基于 ARMA 模型检出时间序列中的离群值*

王振华　吴余海
（中国科学院计算中心）

THE INSPECTION OF OUTLIERS BY ARMA MODEL IN TIME SERIES

Wang Zhen-hua　Wu Yu-hai
(Computing Center, Academia Sinica)

Abstract

The ARMA model in time series is used to check the time outliers that have occurred and to estimate their values. An ARMA model is established after the outliers are obliterated. In looking for the outliers, robust estimation and 3σ revised rules are used to obtain the results in a faster and more reliable computation.

§1. 引　言

时间序列观测过程中有时会受到突然干扰，使得数据序列出现离群值（Outliers），而且这种干扰出现的时刻往往未知．例如，观测仪器的偶然故障、人为的观测差错，用磁带记录信号时，出现代码错误等现象都会造成离群值的发生．如果不检出离群值而直接用观测序列作时间序列分析，如周期识别、参数估计、预测等，会产生假像，甚致得出错误的结果．所以，从观测数据中将离群值检出并消除是必要的．同时还需找到出现突然干扰的时刻，以便查出该时刻出现突然干扰或其它事故的原因．本文不仅给出离群值的估计，也给出离群值出现时刻的估计．

§2. 离群值的估计

考虑离群值的两种情况，一种是可加离群值（Additive Outliers），记为 AO；另一种是新息离群值（Innovation Outliers），记为 *AI*．

$$\text{AO:}\quad X_t^* = X_t + AO_T \cdot I_t(T), \tag{1}$$

* 1990 年 1 月 18 日收到．

$$\text{AI}:\ X_t^* = X_t + \phi^{-1}(B)\cdot\theta(B)AI_T\cdot I_t(T), \tag{2}$$

这里 X_t^* 为观测所得数据序列，X_t 为未受干扰冲击的时间序列，AO_T 为时刻 T 出现的可加 Outliers，AI_T 为时刻 T 出现的新息 Outliers，$I_t(T)$ 为示性函数，$I_t(T)=1$，如果 $t=T$，否则为 0. 假定 X_t 是均值为零的平稳 ARMA(p,q) 序列，可表为

$$\phi(B)\cdot X_t = \theta(B)\cdot a_t, \tag{3}$$

其中 $\theta(B)$、$\phi(B)$ 分别满足可逆性及稳定性条件. a_t 为白噪声，$\phi(B)=1-\phi_1B-\phi_2B^2-\cdots-\phi_pB^p$，$\theta(B)=1+\theta_1B+\theta_2B^2+\cdots+\theta_qB^q$，$B$ 为后移算子.

为了得到 Outliers 估计，先假定干扰冲击的时刻 T 和参数 ϕ、θ 已知，即

$$W(B)X_t^* = e_t, \tag{4}$$

其中 e_t 为扰动后的序列.

$$W(B)=\phi(B)\theta^{-1}(B)=1+W_1B+W_2B^2+\cdots=\sum_{i=0}^{\infty}W_iB^i, \tag{5}$$

其中 $W_0\equiv 1$，对于 AO 模型有

$$e_t = AO_TW(B)I_t(T)+a_t = AO_TX_{ot}+a_t \tag{6}$$

对于 AI 模型有

$$e_t = AI_T\cdot I_t(T)+a_t = AI_T\cdot X_{It}+a_t, \tag{7}$$

其中 $X_{ot}=W(B)I_t(T)$，$X_{It}=I_t(T)$. 由熟知的最小二乘法得到 AO_T 和 AI_T 的估计分别为

$$\widehat{AO}_T = \frac{\sum_t e_tX_{ot}}{\sum_t X_{ot}^2}, \tag{8}$$

$$\widehat{AI}_T = \frac{\sum_t e_tX_{It}}{\sum_t X_{It}^2}. \tag{9}$$

(8)式的分子可写为

$$\begin{aligned}\sum_t e_t\cdot X_{ot} &= \sum_t e_t[1+W_1B+W_2B^2+\cdots]I_t(T)\\ &= \sum_t e_t[I_t(T)+W_1I_{t-1}(T)+W_2I_{t-2}(T)+\cdots]\\ &= e_T+W_1e_{T+1}+W_2e_{T+2}+\cdots\\ &= \sum_{t\geq 0} e_{T+K}\cdot W_K = W(B^{-1})e_T\\ &= W(F)e_T.\end{aligned} \tag{10}$$

这里 $F=B^{-1}$ 为前移算子. (8)式的分母为

$$\begin{aligned}\sum_t X_{ot}^2 &= \sum_t (I_t(T)+W_1I_{t-1}(T)+W_2I_{t-2}(T)+\cdots)^2\\ &= \sum_t\Big(\sum_{j,l\geq 0} W_j\cdot W_l\cdot I_{t-j}(T)\cdot I_{t-l}(T)\Big)\end{aligned}$$

$$= \sum_{i \geqslant 0} W_i^2 = \eta^2. \tag{11}$$

由(10)、(11)两式给出 AO_T 的估计

$$\widehat{AO}_T = \frac{W(F)e_T}{\eta^2}. \tag{12}$$

由最小二乘估计，知 AO_T 的方差为

$$V_{ar}(\widehat{AO}_T) = \frac{\sigma^2}{\sum_{i \geqslant 0} W_i^2} = \frac{\sigma^2}{\eta^2}. \tag{13}$$

(9)式的分子为

$$\sum_t e_t \cdot X_{1t} = \sum_t e_t I_t(T) = e_T. \tag{14}$$

(9)式的分母为

$$\sum_t X_{1t}^2 = \sum_t I_t^2(T) = 1. \tag{15}$$

于是 AI_T 的估计及其方差如下给出：

$$\widehat{AI}_T = e_T, \quad V_{ar}(AI_T) = \sigma^2. \tag{16}$$

对于 $\widehat{AO}_T$ 和 $\widehat{AI}_T$ 的显著性检验要考虑下面的两个估计量：

$$\hat{S}_{oT} = \eta\widehat{AO}_T/\sigma \text{ 和 } \hat{S}_{iT} = \widehat{AI}_T/\sigma. \tag{17}$$

§3. 估计离群值的具体步骤

实际上 Outliers 的干扰冲击时刻是未知的，它也是一个估计值，用迭代法实现。

1) 先对 X_t^* 作 ARMA(p,q) 模型分析，得到

$$\Phi(B)X_t^* = \theta(B)e_t. \tag{18}$$

计算

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(\hat{e}_i - \hat{m})^2, \tag{19}$$

式中 $\hat{m}$ 表示 $\{e_i\}$ 的估计均值。考虑到 $\hat{e}_i$ 已受到离群值的影响，对(19)式中的位置参数和尺度参数 $\hat{m}$、$\hat{\sigma}$ 可采用 Robust 估计方法[5]来得到。

令 $\hat{m}^{(0)}$、$\hat{\sigma}^{(0)}$ 分别为位置和尺度的初始估计，例如可取 $\hat{m}^{(0)} = \text{Med}\{\hat{e}_i\}$，$\hat{\sigma}^{(0)} = \text{Med}\{|\hat{e}_i - m^{(0)}|\}$，此处 Med$\{\cdot\}$ 表示取$\{\cdot\}$的中位数。若 $\hat{m}^{(j)}$、$\hat{\sigma}^{(j)}$ 分别为第 j 步迭代得到的位置和尺度参数的估值，则其第 $j+1$ 步迭代为

$$[\hat{\sigma}^{(j+1)}]^2 = \frac{1}{(N-1)\beta}\cdot\sum_{i=1}^{N}\phi^2\left(\frac{\hat{e}_i - m^{(j)}}{\hat{\sigma}^{(j)}}\right)[\hat{\sigma}^{(j)}]^2, \tag{20}$$

$$[\hat{m}^{(j+1)}] = \hat{m}^{(j)} + \frac{\sum_{i=1}^{N}\phi\left(\frac{\hat{e}_i - m^{(j)}}{\hat{\sigma}^{(j)}}\right)\hat{\sigma}^{(j)}}{\sum_{i=1}^{N}\phi'\left(\frac{\hat{e}_i - m^{(j)}}{\hat{\sigma}^{(j)}}\right)}, \tag{21}$$

这里 $\psi(x)=\max[-k,\min(k,x)]$，$k$ 为一给定常数，$\psi'(x)$ 为 $\psi(x)$ 的导数，$\beta=E_\Phi(\psi^2(x))$，Φ 表示标准正态分布。利用(20)、(21)两式，一般迭代几次即可得到位置和尺度参数的 Robust 估计。

2) 为了减少计算量。先找出 Outliers 出现的嫌疑时刻，然后对这些可疑点再作进一步判断，即对 $t=1,2,\cdots,N$ 作如下检查：

若
$$\left|\frac{\hat{e}_t-\hat{m}}{\hat{\sigma}}\right|>\alpha, \tag{22}$$

则使(22)式成立的时刻记作 $T_1,T_2,\cdots T_l$。这些时刻就是 Outliers 出现的可疑点。此处 α 为一给定的正数(如取 $\alpha=0.9$)。

3) 计算

$$\eta=\left(\sum_{j\geqslant 0}W_j^2\right)^{1/2}$$

注意$\{W_j\}$实际上是逆函数，可表为 $W(B)=\sum_{i=0}^{\infty}W_iB^i$ 其中 $W_0\equiv -1$。用递推法由参数 ϕ_i 和 θ_i 算出[6]、[7]。

$$W_j=\phi_j^*+\sum_{i=1}^{j}\theta_i^*W_{j-i},\quad W_0\equiv -1, \tag{23}$$

其中

$$\phi_i^*=\begin{cases}\phi_i, & 1\leqslant i\leqslant p,\\ 0, & i>p,\end{cases} \tag{24}$$

$$\theta_i^*=\begin{cases}\theta_i, & 1\leqslant i\leqslant q,\\ 0, & i>q.\end{cases} \tag{25}$$

由(12)、(16)、(17)式得

$$\widehat{AO}_t=\left(\sum_{k\geqslant 0}\hat{e}_{k+t}\cdot W_k\right)\Big/\eta^2, \tag{26}$$

$$\widehat{AI}_t=\hat{e}_t, \tag{27}$$

$$\hat{S}_{ot}=\widehat{AO}_t\cdot\eta/\hat{\sigma}, \tag{28}$$

$$\hat{S}_{it}=\widehat{AI}_t/\hat{\sigma}. \tag{29}$$

对于 $t=T_1,T_2,\cdots,T_l$，计算(26)—(29)式，再求

$$S_T^*=\max_t\cdot\max_{O,I}(|\hat{S}_{ot}|,|\hat{S}_{it}|). \tag{30}$$

若 $S^*>C$（如取 $C=3$），则 S_T^* 作为 AO_T 或 AI_T 的估计值，T 为 Outlier 出现的时刻。

4) 计算

$$\tilde{e}_T=\hat{e}_T-\widehat{AO}_T, \tag{31}$$

或

$$\tilde{e}_T=\hat{e}_T-\widehat{AI}_T. \tag{32}$$

5) 由步骤(4)给出的剩余 $\tilde{e}_T$ 按(20)，(21)式算出 σ 的估计，重复步骤 2) 和 4) 直至

Outliers 都找出为止.

6) 对去掉 Outliers 的时间序列作 ARMA 分析,即对

$$X_t = X_t^* - \widehat{AO}_T \cdot I_t(T) \tag{33}$$

或

$$X_t = X_t^* - \phi^{-1}(B) \cdot \theta(B)\widehat{AI}_T \cdot I_t(T). \tag{34}$$

再作 ARMA 分析. 这里的 $\phi^{-1}(B) \cdot \theta(B) = G(B) = 1 + G_1B + G_2B^2 + \cdots$, $\{G_i\}$ 为格林函数. 由参数 ϕ_i 和 θ_i 递推算出 [6]、[7] $G_j = -\theta_j^* + \sum_{i=1}^{j} \phi_i^* \cdot G_{j-i}$, $G_0 \equiv 1$, 式中 ϕ_i^*、θ_j^* 分别由(24)、(25)两式给出.

§4. 算 例

用基于上述算法所编制的程序对下面具有 Outliers 的数据作了分析,这些数据 X_t^* 为

80.9, 83.4, 47.7, 47.6, 30.7, 12.2, 9.6, 10.6, 32.4, 47.6,
54.0, 62.9, 85.9, 61.2, 45.1, 36.4, 20.9, 71.4, 37.8, 69.8,
106.1, 100.8, 81.6, 66.5, 34.8, 30.6, 7.0, 19.8, 92.5, 154.4,
125.9, 84.8, 68.1, 38.5, 22.8, 10.2, 24.1, 82.9, 132.0, 130.9,
118.1, 89.9, 66.6, 60.0, 46.9, 41.0, 21.3, 16.0, 6.4, 4.1,
6.8, 14.5, 34.0, 45.0, 43.1, 47.5, 42.2, 28.1, 10.1, 8.1,
2.5, 0., 1.4, 5., 12.2, 13.9, 35.4, 45.8, 41.1, 30.4,
23.9, 15.7, 6.6, 4.0, 1.8, 8.5, 16.6, 36.3, 49.7, 62.5,
67., 71.0, 47.8, 27.5, 8.5, 13.2, 56.9, 121.5, 138.3, 103.2,
85.8, 63.2, 36.8, 24.2, 10.7, 15.0, 40.1, 61.5, 98.5, 124.3,
95.9, 66.5, 64.5, 54.2, 39.0, 20.6, 6.7, 4.3, 22.8, 54.8,
93.8, 95.7, 77.2, 59.1, 44., 47.0, 30.5, 163., 7.3, 37.3,
73.9, 139.1, 111.2, 101.7, 66.3, 44.7, 17.1, 11.3, 12.3, 3.4,
6.0, 32.3, 54.3, 59.7, 63.7, 63.5, 52.2, 25.4, 13.1, 6.8,
6.8, 6.3, 7.1, 35.6, 70.3, 84.9, 78.0, 64.0, 41.8, 26.2,
26.7, 12.1, 9.5, 2.7, 5.0, 24.4, 42.0, 63.5, 53.8, 62.0,
48.5, 43.9, 18.6, 5.7, 3.6, 1.4, 9.6, 47.4, 57.1, 103.9,
80.6, 43.6, 37.6, 26.1, 14.2, 5.8, 16.7

未去除 Outliers 的 ARMA 模型为:

$$X_t^* - 0.81583X_{t-1}^* + 0.127569 * 10^{-2}X_{t-2}^* + 0.25576X_{t-3}^* = e_t$$

经程序计算找出第 118 个数据有可加 Outliers, Outliers 估值为 $\widehat{AO}_{118} = 147.39$.

$$X_{118} = X_{118}^* - \widehat{AO}_{118} = 163 - 147.39 = 15.61.$$

X_{118} 的实际值为 16.3; 去除 Outlier 影响后 ARMA 分析结果为

$$X_t - 1.43575X_{t-1} + 0.70524X_{t-2} = -0.220459 + e_t - 0.346022e_{t-1}.$$

参　考　文　献

[1] BOX, G. E. P, G. C. Tiao, Intervention analysis with applications to economic and environmental applications, J. Amer. Statist. Assoc., 70(1975), 70—79.

[2] Fox, A. J. (1972). Outliers in time series, J. Roy. Statist. Soc., Ser. B. 34, 350—363.

[3] Hillmer, S. C., G. C. Tiao An ARIMA-model-based approach toseasonal adjustment, J. Amer. Statist. Assoc., 77(1982), 63—70.

[4] B. Abraham and J. Ledolter, Statistical methods for forecasting. John Wiley & Sons, 1983.

[5] P. J. Huber, Robust Statistics, John Wiley & Sons, 1981.

[6] 项静恬，杜金观，史久恩，动态数据处理——时间序列分析，气象出版社，1986.

[7] 杨位钦，顾　岚，时间序列分析与动态数据建模，北京工业学院出版，1986.