Modelling count time series in a state-dependent under-reporting scheme

A. Fernández, **A. Cabaña**, H. Joe, D. Moriña, P.Puig [1]

Since the introduction of the Integer-Valued AutoRegressive (INAR) models in [1], the interest in the analysis of count time series has been growing. The main reason for this increasing popularity is the limited performance of the classical time series analysis approach when dealing with discrete valued time series. With the introduction of discrete time series analysis techniques, several challenges appeared such as unobserved heterogeneity, periodicity, under-reporting, ... Many efforts have been devoted in order to introduce seasonality in these models [2] and also coping with unobserved heterogeneity. However, the problem of under-reported data is still in a quite early stage of study in many different fields. This phenomenon is very common in many contexts such as epidemiological and biomedical research. It might lead to potentially biased inference and may also invalidate the main assumptions of the classical models. For instance, in [3] the author explores a Markov chain Monte Carlo based method to study worker absenteeism where sources of under-reporting are detected. Also in the context of public health , it is well known that several diseases have been traditionally under-reported (occupational related diseases, food exposures diseases, ...).

The model we will present here considers two discrete time series: the observed series of counts Y_t which may be under-reported, and the underlying series X_t with an INAR(1) structure

$$X_t = \alpha \circ X_{t-1} + W_t$$

where $0 < \alpha < 1$ is a fixed parameter and W_t is Poisson(λ). The *binomial thinning* operator is defined as $\alpha \circ X_{t-1} = \sum_{i=1}^{X_{t-1}} Z_i$, where Z_i are i.i.d Bernoulli random variables with probability of success equal to α.

The way we allow Y_t to be under-reported is by defining that $Y_t = X_t$ with probability $1 - \omega$ or it is $q \circ X_t$ with probability ω. This process $\{Y_n\}$ represents an under-reported phenomenon coming from the latent INAR(1) process, where parameters ω and $q$are the frequency and the intensity of under-reporting, respectively.

For a more general model, suppose now that the states of under-reporting follow a binary discrete-time Markov chain. This new situation only adds one further parameter to the previous situation, and is more flexible for modelling.

We derive the autocorrelation structure of both models, and compute maximum likelyhood estimators for the parameters via a modification of Viterbi algorithm.

Several examples of application of the models in the field of public health will be discussed, using real data.

[1] Department de Matemàtiques, Universitat Autnoma E-mail: `acabana@mat.uab.cat`

References

[1] Al-Osh, M. A. and Alzaid, A. A. (1987). First-order integer-valued autoregressive INAR(1) process. *Journal of Time Series Analysis*, **8**, 261–275.

[2] Heiligers B. *E-optimal designs in weighted polynomial regression* // Ann. Stat., 1994, v. 22, p. 917–929.

[3] Moriña, D., Puig, P., Ríos, J., Vilella, A. and Trilla, A. (2011). A statistical model for hospital admissions caused by seasonal diseases. *Statistics in Medicine*, **30**, 3125–3136.

[4] Winkelmann, R (1996). Markov chain Monte Carlo analysis of underreported count data with an application to worker absenteeism. *Empirical Economics Dec*; **21(4)**, 575–587.