Penalized Estimation for Integer Autoregressive Models

Konstantinos Fokianos

Abstract The integer autoregressive model of order p can be employed for the analysis of discrete–valued time series data. It can be shown, under some conditions, that its correlation structure is identical to that of the usual autoregressive process. The model is usually fitted by the method of least squares. However, consider an alternative estimation scheme, which is based on minimizing the least squares criterion subject to some constraints on the parameters of interest. The ridge type of constraints are used in this article and it is shown that under some reasonable conditions on the penalty parameter, the resulting estimates have less mean square error than that of the ordinary least squares. A real data set and some limited simulations support further the results.

1 Introduction

Ludwig Fahrmeir, whom this volume honors, has made seminal contributions to the statistical analysis of integer valued time series by promoting the idea of generalized linear models for inference. In particular, I would like to mention the articles Fahrmeir & Kaufmann (1985) and Fahrmeir & Kaufmann (1987) and the text Fahrmeir & Tutz (2001) which deal respectively with the following;

- the development of maximum likelihood estimation for the regression parameters of a generalized linear model with independent data for both canonical and non-canonical link functions,
- the extension of these results to categorical time series,
- the presentation of the above in a coherent piece of work.

Konstantinos Fokianos
Department of Mathematics & Statistics, University of Cyprus, PO BOX 20537, Nicosia, Cyprus,
URL: http://www.ucy.ac.cy/goto/mathstatistics/en-US/HOME.aspx,
e-mail: fokianos@ucy.ac.cy

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_18,

The results of these references have influenced considerably my research on time series, see e.g. Kedem & Fokianos (2002). On a more personal level, I wish to express my gratitude to Ludwig Fahrmeir for inviting me to Munich on a number of occasions and for giving me the opportunity to discuss with him several issues of mutual interest and to gain important insight.

Integer valued time series occur in diverse applications and therefore statistical methodology should be developed to take into account the discrete nature of the data. In this work, attention is focused on the so called integer autoregressive models of order p–denoted by INAR(p). These processes provide a class of models whose second order structure is identical to that of the standard AR(p) models and estimation can be carried out by standard least squares techniques.

The question of interest in this manuscript is whether the least squares estimators can become more efficient and under what conditions. It is shown that increase in efficiency can be achieved by introducing the so–called penalized least squares criterion (7) for estimation. In particular, it is shown that there are two cases that need to be considered. The first is when the true parameter vector that generates the process assumes "large" values componentwise; then minimization of (7) does not offer any improvement over the standard least squares estimators. On the other hand, when the true parameter vector values are assumed to be "small" , then it is possible to gain in efficiency. Here, the term efficiency, refers to mean square error improvement, since it is well known that penalized estimators are usually biased. The same phenomenon occurs in linear models theory, namely the method of ridge regression. It is well known that the mean square error of ridge estimators is less than the mean square error of the ordinary least squares estimators for some values of the ridge parameter. It is conjectured that the results carry over to the dependent data case under some reasonable assumptions. Some research advocating the use of shrinkage estimators in time series can be found in the recent article by Taniguchi & Hirukawa (2005).

When using penalized criteria for inference, there is an extra complexity introduced, that is the choice of the penalty parameter–see (7). It is a common practice to use cross–validation methods but their performance is questionable, especially in the time series context. Therefore, it is proposed to estimate the regularization parameter by using the AIC. Real data show-see Section 4–that a unique minimizer exists but the method requires more research.

The paper starts with Section 2 where INAR(p) processes are briefly reviewed and the least squares approach to the problem of estimation is discussed. The asymptotic distribution of least squares estimators is also stated. Section 3 introduces the penalized least squares estimator and discuss their asymptotic properties, see Theorems 2 and 3, which constitute the main results. Section 4 complements the presentation by some simulated and real data examples. The article concludes with some comments and an appendix.

2 Integer Autoregressive Processes and Inference

This section reviews briefly some probabilistic properties of the integer autoregressive processes and discuss estimation of unknown parameter by conditional least squares inference.

2.1 Integer Autoregressive Processes

Integer autoregressive processes have been introduced by Al-Osh & Alzaid (1987) and Alzaid & Al-Osh (1990) as a convenient way to transfer the usual autoregressive structure to discrete valued time series. The main concept towards this unification is given by the notion of thinning which is defined by the following:

Definition 1. Suppose that X is a non–negative integer random variable and let $\alpha \in [0,1]$. Then, the thinning operator, denoted by $\circ$, is defined as

$$\alpha \circ X = \begin{cases} \sum_{i=1}^{X} Y_i, & \text{if } X > 0, \\ 0, & \text{otherwise,} \end{cases}$$

where $\{Y_i\}$ is a sequence of independent and identically distributed Bernoulli random variables–independent of X–with success probability α. The sequence $\{Y_i\}$ is termed as a counting series.

Definition 1 allows for specification of the integer autoregressive process of order p. More specifically, suppose that for $i = 1,2,\ldots,p$, $\alpha_i \in [0,1)$ and let $\{\varepsilon_t\}$ be a sequence of independent and identically distributed nonnegative integer valued random variables with $\mathrm{E}[\varepsilon_t] = \mu$ and $\mathrm{Var}[\varepsilon_t] = \sigma^2$. The following process

$$X_t = \sum_{i=1}^{p} \alpha_i \circ X_{t-i} + \varepsilon_t, \tag{1}$$

is called integer autoregressive process of order p and is denoted by INAR(p). It should be noted that the Bernoulli variables used for defining the random variable $\alpha_1 \circ X_{t-1}$ are independent of those involved in the definition of $\alpha_2 \circ X_{t-2}$, and so on. This assumption guarantees that the INAR(p) process has the classical AR(p) correlation structure, see Du & Li (1991). A unique stationary and ergodic solution of (1) exists if

$$\sum_{j=1}^{p} \alpha_j < 1. \tag{2}$$

Various other authors have studied the above model, including Al-Osh & Alzaid (1987), Alzaid & Al-Osh (1990), McKenzie (1985), McKenzie (1986) and McKenzie (1988). Some very recent work extending the model in different directions can be found in the papers by Ferland et al. (2006), Neal & Subba Rao (2007), Zheng et al. (2006) and Zhu & Joe (2006).

2.2 Conditional Least Squares Inference

In what follows consider the INAR(p) model defined by (1). The $(p+1)$–parameter vector $\beta = (\mu, \alpha_1, \ldots, \alpha_p)'$ belongs to the $[0,\infty] \times [0,1)^p$ and it is usually estimated by conditional least squares method. Suppose that $\mathscr{F}_t$ is the σ–field generated by the past information, say $X_1, X_2, \ldots, X_t$. The conditional least squares estimator of β is calculated ny minimizing the following sum of squares:

$$S(\beta) = \sum_{t=p+1}^{N} (X_t - E(X_t|\mathscr{F}_{t-1}))^2 = \sum_{t=p+1}^{N} (X_t - \mu - \sum_{i=1}^{p} \alpha_i X_{t-i})^2. \tag{3}$$

Denote by $\hat{\beta}$ the value that minimizes the above expression and notice that standard arguments show that (see Brockwell & Davis (1991), for example)

$$\hat{\beta} = Q^{-1} r \tag{4}$$

where the $(p+1) \times (p+1)$ matrix Q is equal to

$$Q = \begin{bmatrix} N-p & \sum_{t=p+1}^{N} X_{t-1} & \cdots & \sum_{t=p+1}^{N} X_{t-p} \\ \sum_{t=p+1}^{N} X_{t-1} & \sum_{t=p+1}^{N} X_{t-1}^2 & \cdots & \sum_{t=p+1}^{N} X_{t-p} X_{t-1} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{t=p+1}^{N} X_{t-p} & \sum_{t=p+1}^{N} X_{t-1} X_{t-p} & \cdots & \sum_{t=p+1}^{N} X_{t-p}^2 \end{bmatrix},$$

and the $(p+1)$–dimensional vector r is defined by

$$r = \left(\sum_{t=p+1}^{N} X_t, \sum_{t=p+1}^{N} X_t X_{t-1}, \ldots, \sum_{t=p+1}^{N} X_t X_{t-p} \right)'.$$

Then the following theorem holds true for the estimator $\hat{\beta}$:

Theorem 1. (Du & Li 1991) Suppose that $\hat{\beta}$ is the conditional least squares estimator defined by means of minimizing (3) for the INAR(p) model (1). In addition, assume that the error process has $\mathrm{E}[\varepsilon_t] = \mu$, $\mathrm{Var}[\varepsilon_t] = \sigma^2$ and $\mathrm{E}[\varepsilon_t^3] < \infty$. Suppose that condition (2) is satisfied and let μ_x to denote the mean of the stationary distribution of the INAR(p) model (1). Then

$$\sqrt{N}\left(\hat{\beta} - \beta\right) \to N_{p+1}(0, V^{-1} W V^{-1}),$$

where the $(p+1) \times (p+1)$ matrix $V = [v_{ij}]$ is defined by

$$v_{ij} = \begin{cases} 1, & i = 1, j = 1, \\ \mu_x, & i = 1, j > 1 \text{ or } i > 1, j = 1, \\ \mathrm{E}[X_{p+1-i} X_{p+1-j}] + \mu_x^2, & i, j \geq 2. \end{cases}$$

Furthermore, the $(p+1) \times (p+1)$ matrix $W = [w_{ij}]$ is given by

$$w_{ij} = \begin{cases} \mathrm{E}[(X_{p+1} - \sum_{k=1}^{p} \alpha_k X_{k-p} - \mu)^2], & i = 1, j = 1, \\ \mathrm{E}[X_{p+1-i}(X_{p+1} - \sum_{k=1}^{p} \alpha_k X_{k-p} - \mu)^2] & j = 1, i > 1, \\ \mathrm{E}[X_{p+1-j}(X_{p+1} - \sum_{k=1}^{p} \alpha_k X_{k-p} - \mu)^2] & i = 1, j > 1, \\ \mathrm{E}[X_{p+1-i}X_{p+1-j}(X_{p+1} - \sum_{k=1}^{p} \alpha_k X_{k-p} - \mu)^2], & i, j \geq 2, \end{cases}$$

where expectation is taken with respect to the stationary distribution.

In addition, it can be shown that the estimator $\hat{\beta}$ is strongly consistent. Theorem 1 is proved by standard arguments from martingale theory, see Klimko & Nelson (1978) and Hall & Heyde (1980), for more. A consistent estimator of the matrix V is given by

$$\hat{V} = \frac{1}{N} Q. \tag{5}$$

Indeed, $\lim_{N\to\infty} \hat{V} = V$, in probability, because of the ergodicity of the process. Similarly, the matrix W is estimated by means of

$$\hat{W} = \frac{1}{N} \sum_{t=p+1}^{N} \left((X_t - \sum_{i=1}^{p} \alpha_i X_{t-i} - \mu)^2 \begin{bmatrix} 1 & X_{t-1} & \dots & X_{t-p} \\ X_{t-1} & X_{t-1}^2 & \dots & X_{t-p}X_{t-1} \\ \dots & \dots & \dots & \dots \\ X_{t-p} & X_{t-1}X_{t-p} & \dots & X_{t-p}^2 \end{bmatrix} \right). \tag{6}$$

Therefore, a consistent estimator of the asymptotic covariance matrix of $\hat{\beta}$ is given by $\hat{V}^{-1}\hat{W}\hat{V}^{-1}$–see theorem 1.

3 Penalized Conditional Least Squares Inference

We suggest estimation of the unknown parameter vector β of the INAR(p) by penalizing the conditional least square criterion with a quadratic penalty. As it is the case with the ridge regression, see Hoerl & Kennard (1970a), Hoerl & Kennard (1970b), it is anticipated that the mean square error of the estimates is minimized by some value of the ridge parameter. Therefore, the choice of the ridge (or regularization) parameter is important and its selection is taken up in Section 4.2 where a proposal is made by using the so–called AIC criterion; Akaike (1974). In the following, the first issue is to show how ridge inference proceeds and then apply the resulting estimators to the problem of prediction.

Ridge coefficients are defined by minimization of the following penalized sum of squares

$$\begin{aligned} S_p(\beta) &= S(\beta) + \lambda_N \sum_{j=1}^{p} \alpha_j^2 \\ &= \sum_{t=p+1}^{N} (X_t - \sum_{j=1}^{p} \alpha_j X_{t-j} - \mu)^2 + \lambda_N \sum_{j=1}^{p} \alpha_j^2 \end{aligned} \tag{7}$$

where $\lambda_N \geq 0$, is the so called regularization parameter. When $\lambda_N = 0$, the ordinary CLS estimator is obtained while if $\lambda_N \to \infty$ then all the coefficients shrink towards zero. An alternative way of obtaining the above penalized sum of squares is to postulate the constraint $\sum_{j=1}^{p} \alpha_j^2 \leq t$. Obviously the parameter t is inversely related with λ_N but both constraints are equivalent.

The penalized CLS estimator of $\hat{\beta}$ will be denoted by $\hat{\beta}^{\lambda}$ and it is easily obtained by

$$\hat{\beta}^{\lambda} = (Q + \lambda_N D_{p+1})^{-1} r. \tag{8}$$

The matrix Q and the vector r have been defined immediately after (4) and the $(p+1) \times (p+1)$ matrix D_{p+1} is given by

$$D_{p+1} = \begin{bmatrix} 0 & 0 \\ 0 & I_p \end{bmatrix},$$

where I_p is the diagonal matrix of order p. It is recognized that the penalized CLS estimator is of the same form as the ordinary ridge regression estimator. It is expected therefore that for a suitably chosen value of the regularization parameter, the mean square error of $\hat{\beta}^{\lambda}$ will be less or equal than that of $\hat{\beta}$. In what follows, it is shown that when the true parameter values are small, then a more efficient estimator–in the sense of mean square error–is obtained by means of minimizing (7) provided that the regularization parameter λ_N is of order N.

We study the asymptotic properties of $\hat{\beta}^{\lambda}$ in the following theorem whose proof is postponed in the appendix.

Theorem 2. Assume the same conditions as in Theorem 1. Assume further that λ_N is such that $\lambda_N/\sqrt{N} \to \lambda_0 \geq 0$. Then

$$\sqrt{N}(\hat{\beta}^{\lambda} - \beta) \to N_{p+1}(-\lambda_0 V^{-1} b, V^{-1} W V^{-1})$$

in distribution, as $N \to \infty$. The matrices V and W have been defined in Theorem 1 and the $(p+1)$–dimensional vector b is given by $b = (0, \alpha_1, \ldots, \alpha_p)'$.

The above theorem shows that when $N \to \infty$, then the penalized CLS (8) are asymptotically normal but biased while their asymptotic covariance matrix is given by the same formula that corresponds to the ordinary CLS estimators–see Theorem 1. Hence, there seems to be no particular improvement when using the penalized CLS estimator unless $\lambda_N = o(\sqrt{N})$, and this is in agreement with the asymptotic results for least squares regression with independent data obtained by Knight & Fu (2000, Th. 2). Theorem 2 implies that when the true parameter values are large and $\lambda_0 > 0$, then the bias of the restricted estimators might be of considerable magnitude.

Suppose now that the data are generated by the INAR(p) process (1) where the vector of unknown parameters satisfies

$$\beta_N = \beta + \frac{c}{\sqrt{N}},$$

for some vector of the form $c = (0, c_1, \ldots, c_p)'$, such that condition (2) is satisfied. Then the second part of the following theorem shows that for small α_j's there is a gain when using the ridge regression. The proof of the theorem is along the lines of Theorem 2 and therefore it is omitted.

Theorem 3. Assume the same conditions as in Theorem 1. Assume further that $\beta_N = \beta + c/\sqrt{N}$ where c is of the form $c = (0, c_1, \ldots, c_p)'$ such that condition (2) holds true. Let $\hat{\beta}^{\lambda}$ be the penalized CLS (8). Then

1. If $\lambda_N/\sqrt{N} \to \lambda_0 \geq 0$, then

$$\sqrt{N}(\hat{\beta}^{\lambda} - \beta) \to N_{p+1}(-\lambda_0 V^{-1} b, V^{-1} W V^{-1}),$$

   in distribution, as $N \to \infty$.
2. If $\alpha_i = 0$ for $i = 1, 2, \ldots, p$ so that $\beta = (\mu, 0, \ldots, 0)'$ and $\lambda_N/N \to \lambda_0 \geq 0$, then

$$\sqrt{N}(\hat{\beta}^{\lambda} - c/\sqrt{N}) \to N_{p+1}(-\lambda_0 \tilde{V}^{-1} c, \tilde{V}^{-1} W \tilde{V}^{-1}),$$

   in distribution, as $N \to \infty$, where $\tilde{V} = V + \lambda_0 D_{p+1}$.

The above notation is the same as that of Theorem 2.

The second part of the above theorem shows that for large sample sizes, the asymptotic distribution of $\hat{\beta}^{\lambda}$ is a multivariate normal provided that the choice of λ_N is of order N and the true parameter value is relatively small. In particular, certain choices of λ_N yield to consistent estimators which are asymptotically normally distributed. However, other choices of λ_N yield to biased estimators. We anticipate though that the bias will be small and regularization will provide estimates with smaller mean square error.

An estimator of the asymptotic covariance matrix is given by

$$\left(\hat{V} + \frac{\lambda_N}{N} D\right)^{-1} \hat{W} \left(\hat{V} + \frac{\lambda_N}{N} D\right)^{-1} \tag{9}$$

where all the matrices are evaluated at $\hat{\beta}^{\lambda}$. The matrices $\hat{V}$, $\hat{W}$ have been defined by (5) and (6), respectively. For large N, and if $\lambda_N = o(\sqrt{N})$, formula (9) reduces to that used for the asymptotic variance estimator of the conditional LS estimator $\hat{\beta}$-see Theorem 1.

4 Examples

A limited simulation study and a real data example are presented to complement the theoretical findings.

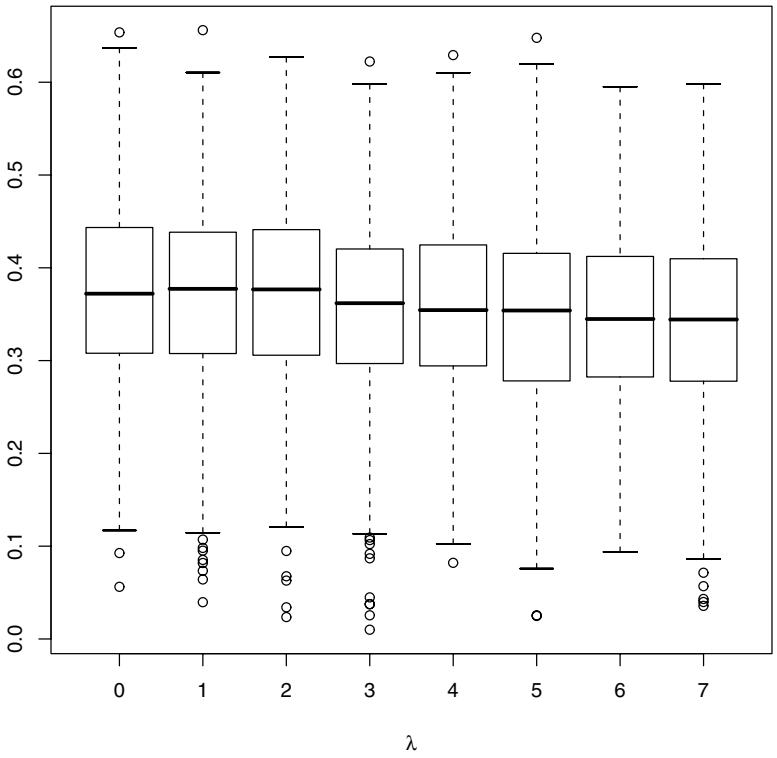

Fig. 1 Boxplots of the distribution of $\hat{\alpha}_1^{\lambda}$ for various values of the penalty parameter.

4.1 Simulations

To study the empirical performance of the penalized LS estimators for the INAR(p) model, a limited simulation study is presented. First, data are generated by the INAR(1) process

$$X_t = \alpha_1 \circ X_{t-1} + \varepsilon_t,$$

for $t = 1, 2, \ldots, N$, where the error sequence is assumed to be i.i.d. Poisson with mean μ. The computation were carried out by the statistical language `R` and all simulation output is based on 1000 runs.

The asymptotic normality of the restricted estimators is demonstrated for various values of the penalty parameter–see Figure 1–where the boxplots of the distribution of $\hat{\alpha}_1^{\lambda}$ are shown for $\lambda = 0, 1, 2, \ldots, 7$. The sample size is $N = 100$, $\alpha_1 = 0.4$ and $\mu = 1$. The asserted asymptotic normality is in agreement with the simulation findings.

Table 1 Penalized estimators for 100 observations from the INAR(2) model with true parameters $\alpha_1 = 0.1$ and $\alpha_2 = 0.2$ and for different values of the Poisson mean μ. The regularization parameter varies from 0 to 10 by 0.5 and the number of simulations is 1000.

λ	$\mu = 0.50$				$\mu = 1.00$			
	$\hat{\mu}^\lambda$	$\hat{\alpha}_1^\lambda$	$\hat{\alpha}_2^\lambda$	$e_1(\lambda)$	$\hat{\mu}^\lambda$	$\hat{\alpha}_1^\lambda$	$\hat{\alpha}_2^\lambda$	$e_1(\lambda)$
0.0	0.531	0.0852	0.171	1.000	1.07	0.0821	0.170	1.000
0.5	0.531	0.0831	0.172	0.947	1.06	0.0816	0.171	0.906
1.0	0.535	0.0835	0.159	0.958	1.05	0.0927	0.167	0.893
1.5	0.537	0.0802	0.166	0.947	1.05	0.0865	0.169	0.886
2.0	0.535	0.0808	0.163	0.944	1.08	0.0805	0.165	0.962
2.5	0.545	0.0779	0.159	0.998	1.08	0.0766	0.164	0.955
3.0	0.539	0.0789	0.158	0.918	1.07	0.0792	0.166	0.904
3.5	0.549	0.0788	0.151	0.946	1.08	0.0756	0.163	0.893
4.0	0.553	0.0785	0.152	0.915	1.08	0.0798	0.164	0.977
4.5	0.549	0.0738	0.150	0.926	1.07	0.0820	0.161	0.916
5.0	0.555	0.0731	0.147	0.948	1.07	0.0835	0.158	0.825
5.5	0.549	0.0715	0.145	0.905	1.08	0.0793	0.158	0.884
6.0	0.548	0.0723	0.147	0.893	1.08	0.0773	0.163	0.848
6.5	0.548	0.0757	0.144	0.891	1.09	0.0813	0.158	0.906
7.0	0.556	0.0721	0.143	0.900	1.10	0.0730	0.159	0.932
7.5	0.554	0.0757	0.144	0.896	1.08	0.0826	0.154	0.837
8.0	0.563	0.0740	0.135	0.960	1.09	0.0798	0.155	0.830
8.5	0.560	0.0699	0.140	0.897	1.10	0.0751	0.151	0.902
9.0	0.561	0.0738	0.130	0.915	1.10	0.0758	0.148	0.922
9.5	0.566	0.0688	0.138	0.926	1.10	0.0763	0.150	0.936
10.0	0.562	0.0697	0.133	0.901	1.10	0.0768	0.150	0.894

Notice that $\lambda = 0$ corresponds to the ordinary CLS estimators while for large values of λ, the resulting estimator is more biased compared to the CLS estimator.

Furthermore, consider the INAR(2) model

$$X_t = \alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + \varepsilon_t$$

where ε_t are assumed to be Poisson random variables with mean μ, as before. Table 1 shows the results of 1000 simulations when there are 100 observations available from the process at hand. Note that the resulting penalized estimators are biased as it was claimed before. However, their relative efficiency to the ordinary least squares estimators is superior in both cases considered. The quantity $e_1(\lambda)$–that is the efficiency–has been defined by

$$e_1(\lambda) = \frac{\text{MSE}(\hat{\beta}^\lambda)}{\text{MSE}(\hat{\beta})},$$

and it is the ratio of the mean square error of the constrained estimator to the mean square error of the unconstrained estimator. Table 2 shows the same results but for $N = 500$. Here, most of the values of $e_1(\lambda)$ fluctuate around unity showing that there is no any improvement by penalization.

Table 2 Penalized estimators for 500 observations from the INAR(2) model with true parameters $\alpha_1 = 0.1$ and $\alpha_2 = 0.2$ and for different values of the Poisson mean μ. The regularization parameter varies from 0 to 10 by 0.5 and the number of simulations is 1000.

λ	$\mu = 0.50$				$\mu = 1.00$			
	$\hat{\mu}^\lambda$	$\hat{\alpha}_1^\lambda$	$\hat{\alpha}_2^\lambda$	$e_1(\lambda)$	$\hat{\mu}^\lambda$	$\hat{\alpha}_1^\lambda$	$\hat{\alpha}_2^\lambda$	$e_1(\lambda)$
0.0	0.508	0.0945	0.193	1.000	1.01	0.0982	0.193	1.000
0.5	0.507	0.0970	0.194	0.995	1.01	0.0969	0.195	1.018
1.0	0.507	0.0963	0.192	1.111	1.01	0.0981	0.192	0.971
1.5	0.508	0.0968	0.193	1.034	1.01	0.0979	0.193	1.032
2.0	0.508	0.0971	0.191	1.022	1.01	0.0957	0.193	1.055
2.5	0.509	0.0949	0.191	1.044	1.01	0.0964	0.193	0.993
3.0	0.508	0.0962	0.192	1.084	1.01	0.0946	0.194	0.973
3.5	0.509	0.0978	0.188	1.067	1.02	0.0954	0.192	1.008
4.0	0.509	0.0941	0.189	1.018	1.02	0.0977	0.190	1.013
4.5	0.513	0.0921	0.189	1.019	1.02	0.0943	0.191	1.074
5.0	0.511	0.0955	0.191	1.072	1.02	0.0964	0.192	1.044
5.5	0.510	0.0972	0.187	1.036	1.02	0.0977	0.192	1.068
6.0	0.511	0.0955	0.186	1.003	1.02	0.0962	0.191	0.950
6.5	0.517	0.0921	0.185	1.084	1.02	0.0965	0.191	0.984
7.0	0.514	0.0948	0.187	1.009	1.03	0.0934	0.188	1.008
7.5	0.513	0.0936	0.187	1.004	1.02	0.0962	0.191	1.061
8.0	0.514	0.0924	0.186	0.978	1.02	0.0945	0.188	0.964
8.5	0.517	0.0917	0.183	1.002	1.02	0.0967	0.187	0.990
9.0	0.512	0.0936	0.187	0.955	1.02	0.0975	0.190	0.970
9.5	0.516	0.0944	0.182	1.044	1.02	0.0975	0.189	1.007
10.0	0.519	0.0910	0.183	1.054	1.02	0.0919	0.188	1.041

4.2 Data Example

The Westgren gold particle data is used to demonstrate the penalized least squares estimation method. The data consists of consecutive count measurements of gold particles in a well defined colloidal solution of equally spaced points in time. These data have been analyzed by various authors, including Guttorp (1991), Grunwald et al. (2000) and more recently by Jung & Tremayne (2006). In particular, the first 370 observations are used throughout the subsequent analysis, along the lines of Jung & Tremayne (2006).

To analyze the data, consider the INAR(p) model (1) for $p = 1,2,3,4$. For comparison purposes, which are described below, the first four observations are removed and all models were fitted based on the 366 observations. Figure 2 shows the values of AIC for each INAR model fitted to the data, defined as

$$\mathrm{AIC}(p,\lambda) = 366 \log\left(\frac{\sum_{t=5}^{370}(X_t - \hat{\mu}^\lambda - \sum_{i=1}^{p}\hat{\alpha}_i^\lambda X_{t-i})^2}{366}\right) + 2\mathrm{df}_\lambda \qquad (10)$$

where the quantity df_λ is called the effective degrees of freedom as in the ordinary ridge regression. In other words, recall the definition of Q from (4) and set

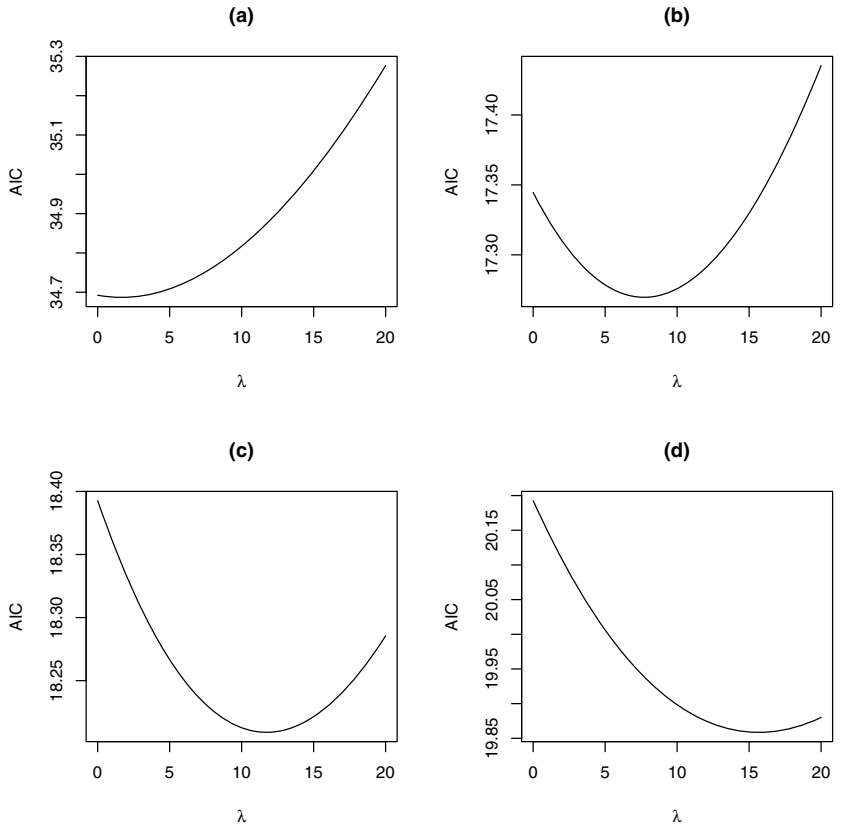

Fig. 2 Selection of λ by AIC for the gold particle data. (a) INAR(1), (b) INAR(2), (c) INAR(3), (d) INAR(4).

$$X = \begin{bmatrix} 1 & X_p & X_{p-1} & \dots & X_1 \\ 1 & X_{p+1} & X_p & \dots & X_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n-1} & X_{n-2} & \dots & X_{n-p} \end{bmatrix}.$$

it is clear that $Q = X'X$ and therefore the effective degrees of freedom are defined by

$$\mathrm{df}_\lambda = \mathrm{tr}\left(X(Q + \lambda D_{p+1})^{-1}X'\right),$$

where $\mathrm{tr}(.)$ denotes the trace of a matrix. Although estimation proceeds from the least squares, it can be argued that the AIC is the expected Kullback–Leibler distance of the maximum Gaussian likelihood model relative to the true distribution of the process, see Brockwell & Davis (1991, p. 306).

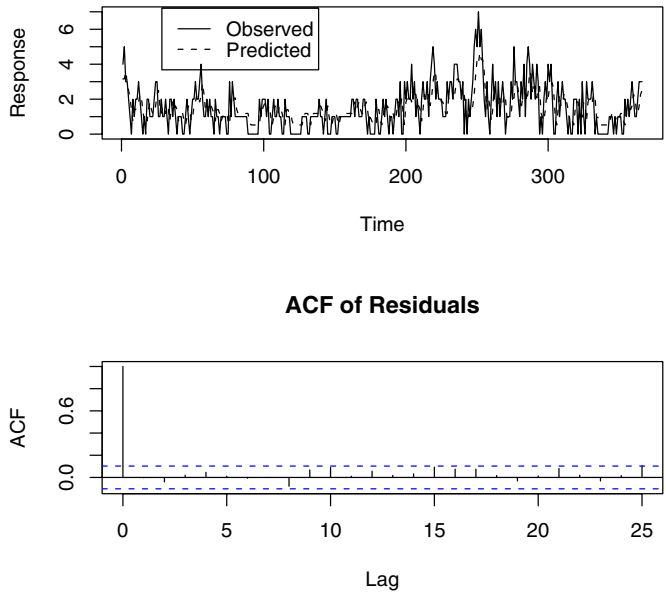

Fig. 3 Diagnostics for the INAR(2) model fitted to the gold particle data by minimizing (7) using $\lambda = 7.65$.

Using the above definition, and turning back to Figure 2, we note that the plot suggests the existence of a value of λ such that (10) attains a minimum. Notice that the values of the penalty parameter λ varies between 0 and 20 for a fine grid of values. When comparing the AIC from all different models, the INAR(2) yields its minimum value–in fact for all λ. Therefore the point that was made by previous authors that the INAR(2) model fits these data well is iterated further–see Jung & Tremayne (2006).

Hence this model is used for data fitting at the value of λ that minimizes (10). It turns out that $\lambda_{\text{opt}} = 7.65$ and the corresponding estimators are given by $\hat{\alpha}_1^{7.65} = 0.43082$, $\hat{\alpha}_2^{7.65} = 0.22685$ and $\hat{\mu}^{7.65} = 0.52510$. Figure 3 shows some further diagnostics for the model at hand. The upper panel shows a plot of the observed versus the predicted data while the lower plot shows the autocorrelation function of residuals. Both graphs indicate the adequacy of the INAR(2) model.

5 Discussion

This article introduces the ridge regression idea to the INAR processes. It was shown by theory and some supporting simulations that improvement over ordinary CLS is possible given a good choice of the regularization parameter. The choice of the regularization parameter is based on the minimization of the AIC and it was shown that for the Westgren gold particle data the method appears to work nicely. However, further investigation is needed to understand the results obtained from such procedure.

Integer autoregressive models have been generalized in different directions by several authors. For instance, Latour (1998) studies generalized integer valued autoregressive models of order p. This class of models is based on generalization of the thinning operator but their second order properties are similar to those of INAR(p) models. Hence the results reported here should be applicable in this class of models as well.

Another interesting class of models is that of conditional linear AR(1) models (see Grunwald et al. (2000)) specified by the following

$$m(X_t) = \alpha_1 X_{t-1} + \mu,$$

where $m(X_t) = \mathrm{E}[X_t \mid X_{t-1}]$, with X_t a time–homogeneous first-order Markov process. This class of model includes several AR(1) models proposed in the literature for non–Gaussian data. Inference is carried out either by maximum likelihood or by least squares. Therefore, the proposed ridge methods should apply to those models as well.

In a different direction, the recent contribution of Zhu & Joe (2006) extends the INAR(p) to include covariates. Estimation of regression coefficients is based on maximum likelihood and therefore the ridge constraints can be easily incorporated so that (7) is of the form of maximizing a penalized log likelihood function.

As a final remark, alternative penalties can be used so that model selection can be combined with estimation. For instance, consider penalty function of the following form

$$J(\beta) = \sum_{j=1}^{p} |\beta_j|^q,$$

where $q > 0$. The choices of $q = 1, 2$ yield to the Lasso (Tibshirani 1996) and ridge estimators respectively. In general these estimators were introduced by Frank & Friedman (1993) and were termed as Bridge estimators. When $q \leq 1$, the penalty function has the neat property to set some of the regression coefficient equal to 0, that is it can be used for model selection and estimation.

References

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **AC–19**: 716–723.

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

Alzaid, A. A. & Al-Osh, M. (1990). An integer-valued *p*th-order autoregressive structure (INAR(*p*)) process, *Journal of Applied Probability* **27**: 314–324.

Brockwell, P. J. & Davis, R. A. (1991). *Time Series: Data Analysis and Theory*, 2nd edn, Springer, New York.

Du, J. G. & Li, Y. (1991). The integer-valued autoregressive INAR(p) model, *Journal of Time Series Analysis* **12**: 129–142.

Fahrmeir, L. & Kaufmann, H. (1985). Consistency and asymptotic Normality of the maximum likelihood estimates in generalized linear models, *Annals of Statistics* **13**: 342–368.

Fahrmeir, L. & Kaufmann, H. (1987). Regression Models for Nonstationary Categorical Time Series, *Journal of Time Series Analysis* **8**: 147–160.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd edn, Springer, New York.

Ferland, R., Latour, A. & Oraichi, D. (2006). Integer–valued GARCH processes, *Journal of Time Series Analysis Analysis* **27**: 923–942.

Frank, I. E. & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools, *Technometrics* **35**: 109–148. (with discussion).

Grunwald, G. K., Hyndman, R. J., Tedesco, L. & Tweedie, R. L. (2000). Non-Gaussian conditional linear AR(1) models, *Australian & New Zealand Journal of Statistics* **42**: 479–495.

Guttorp, P. (1991). *Statistical Inference for Branching Processes*, Wiley, New York.

Hall, P. & Heyde, C. C. (1980). *Martingale Limit Theory and its Applications*, Academic Press, New York.

Hoerl, A. E. & Kennard, R. W. (1970a). Ridge regression:Applications to non–orthogonal problems, *Technometrics* **12**: 69–82.

Hoerl, A. E. & Kennard, R. W. (1970b). Ridge regression:Biased estimation for non–orthogonal problems, *Technometrics* **12**: 55–67.

Jung, R. C. & Tremayne, A. R. (2006). Coherent forecasting in integer time series models, *International Journal of Forecasting* **22**: 223–238.

Kedem, B. & Fokianos, K. (2002). *Regression Models for Time Series Analysis*, Wiley, Hoboken, NJ.

Klimko, L. A. & Nelson, P. I. (1978). On conditional least squares estimation for stochastic processes, *The Annals of Statistics* **6**: 629–642.

Knight, K. & Fu, W. (2000). Asymptotics for lasso–type estimators, *Annals of Statistics* **28**: 1356–1378.

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

McKenzie, E. (1985). Some simple models for discrete variate time series, *Water Resources Bulletin* **21**: 645–650.

McKenzie, E. (1986). Autoregressive moving-average processes with negative-binomial and geometric marginal distributions, *Advances in Applied Probability* **18**: 679–705.

McKenzie, E. (1988). Some ARMA models for dependent sequences of Poisson counts, *Advances in Applied Probability* **20**: 822–835.

Neal, P. & Subba Rao, T. (2007). MCMC for integer–valued ARMA processes, *Journal of Time Series Analysis* **28**: 92–100.

Taniguchi, M. & Hirukawa, J. (2005). The Stein-James estimator for short- and long-memory Gaussian processes, *Biometrika* **92**: 737–746.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B* **58**: 267–288.

Zheng, H., Basawa, I. V. & Datta, S. (2006). Inference for the *p*th–order random coefficient integer–valued process, *Journal of Time Series Analysis* **27**: 411–440.

Zhu, R. & Joe, H. (2006). Modelling count data time series with Markov processes based on binomial thinning, *Journal of Time Series Analysis* **27**: 725–738.

Appendix

Suppose that $M_N^0 = -2^{-1}(\partial S_p(\beta)/\partial\mu) = \sum_{t=1}^{N}(X_t - \sum_{i=1}^{p}\alpha_i X_{t-i} - \mu)$ and put $M_0^0 = 0$. Then

$$\begin{aligned}\mathrm{E}(M_N^0|\mathscr{F}_{N-1}) &= \mathrm{E}\left(M_{N-1}^0 + X_N - \sum_{i=1}^{p}\alpha_i X_{N-i} - \mu|\mathscr{F}_{N-1}\right)\\ &= M_{N-1}^0 + \mathrm{E}\left(X_N - \sum_{i=1}^{p}\alpha_i X_{N-i} - \mu|\mathscr{F}_{N-1}\right) = M_{N-1}^0,\end{aligned}$$

from the properties of the INAR(p) processes. Thus, the sequence $\{M_N^0, \mathscr{F}_N, N \geq 0\}$ forms a martingale which is square integrable. Furthermore, if condition (2) is fulfilled, then the sequence X_t is stationary and ergodic. Hence, from the ergodic theorem,

$$\frac{1}{N}\sum_{t=p+1}^{N}\left(X_t - \sum_{i=1}^{p}\alpha_i X_{t-i} - \mu\right)^2 \to \mathrm{E}\left(X_{p+1} - \sum_{i=1}^{p}\alpha_i X_{p-i} - \mu\right)^2 \equiv \sigma_1^2,$$

almost surely, as $N \to \infty$. Therefore, by (Hall & Heyde 1980, Cor. 3.2) we obtain that

$$\frac{1}{\sqrt{N}}M_N^0 \to N(0, \sigma_1^2),$$

in distribution, as $N \to \infty$. Along the same lines, it can be shown that if $M_N^j = -2^{-1}(\partial S_p(\beta)/\partial\alpha_j) = \sum_{t=p+1}^{N} X_{t-j}(X_t - \sum_{i=1}^{p}\alpha_i X_{t-i} - \mu) - \lambda_N\alpha_j$, for $j = 1, 2, \ldots, p$, then $\tilde{M}_N^j = M_N^j + \lambda_N\alpha_j$ is a martingale that satisfies

$$\frac{1}{N}\sum_{t=p+1}^{N} X_{t-j}^2\left(X_t - \sum_{i=1}^{p}\alpha_i X_{t-i} - \mu\right)^2 \to \mathrm{E}\left(X_{p+1-j}^2(X_{p+1} - \sum_{i=1}^{p}\alpha_i X_{p-i} - \mu)^2\right) \equiv \sigma_j^2,$$

almost surely, and

$$\frac{1}{\sqrt{N}}\tilde{M}_N^j \longrightarrow N(0, \sigma_j^2)$$

for all $j = 1, \ldots, p$. Using the assumption that $\lambda_N/\sqrt{N} \to \lambda_0 \geq 0$ then

$$\frac{1}{\sqrt{N}} M_N^j \longrightarrow N(-\lambda_0 \alpha_j, \sigma_j^2).$$

By the Cramer-Wold device and the properties of the INAR(p) process, it can be shown that

$$\frac{1}{\sqrt{N}} \begin{pmatrix} M_N^0 \\ M_N^1 \\ \vdots \\ M_N^p \end{pmatrix} \rightarrow N_{p+1}\left(-\lambda_0 b, W\right),$$

in distribution, as $N \rightarrow \infty$.

Recall the penalized conditional least squares estimators, given by (8). It can be shown that

$$\sqrt{N}(\hat{\beta}^{\lambda} - \beta) = \left(\frac{1}{N} Q + \frac{\lambda_N}{N} D_{p+1}\right)^{-1} \frac{1}{\sqrt{N}} \begin{pmatrix} M_N^0 \\ M_N^1 \\ \vdots \\ M_N^p \end{pmatrix} \rightarrow N_{p+1}(-\lambda_0 V^{-1} b, V^{-1} W V^{-1}),$$

in distribution, as $N \rightarrow \infty$. The theorem is proved.