Parameter Estimation for INAR Processes Based on High-Order Statistics

Isabel Silva[1] and M. Eduarda Silva[2]

[1] Faculdade de Engenharia & CEC, Universidade do Porto
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal, *ims@fe.up.pt*
[2] Faculdade de Economia, Universidade do Porto & UIMA
Rua Dr. Roberto Frias, 4200-464 Porto, Portugal, *mesilva@fep.up.pt*

Abstract. The high-order statistics (moments and cumulants of order higher than two) have been widely applied in several fields, specially in problems where it is conjectured a lack of Gaussianity and/or non-linearity. Since the INteger-valued AutoRegressive, INAR, models are non-Gaussian, the high-order statistics can provide additional information that allows a better characterization of these processes. Thus, an estimation method for the parameters of an INAR model, based on Least Squares applied on third-order moments is proposed. The results of a Monte Carlo study, to investigate the performance of the estimator, are presented and the method is applied to a set of real data.

Keywords: INAR process, estimation, high-order statistics

1 Introduction

In the recent past, the high-order statistics (HOS) have been widely applied in several fields, specially in problems where is conjectured a lack of Gaussianity and/or non-linearity. By HOS it is meant the moments and cumulants of order higher than two, in the time domain, and the corresponding multidimensional Fourier transform (polyspectrum), in the frequency domain. In this work, the time domain approach is considered.

Let $\{X_t\}$ be a kth-order stationary stochastic process. The kth-order joint moment of $X_t, X_{t+s_1}, \ldots, X_{t+s_{k-1}}$, for $s_1, \ldots, s_{k-1} \in \mathbb{R}$, is a function of $k-1$ variables defined by $\mu_X(s_1, \ldots, s_{k-1}) = \mathrm{E}[X_t X_{t+s_1} \ldots X_{t+s_{k-1}}]$, with $\mu_X = \mathrm{E}[X_t]$.

Recently, the integer-valued autoregressive process has been proposed in the literature to model time series of counts. The pth-order integer-valued autoregressive, INAR(p), process is defined as a discrete time non-negative integer-valued stochastic process, $\{X_t\}$, that satisfies the following equation (Latour (1998)):

$$X_t = \alpha_1 * X_{t-1} + \alpha_2 * X_{t-2} + \cdots + \alpha_p * X_{t-p} + e_t, \qquad (1)$$

where

(i) $\{e_t\}$, designated the innovation process, is a sequence of independent and identically distributed (i.i.d.) non-negative integer-valued random variables with $\mathrm{E}[e_t] = \mu_e$, $\mathrm{Var}[e_t] = \sigma_e^2$ and $\mathrm{E}[e_t^3] = \gamma_e$;

(ii) the symbol $*$ represents the thinning operation (Steutel and Van Harn (1979), Gauthier and Latour (1994)), defined by

$$\alpha_i * X_{t-i} = \sum_{j=1}^{X_{t-i}} Y_{i,j}, \quad \text{for } i = 1, \ldots, p,$$

where $\{Y_{i,j}\}$, designated the counting series, is a set of i.i.d. non-negative integer-valued random variables such that $\mathrm{E}[Y_{i,j}] = \alpha_i$, $\mathrm{Var}[Y_{i,j}] = \sigma_i^2$ and $\mathrm{E}[Y_{i,j}^3] = \gamma_i$. All the counting series are assumed independent of $\{e_t\}$;

(iii) $0 \leq \alpha_i < 1$, $i = 1, \ldots, p-1$, and $0 < \alpha_p < 1$. Note that the stationarity condition for the INAR(p) process is that $\sum_{k=1}^{p} \alpha_k < 1$.

A special case is the Poisson INAR process with binomial thinning operation, where $\{e_t\}$ has a Poisson distribution with parameter λ and the counting series, $\{Y_j^{(i)}\}$, are a set of Bernoulli random variables with $P(Y_j^{(i)} = 1) = 1 - P(Y_j^{(i)} = 0) = \alpha_i$.

Since the INAR models are non-Gaussian, the HOS can provide additional information in the characterization of these processes. Thus, an estimation method for the parameters of an INAR model that uses HOS is proposed in this work. This method applies the Least Squares estimation method to minimize the errors between the third-order moment of the observations and of the fitted model.

This work is organized as follows: in Section 2 the third-order characterization of INAR(p) models is provided and the proposed Least Squares Estimation method using HOS is described. In Section 3 the results of a simulation study to assess the small sample properties of the proposed estimator are given and the method is applied to a set of observations concerning the number of plants within the industrial sector. Finally, some remarks are presented in Section 4.

2 Least squares estimation using HOS

The third-order characterization, in terms of moments and cumulants, of INAR models has been obtained by Silva and Oliveira (2004, 2005) and Silva (2005). In particular, the third-order moments of an INAR(p) process, defined by (1), satisfy a set of Yule-Walker type equations similar to those satisfied by the bilinear process, that can be written as:

$$\begin{aligned}\mu_X(0,0) = &\textstyle\sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\alpha_i\alpha_j\alpha_k\,\mu_X(i-j,i-k)\\ &+3\textstyle\sum_{i=1}^{p}\sum_{j=1}^{p}\alpha_j{\sigma_i}^2\mu_X(i-j)+3\mu_X(\sigma_e^2+{\mu_e}^2)\sum_{i=1}^{p}\alpha_i\\ &+3\mu_e\textstyle\sum_{i=1}^{p}\sum_{j=1}^{p}\alpha_i\alpha_j\mu_X(i-j)+3\mu_X\mu_e\sum_{i=1}^{p}{\sigma_i}^2\\ &+\mu_X\textstyle\sum_{i=1}^{p}(\gamma_i-3\alpha_i{\sigma_i}^2-\alpha_i^3)+\gamma_e,\end{aligned}\tag{2}$$

$$\mu_X(0,k)=\textstyle\sum_{i=1}^{p}\alpha_i\mu_X(0,k-i)+\mu_e\mu_X(0),\quad k>0,\tag{3}$$

$$\begin{aligned}\mu_X(k,k)=&\textstyle\sum_{i=1}^{p}\sum_{j=1}^{p}\alpha_i\alpha_j\,\mu_X(k-i,k-j)+\sum_{i=1}^{p}{\sigma_i}^2\mu_X(k-i)\\ &+2\mu_e\mu_X(k)-\mu_X({\mu_e}^2-{\sigma_e}^2),\quad k>0,\end{aligned}\tag{4}$$

$$\mu_X(k,m)=\textstyle\sum_{i=1}^{p}\alpha_i\mu_X(k,m-i)+\mu_e\mu_X(k),\quad m>k>0,\tag{5}$$

where $\mu_X(0)=\sum_{i=1}^{p}\alpha_i\mu_X(i)+\mu_e\mu_X+V_p$, is the second-order moment of $\{X_t\}$, with $V_p={\sigma_e}^2+\mu_X\sum_{i=1}^{p}{\sigma_i}^2$, which represents the variance of the one-step-ahead prediction error (Silva (2005)).

These equations indicate that the INAR processes have a non-linear structure, therefore the first- and second-order moments are not sufficient to describe the dependence structure of the process.

Let $\{x_1,x_2,\dots,x_n\}$ be a realization of a non-negative integer-valued stationary stochastic process with third-order moments $\mu(0,k)$, $k>0$. The approximating model considered is an INAR(p) process (order known) with parameters $\alpha_1,\dots,\alpha_p,\mu_e,\sigma_e^2$ and third-order moments $\mu_X(0,k)$, $k>0$, satisfying (3), which can be represented in the following matrix form

$$\boldsymbol{\mu}_{3,X}=\mathbf{M}_{3,X}\boldsymbol{\alpha}+\mu_e\mu_X(0)\mathbf{1}_p,\tag{6}$$

where $\boldsymbol{\mu}_{3,X}$ is defined as

$$\boldsymbol{\mu}_{3,X}=[\,\mu_X(0,1)\,\cdots\,\mu_X(0,p)\,]^T,$$

$\mathbf{M}_{3,X}$ is the $p\times p$ non-symmetric Toeplitz matrix of the third-order moments of the INAR(p) process

$$\mathbf{M}_{3,X}=\begin{bmatrix}\mu_X(0,0) & \mu_X(1,1) & \dots\,\mu_X(p-1,p-1)\\ \mu_X(0,1) & \mu_X(0,0) & \dots\,\mu_X(p-2,p-2)\\ \vdots & \vdots & \ddots\quad\vdots\\ \mu_X(0,p-1) & \mu_X(0,p-2) & \dots\quad\mu_X(0,0)\end{bmatrix},$$

with $\mu_X(\cdot,\cdot)$ given in (2) to (5), $\boldsymbol{\alpha}=[\,\alpha_1\,\cdots\,\alpha_p\,]^T$ is the vector of coefficients, $\mu_X(0)$ is the second-order moment of the INAR(p) process and $\mathbf{1}_p$ is a $p\times 1$ vector of ones.

Defining

$$\mathbf{H} = [\mathbf{M}_{3,X} \qquad \mu_X(0)\mathbf{1}_p] \qquad \text{and} \qquad \boldsymbol{\theta} = [\,\alpha_1 \,\cdots\, \alpha_p \;\mu_e\,]^T,$$

equation (6) can be rewritten as

$$\boldsymbol{\mu}_{3,X} = \mathbf{H}\boldsymbol{\theta},$$

suggesting that $\boldsymbol{\theta}$ may be estimated by least squares, i.e., minimizing the squared error between the third-order moments of the fitted INAR(p) model, $\boldsymbol{\mu}_{3,X}$, and the third-order moments of the data,

$$\boldsymbol{\mu}_3 = [\,\mu(0,1) \,\cdots\, \mu(0,p)\,]^T.$$

Thus, $\hat{\boldsymbol{\theta}}$, the Least Squares estimator of $\boldsymbol{\theta}$ based on HOS (LS_HOS) satisfies

$$\hat{\boldsymbol{\theta}} = \min_{\boldsymbol{\theta}} \{L^*(\boldsymbol{\theta})\}$$

where

$$L^*(\boldsymbol{\theta}) = (\boldsymbol{\mu}_3 - \mathbf{H}\boldsymbol{\theta})^T(\boldsymbol{\mu}_3 - \mathbf{H}\boldsymbol{\theta}).$$

In practice, the estimator is calculated by substituting the moments in $\boldsymbol{\mu}_3$ and $\mathbf{H}$ by their sample counterparts.

Thus,

$$\hat{\boldsymbol{\theta}} = \min_{\boldsymbol{\theta}} \{\hat{L}^*(\boldsymbol{\theta})\} = \min_{\boldsymbol{\theta}} \{(\hat{\boldsymbol{\mu}}_3 - \hat{\mathbf{H}}\boldsymbol{\theta})^T(\hat{\boldsymbol{\mu}}_3 - \hat{\mathbf{H}}\boldsymbol{\theta})\}.$$

Note that an estimator for σ_e^2 can be obtained by $\hat{\sigma}_e^2 = \hat{V}_p - \overline{X}\sum_{i=1}^p \hat{\sigma}_i^2$, where $\overline{X}$ is the sample mean of the observations, $\hat{\sigma}_i^2$ is an estimator of the counting series variance for the i-th thinning operation, $\alpha_i * X_{t-i}, i = 1, \ldots, p$, and $\hat{V}_p = \hat{R}(0) - \sum_{i=1}^p \hat{\alpha}_i \hat{R}(i)$, with $\hat{R}(i) = \frac{1}{N}\sum_{t=1}^{N-i}(X_t - \overline{X})(X_{t+i} - \overline{X})$, representing the sample autocovariance function. The estimation of $\hat{\sigma}_i^2$ depends on the distribution of the counting series, for instance, in the case of the binomial thinning operation (when the counting series are Bernoulli distributed), $\hat{\sigma}_i^2 = \hat{\alpha}_i(1 - \hat{\alpha}_i)$, for $i = 1, \ldots, p$.

3 Monte Carlo results and application to real data

The aim of the simulation study presented in this section is twofold: to examine the small sample properties of the estimator previously described and compare its performance with other estimation methods for the parameters of an INAR process.

Thus, 1000 realizations of Poisson INAR(p) processes ($e_t \sim \mathcal{P}o(\lambda)$), with binomial thinning operation, are generated, for $p = 0, \ldots, 3$. The sample sizes used are $N = 50, 200, 500$ and 1000 and parameter values considered

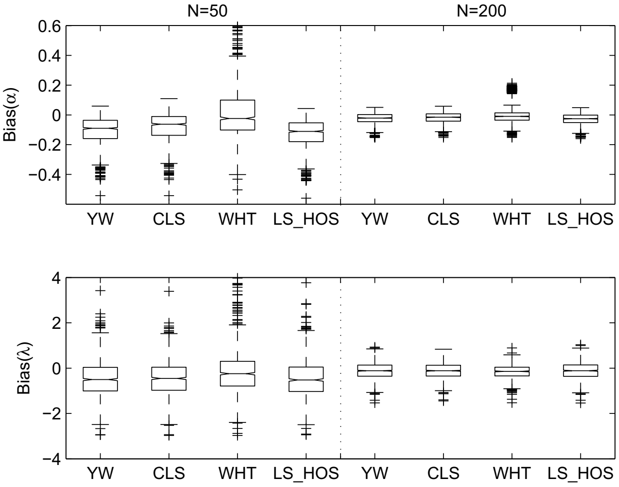

Fig. 1. Boxplots of the sample bias for the estimates obtained in 1000 realizations of 50 and 200 observations of the INAR(1) model: $X_t = 0.9 * X_{t-1} + e_t$, where $e_t \sim \mathcal{P}o(1)$.

are: $\lambda \in \{1.0, 3.0\}$, for $p = 1$, $\alpha_1 \in \{0.1, 0.4, 0.6, 0.9\}$, for $p = 2$, $(\alpha_1, \alpha_2) \in \{(0.1, 0.6), (0.6, 0.1), (0.3, 0.4), (0.4, 0.3), (0.1, 0.1), (0.4, 0.4)\}$, and for $p = 3$, $(\alpha_1, \alpha_2, \alpha_3) \in \{(0.1, 0.1, 0.4), (0.1, 0.4, 0.1), (0.4, 0.1, 0.1), (0.3, 0.3, 0.3)\}$.

For each realization, the estimation methods used to obtain $\hat{\boldsymbol{\theta}} = [\hat{\alpha}_1, \ldots, \hat{\alpha}_p, \hat{\mu}_e]^T$ are Yule-Walker (YW), Conditional Least Squares (CLS), Whittle (WHT) and Least Squares using HOS (LS_HOS). For a detailed description of the YW, CLS and WHT estimation methods see Silva (2005). The minimizations necessary in the methods CLS, WHT and LS_HOS are performed through the MATLAB function *fminunc*, which finds a minimum of a scalar unconstrained multivariable function by using the BFGS Quasi-Newton method with a mixed quadratic and cubic line search procedure (MathWorks (2004)). The initial values of the iterative methods (CLS, WHT and LS_HOS) are the YW estimates. For each case, the mean bias, variance and mean square error are evaluated.

With respect to the small sample properties of the LS_HOS estimator, the following conclusions can be drawn from the analysis of all the simulations. In general, the sample bias, variance and mean square error decrease as the sample size increases, indicating that the distribution of the estimators is consistent and symmetric. However, for a small sample size there is evidence of departure from symmetry in the marginal distributions, specially for values of the parameter near the non-stationary region.

When the several estimation methods are compared it is found that the LS_HOS provides similar results, in terms of the smallest values of sample bias, variance and mean square error, to the other methods. It is also verified that, in general, the proportion of non-admissible estimates of the methods is less for LS_HOS, followed by WHT and CLS. In order to illustrate some of these conclusions, Figure 1 shows the boxplots of the sample bias for the estimates obtained from 50 and 200 observations of the INAR(1) process with parameter values $(\alpha_1, \lambda) = (0.9, 1.0)$. Note that the value of α is near the non-stationary region, however, even for $N = 50$ observations, the LS_HOS estimates presents the best results.

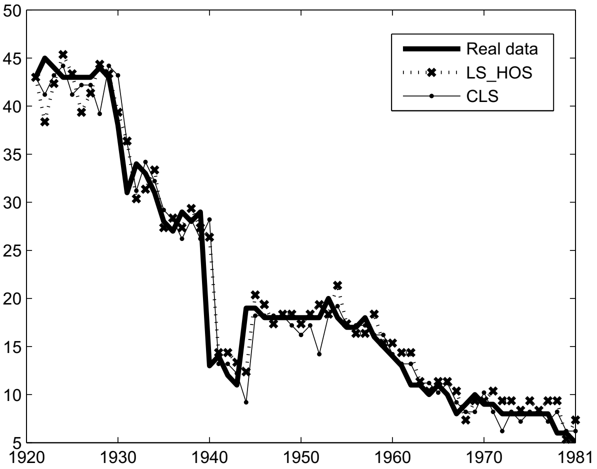

Fig. 2. The number of Swedish mechanical paper and pulp mills, from 1921 to 1981, and the fitted values considering the LS_HOS and CLS estimates.

Figure 2 presents the number of Swedish mechanical paper and pulp mills, from 1921 to 1981, used by Brännäs (1995) and Brännäs and Hellström (2001). These authors fitted an INAR(1) process to this dataset using some explanatory variables. Here, an INAR(1) process where the innovations are i.i.d. with mean μ_e and variance σ_e^2 is considered. Since the mean of the data is 20.40 and its variance is 155.16, a Poisson innovation process is not assumed but then the method does not require that or anyother assumption on the distribution of the innovations. Table 1 presents the parameter estimates obtained by CLS and LS_HOS methods. The fit of both models, based on LS_HOS and CLS estimates, are also shown in Figure 2. The mean square errors (MSE) between the observations and the fitted values are exhibited in Table 1. It can be seen that the MSE is slightly smaller for the LS_HOS fit than for CLS fit. The last two columns of the table present the mean and variance of the estimated models:

$$\hat{\mu}_x = \frac{\hat{\mu}_e}{1-\hat{\alpha}} \quad \text{and} \quad \hat{\sigma}_x^2 = \frac{(1-\hat{\alpha})(\hat{\mu}_e\hat{\alpha}+\hat{\sigma}_e^2)}{(1-\hat{\alpha})^2(1+\hat{\alpha})}.$$

It is noticeable that the model estimated by LS_HOS presents mean and variance closer to the sample values. The residuals from both fitted models are uncorrelated.

Method	$\hat{\alpha}$	$\hat{\mu}_e$	$\hat{\sigma}_e^2$	MSE	$\hat{\mu}_x$	$\hat{\sigma}_x^2$
CLS	0.9591	0.2017	15.2268	8.5494	4.9315	192.2764
LS_HOS	0.9269	1.3635	19.2253	7.4465	18.6525	145.4513

Table 1. The parameter estimates of the number of Swedish mechanical paper and pulp mills, from 1921 to 1981.

4 Final remarks

The principal advantage of HOS is the capability to detect and characterize the deviations from Gaussianity and non-linearity of the processes. Thus in this work a new estimation method for the parameters of INAR processes based on HOS is proposed. This method uses the Least Squares estimation to minimize the errors between the third-order moment of the observations and of the fitted model. A Monte Carlo study indicates that this estimation method provides good results in small samples, in terms of sample bias, variance and mean square error. Moreover, when used in the context of a non-Poisson real dataset the LS_HOS estimates provide a model with mean, variance and autocorrelations closer to the sample values.

Acknowledgments

For the first author, this work reports research developed under financial support provided by FCT - Fundação para a Ciência e Tecnologia, Portugal.

References

BRÄNNÄS, K. (1995): Explanatory variables in the AR(1) count data model. *Umeå Economic Studies 381.*

BRÄNNÄS, K. and HELLSTRÖM, J. (2001): Generalized integer-valued autoregression. *Econometric Reviews 20 (4), 425-443.*

GAUTHIER, G. and LATOUR, A. (1994): Convergence forte des estimateurs des paramtres dun processus GENAR(p). *Annales des Sciences Mathématiques du Québec 18, 49-71.*

LATOUR, A. (1998): Existence and stochastic structure of a non-negative integer-valued autoregressive process. *Journal of Time Series Analysis 19, 439-455.*

MATHWORKS (2004): Optimization toolbox user's guide for MATLAB. Available from `http://www.mathworks.com/access/helpdesk/help/pdf_doc/optim/optim_tb.pdf`

SILVA, I. (2005): *Contributions to the analysis of discrete-valued time series.* PhD Thesis. Universidade do Porto, Portugal.

SILVA, M.E. and OLIVEIRA, V.L. (2004): Difference equations for the higher-order moments and cumulants of the INAR(1) model. *Journal of Time Series Analysis 25, 317-333.*

SILVA, M.E. and OLIVEIRA, V.L. (2005): Difference equations for the higher-order moments and cumulants of the INAR(p) model. *Journal of Time Series Analysis 26, 17-36.*

STEUTEL, F.W. and VAN HARN, K. (1979): Discrete analogues of self-decomposability and stability. *The Annals of Probability 7, 893-899.*