Research article Available online www.ijsrr.org ISSN: 2279–0543

International Journal of Scientific Research and Reviews

Bayesian Analysis to the Detection of Outliers in an Autoregressive Model with Exponential White Noise

R. Chinnadurai[1*] and P. ARUMUGAM [2]

[1]Department of Statistics, Manonmaniam Sundranar University, Tirunelveli, Tamil Nadu, India
[2] Department of Statistics, Annamalai University, Tamil Nadu, India
Email:aru_stat1976@hotmail.com

ABSTRACT

In this paper, we develop a Model to detect the presence of outliers in an autoregressive model with exponential white noise through the Bayesian methodology. The developed model is illustrated with a simulation by adopting Gibbs sampling.

KEYWORDS: Autoregressive model, Bayesian analysis, outliers, posterior distribution, Gibbs Samples.

***Corresponding author**

R. Chinnadurai

Department of Statistics, Manonmaniam Sundranar University, Tirunelveli,

Tamil Nadu, India

E-mail: mrchinnadurai@yahoo.com

INTRODUCTION

The concept of outliers in a data set is considered to be as the subset of statistics. To quote [6] in almost every true series of observations some are found ,which differ so much from the other as to indicate some abnormal source of ever not contemplated in the theoretical discussions, and the introduction of which in the investigates can only serve to perplex and mislead the inquirer. Barnet and Lewis[2] pointed out that even before the development of formal statistical method argument raged over whether, and an what basis, we should discard observations from a set of data as the grounds that they are unrepresentative, 'spurious', or 'mavericks' or 'rogues'. But it is now evident that outliers do not inevitably 'perplex' or mislead; they are not necessarily 'bad' or 'erroneous'. According to Barnet and Lewis[2]. We shall define outliers in a set of data to be an observation (subset of observations) which appears to be inconsistent with remainder of that set of data.

Abraham and Box[1] rightly pointed out that the time series often contain discrepant observations, it is appropriate to employ models which reflect this fact. In a different context, Dixon[4, 8],suggested set- ups in which a small probability α exists that any observation is bad. In applications we are frequently faced with time series data which, for a variety of different reasons, have characteristics not compatible with the usual assumptions of linearity or/and Gaussian errors. One of the many ways the assumption of linearity may fail is the presence of limit cycle (see for example Tong)[7].Process with non-Gaussian white noise are useful for modeling a wide range of phenomena that do most support negative values or have a highly skewed distribution.

In this paper we represent the problem of detecting outliers in an autoregressive model with an exponential white noise, which is a very special case of the general autoregressive model for non-negative variables. Due to the nature of the model considered, the explicit Bayesian solutions are difficult to reach. We analyse a simple model to highlight the sort of problems that arise. However, contrary to classical analysis, Bayesian methodology can be applied with success through the use of Gibbs sampler. The paper is organized as follows: the AR (p) model with exponential white noise along with aberrant innovation is specified, the Bayesians Inference and detection of the outliers for this model. An exact analysis, although possible to perform, becomes quite demanding computationally due to the fact that we are in the presence of a constrained parameter model. Integration needed to perform Bayesian analysis cannot be calculated since an explicit form for the support of the posterior distribution is difficult to get, particularly for large sample sizes. Hence, we suggest unrigs Gibbs sampling to obtain samples from the posterior distribution. We conduct a simulation study to compare the performance of the new model developed.

SPECIFICATION OF MODEL

The p^{th} order autoregressive aberrant innovation model may be defined as

$$y_t = B^T\phi + \delta x_t + a_t, \cdots \quad (1)$$

Where $B^T = (y_{t-1}, \ldots y_{t-p})$, $\phi^T = (\phi_1, \phi_2 \ldots \phi_p)$, $y_t = 1$ if there is an aberrant innovation at t and $x_t = 0$ otherwise and at are independent exponentially distributed white noise with pdf $fa_t(y; \alpha) = \alpha e^{-\alpha y} I_{(0,\infty)}(y)$

The parameters ϕ, δ and α are unknown. We also define $y^T=(y_1, y_2 \ldots\ldots y_n)$ as a vector of r unities and (n-r) zero where r and hence x are unknown.

Assuming x to be known, Fox[5] considered a likelihood ratio criterion for this setup of the autoregressive model defined. The observations y_t are assumed to be deviations from the mean and in the case when the mean is unknown we can include that also in the expression for the likelihood and often to a sufficient approximations we can take y_t as the deviations from the sample mean.

Due to the practical limitation, the First order autoregressive model is to be considered, the model is given by

$$y_t = \phi\, y_{t-1} + \delta\, x_t + a_t \,, \qquad (2)$$

Where y_t and at are as defined as

The parameter space for the model is

$$\Theta = \{\theta = (\phi, r, \delta, \alpha)\alpha > 0,\ \delta > 0, 0 < \phi < 1\}$$

Stationary of the process is guaranteed by the restriction imposed on the parameterϕ.

The likelihood function based on $y= (y_1, y_2, \ldots\ldots y_n)$is given by

$$L(\theta/y) = \alpha^{n-1} \exp\left\{-\alpha\left[\sum_{t=2}^{n} y_t - \phi\sum_{t=2}^{n} y_{t-1} - \delta N(r)\right]\right\} I_U(\theta) \qquad (3)$$

Where

$N(r) = \sum_{t=2}^{n} I_{(r,+\infty)}(y_{t-1})$ and $U = \{\theta \in \Theta : y_t - \phi y_{t-1} - \delta I_{(r,+\infty)}(x_{t-1}) \geq 0\ t = 2,...n\}$ The likelihood function is a step function in r, with breaks at the observed x_{t-1}.

BAYESIAN INFERENCE AND DETECTION OF OUTLIER

We assume that, a priori, the parameters ϕ and r are independent, uniformly distributed in (0,1) and (0, β) respectively. For the parameter of the exponential error we assume a conjugate prior independent of (ϕ, r), of the form

$$(\alpha, \delta) \sim \exp(\delta \mid \alpha, f)\, Ga(\alpha \mid g, h),\ f,g,h > 0 \quad (4)$$

This prior implies that the change in the error has a conditional mean proportional to the mean of the error. Hence the posterior distribution for θ is

$$P(\theta/Y) \propto \alpha^{n_1 - 1} \exp\{-\alpha[S_1 - \phi S_2 - \delta(N^*(r))]\} I_{\Theta_n(x)}(\theta) \quad (5)$$

Where

$$S_1 = \sum_{t=2}^{n} y_t + h,\ S_2 = \sum_{t=2}^{n} y_{t-1},\ n_1 = n + g,\ N^*(r) = N(r) + f,$$

$$\Theta_n(y) = \left\{\theta \in \Theta; y_t - \phi\, y_{t-1} - \delta I_{(r,+\infty)}(x_{(t-1)}) \geq 0 \quad \forall t = 2,...n, r \leq \beta^* = \min\left(x_{(n-1)}\right), \beta\right\}$$

With $y_{(r)}$ being the r^{th}order statistic. The posterior mean will be obtained from the following

$$E[g(\theta/Y)] = \int_{\Theta_n} g(\theta)\, p(\theta/y)\, d\theta \quad (6)$$

For suitable choicer of g(.).

In the model, we are considering the parameter r affects particularly the parameter δ of the error term. Hence, for the problem we have in hand we consider k=3, with $\theta_1 = (r, \delta)$, $\theta_2 = \alpha$ and $\theta_3 = \phi$. The full conditional posterior densities are, respectively

$$p(r, \delta \mid y, \alpha, \phi) = p(\delta \mid y, r, \alpha, \phi)\, p(r \mid y, \alpha, \phi) \quad (7)$$

With $p(r \mid y, \alpha, \phi)$ defined by

$$p_k(r \mid y, \alpha, \phi) \propto \left[\alpha N_k^*(r)\right]^{-1} \left\{\exp\left[\alpha N_k^*(r)\delta_r^*\right] - 1\right] \quad (8)$$

$r \in (y_{(k-1)}, y_{(k)})k = 1,2\ldots.nd$, when nd is the number of distinct y_t, t =1,2….n-1 and $N_k^*(r)$ the number of observations

$$y_t \geq r, \text{ when } \quad r \in \left(y_{(k-1)}, y_{(k)}\right), \left(y_{(0)} = 0\right)$$

$$\delta \mid y, r, \alpha, \phi \sim \text{Exp}\left(\alpha N^*(r), \delta_r^*\right)$$

Where $\delta_r^* = \min_{y_{t-1} \geq r} \left(y_t - \phi\, y_{t-1}\right)$

$t = 2,\ldots n$

And $y \sim E_{y_t}(a, b)$ means that the distribution of y= {b-y/ $y \leq b$} is exp (a) i.e. With p.d.f

$$p(y \mid a, b) = \frac{a \exp\{-a(b - y)\}}{1 - \exp(-ab)}, I_{(0, b)}(y) \qquad (9)$$

$$\alpha \mid y, r, \delta, \phi \sim \text{Ga}\left(n_1, S_2 - \phi\, S_1 - N^*(r)\, \delta\right)$$

i.e.

$$p(\alpha \mid y, r, \delta, \phi) \propto \alpha^{n_1 - 1} \exp\left\{-\alpha\left[\left(S_2 - \phi\, S_1 - N^*(r)\, \delta\right)\right]\right\} \qquad (10)$$

$$\phi \mid y, r, \delta, \alpha \sim E_{y_t}\left(\alpha S_1, \phi_r^*\right)$$

Where $\phi_r^* = \min_{t=2,\ldots n}\left[1, \frac{y_t - \delta I_{(r, +\infty)}(y_{t-1})}{y_{t-1}}\right]$

Suppose that $\bar{y}_q$ and $\bar{y}_{q-1}$ respectively, denote the average of the suspected observation and that of the observations just previous to the suspected ones. Then

$$D = \sum_{t=3}^{n} y_{t-1}^2, d = \sum_{t=2}^{n} y_{t-1}, y_t, B = D - r\,\bar{y}_{q-1}^2$$

$$\phi^* = B^{-1}\left(d - r\,\bar{y}_q\,\bar{y}_{q-1}\right), \hat{\phi} = D^{-1}d$$

$$\upsilon S^2 = \sum_{t=1}^{n} y_t^2 - r\,\bar{y}_q^2 - \phi^{*2} B$$

$$\upsilon_0 S_0^2 = \sum_{t=1}^{n} y_t^2 - \hat{\phi}^2 D$$

When r=1, the q[th]observation being suspected, say $\phi^* = \left(d - y_q, y_{q-1}\right)\Big/\left(D - y_{q-1}^2\right)$ which is very like the regression estimate of ϕ except that relevant quantities allowing for the effect of the outlier are subtracted from the numerator and denominator.

It can be seen that p(ϕ/y) is a weighted average of scaled t distribution with mean ϕ, scaling factor $SB^{-1/2}$ and degrees of freedom n-2. This distribution gives us all the information aboutϕ. In particular, the posterior mean and variance are given by

$$\bar{\phi}_r = E(\phi/y) = \sum_r w_r \phi^* \qquad (11)$$

$$v(\phi/Y) = w_0\left\{\frac{n-1}{n-3}\frac{S^2}{D} = \left(\hat{\phi} - \bar{\phi}_r\right)^2 + \sum_{r\neq 0} w_{(r)}\frac{n-2}{n-4}\frac{S^2}{B} + \left(\phi^* - \bar{\phi}_r\right)^2\right\} \qquad (12)$$

NUMERICAL STUDY

We have generated 100 observations from the model $y_t = \frac{1}{2}y_{t-1} + 5x_t + a_t$, with different values for the parameter. The following table shows the mean and variance of the posterior distribution mean and variance

Table 1: Posterior mean and variance of ϕ /y

α	**E(ϕ /y)**	**Var(ϕ /y)**
0.0	4.06	1.98
0.001	4.56	1.02
0.01	4.50	0.97
0.03	4.31	0.90
0.05	4.11	0.86

CONCLUSION

For conclusion we show the corresponding mean and variance when it is assumed that there are no outliers. Again it is found that the conclusion are not sensitive to moderate changes in α. However, as might be expected there is a dramatic difference between the assumption of no possibility of outliers (α =0) and the assumption of some such probability even a very remote area (α =0.001).

REFERENCE

1.Abraham B and Box GEP.Bayesian analysis of some outlier's problems in time series. Biometrics. 1976; 66(2): 229-36.

2. Barnett V and Lewis T. "Outliers in Statistical Data". 3rd ed. John Wiley & Sons: New Jersey, USA; 1994.

3. Box GEP andTiaoA. Bayesian approach to some outliers Problems. Biometrics. 1968; 55: 119-29.

4. Dixon WJ. Processing data for outliers. Biometrica.1958; 9(1):74-89.

5. Fox AJ. Outliers in time series. JR Stat Soc Series B.1972; 34(3): 350-63.

6. Pierce B. Criterion for the rejection of doubtful observation.Astron J. 1852;II(21):161.

7. Tong H. Non-linear Time Series: A Dynamical Approach. Oxford University Press: UK; 1990.

8. Tukey JW."A survey of sampling from contaminated distribution". In: Olkin I, Churye SG and Others.(eds.) Contribution to probability and statistics; Essays in honors of Harold Hostelling. Stanford University Press: USA;1960; 448-85.