Poisson autoregressive process modeling via the penalized conditional maximum likelihood procedure

Xinyang Wang[1], Dehui Wang*[1] and Haixiang Zhang[2]

[1]Mathematics School of Jilin University, Changchun 130012, China
[2]Center for Applied Mathematics, Tianjin University, Tianjin 300072, China

Abstract

In this paper, we consider the penalized estimation procedure for Poisson autoregressive model with sparse parameter structure. We study the theoretical properties of penalized conditional maximum likelihood (PCML) with several different penalties. We show that the penalized estimators perform as well as the true model was known. We establish the oracle properties of PCML estimators. Some simulation studies are conducted to verify the proposed procedure. A real data example is also provided.

Keywords: Integer-valued time series ; Penalty function; Poisson autoregressive ; Oracle properties.

1 Introduction

In recent years, integer-valued time series is playing an important role because this kind of data is very popular in practice. For example, the number of patients in a hospital, the daily number of transaction in stock market, the monthly number of insurance claim , and so on. Generally speaking, integer-valued time series includes two groups: (A) 'thinning' models which are based on the thinning operator (Steutel and van Harn, 1979). Al-Osh and Alzaid (1987) proposed the first-order integer-valued autoregressive (INAR) process; Zheng et al. (2006) proposed the random coefficient integer-valued autoregressive (RCINAR) process; Zhang et al. (2010) considered a series of integer-valued autoregressive processes based on the signed generalized power series thinning operator. (B) state-space models. Davis et al. (2003) introduced a general class of observation-driven models for Poisson counts process and derived some important properties; Ferland et al. (2006) proposed an integer-valued GARCH model and studied its maximum likelihood estimation; Fokianos et al. (2009) studied the likelihood-based inference and geometric ergodicity for linear and nonlinear Poisson autoregressive model; Zhu and Wang (2011) studied statistical inference for the Poisson autoregressive model.

*The corresponding address: wangdh@jlu.edu.cn

There are very limited literatures concerning with penalized estimation procedure for integer-valued time series. In practice, some parameters are exactly zero when the order of integer-valued time series model is large. However, many traditional estimation methods (e.g. conditional maximum likelihood) fail to accurately estimate these parameters. To deal with this problem, we use the PCML method to estimate the parameters in Poisson autoregressive model with sparse structure. In time series analysis, Wang et al. (2007a) proposed the modified LASSO to select the significant parameters of linear regression with autoregressive errors model. Nardi and Rinaldo (2011) studied the LASSO procedure for autoregressive model with double asymptotic framework. Yoon et al. (2013) investigated the properties of LASSO estimator in autoregressive regression model. Based on the previous results about penalized estimation for autoregressive model, it is feasible to apply penalized estimation method to model sparse Poisson autoregressive data.

The rest of the paper is organized as follows: In Section 2, we introduce the Poisson autoregressive model and the PCML estimator. We also review some basic properties of the Poisson autoregressive model. In Section 3, we define some notations and give the theoretical properties of PCML estimator. Simulation studies are given in Section 4 and a real data example is presented in Section 5. Some concluding remarks are given in Section 6. All proof details are reported in the Appendix.

2 Penalized conditional maximum likelihood

First, we review the Poisson autoregressive model, which is defined as follows

$$\begin{cases} X_t \big| \mathcal{F}_{t-1} : \mathcal{P}(\gamma_t), \qquad \forall t \in \mathbb{Z}, \\ \gamma_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}, \end{cases} \tag{1}$$

where $\mathcal{F}_{t-1}$ is the σ-field generated by $\{X_{t-1}, X_{t-2}, \ldots\}$ and $\alpha_0 > 0$, $\alpha_i \geq 0$, $i = 1, \ldots p$. For convenience, let $\boldsymbol{\theta} = (\alpha_0, \alpha_p, \ldots, \alpha_1)^{\mathrm{T}}$, where 'T' denotes the transpose. Following Zhu and Wang (2011), the conditional log-likelihood function for model (1) is

$$L_n(\boldsymbol{\theta}) = \sum_{t=1}^{n} l_n(\boldsymbol{\theta}) = \sum_{t=1}^{n} \Big(X_t \ln \gamma_t - \gamma_t - \ln(X_t!) \Big). \tag{2}$$

Motivated by (3.14) in Fan and Lv (2010), we propose the PCML function as

$$Q_n(\boldsymbol{\theta}) = L_n(\boldsymbol{\theta}) - n \sum_{i=1}^{p+1} P_\lambda(|\theta_i|), \tag{3}$$

where $L_n(\boldsymbol{\theta})$ is defined in (2), and $P_\lambda(\cdot)$ is a penalty function. Here we pay attention to the following four penalty functions, which have the oracle property (Fan and Li, 2001):

(P.1) The SCAD penalty function (Fan and Li, 2001) is defined by

$$P_{\lambda,a}(|\theta_i|) = \begin{cases} \lambda|\theta_i|, & |\theta_i| \leq \lambda, \\ -(\theta_i^2 - 2a\lambda|\theta_i| + \lambda^2)/[2(a-1)], & \lambda < |\theta_i| \leq a\lambda, \\ (a+1)\lambda^2/2, & |\theta_i| > a\lambda. \end{cases}$$

where $\lambda > 0$ is the tuning parameter, and $a > 2$ is the shape parameter.

(P.2) The adaptive LASSO (Zou, 2006) is defined as $P_\lambda(|\theta_i|) = \lambda w_i|\theta_i|$, where $\lambda > 0$ is the tuning parameter. Of note that the weight w_i is defined as $w_i = 1/|\bar{\theta}_i|$, where $\bar{\theta}_i$ is the CML estimator.

(P.3) The MCP (Zhang, 2010) is defined as follows,

$$P_{\lambda,\delta}(|\theta_i|) = \lambda\big\{|\theta_i| - \frac{|\theta_i|^2}{2\delta\lambda}\big\}I(0 \leq |\theta_i| < \delta\lambda) + \frac{\lambda^2\delta}{2}I(|\theta_i| \geq \delta\lambda),$$

where $\lambda > 0$ is the tuning parameter, and $\delta > 0$ is the shape parameter.

(P.4) Dicker et al. (2013) gave the definition of SELO function,

$$P_{\lambda,\tau}(|\theta_i|) = \frac{\lambda}{\log(2)}\log\big(\frac{|\theta_i|}{|\theta_i| + \tau} + 1\big),$$

where $\lambda > 0$ is the tuning parameter, and $\tau > 0$ is the shape parameter.

The PCML estimator is obtain by maximizing $Q_n(\theta)$, which is defined as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q_n(\boldsymbol{\theta}),$$

where $Q_n(\theta)$ is defined in (3). In the next section, we will study some theoretical properties of the PCML estimator, which include the consistence and oracle properties.

3 Consistence and oracle property

Without loss of generality, we assume that $\boldsymbol{\theta_0} = (\boldsymbol{\theta_{10}}, \boldsymbol{\theta_{20}})^{\mathrm{T}}$ with $\boldsymbol{\theta_{20}} = \mathbf{0}$, where $\boldsymbol{\theta_0} = (\alpha_0^0, \alpha_p^0, \ldots, \alpha_1^0)^{\mathrm{T}}$ is the true value for the parameter of interest in model (1). Define $\mathbf{b} = (\dot{P}_{\lambda_n}(|\theta_1^0|)\mathrm{sgn}(\theta_1^0), \ldots, \dot{P}_{\lambda_n}(|\theta_s^0|)\mathrm{sgn}(\theta_s^0))^{\mathrm{T}}$, and $\boldsymbol{\Lambda} = \mathrm{diag}\{\ddot{P}_{\lambda_n}(|\theta_1^0|), \ldots \ddot{P}_{\lambda_n}(|\theta_s^0|)\}$, where s is the number of components in $\boldsymbol{\theta_{10}}$, $\dot{P}(\cdot)$ and $\ddot{P}(\cdot)$ denote the first and second derivative of the penalty function $P(\cdot)$, respectively. Furthermore, we introduce some notations as $a_n = \max_{1\leq i\leq p+1}\{\dot{P}_{\lambda_n}(|\theta_i^0|), \theta_i^0 \neq 0\}$ and $b_n = \max_{1\leq i\leq p+1}\{\ddot{P}_{\lambda_n}(|\theta_i^0|), \theta_i^0 \neq 0\}$. To study the theoretical properties of $\hat{\boldsymbol{\theta}}$, we need the following regularity conditions:

(C.1) $0 < \sum_{i=1}^{p} \alpha_i < 1$; (C.2) $a_n = O(n^{-1/2})$; (C.3) $b_n = o(1)$.

Here (C.1) ensures that $\{X_t\}$ is strictly stationary and ergodic (Doukhan et al. 2012), which is used for the asymptotic properties of $\hat{\boldsymbol{\theta}}$. Zhu and Wang (2011) proved that for any positive integer m, $E(X_t^m) < \infty$ if and only if (C.1) holds. (C.2) is to ensure that the estimator is $\sqrt{n}$-consistent. (C.3) is used to make sure that the influence of penalty function does not exceed that of CML criterion function on the resulting estimator. To check the rationality of (C.2) and (C.3) with SCAD penalty (other penalties are similar), by some calculation we can derive that $a_n = \max_{1\le i\le p+1}\{\lambda_n I(|\theta_i| \le \lambda_n) + \frac{(a\lambda_n - |\theta_i|)_+}{a-1} I(|\theta_i| > \lambda_n), \theta_i \ne 0\}$ and $b_n = \max_{1\le i\le p+1}\{-(a-1)^{-1} I(\lambda_n < |\theta_i| < a\lambda_n), \theta_i \ne 0\}$. Then, the classical condition for penalty-based procedure (Fan and Li, 2001) with $\lambda_n = O(n^{-1/2})$ can ensure the (C.2) and (C.3) hold. Of note the notation λ_n is used to indicate that the tuning parameter λ depends on the sample size n. The performance of the tuning parameter λ_n will be studied via simulation in the Section 4. Now we focus on the properties of PCML estimator, which are given below.

Theorem 1. *Under the conditions (C.1)-(C.3), there exists a local maximizer $\hat{\boldsymbol{\theta}}$ of $Q_n(\boldsymbol{\theta})$ such that $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}\| = O_p(n^{-1/2} + a_n)$.*

The above Theorem 1 implies that there exists a $\sqrt{n}$-consistent estimator for $\boldsymbol{\theta_0}$. To establish the sparsity of PCML estimator, we need the following lemma.

Lemma 1. *We assume that $\liminf_{n\to\infty} \liminf_{\boldsymbol{\theta}\to 0^+} \lambda_n^{-1} \dot{P}_{\lambda_n}(|\theta_j|) > 0$ and the conditions (C.1)-(C.3) hold, so with probability tending to 1, for any given $\boldsymbol{\theta_1}$ satisfying $\|\boldsymbol{\theta_1} - \boldsymbol{\theta_{10}}\| = O_p(n^{-1/2})$ and any constant $\eta > 0$, we have*

$$Q_n\left\{\begin{pmatrix}\boldsymbol{\theta_1}\\ \boldsymbol{0}\end{pmatrix}\right\} = \max_{\|\boldsymbol{\theta_2}\|\le \eta n^{-1/2}} Q_n\left\{\begin{pmatrix}\boldsymbol{\theta_1}\\ \boldsymbol{\theta_2}\end{pmatrix}\right\}.$$

Note that the assumption $\liminf_{n\to\infty} \liminf_{\boldsymbol{\theta}\to 0^+} \lambda_n^{-1} \dot{P}_{\lambda_n}(|\theta_j|) > 0$ is mild, since $\liminf_{n\to\infty} \liminf_{\boldsymbol{\theta}\to 0^+} \lambda_n^{-1} \dot{P}_{\lambda_n}(|\theta_j|) = 1$ (SCAD penalty; other cases are similar). We discuss the oracle property of PCML estimator. The oracle property, proposed by Fan and Li (2001), means that the penalized estimation method performs as well as if the true model was known in advance. Specifically, it can identify the right subset model and has the optimal estimation rate (Zou, 2006), which has been argued (Fan and Li 2001; Fan and Peng, 2004; Zhang, 2010) that a good procedure should have this oracle property. The following theoretical result provides the oracle property of PCML estimator $\hat{\boldsymbol{\theta}}$.

Theorem 2. (Oracle Property) *Under the conditions (C.1)-(C.3), with probability tending to 1, the root-n consistent estimate in Theorem 1 satisfies:*

(i) Sparsity : $\hat{\boldsymbol{\theta}}_{\boldsymbol{2}} = \boldsymbol{0}$;

(ii) Asymptotic normality :

$$\sqrt{n}\left(\boldsymbol{\Sigma}^s(\boldsymbol{\theta_0}) + \boldsymbol{\Lambda}\right)\left\{\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{1}} - \boldsymbol{\theta_{10}}\right) + \left(\boldsymbol{\Sigma}^s(\boldsymbol{\theta_0}) + \boldsymbol{\Lambda}\right)^{-1}\boldsymbol{b}\right\} \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{\Sigma}^s(\boldsymbol{\theta_0})),$$

where '$\xrightarrow{D}$' denotes convergence in distribution, $\boldsymbol{\Sigma}^s(\boldsymbol{\theta_0})$ represents the Fisher information ($\boldsymbol{\Sigma}(\boldsymbol{\theta_0})$ is defined in Lemma A.1) with $\boldsymbol{\theta_{20}} = \boldsymbol{0}$.

Note that $\{X_t\}$ is a strictly stationary and ergodic process, then by ergodic theorem, we have that $\hat{\mathbf{\Sigma}}^s(\boldsymbol{\theta_0}) = \frac{1}{n}\sum_{t=1}^{n}\frac{1}{\gamma_t}\boldsymbol{Y}_t^s\boldsymbol{Y}_t^{s\mathrm{T}} \xrightarrow{a.s.} \mathbf{\Sigma}^s(\boldsymbol{\theta_0})$, where $\boldsymbol{Y}_t^s = (1, X_{t-p}, \ldots, X_{t-p+s-2})^{\mathrm{T}}$. Therefore, the consistent covariance estimate for $\hat{\boldsymbol{\theta}}_{\mathbf{1}}$ is $\frac{1}{n}(\hat{\mathbf{\Sigma}}^s(\boldsymbol{\theta_0}) + \mathbf{\Lambda})^{-1}\hat{\mathbf{\Sigma}}^s(\boldsymbol{\theta_0})(\hat{\mathbf{\Sigma}}^s(\boldsymbol{\theta_0}) + \mathbf{\Lambda})^{-1}$.

4 Simulation studies

In this section, we report some numerical results to check the performance of PCML estimator for the Poisson autoregressive model with the help of R software. In our simulation, we use the local quadratic approximation (Fan and Li, 2001) to derive the PCML estimator. Because the penalty functions are singular at the origin and non-differentiable at the origin respect to $\boldsymbol{\theta}$. Suppose that a given initial $\boldsymbol{\theta^{(0)}}$ is close to the true value $\boldsymbol{\theta_0}$ (e.g. the CML estimator), Fan and Li (2001) proposed that the first order derivative of the penalty function can be locally approximated by $\dot{P}_\lambda(|\theta_i|) = \dot{P}_\lambda(|\theta_i|)\mathrm{sgn}(\theta_i) \approx \{\dot{P}_\lambda(|\theta_i^{(0)}|)/|\theta_i^{(0)}|\}\theta_i$. In other words,

$$P_\lambda(|\theta_i|) \approx P_\lambda(|\theta_i^{(0)}|) + \frac{1}{2}\left\{\dot{P}_\lambda(|\theta_i^{(0)}|)/|\theta_i^{(0)}|\right\}\left(\theta_i^2 - (\theta_i^{(0)})^2\right), \quad \text{for } \theta_i \approx \theta_i^{(0)}.$$

Then, for $k = 1, 2, \ldots$, we can repeatedly solve

$$\boldsymbol{\theta^{(k+1)}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left\{L_n(\boldsymbol{\theta}) - n\sum_{i=1}^{p+1}\frac{\dot{P}_\lambda(|\theta_i^{(k)}|)}{2|\theta_i^{(k)}|}\theta_i^2\right\} \tag{4}$$

until the sequence of $\{\boldsymbol{\theta^{(k)}}\}$ converges. We use BIC to determine the optimal tuning parameter λ in our procedure. The BIC criterion (Wang et al. 2007b) is

$$\mathrm{BIC}(\lambda) = \log\left(\frac{\sum_{t=1}^{n} S(\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}})^2}{n - \hat{\mathrm{df}}}\right) + \frac{\log(n)}{n}\hat{\mathrm{df}}, \tag{5}$$

where $S(\boldsymbol{\theta}) = X_t - \sum_{i=1}^{p}\alpha_i X_{t-i} - \alpha_0$, $\hat{\mathrm{df}}$ is the number of non-zero components of $\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}$. Then we choose the tuning parameter $\hat{\lambda}$ which minimizes $\mathrm{BIC}(\lambda)$ and the PCML estimator is $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\hat{\boldsymbol{\lambda}}}$.

To investigate the performance of the PCML estimator, as suggested by one referee, we consider two simulation studies:

Case I: $\boldsymbol{\theta_0} = (0.5, 0.2, 0, 0, 0.2, 0, 0, 0, 0.2)^{\mathrm{T}}$.

Case II: $\boldsymbol{\theta_0} = (0.5, 0.3, 0, 0, 0.3, 0, 0, 0, 0.3)^{\mathrm{T}}$.

Here Case I has parameter values putting the process well inside the stationarity region, and Case II covers situation where the stationarity condition is nearly not satisfied. We consider the adaptive LASSO (ALASSO), SCAD, MCP and SELO in our procedure, the Bias and mean squared errors (MSE) of the PCML estimators are reported in Tables 1 and 2, respectively. Set $\Psi = \{i; \alpha_i^0 \neq 0, i = 0, \ldots, p\}$, the performances of each penalty functions are presented in Tables 3 and 4, which include the estimated average model size (MS) $\hat{\Psi} = \{i; \hat{\alpha}_i \neq 0, i = 1, \ldots, p\}$; the proportion of selecting the correct model $I\{\hat{\Psi} = \Psi\}$ (CMR); the false positive rate $|\hat{\Psi}\backslash\Psi|/|\hat{\Psi}|$

($F+$), where '$|\Psi|$' denotes cardinality of the set and '$\backslash$' denotes the difference of set; and the false negative rate $|\Psi \backslash \hat{\Psi}|/(p - |\hat{\Psi}|)$ ($F-$). Figure 1 reports the optimal tuning parameter λ with $n = 500$ in Case I (other cases are similar and omitted here). All the results are based on 1000 replications, with sample size $n = 150$, 300 and 500, respectively.

It can be seen from the results that the PCML performs well in both studies. Specifically, the Bias and MSE become smaller as the increasing of sample size. Although all the penalty functions select a larger model, the results of average model size tend to the true value as the sample size increases. The results based on SELO have highest accuracy and smallest model size, which indicate that SELO procedure performs better than the other three penalty functions in practice.

5 An application

In this section, we will apply the proposed methodology to the monthly counts of burglaries in the 25 police car beat in Pittsburgh from January 1990 to December 2001 (http://www.forecastingprinciples.com). The data set totally consists of 144 monthly observations with ACF and PACF are presented in Figure 2. It is easy to see that we may fit the data with Poisson autoregressive model with order $p = 9$ (BIC = 634.547). In Table 5, we report the estimate, standard errors (SE) and p-value, which indicate that the Poisson autoregressive is

$$\begin{cases} X_t \big| \mathcal{F}_{t-1} : \mathcal{P}(\gamma_t); \\ \gamma_t = \alpha_0 + \alpha_2 X_{t-2} + \alpha_3 X_{t-3} + \alpha_9 X_{t-9} \end{cases} \tag{6}$$

with BIC = 614.588, which shows that it is more appropriate to use the Poisson autoregressive model with sparse structure for this data. In Figure 3, we present the predicted value $\hat{X}_t = \gamma_t(\hat{\boldsymbol{\theta}})$ (Fokianos, 2009) with the SELO procedure (other cases are similar), which can reasonably approximates the tendency of the observations.

To check the adequacy of model (6), we consider the Pearson residuals which is defined by $e_t = (X_t - \gamma_t)/\sqrt{\gamma_t}$. The cumulative periodogram plot (Brockwell et al. 2001) of $\{e_t\}$ is reported in Figure 4, which indicates the sequence $\{e_t\}$ is white noise sequence. Thus, by Kedem et al. (2005), it is suitable to fit the real data using model (6).

6 Conclusion

We consider the penalized approach to modeling the Poisson autoregressive process. The oracle properties of the PCML estimator are established. To illustrate the effectiveness of the proposed procedure, some simulation results and a real data example are also provided. The PCML method can also be extended to other kinds of state-space models, such as INGARCH (Ferland et al. 2006) and nonlinear Poisson autoregressive model (Fokianos et al. 2009), which will be studied in the future of our research.

Acknowledgement

We thank the Editor and two reviewers for their valuable suggestions and comments which greatly improved the article. This work is supported by National Natural Science Foundation of China (No. 11271155, 11371168, J1310022, 11571138, 11501241, 11571051, 11301137, 11301212 and 11401146), National Social Science Foundation of China (16BTJ020), Science and Technology Research Program of Education Department in Jilin Province for the 12th Five-Year Plan (440020031139) and Jilin Province Natural Science Foundation (20150520053JH).

Appendix

To prove Theorem 1, we need the following lemma.
Lemma A.1. *Under condition (C.1), as $n \to \infty$ we have*

$$\frac{1}{\sqrt{n}} B(\boldsymbol{\theta_0}) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta_0})),$$

where $B(\boldsymbol{\theta_0}) = \sum_{t=1}^{n} \frac{\partial l_n(\boldsymbol{\theta_0})}{\partial \boldsymbol{\theta}}$*; the Fisher information matrix* $\boldsymbol{\Sigma}(\boldsymbol{\theta_0}) = E(\frac{\boldsymbol{Y}_t \boldsymbol{Y}_t^{\mathrm{T}}}{\gamma_t})$ *with* $\boldsymbol{Y}_t = (1, X_{t-p}, \ldots, X_{t-1})^{\mathrm{T}}$.

Proof of Lemma A.1. Let

$$\begin{aligned} T_{n1} &= \sum_{t=1}^{n} \left(\frac{X_t}{\gamma_t} - 1 \right), \\ T_{ni} &= \sum_{t=1}^{n} \left(\frac{X_t}{\gamma_t} - 1 \right) X_{t-(p+2-i)}, \ 2 \le i \le p+1, \end{aligned}$$

Through some calculation, we can derive that

$$\begin{aligned} &E\left(\left(\frac{X_n}{\gamma_n} - 1 \right) \Big| \mathscr{F}_{n-1} \right) = 0 \ , \\ &E\big(T_{n1} | \mathscr{F}_{n-1}\big) = E\left(T_{(n-1)1} + \left(\frac{X_n}{\gamma_n} - 1 \right) \Big| \mathscr{F}_{n-1} \right) = T_{(n-1)1}, \end{aligned}$$

which implies that $\{T_{n1}, \mathscr{F}_n, n \ge 1\}$ is a martingale with $\mathscr{F}_n = \sigma\left(X_n, X_{n-1}, \ldots, X_0\right)$. By $E|X_t|^4 < \infty$, the strict stationarity of $\{X_t\}$, and the ergodic theorem, we obtain that

$$E\left(\frac{X_n}{\gamma_n} - 1 \right)^2 < \infty,$$

$$\frac{1}{n} \sum_{t=1}^{n} \left(\frac{X_t}{\gamma_t} - 1 \right)^2 \xrightarrow{a.s.} E\left(\frac{X_n}{\gamma_n} - 1 \right)^2 = E(\frac{1}{\gamma_n}) = \sigma_{11}.$$

Using the martingale central limit theorem, we get that

$$\frac{1}{\sqrt{n}} T_{n1} \xrightarrow{D} N(0, \sigma_{11}).$$

Similarly, we can prove $\{T_{ni}, \mathscr{F}_n, n \geq 1\}$, $i = 2, \ldots, p+1$ is a martingale and

$$\frac{1}{\sqrt{n}} T_{ni} \xrightarrow{D} N(0, \sigma_{ii}).$$

For any $\mathbf{c} = (c_1, \ldots, c_{p+1})^{\mathrm{T}} \in \mathbb{R}^{p+1} \backslash (0, \ldots, 0)^{\mathrm{T}}$, we get

$$\begin{aligned} \frac{1}{\sqrt{n}} \mathbf{c}^{\mathrm{T}} \begin{pmatrix} T_{n1} \\ T_{n2} \\ \vdots \\ T_{n(p+1)} \end{pmatrix} &= \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (c_1 + c_2 X_{t-p} + \cdots + c_{p+1} X_{t-1}) \left(\frac{X_t}{\gamma_t} - 1 \right) \\ &\xrightarrow{D} N \left(\mathbf{0}, E\left(c_1 + c_2 X_0 + \cdots + c_{p+1} X_{p-1}\right)^2 \left(\frac{X_p}{\gamma_p} - 1 \right)^2 \right). \end{aligned}$$

Thus, by the Cramer-Wold device,

$$\frac{1}{\sqrt{n}} \begin{pmatrix} T_{n1} \\ T_{n2} \\ \vdots \\ T_{n(p+1)} \end{pmatrix} = \frac{1}{\sqrt{n}} B(\boldsymbol{\theta_0}) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta_0})).$$

This end this proof. □

Proof of Theorem 1. Let $\beta_n = (n^{-1/2} + a_n)$, following Fan and Li (2001), we need to show that for any $\varepsilon > 0$, there exists a constant d, such that

$$\mathbf{P} \left[\sup_{\|\mathbf{u}\|=d} \{Q_n(\boldsymbol{\theta_0} + \beta_n \mathbf{u})\} < Q_n(\boldsymbol{\theta_0}) \right] \geq 1 - \varepsilon, \tag{7}$$

which implies that there exists a local maximum in the ball $\{\boldsymbol{\theta_0} + \beta_n \mathbf{u} : \|\mathbf{u}\| \leq d\}$ with probability at least $1 - \varepsilon$, then there exists a local maximizer with $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}\| = O_p(\beta_n)$. Note that

$$\begin{aligned} D_n(\mathbf{u}) &= Q_n(\boldsymbol{\theta_0} + \beta_n \mathbf{u}) - Q_n(\boldsymbol{\theta_0}) \\ &= L_n(\boldsymbol{\theta_0} + \beta_n \mathbf{u}) - L_n(\boldsymbol{\theta_0}) - n \sum_{i}^{p+1} \left(P_{\lambda_n}(|\theta_i^0 + \beta_n u_i|) - P_{\lambda_n}(|\theta_i^0|) \right) \\ &\leq L_n(\boldsymbol{\theta_0} + \beta_n \mathbf{u}) - L_n(\boldsymbol{\theta_0}) - n \sum_{i}^{s} \left(P_{\lambda_n}(|\theta_i^0 + \beta_n u_i|) - P_{\lambda_n}(|\theta_i^0|) \right). \end{aligned} \tag{8}$$

By Taylor series expansion, we obtain

$$\begin{aligned}
&\beta_n \mathbf{u}^{\mathrm{T}} B(\boldsymbol{\theta_0}) + \frac{1}{2}\beta_n^2 \mathbf{u}^{\mathrm{T}} \frac{\partial^2 L_n(\boldsymbol{\theta_0})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \mathbf{u}\big\{1 + o(1)\big\} \\
&\qquad - \sum_{i}^{s} \Big\{ n\beta_n \dot{P}_{\lambda_n}(|\theta_i^0|) sgn(\theta_i^0) u_i + n\beta_n^2 \ddot{P}_{\lambda_n}(|\theta_i^0|) u_i^2 \left[1 + o(1)\right] \Big\} \\
&\qquad = A_1 + A_2 + A_3,
\end{aligned}$$

where

$$\begin{aligned}
A_1 &= \beta_n \mathbf{u}^{\mathrm{T}} B(\boldsymbol{\theta_0}), \\
A_2 &= \frac{1}{2}\beta_n^2 \mathbf{u}^{\mathrm{T}} \frac{\partial^2 L_n(\boldsymbol{\theta_0})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \mathbf{u}\big\{1 + o(1)\big\}, \\
A_3 &= -\sum_{i}^{s} \Big\{ n\beta_n \dot{P}_{\lambda_n}(|\theta_i^0|) \mathrm{sgn}(\theta_i^0) u_i + n\beta_n^2 \ddot{P}_{\lambda_n}(|\theta_i^0|) u_j^2 \left[1 + o(1)\right] \Big\}.
\end{aligned}$$

From Lemma A.1, we know that $n^{-1/2}B(\boldsymbol{\theta_0}) = O_p(1)$, then $A_1 = O_p(n^{1/2}\beta_n) = O_p(n\beta_n^2)$. By ergodicity, we get $A_2 = -n\beta_n^2 \mathbf{u}^{\mathrm{T}} \boldsymbol{\Sigma}(\boldsymbol{\theta_0})\mathbf{u}$, as $n \to \infty$. From conditions (C.2) and (C.3), we have A_3 is bounded by $\sqrt{s}\beta_n a_n \|\mathbf{u}\| + n\beta_n^2 b_n \|\mathbf{u}\|^2$. By choosing a sufficient large d, both A_1 and A_3 are dominated by A_2. The proof is completed. □

Proof of Lemma 1. We need to prove that with probability tending to one, as $n \to \infty$ for any $\boldsymbol{\theta_1}$ satisfying $\|\boldsymbol{\theta_1} - \boldsymbol{\theta_{10}}\| = O_p(n^{-1/2})$ and for some small $\epsilon_n = \eta n^{-1/2}$ and $j = s+1, \ldots, p+1$

$$\frac{\partial Q_n(\boldsymbol{\theta})}{\partial \theta_j} < 0, \quad for \ 0 < \theta_j < \epsilon_n, \tag{9}$$

$$\frac{\partial Q_n(\boldsymbol{\theta})}{\partial \theta_j} > 0, \quad for \ -\epsilon_n < \theta_j < 0. \tag{10}$$

To show (9), by Taylor's expansion,

$$\begin{aligned}
\frac{\partial Q_n(\boldsymbol{\theta})}{\partial \theta_j} =& \frac{\partial L_n(\boldsymbol{\theta})}{\partial \theta_j} - n\dot{P}_{\lambda_n}(|\theta_j|)\mathrm{sgn}(\theta_j) \\
=& \frac{\partial L_n(\boldsymbol{\theta_0})}{\partial \theta_j} + \sum_{i=1}^{p+1} \frac{\partial^2 L_n(\boldsymbol{\theta_0})}{\partial \theta_i \partial \theta_j}(\theta_i - \theta_i^0)\left\{1 + o(1)\right\} \\
& - n\dot{P}_{\lambda_n}(|\theta_j|)\mathrm{sgn}(\theta_j).
\end{aligned} \tag{11}$$

From Lemma A.1, we know that $\frac{\partial L_n(\boldsymbol{\theta_0})}{\partial \theta_j} = O_p(n^{1/2})$. By law of large numbers, strict stationarity and $\|\boldsymbol{\theta_1} - \boldsymbol{\theta_{10}}\| = O_p(n^{-1/2})$, we have

$$\sum_{i=1}^{p+1} \frac{\partial^2 L_n(\boldsymbol{\theta_0})}{\partial \theta_i \partial \theta_j}(\theta_i - \theta_{i0})\left\{1 + o(1)\right\} = O_p(n^{1/2}).$$

Thus, $\frac{\partial Q_n(\boldsymbol{\theta})}{\partial \theta_j} = n\lambda_n \left\{ O_p(n^{-1/2}/\lambda_n) - \lambda_n^{-1}\dot{P}_{\lambda_n}(|\theta_j|)\text{sgn}(\theta_j) \right\}$. Since $n^{-1/2}/\lambda_n \to 0$ and $\lambda_n^{-1}\dot{P}_{\lambda_n}(|\theta_j|) > 0$ as $n \to \infty$. The sign of (11) is dominated by that of θ_j. Hence, (10) follows. This completes the proof. □

Proof of Theorem 2. Part (i) holds by Lemma 1. We only need to prove (ii). From part (i), we know that $\hat{\boldsymbol{\theta}}_{\mathbf{2}} = \mathbf{0}$ with probability tending to 1. Thus, there exists a root-n consistent local maximum estimator $\hat{\boldsymbol{\theta}}_{\mathbf{1}}$ satisfies the following equation

$$\left. \frac{\partial Q_n(\boldsymbol{\theta})}{\partial \theta_j} \right|_{\boldsymbol{\theta}=\begin{pmatrix} \hat{\boldsymbol{\theta}}_{\mathbf{1}} \\ \mathbf{0} \end{pmatrix}} = 0, \quad for\ j = 1, \ldots, s.$$

By the Taylor expansion, we have

$$\begin{aligned} 0 =& \frac{\partial L_n(\boldsymbol{\theta_0})}{\partial \theta_j} - n\dot{P}_{\lambda_n}(|\hat{\theta}_j|)\text{sgn}(\hat{\theta}_j) \\ =& \frac{\partial L_n(\boldsymbol{\theta_0})}{\partial \theta_j} + \sum_{l=1}^{s} \left\{ \frac{\partial^2 L_n(\boldsymbol{\theta_0})}{\partial \theta_l \partial \theta_j} + o_p(1) \right\} (\hat{\theta}_l - \theta_l^0) \\ & - n\left\{ \dot{P}_{\lambda_n}(|\theta_j^0|)\text{sgn}(\theta_j^0) + \left(\ddot{P}_{\lambda_n}(|\theta_j^0|) + o_p(1) \right)(\hat{\theta}_j - \theta_j^0) \right\}. \end{aligned}$$

This indicates

$$\sqrt{n}(\boldsymbol{\Sigma}^s(\boldsymbol{\theta_0}) + \boldsymbol{\Lambda})\{(\hat{\boldsymbol{\theta}}_{\mathbf{1}} - \boldsymbol{\theta}_{\mathbf{10}}) + (\boldsymbol{\Sigma}^s(\boldsymbol{\theta_0}) + \boldsymbol{\Lambda})^{-1}\mathbf{b})\} = \frac{1}{\sqrt{n}} B^s(\boldsymbol{\theta}_0) + o_p(1),$$

where $B^s(\boldsymbol{\theta_0}) = \sum_{t=1}^{n} \frac{1}{\gamma_t}(X_t - \gamma_t)\, \boldsymbol{Y}_t^s$ and $\boldsymbol{Y}_t^s = (1, X_{t-p}, \ldots, X_{t-p+s-2})^{\mathrm{T}}$. From the Slutskys theorem and the martingale central limit theorem, we complete the proof. □

References

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

[2] Al-Osh M A, Alzaid A A (1988) Integer-valued moving average (INMA) process. Statistical Papers, 29(1): 281-300.

[3] Alzaid A A, Al-Osh M A (1990) An integer-valued pth-order autoregressive structure (INAR (p)) process. Journal of Applied Probability, 314-324.

[4] Brillinger D R (2001) Time series: data analysis and theory. Siam.

[5] Brockwell P J, Davis R A (2013) Time series: theory and methods. Springer Science & Business Media.

[6] Davis R A, Dunsmuir W T M, Streett S B (2003) Observation-driven models for Poisson counts. Biometrika, 90(4):777-790.

[7] Dicker L, Huang B, Lin X (2013) Variable selection and estimation with the seamless-L0 penalty. Statistica Sinica, 929-962.

[8] Doukhan P, Fokianos K, Tjøtheim D (2012) On weak dependence conditions for Poisson autoregressions. Statistics&Probability Letters, 82: 942-48.

[9] Efron B, Hastie T, Johnstone I, et al. (2004) Least angle regression. The Annals of statistics, 32(2): 407-499.

[10] Engle R F (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica: Journal of the Econometric Society, 987-1007.

[11] Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association, 96(456):1348-1360.

[12] Fan J, Li R (2002) Variable selection for Cox's proportional hazards model and frailty model. The Annals of Statistics, 74-99.

[13] Fan J, Lv J (2010). A selective overview of variable selection in high dimensional feature space. Statistica Sinica, 20: 101-148.

[14] Fan J., Peng H (2004) On nonconcave penalized likelihood with diverging number of parameters. The Annals of Statistics, 32: 928-961.

[15] Ferland R, Latour A, Oraichi D (2006) Integer-valued GARCH process. Journal of Time Series Analysis, 27(6): 923-942.

[16] Fokianos K, Rahbek A, Tjøstheim D (2009) Poisson autoregression. Journal of the American Statistical Association, 104(488):1430-1439.

[17] Franke J, Seligmann T (1993) Conditional maximum likelihood estimates for INAR (1) processes and their application to modelling epileptic seizure counts. Developments in time series analysis, 310-330.

[18] Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. Journal of statistical software, 33(1):1.

[19] Hall P, Heyde C C (2014) Martingale limit theory and its application. Academic press.

[20] Kedem B, Fokianos K (2005) Regression models for time series analysis. John Wiley & Sons.

[21] Khoo W C, Ong S H, Biswas A (2017) Modeling time series of counts with a new class of INAR (1) model. Statistical Papers, 58: 393-416.

[22] Klimko L A, Nelson P I (1978) On conditional least squares estimation for stochastic processes. The Annals of Statistics, 629-642.

[23] Knight K, Fu W (2000) Asymptotics for LASSO-type estimators. The Annals of Statistics, 1356-1378.

[24] Nardi Y, Rinaldo A (2011) Autoregressive process modeling via the LASSO procedure. Journal of Multivariate Analysis, 102(3):528-549.

[25] Steutel F W, Van Harn K (1979) Discrete analogues of self-decomposability and stability. The Annals of Probability, 893-899.

[26] Tibshirani R (1996) Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

[27] Tong X., He X., Sun L., and Sun J. (2009). Variable selection for panel count data via non-concave penalized estimating function. Scandinavian Journal of Statistics, 36: 620 - 635.

[28] Wang H, Li G, Tsai C L (2007a) Regression coefficient and autoregressive order shrinkage and selection via the LASSO. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(1):63-78.

[29] Wang H, Li R, Tsai C L (2007b) Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika, 94(3): 553-568.

[30] Yang K, Wang D, Jia B, Li H. (2016) An integer-valued threshold autoregressive process based on negative binomial thinning. Statistical Papers, DOI: 10.1007/s00362-016-0808-1.

[31] Yoon Y J, Park C, Lee T (2013) Penalized regression models with autoregressive error terms. Journal of Statistical Computation and Simulation, 83(9):1756-1772.

[32] Zhang H H, Lu W (2007) Adaptive Lasso for Cox's proportional hazards model. Biometrika, 4(3): 691-703.

[33] Zhang C H (2010) Nearly unbiased variable selection under minimax concave penalty. The Annals of statistics, 894-942.

[34] Zhang H, Wang D, Zhu F (2010) Inference for INAR (p) processes with signed generalized power series thinning operator. Journal of Statistical Planning and Inference, 140(3):667-683.

[35] Zhang H, Sun J, Wang D (2013) Variable selection and estimation for multivariate panel count data via the seamless-L0 penalty. The Canadian Journal of Statistics, 41: 368 - 385.

[36] Zheng H, Basawa I V, Datta S (2006) Inference for pth-order random coefficient integer-valued autoregressive processes. Journal of Time Series Analysis, 27(3):411-440.

[37] Zhu F, Wang D (2011) Estimation and testing for a Poisson autoregressive model. Metrika, 73(2):211-230.

[38] Zou H (2006) The adaptive LASSO and its oracle properties. Journal of the American statistical association, 101(476):1418-1429.

Table 1: Bias and MSE (in parentheses) of the estimators in Case I.

Sample size		ALASSO	SCAD	MCP	SELO
$n = 150$	$\hat{\alpha}_0$	$-0.0552(0.0486)$	$-0.0576(0.0501)$	$-0.0571(0.0493)$	$-0.0570(0.0491)$
	$\hat{\alpha}_1$	$-0.0316(0.0071)$	$-0.0339(0.0081)$	$-0.0328(0.0076)$	$-0.0321(0.0073)$
	$\hat{\alpha}_2$	$0.0151(0.0007)$	$0.0216(0.0013)$	$0.0221(0.0013)$	$0.0146(0.0008)$
	$\hat{\alpha}_3$	$0.0093(0.0009)$	$0.0109(0.0012)$	$0.0102(0.0014)$	$0.0109(0.0008)$
	$\hat{\alpha}_4$	$-0.0271(0.0083)$	$-0.0296(0.0092)$	$-0.0295(0.0089)$	$-0.0283(0.0086)$
	$\hat{\alpha}_5$	$0.0189(0.0011)$	$0.0226(0.0012)$	$0.0218(0.0008)$	$0.0175(0.0010)$
	$\hat{\alpha}_6$	$0.0101(0.0005)$	$0.0173(0.0013)$	$0.0176(0.0010)$	$0.0108(0.0008)$
	$\hat{\alpha}_7$	$0.0123(0.0010)$	$0.0187(0.0015)$	$0.0183(0.0012)$	$0.0128(0.0009)$
	$\hat{\alpha}_8$	$-0.0293(0.0067)$	$-0.0306(0.0086)$	$-0.0310(0.0079)$	$-0.0296(0.0072)$
$n = 300$	$\hat{\alpha}_0$	$-0.0311(0.0136)$	$-0.0329(0.0148)$	$-0.0325(0.0142)$	$-0.0320(0.0139)$
	$\hat{\alpha}_1$	$-0.0182(0.0034)$	$-0.0194(0.0041)$	$-0.0186(0.0036)$	$-0.0179(0.0034)$
	$\hat{\alpha}_2$	$0.0116(0.0004)$	$0.0126(0.0007)$	$0.0120(0.0006)$	$0.0114(0.0003)$
	$\hat{\alpha}_3$	$0.0082(0.0005)$	$0.0108(0.0009)$	$0.0106(0.0009)$	$0.0078(0.0005)$
	$\hat{\alpha}_4$	$-0.0162(0.0037)$	$-0.0189(0.0045)$	$0.0176(0.0042)$	$-0.0165(0.0040)$
	$\hat{\alpha}_5$	$0.0110(0.0004)$	$0.0139(0.0009)$	$0.0146(0.0007)$	$0.0107(0.0003)$
	$\hat{\alpha}_6$	$0.0083(0.0003)$	$0.0093(0.0006)$	$0.0086(0.0006)$	$0.0081(0.0005)$
	$\hat{\alpha}_7$	$0.0103(0.0007)$	$0.0109(0.0011)$	$0.0097(0.0010)$	$0.0075(0.0005)$
	$\hat{\alpha}_8$	$-0.0169(0.0037)$	$-0.0278(0.0041)$	$-0.0261(0.0040)$	$-0.0172(0.0039)$
$n = 500$	$\hat{\alpha}_0$	$-0.0227(0.0078)$	$-0.0236(0.0086)$	$-0.0233(0.0081)$	$-0.0229(0.0080)$
	$\hat{\alpha}_1$	$-0.0153(0.0018)$	$-0.0157(0.0022)$	$-0.0152(0.0020)$	$-0.0149(0.0017)$
	$\hat{\alpha}_2$	$0.0060(0.0002)$	$0.0072(0.0007)$	$0.0066(0.0004)$	$0.0055(0.0001)$
	$\hat{\alpha}_3$	$0.0072(0.0003)$	$0.0098(0.0006)$	$0.0083(0.0004)$	$0.0059(0.0002)$
	$\hat{\alpha}_4$	$-0.0092(0.0020)$	$-0.0103(0.0023)$	$-0.0101(0.0023)$	$-0.0095(0.0022)$
	$\hat{\alpha}_5$	$0.0063(0.0002)$	$0.0076(0.0010)$	$0.0073(0.0006)$	$0.0055(0.0002)$
	$\hat{\alpha}_6$	$0.0063(0.0001)$	$0.0078(0.0005)$	$0.0074(0.0004)$	$0.0056(0.0001)$
	$\hat{\alpha}_7$	$0.0058(0.0002)$	$0.0109(0.0009)$	$0.0063(0.0003)$	$0.0050(0.0001)$
	$\hat{\alpha}_8$	$-0.0140(0.0022)$	$-0.0153(0.0027)$	$-0.0148(0.0030)$	$-0.0140(0.0023)$

Table 2: Bias and MSE (in parentheses) of the estimators in Case II.

Sample size		ALASSO	SCAD	MCP	SELO
$n = 150$	$\hat{\alpha}_0$	0.1156(0.1002)	0.1289(0.1211)	0.1256(0.1123)	0.1136(0.0965)
	$\hat{\alpha}_1$	−0.0421(0.0082)	−0.0539(0.0087)	−0.0521(0.0084)	−0.0420(0.0079)
	$\hat{\alpha}_2$	0.0122(0.0011)	0.0166(0.0018)	0.0150(0.0018)	0.0119(0.0009)
	$\hat{\alpha}_3$	0.0090(0.0012)	0.0125(0.0029)	0.0113(0.0018)	0.0102(0.0011)
	$\hat{\alpha}_4$	−0.0357(0.0088)	−0.0363(0.0092)	−0.0360(0.0091)	−0.0359(0.0090)
	$\hat{\alpha}_5$	0.0158(0.0018)	0.0201(0.0022)	0.0193(0.0020)	0.0152(0.0012)
	$\hat{\alpha}_6$	0.0095(0.0009)	0.0125(0.0016)	0.0111(0.0012)	0.0090(0.0008)
	$\hat{\alpha}_7$	0.0113(0.0013)	0.0149(0.0017)	0.0135(0.0011)	0.0113(0.0011)
	$\hat{\alpha}_8$	−0.0356(0.0077)	−0.0425(0.0090)	−0.0380(0.0081)	−0.0351(0.0075)
$n = 300$	$\hat{\alpha}_0$	0.0596(0.0498)	0.0652(0.0593)	0.0638(0.0549)	0.0615(0.0501)
	$\hat{\alpha}_1$	−0.0339(0.0031)	−0.0357(0.0039)	−0.0342(0.0035)	−0.0335(0.0030)
	$\hat{\alpha}_2$	0.0101(0.0005)	0.0172(0.0010)	0.0168(0.0012)	0.0098(0.0006)
	$\hat{\alpha}_3$	0.0072(0.0006)	0.0153(0.0011)	0.0152(0.0010)	0.0063(0.0006)
	$\hat{\alpha}_4$	−0.0245(0.0041)	−0.0251(0.0048)	−0.0249(0.0046)	−0.0241(0.0041)
	$\hat{\alpha}_5$	0.0130(0.0008)	0.0159(0.0014)	0.0157(0.0012)	0.0127(0.0006)
	$\hat{\alpha}_6$	0.0052(0.0007)	0.0106(0.0008)	0.0118(0.0010)	0.0090(0.0006)
	$\hat{\alpha}_7$	0.0091(0.0007)	0.0113(0.0013)	0.0101(0.0007)	0.0099(0.0004)
	$\hat{\alpha}_8$	−0.0215(0.0037)	−0.0220(0.0041)	−0.0217(0.0040)	−0.0216(0.0039)
$n = 500$	$\hat{\alpha}_0$	0.0338(0.0226)	0.0396(0.0286)	0.0378(0.0269)	0.0341(0.0229)
	$\hat{\alpha}_1$	−0.0215(0.0018)	−0.0238(0.0023)	−0.0227(0.0021)	−0.0214(0.0017)
	$\hat{\alpha}_2$	0.0060(0.0003)	0.0102(0.0012)	0.0106(0.0009)	0.0055(0.0002)
	$\hat{\alpha}_3$	0.0061(0.0002)	0.0096(0.0011)	0.0095(0.0011)	0.0059(0.0002)
	$\hat{\alpha}_4$	−0.0163(0.0022)	−0.0177(0.0028)	−0.0175(0.0025)	−0.0161(0.0022)
	$\hat{\alpha}_5$	0.0082(0.0003)	0.0110(0.0008)	0.0107(0.0011)	0.0090(0.0003)
	$\hat{\alpha}_6$	0.0058(0.0003)	0.0103(0.0006)	0.0099(0.0007)	0.0052(0.0001)
	$\hat{\alpha}_7$	0.0071(0.0004)	0.0097(0.0011)	0.0085(0.0006)	0.0070(0.0002)
	$\hat{\alpha}_8$	−0.0162(0.0022)	−0.0175(0.0027)	−0.0182(0.0030)	−0.0166(0.0023)

Table 3: Simulation results for model selection in Case I.

Sample size	Penalty function	MS	CMR	F+	F−
$n = 150$	ALASSO	4.1780	0.5640	0.0954	0.0395
	SCAD	4.2240	0.4640	0.1211	0.0554
	MCP	4.1880	0.5020	0.1064	0.0441
	SELO	4.1340	0.6280	0.0829	0.0294
$n = 300$	ALASSO	4.1260	0.6440	0.0746	0.0216
	SCAD	4.1700	0.5680	0.0869	0.0385
	MCP	4.1660	0.6020	0.0907	0.0260
	SELO	4.1080	0.7300	0.0506	0.0192
$n = 500$	ALASSO	4.0500	0.7980	0.0338	0.0151
	SCAD	4.1400	0.6120	0.0770	0.0305
	MCP	4.0940	0.7320	0.0544	0.0185
	SELO	4.0440	0.8980	0.0194	0.0054

Table 4: Simulation results for model selection in Case II.

Sample size	Penalty function	MS	CMR	F+	F−
$n = 150$	ALASSO	4.3180	0.6080	0.1035	0.0118
	SCAD	4.3920	0.4860	0.1344	0.0280
	MCP	4.3760	0.5360	0.1190	0.0213
	SELO	4.3060	0.6500	0.0895	0.0080
$n = 300$	ALASSO	4.2300	0.7440	0.0686	0.0039
	SCAD	4.3120	0.6100	0.0984	0.0126
	MCP	4.2920	0.6580	0.0921	0.0103
	SELO	4.2120	0.7500	0.0655	0.0041
$n = 500$	ALASSO	4.1620	0.8020	0.0521	0.0010
	SCAD	4.3040	0.6560	0.0915	0.0080
	MCP	4.2620	0.6960	0.0818	0.0036
	SELO	4.1160	0.8400	0.0414	0.0007

Table 5: Estimated regression coefficients, standard errors(SE) and p-value.

	ALASSO	SCAD	MCP	SELO
$\hat{\alpha}_0$	1.2136	1.2126	1.2117	1.2130
(SE, p-value)	(0.3534, 0.0006)	(0.3541, 0.0006)	(0.3539, 0.0006)	(0.3536, 0.0006)
$\hat{\alpha}_1$	0	0	0	0
(SE, p-value)	$(-, -)$	$(-, -)$	$(-, -)$	$(-, -)$
$\hat{\alpha}_2$	0.1563	0.1577	0.1572	0.1565
(SE, p-value)	(0.0641, 0.0147)	(0.0643, 0.0141)	(0.0642,0.0143)	(0.0641, 0.0147)
$\hat{\alpha}_3$	0.1270	0.1286	0.1281	0.1275
(SE, p-value)	(0.0635, 0.0454)	(0.0636, 0.0433)	(0.0636, 0.0439)	(0.0635, 0.0447)
$\hat{\alpha}_4$	0	0	0	0
(SE, p-value)	$(-, -)$	$(-, -)$	$(-, -)$	$(-, -)$
α_5	0	0	0	0
(SE, p-value)	$(-, -)$	$(-, -)$	$(-, -)$	$(-, -)$
$\hat{\alpha}_6$	0	0	0	0
(SE, p-value)	$(-, -)$	$(-, -)$	$(-, -)$	$(-, -)$
$\hat{\alpha}_7$	0	0	0	0
(SE, p-value)	$(-, -)$	$(-, -)$	$(-, -)$	$(-, -)$
$\hat{\alpha}_8$	0	0	0	0
(SE, p-value)	$(-, -)$	$(-, -)$	$(-, -)$	$(-, -)$
$\hat{\alpha}_9$	0.2196	0.2208	0.2205	0.2201
(SE, p-value)	(0.0615, 0.0004)	(0.0616, 0.0003)	(0.0616, 0.0003)	(0.0615, 0.0003)

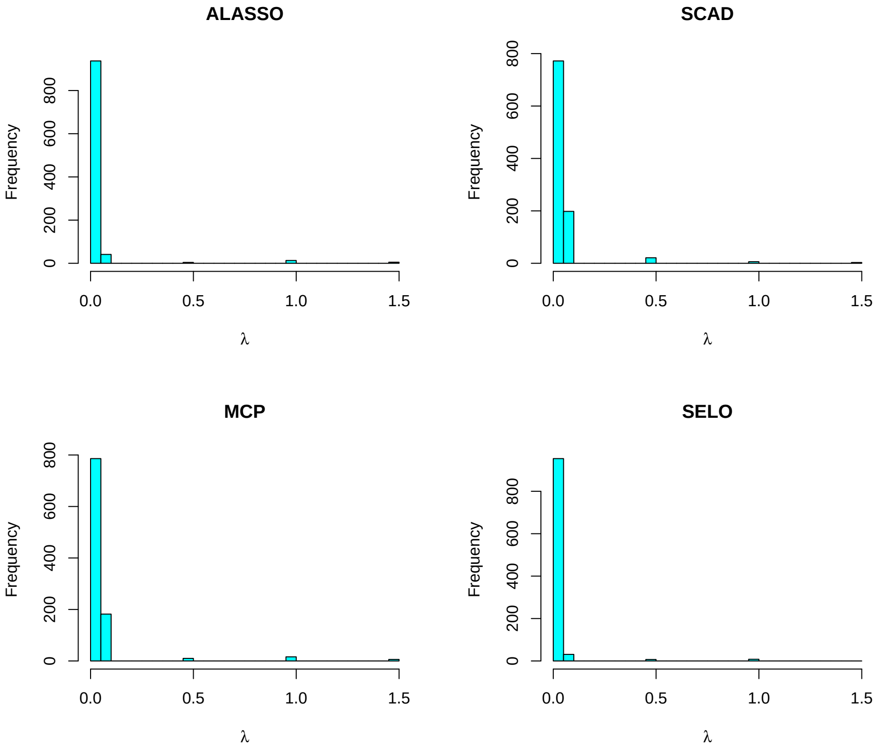

Fig. 1: The optimal tuning parameter λ with $n = 500$ (Case I).

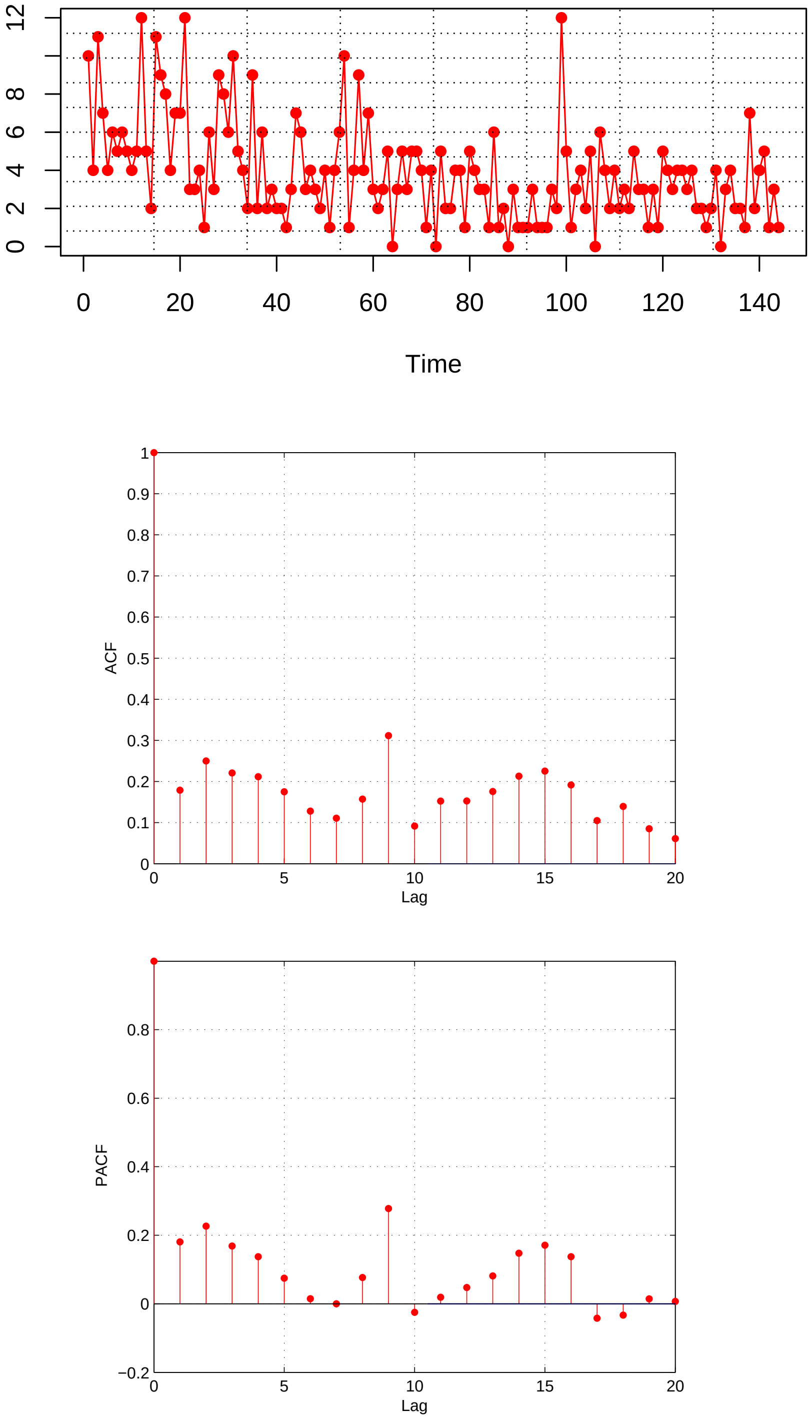

Fig. 2: Monthly counts of burglaries in the 25 police car beat in Pittsburgh and their ACF and PACF plots.

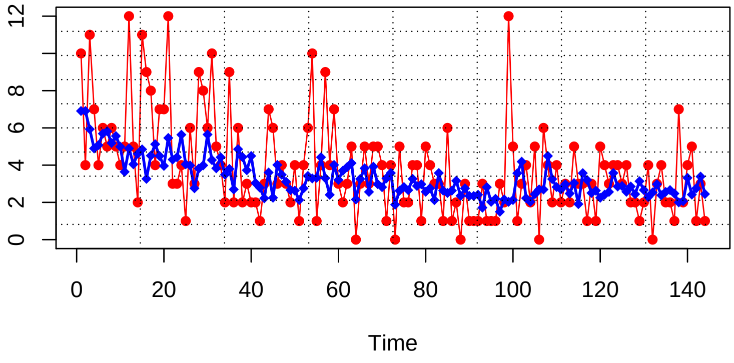

Fig. 3: The observed (red) and predicted (blue) counts of monthly burglaries in Pittsburgh.

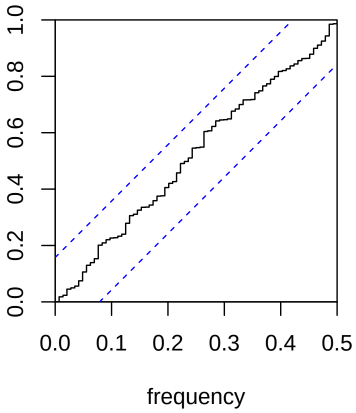

Fig. 4: The cumulative periodogram plot of the Pearson residuals.