Hidden Markov models for time series of counts with excess zeros

Madalina Olteanu, James Ridgway

Hidden Markov models for time series of counts with excess zeros

Madalina Olteanu and James Ridgway *

University Paris 1 Pantheon-Sorbonne - SAMM, EA4543
90 Rue de Tolbiac, 75013 Paris - France

Abstract. Integer-valued time series are often modeled with Markov models or hidden Markov models (HMM). However, when the series represents count data it is often subject to excess zeros. In this case, usual distributions such as binomial or Poisson are unable to estimate the zero mass correctly. In order to overcome this issue, we introduce zero-inflated distributions in the hidden Markov model. The empirical results on simulated and real data show good convergence properties, while excess zeros are better estimated than with classical HMM.

1 Introduction

Time series of count data occur quite often and in various fields such as history, economy or biology. During the past fifty years, several models were proposed in order to deal with integer-valued time series, although preferred methods have not yet been established [2]. One solution for dealing with time series of count data are hidden Markov models (HMM hereafter). Originally introduced for speech recognition [1], they are especially interesting in the context of the presumed existence of several regimes controlling the parameters of the model (coding vs. non-coding regions for DNA data, crisis vs. stable periods for financial data,...). However, count data are often subject to excess zeros, while the available software is generally implemented for usual distributions such as binomial or Poisson. For example, to our knowledge there is no available R-package proposing to model count data subject to zero-overdispersion with HMM.

This paper introduces zero-inflated Poisson distributions (ZIP hereafter) in the HMM models. Introduced in the late 60's [3], ZIP distributions allow for excess zeros according to the following definition:

$$\mathbb{P}(X=x)=\begin{cases} \omega+(1-\omega)\exp(-\lambda) & , \quad x=0 \\ (1-\omega)\frac{\exp(-\lambda)\lambda^x}{x!} & , \quad x>0 \end{cases}$$

where $\omega \in]0,1[$ and $\lambda > 0$. Thus, a ZIP distribution is written as a mixture between a Dirac in zero and a Poisson. ZIP distributions allow for a certain amount of dispersion in the data[9], in the sense that in the Poisson distribution the mean and the variance are both equal to λ, whereas for $Y \sim ZIP(\omega,\lambda)$:

$$\mathbb{E}(Y)=(1-\omega)\lambda=\mu \ , \ \mathbb{V}(Y)=\mu+\left(\frac{\omega}{1-\omega}\right)\mu^2$$

*James Ridgway is currently 3rd year ENSAE student. His internship at SAMM was funded by the University Paris 1 through the "Analyse" project (http://analyseshs.hypotheses.org/). The authors are acknowledging Julien Alerini for his useful comments and support.

This can be an interesting aspect of the distribution since it allows more flexibility in the modelisation.

ZIP distributions were mainly used for regression purposes ([8], [7]), but one can extend their use in the context of hidden Markov models. Let us mention that the idea of mixing ZIP and HMM is not completely new. Recently, [5] and [10] proposed close versions of the same model by introducing a partially observed Markov chain in which the two components in the zero-inflated Poisson occur according to the states of the partially hidden process. Both articles use the model in a regression context and suppose the existence of covariates. We introduce a modified ZIP-HMM model, in which both the Poisson parameter λ and the mixing weight ω depend on the hidden states.

The rest of the paper is organized as follows : section 2 is devoted to a detailed presentation of the model and the associated EM algorithm. Section 3 presents some results on simulated and real-life data, while section 4 concludes the paper.

2 Hidden Markov models with zero-inflated Poisson distribution (ZIP-HMM)

As we mentioned earlier, we are interested in modelling switching time-series for which each regime is composed of an important number of zeroes. Let X_t be the observed series and S_t a homogeneous, unobserved Markov chain, with state-space $E = \{e_1, \cdots, e_q\}$. We suppose that, conditionally to $S_t = e_i$, X_t is distributed according to a ZIP of parameters (ω_i, λ_i). The ZIP-HMM model is estimated through the EM algorithm, [4]. Let us define the parameter space:

$$\Theta = \{\theta = (\omega, \pi, \lambda) \in]0,1[^q \times]0,1[^{q^2} \times (\mathbb{R}^+)^q, \forall j \in \{1 \cdots q\} \sum_{i=1}^{q} \pi_{ji} = 1\}$$

In order to write the complete likelihood, we introduce an auxiliary variable. Let us define Z_t an underlying random process such that $Z_t = 1$ leads to a structural zero, with $Z_t | S_t = e_i \sim \mathcal{B}er(\omega_i)$. By structural zero we mean a zero induced by the Dirac rather than the Poisson. Then, the complete likelihood is given by:

$$L(Z, X, S; \theta) = \prod_{t=1}^{T} \prod_{i=1}^{q} f(X_t, Z_t | S_t; \theta) \prod_{t=1}^{T-1} \prod_{i,j=1}^{q} \pi_{ij}^{\mathbf{1}_{e_i,e_j}(S_t, S_{t+1})} \times C$$

where $f(X_t, Z_t | S_t; \theta) = \omega_i^{\mathbf{1}_{(1,e_i)}(Z_t, X_t)} (1-\omega_i)^{\mathbf{1}_{(0,e_i)}(Z_t, X_t)} \left(\frac{e^{-\lambda_i} \lambda_i^{X_t}}{X_t!} \right)^{\mathbf{1}_{(0,e_i)}(Z_t, X_t)}$

The constant stands for the initial probability and does not intervene in the algorithm. The algorithm consists in maximizing $\mathbb{E}_{\theta^\star} \left[\ln(L(Z, X, S; \theta) | X_1^T \right]$ with respect to θ and updating $\theta^\star$ at each step.

The expectation step is given by:

$$\begin{aligned} Q(\theta|\theta^\star) = & \quad \mathbb{E}_{\theta^\star}\left[\ln(L(Z,X,S;\theta))|X_1^T\right] \\ = & \quad \sum_{t=1}^T \sum_{i=1}^q \{\mathbb{P}_{\theta^\star}(S_t = e_i, Z_t = 1|X_1^T)\ln(\omega_i) \\ + & \quad \mathbb{P}_{\theta^\star}(S_t = e_i, Z_t = 0|X_1^T)(\ln(1-\omega_i) - \lambda_i + X_t\ln(\lambda_i) - \ln(X_t!))\} \\ + & \quad \sum_{t=1}^T \sum_{i,j=1}^q \mathbb{P}_{\theta^\star}(S_{t-1} = e_i, S_t = e_j|X_1^T)\ln(\pi_{ij}) \end{aligned} \tag{1}$$

We take further interest in the part of (1) containing the parameters associated to the ZIP; we denote it ν_θ (the rest of the equation will be dealt with separately). This equation contains two joint probabilities that can be expressed as:

$$\mathbb{P}_{\theta^\star}(S_t = e_i, Z_t = 1|X_1^T) = \mathbb{P}_{\theta^\star}(Z_t = 1|X_1^T, S_t = e_i)\mathbb{P}_{\theta^\star}(S_t = e_i|X_1^T) \tag{2}$$

The latter probability of the right hand-side of (2) is obtained by the Baum-Welch forward-backward algorithm, the first is given by:

$$\mathbb{P}_{\theta^\star}(Z_t = 1|X_1^T, S_t = e_i) = \begin{cases} 0 \text{ if } X_t > 0 \\ \frac{\mathbb{P}_{\theta^\star}(X_t=0|S_t=e_i, Z_t=1)\mathbb{P}_{\theta^\star}(Z_t=1|S_t=e_i)}{\mathbb{P}_{\theta^\star}(X_t=0|S_t=e_i)} \text{ if } X_t = 0 \end{cases}$$

By denoting $\alpha_i^\star := \frac{\omega_i^\star}{\omega_i^\star + (1-\omega_i^\star)e^{\lambda_i^\star}}$, we get:

$$\mathbb{P}_{\theta^\star}(S_t = e_i, Z_t = 1|X_1^T) = \begin{cases} 0 \text{ if } X_t > 0 \\ \alpha_i^\star \mathbb{P}_{\theta^\star}(S_t = e_i|X_1^T) \text{ if } X_t = 0 \end{cases}$$

$$\mathbb{P}_{\theta^\star}(S_t = e_i, Z_t = 0|X_1^T) = \begin{cases} \mathbb{P}_{\theta^\star}(S_t = e_i|X_1^T) \text{ if } X_t > 0 \\ (1-\alpha_i^\star)\mathbb{P}_{\theta^\star}(S_t = e_i|X_1^T) \text{ if } X_t = 0 \end{cases}$$

We can therefore express ν_θ:

$$\begin{aligned} \nu_\theta = & \sum_{t:X_t>0} \sum_{i=0}^q p_t(e_i)\{\ln(1-\omega_i) - \lambda_i + X_t\ln(\lambda_i) - \ln(X_t!)\} \\ & + \sum_{t:X_t=0} \sum_{i=0}^q \alpha_i^\star \ln(\omega_i) + (1-\alpha_i^\star)p_t(e_i)\{\ln(1-\omega_i) - \lambda_i\} \end{aligned} \tag{3}$$

The maximization step can be carried analytically. The fact that the updates may be computed entirely analytically leads to a very fast algorithm. The transition probabilities updates are computed using the forward-backward algorithm. For the ZIP parameters, we obtain the following updates :

$$\pi_{ij} = \frac{\sum_{t=2}^T \mathbb{P}_{\theta^\star}(S_{t-1} = e_i, S_t = e_j|X_1^T)}{\sum_{t=1}^T \mathbb{P}_{\theta^\star}(S_t = e_i|X_1^T)}$$

$$\omega_i = \frac{\alpha_i^\star \sum_{X_t=0} \mathbb{P}_{\theta^\star}(S_t = e_i|X_1^T)}{\sum_{t:t=1}^T \mathbb{P}_{\theta^\star}(S_t = e_i|X_1^T)}$$

$$\lambda_i = \frac{\sum_{t:X_t>0} \mathbb{P}_{\theta^\star}(S_t = e_i|X_1^T) * X_t}{\sum_{t:X_t=0} \mathbb{P}_{\theta^\star}(S_t = e_i|X_1^T)(1-\alpha_i^\star) + \sum_{t:X_t>0} \mathbb{P}_{\theta^\star}(S_t = e_i|X_1^T)}$$

3 Examples

3.1 Simulations

We shall evaluate the quality of the estimates by using the mean squared error (MSE) on simulated data. The simulations were repeated on different sample sizes ranging from 500 to 10000 and on different parameter values. For each parameter configuration, 10.000 samples were simulated. In Table 1 all parameters except a transition probability were kept constant. In Table 2 all parameters except a Poisson parameter were kept constant. The results show relatively low MSE, decreasing with the sample size.

N \ π_{11}	0.1	0.3	0.4	0.5	0.6	0.8	0.9
500	0.0247	0.0286	0.0317	0.035	0.0415	0.0541	0.055
1000	0.0021	0.0054	0.0084	0.0131	0.0081	0.021	0.026
5000	0.0003	0.0015	0.0026	0.0058	0.011	0.0019	0.0008
10000	0.0001	0.0008	0.0018	0.005	0.0105	0.0012	0.0008

Table 1: MSE ($\pi_{22} = 0.6, \omega_1 = 0.2, \lambda_1 = 0.5, \omega_2 = 0.2, \lambda_2 = 3$)

N \ λ_1	0.1	0.5	1	5	10	14
500	0.0498	0.0417	0.0732	0.0199	0.0028	0.0397
1000	0.0085	0.0190	0.032	0.0103	0.0154	0.028
5000	0.0133	0.0193	0.0036	0.0018	0.003	0.0039
10000	0.00195	0.0094	0.001	0.001	0.001	0.0019

Table 2: MSE ($\pi_{11} = 0.4, \pi_{22} = 0.6, \omega_1 = 0.2, \omega_2 = 0.2, \lambda_2 = 3$)

3.2 Real data application

Next, the ZIP-HMM model was applied to a real-life data set. The observations come from the recordings of people flow out of a building on the UCI campus over 15 weeks, 48 time slices per day, [6]. Two models were estimated and compared: HMM with Poisson distribution and ZIP-HMM. The results are quite similar and both models are able to separate the activity periods during the week days and the inactivity periods during night-time and weekends. However, when closely analyzing the results, the ZIP-HMM model seems fit the data better. The estimated transition matrices of the two models are very close :

$$\pi^{HMM} = \begin{pmatrix} 0.972 & 0.028 \\ 0.071 & 0.929 \end{pmatrix}, \ \pi^{ZIP-HMM} = \begin{pmatrix} 0.975 & 0.025 \\ 0.072 & 0.928 \end{pmatrix}$$

The Poisson parameters estimated by the HMM are $\lambda_1^{HMM} = 0.40$ for the first regime, corresponding to the inactivity periods, and $\lambda_2^{HMM} = 12.58$ for the

second regime, corresponding to the activity periods. The parameters estimated by the ZIP-HMM are $\omega_1^{ZIP-HMM} = 0.70$, $\lambda_1^{ZIP-HMM} = 1.92$ for the first regime and $\omega_2^{ZIP-HMM} = 0$, $\lambda_2^{ZIP-HMM} = 13.53$ for the second regime. The second model seems more flexible, especially for modelling the inactivity periods with an important zero overdispersion. In Figure 1 we represent the histograms and the estimated distributions conditionally to the regimes, for both models. Both models select the same periods of inactivity (time series valued 0 or 1). ZIP-HMM is however more flexible and selects more values translating low activity in the building. Moreover, we clearly see that a Poisson distribution alone is not sufficient to estimate the zeros in the first regime. Although the Poisson parameter estimated by the HMM model for the first regime is very low, the zeros are underestimated, while the 1's are overestimated.

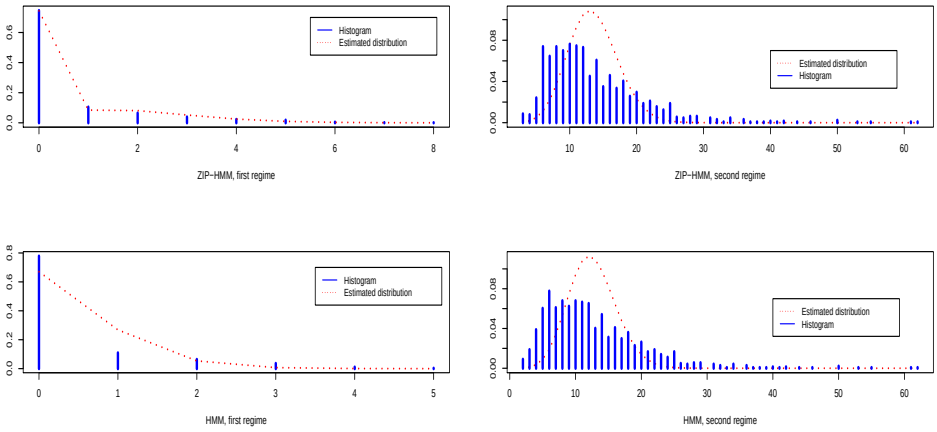

Fig. 1: Conditional distributions

In Figure 2, the conditional probabilities of being in the second regime, corresponding to the activity periods in the building, are estimated with the ZIP-HMM model. In order to facilitate visualization, 20% of the data was represented. The second regime identifies the office hours during week days, but also specific events that took place either in the evening (the peak after the third day) or during weekend (the peak before the second week).

4 Conclusion and future work

We proposed to deal with excess zeros in time series of count data by introducing a zero-inflated Poisson distribution in a hidden Markov model. The estimation was done through the EM algorithm. The method was implemented in R and is available as an R-package. Simulations illustrated the convergence properties, while the real-life data example showed that ZIP-HMM performs better than

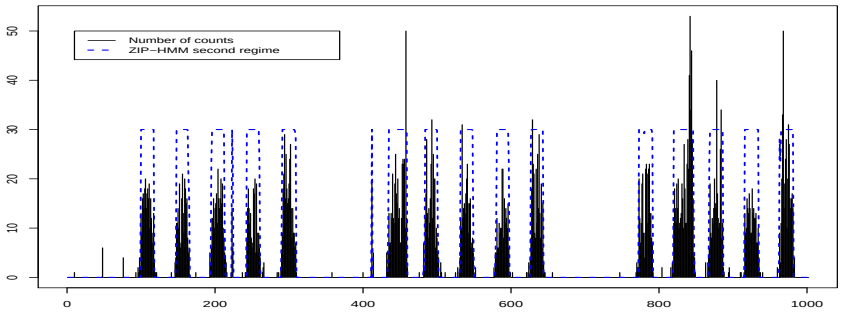

Fig. 2: Zoom on the results for the ZIP-HMM

HMM when there is a strong overdispersion in zero. However, both models had poorer results when estimating the activity periods and more specifically the peaks of activity during office hours. ZIP-HMM managed to identify specific evens but only in the case where they took place in the evening or during the weekend. In order to adress this issue, we consider to further develop the method by introducing time lags in the model.

References

[1] L.E. Baum and T. Petrie, Statistical inference for probabilistic functions of finite state Markov chains, *The Annals of Mathematical Statistics*, 37:1554-1563, 1966

[2] A.C. Cameron and P.K. Trivedi, *Microeconomics: Methods and Applications*, Cambridge University Press, 2005

[3] A.C. Cohen, Estimation in mixtures of discrete distributions, *Proc. Int. Symp. on Classical and Contagious Discrete Distributions*, Montreal, Pergamon Press, p.373-378, 1963

[4] A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society (B)*, 39:1-38, 1977

[5] S.M. DeSantis and D. Bandyopadhyay, Hidden Markov-models for zero-inflated Poisson counts with an application to substance use, *Statistics in Medicine*, 30:1678-1694, 2011

[6] A. Frank and A. Asuncion, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2010

[7] D. Lambert, Zero-inflated Poisson regression, with an application to defects in manufacturing, *Technometrics*, 34:1-14, 1992

[8] J. Mullahy, Specification and testing of some modified count data models, *Journal of Econometrics*, 33:341-365, 1986

[9] M. Ridout, C. Demetrio and J. Hinde, Models for Count Data with Many Zeros International Biometric Conference, *Proceedings of the XIXth International Biometric Conference*, p.179-192, 1998

[10] P. Wang, Markov zero-inflated Poisson regression models for a time series of counts with excess zeros, *Journal of Applied Statistics*, 28:623-632, 2001