_t = z_{t-1}^\top \gamma \tag{3}$$

where $\beta \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}^q$ are unknown vectors of parameters to be estimated, and x_{t-1} and z_{t-1} denote vectors of past explanatory covariates. Now, defining $\theta = \left(\beta^\top, \gamma^\top\right)^\top$ the paper arrives at the partial log-likelihood given samples $\{Y_t\}_{t=1}^N$

$$\log \mathrm{PL}\left(\theta; \mathbf{y}\right) = \sum_{t=1}^N \log\left(\omega_t y_{0,t} + (1-\omega_t)\frac{\exp\left(-\lambda_t\right)\lambda_t^{y_t}}{y_t!}\right). \tag{4}$$

Naturally, $\hat{\theta}_{MPLE} = \mathrm{argmax}_{\theta\in\Theta} \log \mathrm{PL}\left(\theta; \mathbf{y}\right)$
The paper establishes the large sample theory under mild regularity conditions (C1-C3 under asymptotic results) for $\hat{\theta}_{MPLE}$. [2, p.31]; it is shown that $\hat{\theta}_{MPLE}$ is consistent and asymptotically normal with some covariance matrix that depends on the data as well as θ. As stated earlier the minimization is, due to instabilities of the other algorithms such as the Newton-raphson algortihm, done with the EM-algorithm.

Simulation study

The authors present a simulation study with data generated with structure in 2 and 3 as

$$\eta_t = \beta_0 + \beta_1 I_{(y_{t-1}>0)} + \sigma z_t, \qquad \xi_t = \gamma_0 + \gamma_1 I_{(y_t t-1>0)}, \tag{5}$$

with z_t some unobserved noise, $z_t \sim \mathcal{N}(0,1)$. Initially, $\sigma = 0$ i.e. the case for which the methodology has been developed. Here the authors find small biases, asymptotic standard error and empirical standard deviation. Also, they find that confidence interval of 95% appears to have proper coverage. This was true for sample sizes N of $100, 200$ and 500. Although, the coverage of the CI seems to be the same for all sample sizes, the bias, ASE and ESD generally seem to decrease with more samples.[2, p.33, table 1] Finally, they explore the misspecified case, i.e. when $\sigma \neq 0$, then doing model selection, where the true model is 5 amongst candidate models

$$\eta_t = \beta_0 + \sum_{i=1}^{k_1} \beta_i I_{(y_{t-i}>0)}, \qquad \xi_t = \gamma_0 + \sum_{i=1}^{k_2} \gamma_i I_{(y_t t-i>0)}, \tag{6}$$

reveals that for fixed sample size TIC more often than AIC select the correct model (in this case $(k_1, k_2) = (1, 1)$). [2, p.34, table 2]

Own Contribution

We did not manage to implement the zero-inflated Poisson model for time-series ourselves. However, using the ZIM-package in R [1] we can still do a small simulation study to illustrate some key points with regards to estimation of the poisson-intesity parameter, λ_t in the ZIP regression model. These issues extend to the ZIP Markov regression model due how it extends the regular ZIP regression model. We sample $N = (50, 150, 200, 500, 1000, 1500, 2000)$

values from a regular ZIP model with fixed, $\lambda_t = 5$ and $\omega_t = (0.1, 0.25, 0.5, 0.75, 0.8, 0.85)$. This is done $M = 500$ times for each combination, where we also use the EM-algorithm in the ZIM-package to estimate ω. We plot the mean of the bias of λ_t for each combination of ω_t and N along with a ribbon showing each of the 500 runs' 95% smallest biases.

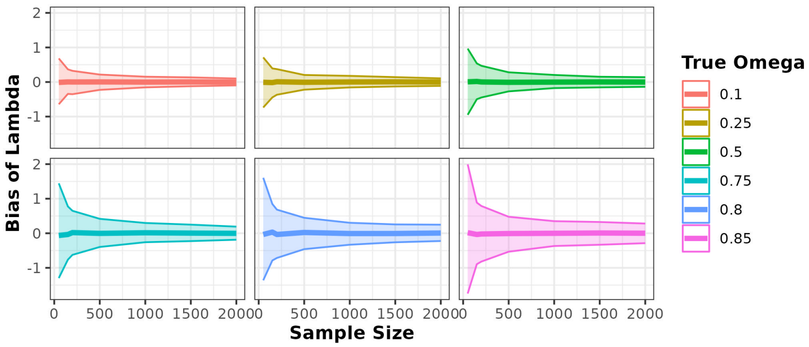

Figure 1: Result of simulation study

As is clear from figure 1 it becomes increasingly more difficult to estimate λ_t for larger values of ω_t, this is of course due to the fact that we more seldomly draw from the Poisson part of the mixture 1. Although improved for larger samples, the issue is not completely remedied. Still, we see that we on average have close to 0 bias in all cases.

Strengths & Weaknesses

We finalize this synopsis with a brief discussion about the paper's strengths and weaknesses. The paper proposes methodology for handling count time series data with excess zeros or overdispersion, integer-valued time-series data is quite common real-life, so the model the paper propose could be applicable in various scenarios. The paper is overall structured and is quite readable. Furthermore, practitioners will most likely find the tools offered by this paper easy to deploy in practical applications. In addition, the paper proposes methods of doing model selection even with misspecified models as well as showcasing how the ZIP auto-regressive model can outperforms the regular poisson auto-regressive model. On the other hand, the EM-algorithm proposed by the paper does not guarantee convergence to the MPLE. Thus, the theoretical results from the paper might not hold in ill-behaved cases. Also, as they mention the EM-algorithm is comparatively slower than Newton's method. Although according to the paper *"One can increase the speed of convergence by implementing an automatic switch to Newton–Raphson once the EM iterates begin to stabilize"* [2, p.31]. How this is done in practise, however, is not addressed. Here it is also worth mentioning that although the simulation study is great. It does not show the reader the limitations of the ZIP markov regression model. This makes it difficult for us to know, when we should not deploy this model.

References

[1] Ming Yang, Gideon Zamba, and Joseph Cavanaugh. *ZIM: Zero-Inflated Models (ZIM) for Count Time Series with Excess Zeros*, 2018. R package version 1.1.0.

[2] Ming Yang, Gideon K.D. Zamba, and Joseph E. Cavanaugh. Markov regression models for count time series with excess zeros: A partial likelihood approach. *Statistical Methodology*, 14:26–38, 2013.