A maximum likelihood and regenerative bootstrap approach for estimation and forecasting of INAR(p) processes with ZI innovations

Aldo M. Garay(*)

Department of Statistics
Federal University of Pernambuco (UFPE)
Recife, Brazil
https://sites.google.com/de.ufpe.br/agaray

ECODEP CONFERENCE - IHP: Room Pierre Grisvard
Paris - France - 2024

(*) Join work with: Patrice Bertail, Francyelle de Lima and Isaac Sales

Summary

Introduction

- A popular approach to deal with count time series having a structure depending on their past observations, is to consider the integer-valued autoregressive (INAR) process, introduced by Steutel and Harn (1979), Al-Osh and Alzaid (1987) and McKenzie (1988).

- The process is established by considering a Poisson distribution for the innovations.

- One frequent manifestation of this overdispersion, in count data, is that the incidence of zero counts is greater than expected.

- In the context of non-negative integer values time series, with excess zeros, Jazi et al. (2012) proposed the ZINAR(1) process, which suppose that the **innovations** of the process follow a Zero Inflated Poisson (ZIP) distribution.

The First-order integer valued autoregressive processes INAR(1)

Stochastic structure

$$Y_t = \alpha \circ Y_{t-1} + V_t, \ t \in \mathbb{Z}, \tag{1}$$

where $\{V_t\}_{t\in\mathbb{Z}}$ is a sequence of non-negative integer-valued random variables iid (*innovations*), independent of Y_{t-1}, for all t.

The *thinning* operator '$\circ$'

Let X be a non negative integer value r.v. and $\alpha \in [\,0, 1]$. Then, the r.v. $\alpha \circ X$ is given by:

$$\alpha \circ X = \sum_{i=1}^{X} Z_i, \tag{2}$$

where $\{Z_i\}_{i\geq 1}$ is a sequence of iid Bernoulli random variables, with $\mathbb{P}(Z_i = 1) = \alpha$, independent of X.

Zero Inflated distributions

Definition

The discrete r.v. V follows a Zero Inflated (ZI) distribution, with parameters ρ and $\boldsymbol{\lambda}$, if it has the following stochastic representation:

$$V = BU, \quad B \perp U,$$

where B is a Bernoulli r.v., with $\mathbb{P}(B = 1) = 1 - \rho$ and $0 \leq \rho < 1$. U is a non negative discrete r.v., with probability mass function (pmf) $h_U(u|\boldsymbol{\lambda})$.

$$\mathbb{P}(V = v) = \begin{cases} \rho + (1-\rho)\, h_U(0|\boldsymbol{\lambda}) & v = 0 \\ (1-\rho)\, h_U(v|\boldsymbol{\lambda}) & v \geq 1, \end{cases} \tag{3}$$

where $h_U(v|\boldsymbol{\lambda}) = \mathbb{P}(U = v)$. We denote $V \sim \mathrm{ZI}(\rho, \boldsymbol{\lambda}; h_U(\cdot))$.

$$E[V] = (1-\rho)\, E[U] \text{ and } Var[V] = (1-\rho)\left(Var[U] + \rho E^2[U]\right). \tag{4}$$

Zero Inflated distributions

- *The Zero Inflated Poisson (ZIP) model:*

$$\mathbb{P}(V = v) = \begin{cases} \rho + (1-\rho)\, e^{-\lambda}, & v = 0 \\ (1-\rho)\, \frac{e^{-\lambda}\lambda^{v}}{v!} & v \geq 1 \end{cases}$$

Notation $V \sim \mathrm{ZIP}(\rho, \lambda)$.

- *The Zero Inflated Negative Binomial (ZINB) model:*

$$\mathbb{P}(V = v) = \begin{cases} \rho + (1-\rho)\left(\frac{\phi}{\mu+\phi}\right)^{\phi}, & v = 0 \\ (1-\rho)\frac{\Gamma(\phi+v)}{\Gamma(v+1)\Gamma(\phi)}\left(\frac{\mu}{\mu+\phi}\right)^{v}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}, & v \geq 1 \end{cases}$$

Notation $V \sim \mathrm{ZINB}(\rho, \mu, \phi)$.

Zero Inflated distributions

- *The Zero Inflated Poisson Inverse Gaussian (ZIPIG) model:*

$$\mathbb{P}(V=v)=\begin{cases}\rho+(1-\rho)e^{\phi-\sqrt{\phi(\phi+2\lambda)}}, & v=0;\\ (1-\rho)\sqrt{\frac{2}{\rho}}\left[\phi(\phi+2\lambda)\right]^{-\frac{(v-1/2)}{2}}\frac{e^{\phi}(\lambda\phi)^{v}}{v!}K_{v-1/2}\left(\sqrt{\phi(\phi+2\lambda)}\right), & v\geq 1\ldots\end{cases}$$

$K_\lambda(t)$ is the modified Bessel function of third kind.
Notation: $V\sim \mathrm{ZIPIG}(\rho,\lambda,\phi)$.

The ZIPIG distribution is a particular case of the Mixed Poisson distribution, with hierarchical representation: $V|Z=z\sim \mathrm{Poisson}(\mu z)$, where Z follows an Inverse Gaussian (IG) distribution, with mean 1 and dispersion parameter ϕ, denoted by $Z\sim \mathrm{IG}(1,\phi)$.

The ZI-INAR(1) process

Definition

The ZI-INAR(1) process is an integer valued first order autoregressive process, with ZI innovations, given by:

$$Y_t = \alpha \circ Y_{t-1} + V_t, \ \ t \in \mathbb{Z}, \tag{6}$$

where $V_t \sim \mathrm{ZI}(\rho, \boldsymbol{\lambda}; h_U(\cdot))$.

Proposition

Let $\{Y_t\}_{t\in\mathbb{Z}}$ be a stationary ZI-INAR(1) process. Then, the expectation and variance of Y_t are given by:

$$E[Y_t] = \frac{(1-\rho)E[U_t]}{1-\alpha} \ \ and \ \ Var(Y_t) = \frac{(1-\rho)\left(\alpha E[U_t] + \rho E[U_t]^2 + Var[U_t]\right)}{1-\alpha^2},$$

where U_t denotes a r.v. with density function (or pmf) $h_U(u|\boldsymbol{\lambda})$, for all $t \in \mathbb{Z}$.

Motivation - INAR(p) processes

$$Y_t = \sum_{i=1}^{p} \alpha_i \circ Y_{t-i} + V_t, \quad t \in \mathbb{Z},$$

- where $\alpha_i \in [0,1]$ $i = 1, \ldots, p$ and $\{V_t\}$ is an i.i.d. sequence of non-negative integer-valued random variables (Jin-Guan and Yuan (1991)).

- $\{V_t\}$ is independent of $Y_{t-1}, \ldots, Y_{t-p}$ and it is stronger than the assumption made by Al-Osh and Alzaid (1987), which define $\{V_t\}$ as a sequence of uncorrelated non-negative integer-valued random variables.

The ZI-INAR(p) process

ZI-INAR(p) processes

$$Y_t = \sum_{i=1}^{p} \alpha_i \circ Y_{t-i} + V_t, \ \ t \in \mathbb{Z},$$

where $V_t \sim \mathrm{ZI}\left(\rho, \boldsymbol{\lambda}; h_U(\cdot)\right)$

- We supposed that U follows different distributions, as Poisson, Negative binomial and Poisson inverse Gaussian.
- We develop an EM-type algorithm for maximum likelihood estimation of the parameters of the ZI-INAR(p) process, that consider the presence of latent variables.
- We also develop a regenerative bootstrap method to construct confidence intervals for the parameters as well as to estimate the forecasting distributions for future values.

Likelihood function of the ZI-INAR(p) processes

- Let $\mathbf{y} = (y_1, \ldots, y_n)^\top$ be realizations of the ZI-INAR(p) process, the likelihood function of $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \pi, \boldsymbol{\lambda})^\top$ with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\top$, given $\mathbf{y}$, is defined by:

$$L(\boldsymbol{\theta}|\mathbf{Y}) = \mathbb{P}(Y_1 = y_1, \ldots, Y_p = y_p) \prod_{t=p+1}^{n} \mathbb{P}(Y_t = y_t | Y_{t-1} = y_{t-1}, \ldots, Y_{t-p} = y_{t-p}),$$

- where

$$\begin{aligned}\mathbb{P}(Y_t = y_t | Y_{t-1} = y_{t-1}, \ldots, Y_{t-p} = y_{t-p}) &= \sum_{k_1=0}^{\min\{y_{t-1}, y_t\}} \binom{y_{t-1}}{k_1} \alpha_1^{k_1} (1-\alpha_1)^{y_{t-1}-k_1} \\ &\times \ldots \sum_{k_p=0}^{\min\{y_{t-p}, y_t - \sum_{i=1}^{p-1} k_i\}} \binom{y_{t-p}}{k_p} \alpha_p^{k_p} (1-\alpha_p)^{y_{t-p}-k_p} \\ &\times \left[\pi \mathbb{I}_{\{0\}}(y_t - \sum_{i=1}^{p} k_i) + (1-\pi) h_U(y_t - \sum_{i=1}^{p} k_i | \boldsymbol{\lambda})\right].\end{aligned}$$

Maximum likelihood estimation

- The likelihood function is expressed in terms of the joint probability $\mathbb{P}(Y_1 = y_1, \ldots, Y_p = y_p)$, which is not available; thus, the exact likelihood function is intractable.

- An approach to estimate the parameters is to use the conditional log-likelihood function, given by:

$$\ell(\boldsymbol{\theta}|\mathbf{y}) \propto \sum_{t=p+1}^{n} \log\left[\mathbb{P}\left(Y_t = y_t | Y_{t-1} = y_{t-1}, \ldots, Y_{t-p} = y_{t-p}\right)\right],$$

- For ZI-INAR(p) processes, the direct maximization of this expression is not easy due to its form.

- An alternative is to consider numerical optimization using the EM algorithm (Dempster et. al. 1977). Its properties ensure the monotone convergence to a stationary point of the log-likelihood function, in contrast to direct maximization.

Maximum likelihood estimation

ZI-INAR(p) processes

$$Y_t = \sum_{i=1}^{p} \alpha_i \circ Y_{t-i} + V_t, \ \ t \in \mathbb{Z},$$

where $V_t \sim \mathrm{ZI}\left(\rho, \boldsymbol{\lambda}; h_U(\cdot)\right)$

- The key to the development of our EM-type algorithm is to consider the presence of latent variables and treat the problem, as if these variables were in fact observed:
  - $S_{t,i} = \alpha_i \circ Y_{t-i}$; where $S_{t,i}|Y_{t-i} = y_{t-i}$ follows a binomial distribution, with parameters y_{t-i} and α_i, when $y_{t-i} > 0$ and if $y_{t-i} = 0$, $S_{t,i}$ follows a degenerate distribution at zero.
  - Considering that $V_t \sim \mathrm{ZI}\left(\rho, \boldsymbol{\lambda}; h_U(\cdot)\right)$, the latent binary random variable W_t exists such that: $W_t \sim \mathrm{Bern}(\pi)$ and

$$V_t|W_t = 0 \sim h_U(\cdot|\boldsymbol{\lambda}),$$
$$V_t|W_t = 1 \text{ follows a degenerate distribution at zero.}$$

ML estimation via the EM algorithm

- Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$, $\mathbf{W} = (W_{p+1}, \ldots, W_n)^\top$, $\mathbf{S} = (\mathbf{S}_{p+1}, \ldots, \mathbf{S}_n)^\top$, with $\mathbf{S}_t = (S_{t,1}, \ldots, S_{t,p})^\top$ and $\mathbf{Y}_c = (\mathbf{Y}_{c_{p+1}}, \ldots, \mathbf{Y}_{c_n})^\top$ with $\mathbf{Y}_{c_i} = (Y_i, W_i, \mathbf{S}_i)$ for i=p+1,. . . ,n.

Complete log-likelihood

$$\begin{aligned}
\ell_c(\boldsymbol{\theta}|\mathbf{y}_c) \quad \propto \quad & \sum_{i=1}^{p} \left\{ \left(\sum_{t=p+1}^{n} s_{t,i} \right) \log(\alpha_i) \right\} + \sum_{i=1}^{p} \left\{ \left[\sum_{t=p+1}^{n} (y_{t-i} - s_{t,i}) \right] \log(1-\alpha_i) \right\} \\
& + \sum_{t=p+1}^{n} w_t \log(\pi) + \sum_{t=p+1}^{n} (1-w_t) \log(1-\pi) \\
& + \sum_{t=p+1}^{n} (1-w_t) \log(h_U(y_t - \sum_{i=1}^{p} s_{t,i}|\boldsymbol{\lambda})).
\end{aligned}$$

ML estimation via the EM algorithm

E-Step: $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = E\left[\ell_c(\boldsymbol{\theta}|\mathbf{y}_c)|\mathbf{y}, \widehat{\boldsymbol{\theta}}^{(k)}\right]$

Given the current estimate $\widehat{\boldsymbol{\theta}}^{(k)}$ at k-th step

$$
\begin{aligned}
Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) \quad \propto \quad & \sum_{i=1}^{p}\left\{\left(\sum_{t=p+1}^{n}\widehat{s}_{t,i}^{(k)}\right)\log(\alpha_i)\right\} + \sum_{i=1}^{p}\left\{\left[\sum_{t=p+1}^{n}\left(y_{t-i} - \widehat{s}_{t,i}^{(k)}\right)\right]\log(1-\alpha_i)\right\} \\
+ \quad & \sum_{t=p+1}^{n}\widehat{w}_t^{(k)}\log(\pi) + \sum_{t=p+1}^{n}\left(1-\widehat{w}_t^{(k)}\right)\log(1-\pi) + \sum_{t=p+1}^{n}Q_t^*\left(\boldsymbol{\lambda}|\widehat{\boldsymbol{\theta}}^{(k)}\right).
\end{aligned}
$$

where

$$
\begin{aligned}
\widehat{s}_{t,i}^{(k)} &= E\left(S_{t,i}|\mathbf{y}, \widehat{\boldsymbol{\theta}}^{(k)}\right), \\
\widehat{w}_t^{(k)} &= E\left(W_t|\mathbf{y}, \widehat{\boldsymbol{\theta}}^{(k)}\right), \\
Q_t^*(\boldsymbol{\lambda}|\widehat{\boldsymbol{\theta}}^{(k)}) &= E\left((1-W_t)\log h_U(y_t - \sum_{i=1}^{p}S_{t,i})|\mathbf{y}, \widehat{\boldsymbol{\theta}}^{(k)}\right).
\end{aligned}
$$

ML estimation via the EM algorithm

- The maximization of $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$ over $\boldsymbol{\theta}$ must be obtained under restrictions on the parameter $\boldsymbol{\alpha}$ in order to ensure the standard condition for stationarity and ergodicity of the ZI-INAR(p) process. Recall that this process will be stationary if and only if $\alpha_i \in [0,1)$, for $i = 1, \ldots, p$, and $\sum_{i=1}^{p} \alpha_i < 1$.

Proposition

Let Δ be the function from the set $\alpha_i \in [0,1), i = 1, \ldots, p, \sum_{i=1}^{p} \alpha_i < 1$ to the set $[0,1)^p$ defined by $\Delta : (\alpha_1, \ldots, \alpha_p)^\top \to (\beta_1, \ldots, \beta_p)^\top$ with $\beta_i = \Delta_i(\boldsymbol{\alpha}) = \alpha_i/(1 - \sum_{j \neq i}^{p} \alpha_j), \quad i = 1, \ldots, p$. Then the Δ transformation admits an inverse given by:

$$\alpha_i = \Delta_i^{-1}(\boldsymbol{\beta}) = \left(1 - \frac{\sum_{i=1}^{p} \frac{\beta_i}{1-\beta_i}}{1 + \sum_{i=1}^{p} \frac{\beta_i}{1-\beta_i}}\right) \frac{\beta_i}{1-\beta_i}, \text{ for } \quad i = 1, \ldots, p, \text{ and } \boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$$

ML estimation via the EM algorithm

M-step

Update $\widehat{\boldsymbol{\theta}}^{(k)}$ by maximizing the Q-function over $\boldsymbol{\theta}$, which leads to the following expressions:

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \underset{\boldsymbol{\beta}\in[0,1)^p}{\arg\max}\left\{Q(\Delta^{-1}(\boldsymbol{\beta}),\pi,\boldsymbol{\lambda}|\widehat{\boldsymbol{\theta}}^{(k)})\right\}, \qquad \widehat{\alpha}_i^{(k+1)} = \Delta_i^{-1}(\widehat{\boldsymbol{\beta}}^{(k+1)})$$

$$\widehat{\pi}^{(k+1)} = \frac{\sum_{t=p+1}^{n}\widehat{w}_t}{n-p} \text{ and } \qquad \widehat{\boldsymbol{\lambda}}^{(k+1)} = \underset{\boldsymbol{\lambda}}{\arg\max}\left\{\sum_{t=p+1}^{n} Q_t^*(\boldsymbol{\lambda}|\widehat{\boldsymbol{\theta}}^{(k)})\right\}, \quad (7)$$

where $\Delta^{-1}(\boldsymbol{\beta}) = (\Delta_1^{-1}(\boldsymbol{\beta}),\ldots,\Delta_p^{-1}(\boldsymbol{\beta}))^\top$.

In the following, we develop the procedure to obtain the expression $Q_t^*(\boldsymbol{\lambda}|\boldsymbol{\theta}^{(k)})$ and $\widehat{\boldsymbol{\lambda}}^{(k+1)}$, considering the three particular cases of the ZI models.

ML estimation via the EM algorithm

M-step

- If $V_t \sim \text{ZIP}(\rho,\lambda)$, then:

$$\widehat{\lambda}^{(k+1)} = \left(\sum_{t=p+1}^{n}(1-\widehat{w}_t^{(k)})y_t - \sum_{t=p+1}^{n}\sum_{i=1}^{p}\widehat{b_t s_{t,i}}^{(k)}\right)\left(\sum_{t=p+1}^{n}(1-\widehat{w}_t^{(k)})\right)^{-1}.$$

- If $V_t \sim \text{ZINB}(\rho,\mu,\phi)$, then $\widehat{\boldsymbol{\lambda}}^{(k+1)} = (\widehat{\mu}^{(k+1)}, \widehat{\phi}^{(k+1)})$ are given by:

$$\begin{aligned}
\widehat{\mu}^{(k+1)} &= \left(\sum_{t=p+1}^{n}(1-\widehat{w}_t^{(k)})y_t - \sum_{t=p+1}^{n}\sum_{i=1}^{p}\widehat{b_t s_{t,i}}^{(k)}\right)\left(\sum_{t=p+1}^{n}(1-\widehat{w}_t^{(k)})\right)^{-1} \\
\widehat{\phi}^{(k+1)} &= \arg\max_{\phi}\left\{\ell\left(\widehat{\boldsymbol{\alpha}}^{(k)}, \widehat{\pi}^{(k)}, \widehat{\mu}^{(k+1)}, \phi|\mathbf{y}\right)\right\},
\end{aligned}$$

$\widehat{\phi}^{(k+1)}$ is obtained using the "*optim*" routine in the $\mathbb{R}$ software.

ML estimation via the EM algorithm

M-step

- If $V_t \sim$ ZIPIG(ρ,μ,ϕ), then $\boldsymbol{\lambda} = (\mu, \phi)$.

$$\widehat{\mu}^{(k+1)} = \left(\sum_{t=p+1}^{n} (1 - \widehat{w}_t^{(k)}) y_t - \sum_{t=p+1}^{n} \sum_{i=1}^{p} \widehat{b_t s_{t,i}}^{(k)} \right) \left(\sum_{t=p+1}^{n} (1 - \widehat{w}_t^{(k)}) \right)^{-1} \text{ and}$$

$$\widehat{\phi}^{(k+1)} = \left(\sum_{t=p+1}^{n} (1 - \widehat{w}_t^{(k)}) \right) \left(\sum_{t=p+1}^{n} \widehat{b_t z_t}^{(k)} + \sum_{t=p+1}^{n} \widehat{b_t / z_t}^{(k)} - 2 \sum_{t=p+1}^{n} (1 - \widehat{w}_t^{(k)}) \right)^{-1}.$$

where
$\widehat{b_t z_t}^{(k)} = E\left[E\left(B_t Z_t | S_t, \mathbf{y}, \widehat{\boldsymbol{\theta}}^{(k)}\right)\right]$ and $\widehat{b_t / z_t}^{(k)} = E\left[E\left(B_t Z_t^{-1} | S_t, , \mathbf{y}, \widehat{\boldsymbol{\theta}}^{(k)}\right)\right]$.

Regenerative Bootstrap method

- In the context of the INAR processes, different bootstrap methods have been developed: residual based bootstrap, Block bootstrap methods, Block of block bootstrap methods.

- However most of the time these methods were only analyzed by simulations. Obtaining the validity of the bootstrap may be complicated or even false in this framework.

- In this manuscript we use a different approach which is simple to implement and which does not require the estimation of the residuals or of any other type of hyper-parameters.

- Since the INAR process, and more generally integer valued processes, can be approximated by Markov chains with atoms (each visited point or sequence of visited points can be seen as an atom), the regenerative approach appears natural in this context.

Regenerative Bootstrap method

Step 1. Count the number of visits l_n to the atom A up to time n. Divide the observed sample path $\mathbf{Y}^{(n)} = (\mathbf{Y}_0, \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ into $l_n + 1$ blocks, $\mathcal{B}_0$, $\mathcal{B}_1, \ldots, \mathcal{B}_{l_n-1}$, $\mathcal{B}_{l_n}^{(n)}$. Drop the first and last (non-regenerative) blocks.

Step 2. Draw sequentially bootstrap data blocks $\mathcal{B}_{1,n}^*, \ldots, \mathcal{B}_{k,n}^*$ independently from the empirical distribution $F_n = (l_n - 1)^{-1} \sum_{j=1}^{l_n-1} \delta_{\mathcal{B}_j}$ of the blocks $\{\mathcal{B}_j\}_{1 \leqslant j \leqslant l_n-1}$ conditional on $\mathbf{Y}^{(n)}$, until the length $l^*(k) = \sum_{j=1}^{k} l(\mathcal{B}_{j,n}^*)$ of the bootstrap data series is larger than n.

Step 3. From the resampled data blocks, construct a pseudo-trajectory by binding the blocks together $\mathbf{Y}^{*(n)} = (\mathcal{B}_{1,n}^*, \ldots, \mathcal{B}_{l_n^*-1,n}^*)$ and truncating the joint blocks to get a time series of length n. Then recompute the *value of the statistics* of interest $\mathcal{T}_n^* = \mathcal{T}_n(Y^{*(n)})$ on these values.

Step 4. If $\mathcal{S}_n = \mathcal{S}(\mathcal{B}_1, ..., \mathcal{B}_{l_n-1})$ is an estimator of the variance (otherwise set it to 1) of the original statistic $\mathcal{T}_n$. Similarly compute $\mathcal{S}_{n,b_n}^* = \mathcal{S}(\mathcal{B}_{1,n}^*, \ldots, \mathcal{B}_{l_n^*-1,n}^*)$.

Step 5. Repeat independently the procedure above in Step 2 to 4, by B times, to compute successive values $\mathcal{S}_{b,n}^{*-1}(\mathcal{T}_{b,n}^* - \mathcal{T}_n)$, $b = 1, \ldots, B$.

Regenerative Bootstrap method

- **Illustrative example** : We give a very simple illustration on the method on a short trajectory taking only three possible values in $\{0, 1, 2\}$ with $p = 2$.

- Let $Y^{(n)} = \{0, 0, 1, 2, 0, 2, 0, 1, 0, 1, 0, 1, 2, 1, 0, 2, 1, 0, 1\}$ By vectorization (with $p = 2$) we get the two variate trajectory

$$\mathbf{Y}^{(n)} = \left\{ \begin{matrix} 0 \\ 0 \end{matrix}, \begin{matrix} 1 \\ 0 \end{matrix}, \begin{matrix} 2 \\ 1 \end{matrix}, \begin{matrix} 0 \\ 2 \end{matrix}, \begin{matrix} 2 \\ 0 \end{matrix}, \begin{matrix} 0 \\ 2 \end{matrix}, \begin{matrix} 1 \\ 0 \end{matrix}, \begin{matrix} 0 \\ 1 \end{matrix}, \begin{matrix} 1 \\ 0 \end{matrix}, \begin{matrix} 0 \\ 1 \end{matrix}, \begin{matrix} 1 \\ 0 \end{matrix}, \begin{matrix} 2 \\ 1 \end{matrix}, \begin{matrix} 1 \\ 2 \end{matrix}, \begin{matrix} 0 \\ 1 \end{matrix}, \begin{matrix} 2 \\ 0 \end{matrix}, \begin{matrix} 1 \\ 2 \end{matrix}, \begin{matrix} 0 \\ 1 \end{matrix}, \begin{matrix} 1 \\ 0 \end{matrix} \right\}.$$

- Notice that $(1, 0)^\top$ is the most visited atom and will be our atom A, then we have

$$\mathcal{B}_0 = \left\{ \begin{matrix} 0 \\ 0 \end{matrix}, \begin{matrix} 1 \\ 0 \end{matrix} \right\}, \mathcal{B}_1 = \left\{ \begin{matrix} 2 \\ 1 \end{matrix}, \begin{matrix} 0 \\ 2 \end{matrix}, \begin{matrix} 2 \\ 0 \end{matrix}, \begin{matrix} 0 \\ 2 \end{matrix}, \begin{matrix} 1 \\ 0 \end{matrix} \right\}, \mathcal{B}_2 = \left\{ \begin{matrix} 0 \\ 1 \end{matrix}, \begin{matrix} 1 \\ 0 \end{matrix} \right\}$$
$$\mathcal{B}_3 = \left\{ \begin{matrix} 0 \\ 1 \end{matrix}, \begin{matrix} 1 \\ 0 \end{matrix} \right\}, \mathcal{B}_4 = \left\{ \begin{matrix} 2 \\ 1 \end{matrix}, \begin{matrix} 1 \\ 2 \end{matrix}, \begin{matrix} 0 \\ 1 \end{matrix}, \begin{matrix} 2 \\ 0 \end{matrix}, \begin{matrix} 1 \\ 2 \end{matrix}, \begin{matrix} 0 \\ 1 \end{matrix}, \begin{matrix} 1 \\ 0 \end{matrix} \right\}.$$

- In turn these blocks may be identify as part of the initial trajectory

$$\mathcal{B}_0 = \text{“}(0, 0, 1)\text{”}, \ \mathcal{B}_1 = \text{“}(2, 0, 2, 0, 1)\text{”}, \mathcal{B}_2 = \text{“}(0, 1)\text{”}$$
$$\mathcal{B}_3 = \text{“}(0, 1)\text{”}, \ \mathcal{B}_4 = \text{“}(2, 1, 0, 2, 1, 0, 1)\text{”}.$$

Prediction in a regenerative framework

- As mentioned by Weiss, (2018), for real-valued processes, the most common type of point forecasting is the conditional mean, as this is known to be optimal in the sense of the mean squared error.

- The main disadvantage of the mean forecast is that it will usually lead to non-integer value predictions, while Y_{n+h} will certainly take an integer value.

- Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \pi, \boldsymbol{\lambda})^\top$, the parameter vector of the ZI-INAR(p) process. We update the values of $\{Y_{n+1}, Y_{n+2}, \ldots\}$ one component at a time, using a procedure similar to that described by Neal and Subba Rao (2007) and Garay et al. (2020) for predictive inference, but including an additional level of the regenerative procedure to get the distribution of the predictor:

Prediction in a regenerative framework

Step 1 Repeat the following procedure B times:

(a) Generate a bootstrap sample using the regenerative bootstrap method.
(b) Estimate the parameters $\boldsymbol{\theta}$, of the model, using the EM procedure, denoted by $\widehat{\boldsymbol{\theta}}^b = (\widehat{\boldsymbol{\alpha}}^b, \widehat{\pi}^b, \widehat{\boldsymbol{\lambda}}^b)^\top$, with $b = 1, \ldots, B$.

Step 2 For $b = 1, \ldots, B = 999$, repeat the following steps $M = 199$ times:

(a) Draw the value of $w_{n+h}^{(m)}$ from the distribution Bern$(\widehat{\pi}^b)$.
(b) Draw the value of $v_{n+h}^{(m)}$ from the distribution of $V_{n+h}|w_{n+h}^{(m)}, \widehat{\boldsymbol{\lambda}}^b$.
(c) For $1 \leq i \leq p$, obtain:

$$s_{n+h,i}^{(m)} = \begin{cases} 0, \text{ if } y_{n+h-i} = 0 \\ \text{Value drawn from Bin}\left(y_{n+h-i}; \widehat{\alpha}_i^b\right), \text{ otherwise.} \end{cases}$$

(d) Set $y_{n+h}^{(m)} = \sum_{i=1}^{p} s_{n+h,i}^{(m)} + v_{n+h}^{(m)}$, for $h \geq 1$ and $m = 1, \ldots, M$.

Step 3 For $b = 1, \ldots, B$, compute: $\widehat{p}_{n+h}^b(j) = \frac{\#\left\{m,\ y_{n+h}^{(m)} = j\right\}}{M}$, $h \geq 1$.

Simulation Study: Robustness of the EM estimates

- The goal of this simulation study is to evaluate the finite-sample performance of the parameter estimates for the ZI-INAR(p) model.

- We generated artificial samples with $p \in \{1, 2, 3\}$, $\pi \in \{0.3, 0.6\}$ and $\alpha = 0.3$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^\top = (0.3, 0.2)^\top$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^\top = (0.25, 0.2, 0.15)^\top$ for ZI-INAR(1) to ZI-INAR(3) processes, respectively.

- The sample sizes were fixed to $n \in \{100, 300, 500, 1000\}$ and the ZI model considered for the innovation were respectively: (i) ZIP, with $\lambda = 2$; and (ii) ZINB and ZIPIG with $\mu = 2$ and $\phi = 1.5$.

- We generated $N = 300$ replicates of size n and analyzed the relative bias (RB) and root relative mean square error (RRMSE):

$$\text{RB}(\widehat{\theta}_i) = \frac{1}{N}\sum_{j=1}^{N}\frac{\widehat{\theta}_{ij} - \theta_i}{\theta_i} \text{ and } \text{RRMSE}(\widehat{\theta}_i) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\frac{\widehat{\theta}_{ij} - \theta_i}{\theta_i}\right)^2},$$

$\widehat{\theta}_{ij}$ is the estimate of parameter θ_i, in the j-th replicate.

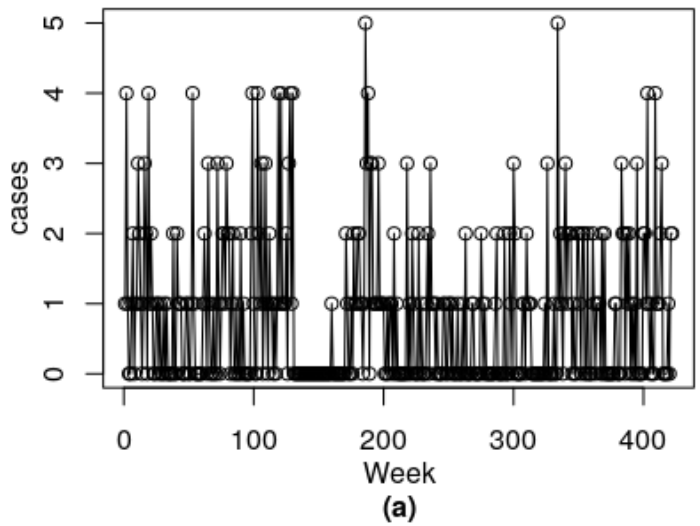
RB and RRMSE (in parentheses) of $\widehat{\theta}$

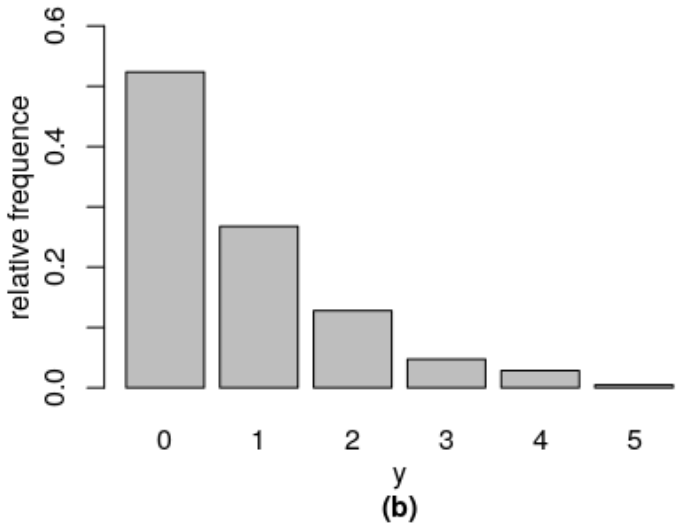
p	π	n	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\alpha}_3$	$\widehat{\pi}$	$\widehat{\lambda}$
1	0.3	100	**-0.067 (0.314)**	-	-	-0.076 (0.344)	-0.004 (0.116)
		300	**-0.017 (0.164)**	-	-	-0.009 (0.194)	0.006 (0.078)
		500	**-0.023 (0.119)**	-	-	-0.013 (0.137)	0.003 (0.058)
		1000	**-0.002 (0.087)**	-	-	-0.015 (0.010)	-0.002 (0.040)
	0.6	100	-0.043 (0.236)	-	-	-0.021 (0.134)	0.001 (0.158)
		300	-0.012 (0.129)	-	-	-0.007 (0.071)	-0.006 (0.090)
		500	-0.012 (0.099)	-	-	-0.005 (0.049)	0.002 (0.065)
		1000	-0.010 (0.075)	-	-	-0.003 (0.039)	-0.004 (0.047)
2	0.3	100	-0.060 (0.428)	-0.106 (0.399)	-	**-0.142 (0.476)**	0.037 (0.176)
		300	-0.003 (0.240)	-0.058 (0.252)	-	**-0.054 (0.279)**	0.013 (0.096)
		500	-0.004 (0.173)	-0.018 (0.183)	-	**-0.026 (0.210)**	-0.001 (0.070)
		1000	-0.010 (0.120)	-0.005 (0.138)	-	**-0.017 (0.150)**	0.006 (0.053)
	0.6	100	-0.020 (0.332)	-0.103 (0.398)	-	-0.056 (0.178)	-0.014 (0.158)
		300	-0.004 (0.195)	-0.032 (0.203)	-	-0.018 (0.097)	-0.010 (0.101)
		500	0.004 (0.146)	-0.029 (0.157)	-	-0.009 (0.076)	-0.003 (0.084)
		1000	-0.009 (0.105)	-0.017 (0.115)	-	-0.009 (0.048)	-0.006 (0.052)

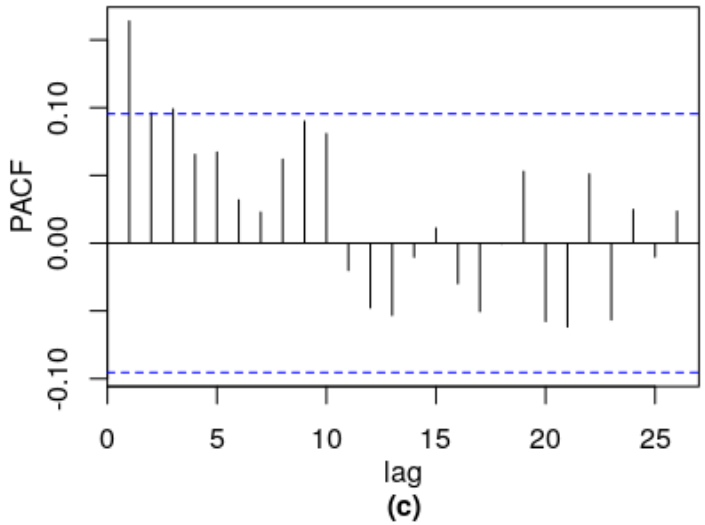
Real dataset: Assaults dataset

- This dataset concerns the weekly number of assaults, recorded from January 2008 to December 2015 at Federal University of Pernambuco (UFPE) – Brazil.
- All occurrences are described, recorded and organized by SSI/UFPE, that in an office in charge of the planning, execution and evaluation of projects and activities related to institutional security at the Federal University of Pernambuco (UFPE) of Brazil.
- 422 observations from January 2008 to December 2015.
- This dataset contains a significant presence of zero values observations (52%).

Real dataset: Assaults dataset







Real dataset: Assaults dataset

Tabela: Model selection criteria of INAR(p) processes with different innovations.

	Innovations								
	Po			NB			PIG		
	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$
AIC_M	1051.72	1041.13	1037.51	1018.09	1009.91	1009.22	1021.62	1013.45	1012.68
BIC_M	1059.80	1053.25	1053.67	1030.21	1026.07	1029.42	1033.74	1029.61	1032.88
	ZIP			ZINB			ZIPIG		
	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$
AIC_M	1014.28	**1004.75**	**1003.52**	1016.06	1006.76	1005.61	1016.09	1006.75	1005.53
BIC_M	1026.40	**1020.91**	1023.72	1032.22	1026.96	1029.85	1032.25	1026.95	1029.77

Assaults dataset

Tabela: Parameter estimates, SE-Boots and CI Boots

Param	ZIP-INAR(2)			ZIP-INAR(3)		
	Estimates	SE Boots	95% CI Boots	Estimates	SE Boots	95% CI Boots
α_1	0.180	0.035	(0.111, 0.248)	0.172	0.036	(0.102, 0.242)
α_2	0.105	0.042	(0.023, 0.187)	0.089	0.043	(0.004, 0.173)
α_3	–	–	—	0.072	0.039	[0.000, 0.148)
π	0.516	0.057	(0.404, 0.628)	0.542	0.060	(0.424, 0.660)
λ	1.173	0.119	(0.939, 1.407)	1.168	0.122	(0.928, 1.407)

Assaults dataset

- Under the ZIP-INAR(2) model, we have that of total values **'zeros'** presents in the sample, the estimated proportion of **'zeros'**provided by the innovation is given by:

$$\frac{1}{B}\sum_{b=1}^{B}\left\{\widehat{\pi}^b + (1-\widehat{\pi}^b)e^{-\widehat{\lambda}^b}\right\} = 0.6677,$$

where $\widehat{\pi}^b$ and $\widehat{\lambda}^b$ are the m.l.e of the parameters, respectively, considering the b-th regenerative bootstrap sample.

Assaults dataset - ZINAR(2) process

- In order to obtain the forecasting distribution for the last two values of the dataset (observations #421 and #422),

- We are more interested in the full distribution of the predicted value of the data points #421 and #422; thus, some summary statistics for the probability $p_{421}(j)$ and $p_{422}(j)$, respectively, defined by:

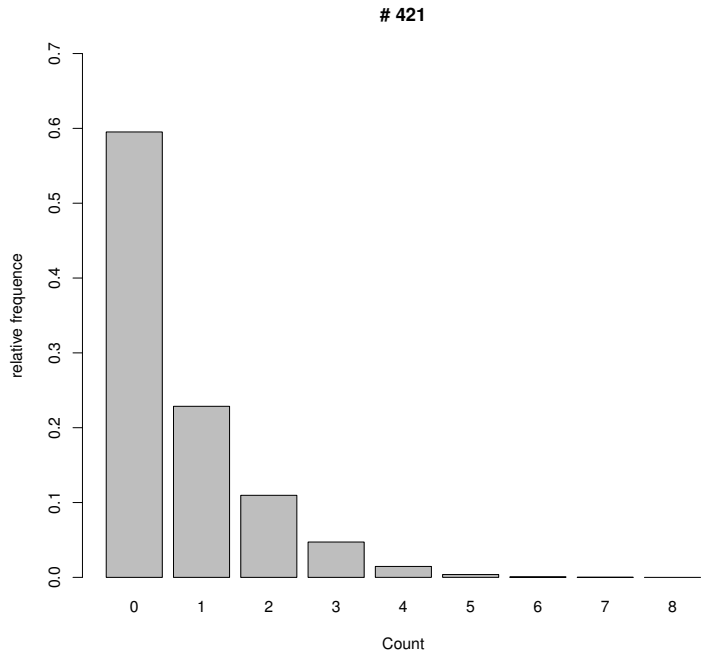
$$p_{421}(j) = \mathbb{P}\left(Y_{421} = j | Y_{420} = y_{420}, Y_{419} = y_{419}\right) \text{ and}$$

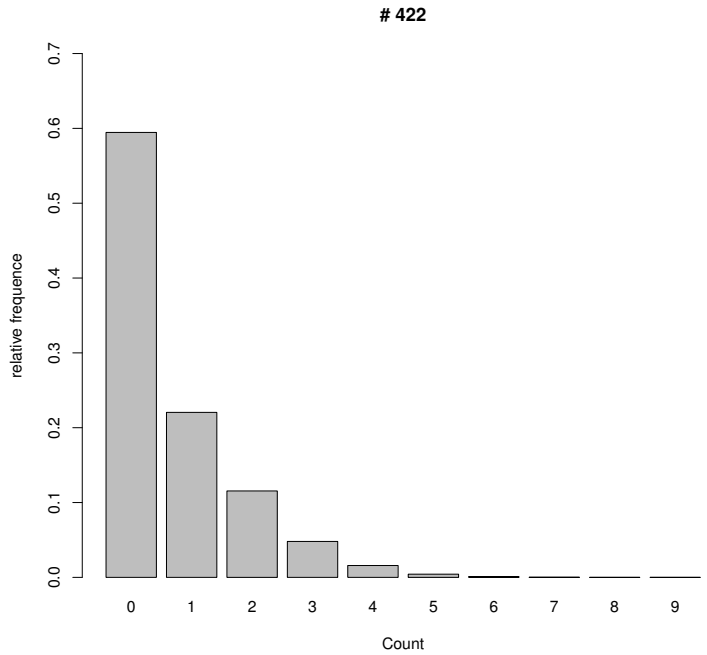
$$p_{422}(j) = \mathbb{P}\left(Y_{422} = j | Y_{421} = y_{421}, Y_{420} = y_{420}\right),$$

  with $j \in \{0, 1, 2, 3, 4, 5\}$

- We present the bar plots of the predictive distributions of the data points #421 and #422.

Real dataset: Assaults dataset





Real dataset: Assaults dataset

Tabela: Summaries of the predictive distributions of Y_{421} and Y_{422} in ZIP-INAR(2) model.

Pred. Val.	# 421				# 422			
j	Mean	SD	$Q_{0.025}$	$Q_{0.975}$	Mean	SD	$Q_{0.025}$	$Q_{0.975}$
0	0.595	0.048	0.497	0.688	0.596	0.048	0.492	0.688
1	0.229	0.036	**0.161**	**0.302**	0.220	0.034	**0.161**	**0.291**
2	0.110	0.025	**0.065**	**0.161**	0.116	0.026	**0.070**	**0.171**
3	0.047	0.017	0.020	0.080	0.048	0.017	0.020	0.085
4	0.015	0.009	0.000	0.035	0.016	0.009	0.000	0.035
5	0.004	0.005	0.000	0.015	0.004	0.005	0.000	0.015

Conclusions

- We study a new class of INAR processes, with innovations following the zero-inflated distribution, a generalization of the ZINAR(1) process proposed by Jazi et al.(2012) and Garay et al.(2021);
- We develop an innovative EM-type algorithm to obtain ML parameter estimates computationally and present a regenerative bootstrap method to construct confidence intervals for the parameters and construct the forecasting distribution for future values.
- Simulation studies and a real data analysis demonstrated the applicability and benefit of the proposed approach for practical cases, where we showed strong evidence of high-order dependence and inflated zero counts.
- Our approach can be further extended, for example, by adding a moving average structure or considering a full Bayesian approach as a basis for inference and prediction.

Main Reference

Taylor & Francis
Taylor & Francis Group

Statistics

A Journal of Theoretical and Applied Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/gsta20

A maximum likelihood and regenerative bootstrap approach for estimation and forecasting of INAR(p) processes with zero-inflated innovations

Patrice Bertail, Aldo M. Garay, Francyelle L. Medina & Isaac C.S. Jales

To cite this article: Patrice Bertail, Aldo M. Garay, Francyelle L. Medina & Isaac C.S. Jales (2024) A maximum likelihood and regenerative bootstrap approach for estimation and forecasting of INAR(p) processes with zero-inflated innovations, Statistics, 58:2, 336-363, DOI: 10.1080/02331888.2024.2344670

To link to this article: https://doi.org/10.1080/02331888.2024.2344670

Bibliography

- Al-Osh, M., Alzaid, A. A., 1987. First-order integer-valued autoregressive (INAR (1)) process. *Journal of Time Series Analysis* 8 (3), 261–275.

- Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*. Series B 39 (1), 1–38.

- Garay, A. M., Hashimoto, E. M., Ortega, E. M., Lachos, V. H., 2011. On estimation and ifluence diagnostics for zero-inflted negative binomial regression models. *Computational Statistics and Data Analysis* 55, 1304–1318.

- Jazi, M. A., Jones, G., Lai, C.-D., 2012. First-order integer valued AR processes with zero inflated Poisson innovations. *Journal of Time Series Analysis* 33 (6), 954–963.

Thank you!!