The Identification of Outliers in ARMAX Models via Genetic Algorithm

Ping Chen[1] and Ying Chen[2]

[1]*Department of Mathematics, Southeast University, Nanjing 210096, China*

[2]*Department of Forensic Science, Jiangsu Police Institute , Nanjing, 210012, China*

Abstract

This paper proposes a procedure to identify additive and innovational outliers by genetic algorithm in autoregressive moving average with exogenous variable(ARMAX) time series models. We use some methods to delete the influence of input process in ARMAX model and then detect outliers in time series based on the previous work, which is an improvement and extension of the detection method on ARMA models. Empirical and simulation studies show that the proposed procedure is effective.

Keywords dynamic systems, innovational outliers, ARMAX model, genetic algorithm

1 Introduction

Outliers in dynamic systems or engineering time series can have adverse effects on model identification and parameter estimation. Several procedures are available in literature to handle outliers in a time series. However, the case of multiple additive outliers and innovational outliers is very difficult to study because of the great number of alternatives and of the masking and swamping effects. Compared to other search algorithms, genetic algorithms allow many candidate solutions to be considered simultaneously at each step. Baragona et al.[1] showed how to use genetic algorithms for outlier detection and classification in ARMA series. Peña and Sánchez[2] presented a new procedure for multifold predictive validation in ARMAX models. Also, Chen et al.[3] and Chen et al.[4] developed some methods for detecting outliers, change point and outlier patches in bilinear time series models. On the other hand, Huang et al.[5] discussed the improved genetic algorithm for vehicle routing problem with time windows. In this paper, a genetic algorithm is proposed to identify additive and innovational outliers in ARMAX series. We are using the standard genetic algorithm with complete replacement of the past population and elitist strategy. The relationship between inverse correlations and outliers is helpful to simplify the fitness function, which may be quickly computed by Trench's algorithm. For the case of large series, it is better to divide the series into several parts so that the two neighbor subseries share a length of same data, and then to detect the subseries respectively. Thus, the problem of large population in genetic algorithms has been avoided, as well as the loss of outliers around the cut point. At last, simulation studies are carried out, which show promising results.

2 Genetic Algorithm and Outliers Models in ARMAX Series

The genetic algorithm(GA) is known to be able to provide us with a powerful optimization tool when the solution space happens to be both discrete and large, and the objective function does not fulfill the usual regularity requirements.The key feature of a GA is the manipulation of a population whose individuals are characterized by possessing a chromosome. This latter can be coded as a string of characters of given length. Each string represents a feasible solution to the optimization problem. The link between the GA and the problem at hand is provided by the fitness function (FF). The FF establishes a mapping from the chromosomes to some set of real numbers. The greater the FF is, the better the adaptation of the individual. The procedure is iterative. It makes use of three evolutionary operators: reproduction, crossover and mutation.

An ARMA model with input process is called ARMAX model, which is defined as

$$Z_t = \sum_{i=1}^{d} \upsilon_i(B) X_{i,t} + n_t,$$

where $\upsilon_i(B) = (\delta_i^{-1}(B) \cdot \omega_i(B))B^{k_i}$ is the transfer function of ith input process, $n_t = (\theta(B)/\phi(B))\varepsilon_t$ is noise process. $\{Z_t\}$ is called response process. And $X_{i,t}$ denotes the ith input process or the difference of ith input process at time t, k_i presents the influence's time delay of ith input process, ε_t is normal white nose process.

$$\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q, \quad \phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$$

where B is the backshift operator.

When $\upsilon_i(B) = 0, i = 1, \cdots, d$, it is ARMA model, when some $\upsilon_i(B)$ is nonzero constant, $i = 1, \cdots, d$, it is regression model with ARMA error.

(1) Additive outliers(AO) model:

Suppose that only the jth point z_j be AO, whose influence magnitude is w_{tj}, then we have

$$Z_t = \sum_{i=1}^{d} \upsilon_i(B) X_{i,t} + w_{tj}\delta_{t,tj} + \frac{\theta(B)}{\phi(B)}\varepsilon_t$$

where $\delta_{t,tj}$ is Kronecker symbol: If $t = tj$, then $\delta_{t,tj} = 1$, else $\delta_{t,tj} = 0$.

(2) Innovational outliers(IO) model

Suppose that only the jth point z_j be IO, whose influence magnitude is w_{tj}, then we have

$$
\begin{aligned}
Z_t &= \sum_{i=1}^{d} \upsilon_i(B)X_{i,t} + \frac{\theta(B)}{\phi(B)}(\varepsilon_t + w_{tj}\delta_{t,tj}) \\
&= \sum_{i=1}^{d} \upsilon_i(B)X_{i,t} + w_{tj}\frac{\theta(B)}{\phi(B)}\delta_{t,tj} + \frac{\theta(B)}{\phi(B)}\varepsilon_t.
\end{aligned}
$$

3 Identification of ARMAX Models

Suppose that the ARMAX model of only one input process is as follows:

$$Z_t = \delta^{-1}(B)\omega(B)X_{t-\mathfrak{b}} + n_t = \upsilon(B)X_t + n_t, \tag{1}$$

Where

$$\delta(B) = 1 - \delta_1(B) - \cdots - \delta_{r1}B^{r1}, \omega(B) = \omega_0 - \omega_1(B) - \cdots - \omega_{r2}B^{r2}$$

and

$$\upsilon(B) = \delta^{-1}(B)\omega(B)B^{\mathfrak{b}}$$

Suppose the input process X_t is stationary and is able to be represented by some member of the general linear class of autoregressive-moving average models. Given a set of data, similar to Box et al.[6], then we can carry out our usual identification and estimation methods to obtain a model for the X_t process $\phi(B)\theta^{-1}(B)X_t = \alpha_t$ which, to a close approximation, transforms the correlated input series X_t to the uncorrelated white nose series α_t. At the same time, we can obtain an estimate s_α^2 of σ_α^2 from the sum of squares of the $\hat{\alpha}'s$. If we now apply this same transformation to Z_t to obtain $\beta_t = \phi(B)\theta^{-1}(B)Z_t$, then the model(1) may be written $\beta_t = \upsilon(B)\alpha_t + \varepsilon_t$, multiplying α_{t-k} on both sides and taking expectations, we obtain $\gamma_{\alpha\beta}(k) = \upsilon_k\sigma_\alpha^2$, where $\gamma_{\alpha\beta}(k) = E[\alpha_{t-k}\beta_t]$is the cross covariance at lag k between α and β. Thus $\upsilon_k = [\rho_{\alpha\beta}(k)\sigma_\beta]/[\sigma_\alpha], k = 0, 1, 2\cdots$.

Hence, after "prewhitening" the input, the cross correlation function between the prewhitened input and correspondingly transformed output is directly proportional to the response function. In practice, we do not know the theoretical function $\rho_{\alpha\beta}(k)$, so we must substitute estimates in υ_k to give

$$\hat{\upsilon_k} = [r_{\alpha\beta}(k)s_\beta]/[s_\alpha], \quad k = 0, 1, 2...$$

where

$$
\begin{aligned}
&r_{\alpha\beta}(k) = c_{\alpha\beta}(k)/[s_\alpha/s_\beta], \\
&c_{\alpha\beta}(k) = \frac{1}{n}\sum_{n=1}^{n-k}(\alpha_t - \bar{\alpha})(\beta_{t+k} - \bar{\beta}), \\
&s_\alpha = \sqrt{c_{\alpha\alpha}(0)}, \\
&s_\beta = \sqrt{c_{\beta\beta}(0)}, \quad k = 0, 1, 2, ...
\end{aligned}
$$

The preliminary estimates $\hat{\upsilon}_k$ can provide a rough basis for selecting suitable transfer function model. First, we may use the estimates $\hat{\upsilon}_k$ so obtained to make guesses of the order r_1 and r_2 of $\delta(B)$ and $\omega(B)$, and of the delay parameter $\mathfrak{b}$. Second, we do not consider the noise n_t now, substituting $Z_t = \hat{\upsilon}(B)X_t$ in the equation $\delta(B)Z_t = \omega(B)B^{\mathfrak{b}}X_t$, based on equating coefficients of B, to obtain initial estimates of the parameters $\delta(B)$ and $\omega(B)$.

4 The Identification of Outliers Via Genetic Algorithm

We let $\{Y_t, t = 0, 1, 2, \cdots\}$ be a zero mean and stationary time series: $Y_t = \sum\limits_{j=0}^{\infty} \psi_j \alpha_{t-j}$, where $\{\alpha_t\}$ is Gaussian zero mean white noise and $Var(\alpha_t) = \sigma^2$, $\{\psi_j, j = 0, 1, 2, \cdots\}$ form a absolutely summable sequence. Let γi_h denote the inverse autocovariance function of the process, for integer h. Also let $\rho i_h = \gamma i_h / \gamma i_0$ denote the inverse autocorrelations. We have

$$\gamma i_k + \psi_1 \gamma i_{k-1} + \psi_2 \gamma i_{k-2} + \cdots = 0, \quad k > 0 \tag{2}$$

When outliers are present, $\{Y_t, t = 0, 1, 2, \cdots\}$ is unobservable. Instead the time series $\{Z_t, t = 0, 1, 2, \cdots\}$ is observed which follows the model: $Z_t = Y_t + d_t$, where d_t is a deterministic perturbation. Let $\psi_j = 0$ if $j < 0$, then we have $d_t = \psi_{t-t_0} w_0$, if time is IO; or $d_t = w_0 \delta_{t,t_0}$, if time t_0 is AO , where w_0 denotes the outlier's magnitude at $t = t_0$. In the present setting, a chromosome ξ is a string of characters of assigned length n that can be evaluated in terms of the FF, where n is the number of observations of the time series, as each locus ξ_j is corresponding to an observation z_j where an outlier may occur. So, $\xi = (\xi_1, \xi_2, \cdots, \xi_n)$. Then a gene $\xi_j = 0$, if the locus is an outlier-free time point, $\xi_j = 1$ if the observation at this time point is an additive outlier, and $\xi_j = 2$ if it is an innovational outlier.

If k outliers are located at $t_1, t_2, \cdots, t_k$ and denoting by $Z = (z_1, z_2, \cdots, z_n)'$ the observed time series and by $Y = (y_1, y_2, \cdots, y_n)'$ the unobserved realization of Y_t, then $Z = \Psi W + Y$, where $W = (w_1, w_2, \cdots, w_k)'$ is the vector of the outliers' magnitudes at $t_1, t_2, \cdots, t_k$ and Ψ is the matrix of the $n \times k$ elements Ψ_{jh} defined as follows: If t_h is the timing of an IO, then $\Psi_{jh} = \psi_{j-t_h}$ if $j > t_h$ and 0, otherwise. If an AO is occurring in t_h, then $\Psi_{jh} = 1$ if $j = t_h$ and 0, otherwise. The idea is to seek for the matrix Ψ which maximizes the likelihood function. When n is large, Γ^{-1} may be replaced by the matrix of inverse autocovariances Γi . Then we have the likelihood function of the time series Z

$$L = p(Z \mid \xi, W) = 2\pi^{-n/2}(det\Gamma i)^{1/2} \exp\{-\frac{1}{2}(Z - \Psi W)'\Gamma i(Z - \Psi W)\} \tag{3}$$

The joint maximum likelihood estimate W^* of W , given the pattern of the outlying observations and Γi, is

$$W^* = (\Psi' \Gamma i \Psi)^{-1} \Psi' \Gamma i Z \quad and \quad Y^* = Z - \Psi W^* \tag{4}$$

In practice, however, Γi is seldom known, so we have to estimate it from the data. Similar to Baragona et al.[1], using the interpolator estimate of Y_t. we may obtain inverse autocorrelations interpolation estimates $\hat{\rho i}_j, j = 1, \cdots, m$ and the estimate $\hat{\gamma i_0}$ of inverse variance and the estimates of inverse autocovariance $\hat{\gamma} i_j = \hat{\gamma} i_0 \cdot \hat{\rho} i_j, j = 1, \cdots, m$. Once the inverse autocovariances have been estimated, the $\psi_j, j = 1, \cdots, q$ may be computed using Equation (2). Nevertheless, these estimates are biased because of the presence of the outliers. So we have to resort to an iterative scheme for (4), which is repeated until convergences. If convergence is attained, and since the likelihood is maximized with respect to the inverse autocorrelations, the argument of the exponential in likelihood's Formula (3) is simply $-n/2$, which is shown by the following proposition(the proof is omited).

Proposition 1 Using the former representation and estimations, then we have

$$-\frac{1}{2}(Z - \Psi W)'\Gamma(Z - \Psi W) = -\frac{1}{2}n$$

Therefore, we have

$$\log L = -\frac{n}{2}\log(2\pi) + \frac{1}{2}\log(det\Gamma i) - \frac{1}{2}n$$

Thus, the likelihood function basically depends on $det\Gamma i$. Because that the fitness function(FF) has to be positive, we let $F(\xi) = a \cdot b^{\log(det\Gamma i) - ck}$, where b is a real constant such that $b > 1$, and the constant terms in the exponent were dropped. We use $a = 10, b = 1.0001$ and $c = 14$ in this paper. We adopted (1) Population size $s = n + 1$; (2) Probability of crossover $p_c = 0.8$; (3) Probability of mutation $p_m = 0.05$; (4) The maximum number of outliers within a chromosome $g = 10$; (5) The number of iterations of the series' adjustment/parameters' estimation procedure was 3;(6) we perform as many iterations of the genetic algorithm as possible within a reasonable time period.

5 Simulation Studies

Example A In the following example, we consider the model

$$\begin{cases} (1 - 0.8B + 0.3B^2)x_t = \varepsilon_t \\ z_t = (1 - 0.5B)x_t - 4\delta_{t,30} + 5\delta_{t,31} - 5\delta_{t,80} - 4\delta_{t,90} + \\ \qquad 6 \times \dfrac{1 - 0.36B + 0.85B^2}{1 - 0.6B}\delta_{t,40} + \dfrac{1 - 0.36B + 0.85B^2}{1 - 0.6B}e_t \end{cases}$$

where $\{\varepsilon_t\}$ and $\{e_t\}$ are all normal white noise, their means are zero and variance $\sigma^2 = 1$.

We create 101 observations $x_0, x_1, \cdots, x_{100}$ of x_t and 100 observations z_1,...,z_{100} of z_t by simulation. Obviously, it is AO at $t = 30, 31, 80, 90$ singly and IO at

$t = 40$, and outlier magnitudes are $w_{30} = -4, w_{31} = 5, w_{80} = -5, w_{90} = -4$ and $w_{40} = 6$, respectively.

Applying our method to the above data and prewhitening the input series. Making $\{x_t\}$ follows an ARMA model:

$$(1 - 0.88088B + 0.32738B^2)x_t = \varepsilon_t.$$

Then we take the same manipulation to prewhiten $\{z_t\}$. By analyzing filtered cross correlation coefficient of $\{z_t\}$ and $\{x_t\}$, we obtain the transfer function $1.03416-0.73967B$ for $\{x_t\}$. Delete the influence of input process $\{x_t\}$ in response process $\{z_t\}$, and let $z_t^* = z_t - (1.03416 - 0.73967B)x_t$. We have that $\{z_t^*\}$ is an ARMA series include outliers. We detect the outliers in $\{z_t^*\}$ by applying the above method. Let $m = 2, q = 9, g = 10, s = 101$. Because the $30th$ point is very close to $31th$ point and they influence each other, it is difficult to identify the outliers. In this case, one needs larger number of iterations. We take 1000 iterations by standard genetic algorithm. and obtain the best individual at $612th$ iterations:

$$t = 30(AO), \quad w_{30}^* = -4.7172;$$
$$t = 31(AO), \quad w_{31}^* = 4.1010;$$
$$t = 40(IO), \quad w_{40}^* = 4.0487;$$
$$t = 80(AO), \quad w_{80}^* = -5.6440;$$
$$t = 90(AO), \quad w_{90}^* = -4.6228$$

The outcome is consistent with our prearrangement. The outliers in $\{z_t\}$ process are detected successfully, and there is no misjudgement.

6 Conclusions

There are AO and IO in our model, and also the outliers(AO) present consecutively. It is quite difficult to detect outliers, however, all of the outliers in the above model have been detected successfully, which shows our method is also efficient for the AO and IO problem in ARMAX model. Some other case studies also show that our method is effective in detecting outliers' location and type and in estimating their size for ARMAX model.

Acknowledgements

The research is supported by National Natural Science Foundation of China (No.11171065)

References

[1] Baragona R, Battaglia F, Calzini C. (2001), "Genetic algorithms for the identification of additive and innovation outliers in time series", *Computational Statistics & Data Analysis,* Vol.37, pp.1-12.

[2] Peña D, Sánchez I. (2005), "Multifold predictive validation in armax time series models", *Journal of the American Statistical Association,* Vol.100, pp.135-146.

[3] Chen P, Li L, Liu Y and Lin J.G. (2010), "Detection of outliers and patches in bilinear time series models", *Mathematical Problems in Engineering*, Vol.2010, pp.1-10.

[4] Chen P, Yang J, Li L.Y. (2013), "Synthetic detection of change point and outliers in bilinear time series models", *International Journal of Systems Science,* http://dx.doi.org/10.1080/00207721.2013.777983. (In press)

[5] Huang L, Pang W, Wang K.P, Zhou C.G, Xiao Y. (2004), "Improved genetic algorithm for vehicle routing problem with time windows", *Advances in Systems Science and Applications,* Vol.4, pp.118-124.

[6] Box G.E.P, Jenkins G.M, Reinsel G.C. (1994), *Time Series Analysis: Forecasting and Control*, third edition, Prentice-Hall, Englewood Cliffs, NJ.

Corresponding author

Author can be contacted at cp18@263.net.cn