第1卷 第5期
2008 年 7 月

www.paper.edu.cn
中国科技论文在线
Sciencepaper Online

Vol.1 NO.5
July 2008

运用极值理论挖掘时间序列模型中的异常点

田玉柱[1]，陈　平[2]

（1. 天水师范学院数学与统计学院，甘肃天水　741001；
2. 东南大学数学系，南京　211189）

摘要：用检验的方法诊断时间序列异常点的关键就是决定检验统计量在一定的显著性水平下是否超越某一临界值，临界值的选取一般文献都描述得很模糊。本文运用极值理论给出诊断 AR（p）模型异常点选取临界值的分布近似方法，这种方法选取的临界值可保证控制在一定的显著性水平下，而且还可以计算出检验的渐近 p 值，比一般仿真选取的临界值更科学、合理。

关键词：概率论与数理统计；时间序列分析；极值分布；AR（p）模型；IO 异常点；AO 异常点

中图分类号：O211.61　　**文献标识码**：A　　**文章编号**：1674-2850(2008)00-0575-7

Outlier detection in time series based on extreme value theory

TIAN Yuzhu[1], CHEN Ping[2]

(1. *Department of Mathematics and Statistics, Tianshui Normal College, Tianshui, Gansu*　741001;
2. *Department of Mathematics, Southeast University, Nanjing*　211189)

Abstract: The detection of outliers in time series hinges on determining whether the test statistic exceeds a critical value for a given significant level. This value is prescribed only vaguely by many authors with no reference to any significant level. This paper will give out an asymptotic critical value in a fixed significant level and an asymptotic p-value for testing by means of extreme value theory in order to detect outliers in AR(p) model. This method is better than the simulation and more scientific.

Key words: probability and mathematical statistics; time series analysis; extreme value distribution; AR(p) model; IO outlier; AO outlier

0　引言

BOX 等[1]指出异常点对时间序列的模型识别、参数估计、诊断检验甚至预测都有重要的影响。自 1972 年 BOX 在时间序列中定义了异常点以来，对时间序列异常点的诊断国内外已有大量的研究，如 FOX[2]，CHANG[3]，CHEN[4]，ALEJANDRO[5] 及 CHAREKA[6] 等。本文运用极值理论给出诊断 AR（p）模型异常点的检验统计量在原假设（无异常点）下的渐近分布，再取相应的渐近临界值来诊断异常点，并用渐近 p 值来比较不同检验统计量的诊断效果，最后通过两个模拟例子加以说明。

1　AR（p）模型及其异常点

1.1　AR（p）模型

设 $\{x_t\}_{t=1}^n$ 表示一个带有截距项的 AR（p）模型的 n 个观测值，即 x_t 可表示为

$$x_t = \mu + \sum_{j=1}^{p} a_j x_{t-j} + \varepsilon_t \tag{1}$$

其中，ε_t 是 iid 的 Gauss 白噪声且 $E(\varepsilon_t)=0$，$E(\varepsilon_t^2)=\sigma^2$，实数 a_1，a_2，…，a_p（$a_p \neq 0$）使得

作者简介：田玉柱（1982—），男，硕士研究生，主要研究方向：时间序列与应用概率统计
通信联系人：陈平，教授，主要研究方向：金融与工程时间序列分析. E-mail：pole1999@163.com

$$\Phi(z)=1-a_1z-a_2z^2-\cdots-a_pz^p\neq 0 \quad |z|\leqslant 1$$

由于$\{x_t\}$的平稳性，令$\Phi(B)=1-\sum_{j=1}^{p}a_jB^j$，$B^kx_t=x_{t-k}$；则

$\Phi(B)^{-1}=\Psi(B)=\sum_{j=0}^{\infty}\varphi_jB^j$，$\varphi_0=1$，故模型（1）可写为：$\Phi(B)x_t=\mu+\varepsilon_t$。

1.2 AR（p）模型的异常点模型及其检验统计量

本文主要讨论常见的IO和AO异常点，也将讨论成片异常点的情形。假设模型（1）的前p个观测中不出现异常点。

1.2.1 孤立异常点情形

设$\{x_t\}_{t=1}^{n}$在时刻q（$p<q$，$p+q\leqslant n$）处有一个大小为ω的异常点，观测$\{y_t\}$可描述为

1）IO（革新异常点）：$\Phi(B)y_t=\mu+\eta_t$，$\eta_t=\varepsilon_t+\omega^{IO}\cdot\delta(t-q)$ （2）

2）AO（革新异常点）：$y_t=x_t+\omega^{AO}\cdot\delta(t-q)$ （3）

上述$\delta(j)=\begin{cases}1, j=0\\0, \text{otherwise}\end{cases}$，IO和AO的诊断就是检验原假设$H_0:\omega_q^{IO}=0$和$H_0':\omega_q^{AO}=0$在一定的显著性水平下是否成立，首先给出两种情形下$\omega_q$的估计。如果模型参数已知，则两种情形下$\omega$的极大似然估计分别为

$$\omega_q^{IO}=\eta_q \tag{4}$$

$$\omega_q^{AO}=\frac{\eta_q-\sum_1^p a_j\eta_{q+j}}{1+\sum_1^p a_j^2} \tag{5}$$

其中$\eta_t=(y_t-\sum_1^p a_j\cdot y_{t-j})-\mu$表示观测残差，且相应地有

$$\mathrm{var}(\omega_q^{IO})=\sigma^2,\ \mathrm{var}(\omega_q^{AO})=\sigma^2/\Delta,\ \Delta=(1+\sum_{j=1}^{p}a_j^2)$$

异常点估计（4）和（5）中η_t可以用$\hat{\eta}_t=(y_t-\sum_{j=1}^{p})a_jy_{t-j}-\hat{\mu}$代替，其中$\hat{\mu}$，$a_1$，…，$a_p$分别表示$\mu$，$a_1$，…，$a_p$的估计，在数值模拟中取最小二乘估计。以下先假设模型所有参数均已知，未知情形后面还将给予适当说明。下面来考察ω_q^{IO}和ω_q^{AO}在原假设下的分布。

1）如果异常点的位置q已知

定理1 （1）在$H_0:\omega_q^{IO}=0$（时刻q无IO）下，IO的似然比检验统计量为$\lambda_q^{IO}=\omega_q^{IO}/\sigma$，则在$H_0$下，$\lambda_q^{IO}\sim N(0,1)$；

（2）在$H_0':\omega_q^{AO}=0$（时刻q无AO）下，AO的似然比检验统计量为$\lambda_q^{AO}=\omega_q^{AO}\cdot\Delta^{1/2}/\sigma$，则在$H_0'$下，$\lambda_q^{AO}\sim N(0,1)$。

2）在实际应用中关于异常点位置的信息是很少的，即q是未知的，因此必须在每一时刻点检查诊断统计量的大小。此时定理1中的λ_q^{IO}和λ_q^{AO}分别换为λ_t^{IO}与λ_t^{AO}后，接下来的定理说明了过程$\{\lambda_t^{IO}\}$和$\{\lambda_t^{AO}\}$分别构成了标准的平稳Gauss过程。

定理2 （1）在$H_0:\omega_t^{IO}=0$（任意t时刻都无IO）下，$\{\omega_t^{IO}\}$是一零均值方差为σ^2的Gauss平稳过程，其自协方差函数为$r_{IO}(h)=\sigma^2\cdot I(h=0)$，其中$I_A(x)$为示性函数；

（2）在$H_0':\omega_t^{AO}=0$（任意t时刻都无AO）下，$\{\omega_t^{AO}\}$是一零均值方差为σ^2/Δ的Gauss平稳过程，且自协方差函数为

$$r_{AO}(h)=\Delta^{-2}\cdot(a_1a_{h+1}+a_2a_{h+2}+\cdots+a_{p-h}a_p)\cdot\sigma^2\cdot I(1\leqslant h\leqslant p)$$

证明 正态分布的证明由定理1可知，下面只给出它们的自协方差函数。

① 在$H_0:\omega_t^{IO}=0$下，$\eta_t=\varepsilon_t$，由式（4）有$r_{IO}(h)=\sigma^2\cdot I(h=0)$

② 在 H_0'：$\omega_t^{AO}=0$ 下，同样 $\eta_t=\varepsilon_t$，且由式（5）有

$$\begin{aligned} r_{AO}(h) &= \mathrm{cov}(\omega_t^{AO},\omega_{t+h}^{AO}) \\ &= \mathrm{cov}\left[\Delta^{-1}(\eta_t-\sum\nolimits_1^p a_j\eta_{t+j}),\Delta^{-1}(\eta_{t+h}-\sum\nolimits_1^p a_j\eta_{t+h+j})\right] \\ &= \Delta^{-2}(a_1a_{h+1}+a_2a_{h+2}+\cdots+a_{p-h}a_p)\cdot\sigma^2\cdot I(1\leqslant h\leqslant p) \end{aligned}$$

1.2.2 成片异常点情形

连续 k 个异常点（$k>1$）的情形，$k=1$ 时退化为孤立异常点的情形。本文仅讨论异常点大小相等的情形。假设模型（1）的 n 个观测值 $\{x_t\}_{t=1}^n$ 从时刻 q 开始在 $q_j=q+j-1$（$j=1,2,\cdots,k$）连续出现 k 个大小均为 ω 的异常点，即 $\omega_q=\cdots=\omega_{q+k-1}\triangleq\omega_q^{AOk}$（或 ω_q^{IOk}），观测序列 $\{y_t\}$ 可描述为：

1）连续 k 个 IO

$$\Phi(B)y_t=\mu+\eta_t,\ \eta_t=\varepsilon_t+\omega_q^{IOk}\cdot\sum\nolimits_{j=1}^k\zeta_t^{(q_j)} \tag{6}$$

上述 $\zeta_t^{(q_j)}=\begin{cases}1, & t=q_j\\ 0, & \text{otherwise}\end{cases}$，类似孤立异常点情形，如果所有的参数 μ，a_1，…，a_p 已知，则此时 ω_q^{IOk} 的回归方程为：$\eta_t=\omega_q^{IOk}\cdot\sum_{j=1}^k\zeta_t^{(q_j)}+\varepsilon_t$，得 ω_q^{IOk} 的极大似然估计为 $\omega_q^{IOk}=k^{-1}\sum_{j=1}^k\eta_{t+j-1}$，在原假设 H_0（不存在成片 IO）下，知 $E(\omega_q^{IOk})=0$，$\mathrm{var}(\omega_q^{IOk})=\sigma^2$，标准化后得似然比统计量且在 H_0 下有

$$\lambda_q^{IOk}\triangleq\sigma^{-1}\omega_q^{IOk}=(k\sigma)^{-1}\sum\nolimits_{j=1}^k\eta_{t+j-1}\sim N(0,1) \tag{7}$$

式（7）说明 q 已知时 λ_q^{IOk} 服从标准正态分布，但实际上 q 一般未知，此时和孤立异常点情形一样，在原假设下 $\{\omega_t^{IOk}\}$ 或 $\{\lambda_t^{IOk}\}$（后面均表示 $\{\omega_t^{IOk}\}$ 的标准化序列）就构成了 Gauss 过程，且 $\{\omega_t^{IOk}\}$ 的自协方差函数为

$\mathrm{cov}(\omega_t^{IOk},\omega_{t+h}^{IOk})=\mathrm{cov}(k^{-1}\sum_{j=1}^k\eta_{t+j-1},k^{-1}\sum_{j=1}^k\eta_{t+h+j-1})=k^{-2}\sigma^2\cdot(k-h)\cdot I(h<k)$，因此 $\{\omega_t^{IOk}\}$ 或 $\{\lambda_t^{IOk}\}$ 就是 *Gauss* 平稳过程。

2）连续 k 个 AO

$$y_t=x_t+\omega_q^{AOk}\cdot\sum\nolimits_{j=1}^k\zeta_t^{(q_j)} \tag{8}$$

上述 $\zeta_t^{(q_j)}$ 同式（6），如果所有的参数 μ，a_1，a_2，…，a_p 已知，则 ω_q^{AOk} 的回归方程分别为 $\eta_t=\omega_q^{AOk}\cdot\Phi(B)(\sum_{j=1}^k\zeta_t^{(q_j)})+\varepsilon_t$，$\omega_q^{AOk}$ 的估计一般形式非常复杂，就简单情形 $p=1$，$k=3$ 时说明即可，ω_q^{AO3} 的极大似然估计为

$$\omega_q^{AO3}=\left[\eta_q+(1-a_1)\eta_{t+1}+(1-a_1)\eta_{t+2}-a_1\eta_{t+3}\right]/D,\quad D=1+2(1-a_1)^2+a_1^2$$

在原假设 H_0'（不存在成片 AO 异常点）下，知 $E(\omega_q^{AO3})=0$，$\mathrm{var}(\omega_q^{AO3})=\sigma^2/D$，标准化后得似然比统计量且在 H_0' 下有：$\lambda_q^{AO3}\triangleq\omega_q^{AO3}\cdot D^{1/2}/\sigma\sim N(0,1)$，这说明当 q 已知时 λ_q^{AO3} 是标准正态分布，而当 q 位置未知时，$\{\omega_t^{AO3}\}$ 或 $\{\lambda_t^{AO3}\}$ 同样构成了 Gauss 过程，且得 $\{\omega_t^{AO3}\}$ 的自协方差函数是

$$\mathrm{cov}(\omega_t^{AO3},\omega_{t+h}^{AO3})=\begin{cases}2\sigma^2\cdot D^{-2}(1-a_1)^2, & h=1\\ \sigma^2\cdot D^{-2}(1-a_1)^2, & h=2\\ \sigma^2\cdot D^{-2}(-a_1), & h=3\\ 0, & \text{otherwise}\end{cases}$$

显然 $\{\omega_t^{AO3}\}$ 或 $\{\lambda_t^{AO3}\}$ 也是平稳的 Gauss 过程。而且 $\{\omega_t^{AO3}\}$ 的自相关函数是 3 步截尾的，即 $\mathrm{cov}(\omega_t^{AO3},\omega_{t+h}^{AO3})=0$，$(h>3)$，当然可以验证 AR（1）模型相应 $\{\omega_t^{AOk}\}$ 的自相关函数是 k 步截尾的，进一步 AR（p）模型相应 $\{\omega_t^{AOk}\}$ 和 $\{\lambda_t^{AOk}\}$（下文中均表示 $\{\omega_t^{AOk}\}$ 的标准化序列）的自相关函数也是有限步截尾的，因此都是平稳的正态过程。

以上定理中都假设 AR（p）模型的所有参数都已知，实际上这在异常点诊断过程中都需要估计，AR（p）模型自回归系数用普通的最小二乘估计，而残差方差在孤立的 IO 异常点时取估计 $\hat{\sigma}_{q,IO}^2=(n-p)^{-1}\{\sum_{t=p+1,t\neq q}^{n}\hat{\eta}_t^2\}$，在 AO 时取 $\hat{\sigma}_{q,AO}^2=(n-p)^{-1}\{\sum_{t=p+1}^{n}\hat{\eta}_t^2-(\omega_q^{AO})^2\sum_{j=1}^{p}a_j^2\}$，关于成片 IOk 取估计 $\hat{\sigma}_{q,IOk}^2=(n-p)^{-1}\{\sum_{\substack{t=p+1\\ t<q,t\geqslant q+k}}^{n}\hat{\eta}_t^2\}$ 代替，这里 $\hat{\eta}_t$ 表示观测残差。

2 极值理论及其 AR（p）模型的异常点诊断

2.1 异常点诊断的关键

由异常点模型（2）和（3）可知，AR（p）模型 IO 和 AO 异常点诊断的关键就是决定统计量 $\max_{t=1}^{n}\{|\lambda_t^{IO}|\}$ 和 $\max_{t=1}^{n}\{|\lambda_t^{AO}|\}$ 在一定的水平下是否超越某一临界值，临界值的选取一般借助于仿真，但一般并不能控制在一定的显著性水平之下，无法计算检验的功效或 p 值。临界值合理的取法应当依据检验统计量在原假设（无异常点）下的分布或渐近分布来决定。定理 1 已经给出了当异常点位置已知时的分布，但是实际异常点位置是未知的，此时定理 2 说明了检验统计量不是序列独立的，它们构成了平稳 Gauss 过程，必须考虑这种相关性对统计量 $\max_{t=1}^{n}\{|\lambda_t^{IO}|\}$ 和 $\max_{t=1}^{n}\{|\lambda_t^{AO}|\}$ 的影响。下面运用极值理论对这个问题加以讨论。

2.2 极值理论及诊断统计量的渐近分布

定义 1 设 $\{X_1,\cdots,X_n\}iid\sim F(x)$，$Z_n\triangleq\max\{X_1,\cdots,X_n\}$ 称为 $\{X_1,\cdots,X_n\}$ 的最大值。

可以通过选取正则化序列 c_n 和 d_n 做仿射变换使得 $d_n^{-1}(Z_n-c_n)$ 的极限分布非退化，而且必为下面三个分布之一。

定义 2 Gumbel 分布：$G(x)=\exp\{-\exp(-x)\}\quad x\in R$

Frechet 分布：$\Phi_{1,\alpha}=\begin{cases}0 & x\leqslant 0\\ \exp(-x^{-\alpha}) & \text{otherwise}\end{cases}$

Weibull 分布：$\Psi_{2,\alpha}=\begin{cases}\exp\{-(-x)^{\alpha}\} & x\leqslant 0\\ 1 & \text{otherwise}\end{cases}$

其中 $\alpha>0$，且正态、指数、对数正态等其极值分布都为 Gumbel 分布。

引理 1 设 $\{X_1\cdots X_n\cdots\}$ 是零均值的 Gauss 平稳序列，记其自相关函数是 $\rho(h)$，如果 Berman 条件成立即 $\lim_{h\to\infty}\rho(h)\cdot\log h=0$，则 $d_n^{-1}(Z_n-c_n)\to\Lambda(x)=\exp\{-\exp(-x)\}$

其中 $Z_n=\max\{X_1,\cdots,X_n\}$，$d_n=(2\log n)^{-1/2}$，$c_n=d_n^{-1}-d_n(\log\log n+\log 4\pi)/2$

引理 2 设 $\{X_1\cdots X_n\cdots\}$ 是零均值的 Gauss 平稳序列，令 $A_n=\max_{t=1}^{n}\{|X_t|\}$，$Z_n=\max_{t=1}^{n}\{X_n\}$ 如果 X_n 对 Berman 条件成立，则 A_n 与 Z_{2n} 有同样的分布。

在异常点诊断中，一般考虑绝对值统计 $\max_{t=1}^{n}\{|\lambda_t^{IO}|\}$ 和 $\max_{t=1}^{n}\{|\lambda_t^{AO}|\}$ 的大小，也可考虑平方统计量 $\max_{t=1}^{n}\{(\lambda_t^{IO})^2\}$ 和 $\max_{t=1}^{n}\{(\lambda_t^{AO})^2\}$ 的分布，下面分别考察它们的分布。

定理 3 在定理 2 的条件下，分别定义 $\eta_{IO}\triangleq\max_{t=1}^{n}\{|\lambda_t^{IO}|\}$ 和 $\eta_{AO}\triangleq\max_{t=1}^{n}\{|\lambda_t^{AO}|\}$，则当 $n\to\infty$ 时有

$$P(\eta_{IO}\leqslant c_{2n}+d_{2n}x)\to\Lambda(x) \tag{9}$$

$$P(\eta_{AO}\leqslant c_{2n}+d_{2n}x)\to\Lambda(x) \tag{10}$$

其中 $c_{2n}=d_{2n}^{-1}-d_{2n}\{\log\log(2n)+\log 4\pi\}/2$，$d_{2n}=\{2\log(2n)\}^{-1/2}$

证明　由定理 2 的（1）可知，在 $H_0:\omega_t^{IO}=0$ 下，$\{\lambda_t^{IO}\}$ 的自相关函数为 $\rho_{IO}(h)=I(h=0)$ 即 $\{\lambda_t^{IO}\}$ 是序列独立的，即显然 $\lim_{h\to\infty}\rho_{IO}(h)\cdot\log h=0$，即 Berman 条件满足。又知 $\{\lambda_t^{IO}\}$ 是标准的 Gauss 平稳过程，故由引理 1 知 $p(\max_{t=1}^{n}\{\lambda_t^{IO}\}\leqslant c_n+d_n x)\to\Lambda(x)$，再应用引理 2 知 $\eta_{IO}\triangleq\max_{t=1}^{n}\{|\lambda_t^{IO}|\}$ 与 $\max_{t=1}^{2n}\{\lambda_t^{IO}\}$ 有同样的分布，因此就有

$$P(\eta_{IO}\leqslant c_{2n}+d_{2n}x)\to\Lambda(x)$$

同样，在 $H_0':\omega_t^{AO}=0$ 下，由定理 2 的（2）知 $\{\lambda_t^{AO}\}$ 的自相关函数为

$$\rho_{AO}(h)=\Delta^{-1}(a_1a_{h+1}+a_2a_{h+2}+\cdots+a_{p-h}a_p)\cdot I(1\leqslant h\leqslant p)$$

故也满足 Berman 条件，同样也有 $P(\eta_{AO}\leqslant c_{2n}+d_{2n}x)\to\Lambda(x)$。

注 1：对于成片异常点，定义 $\eta_{IOk}\triangleq\max_{t=1}^{n}\{|\lambda_t^{IOk}|\}$ 和 $\eta_{AOk}\triangleq\max_{t=1}^{n}\{|\lambda_t^{AOk}|\}$，前面的讨论知 $\{\lambda_t^{IOk}\}$ 和 $\{\lambda_t^{AOk}\}$ 也满足 Berman 条件，定理 3 中的 η_{IO} 和 η_{AO} 换为 η_{IOk} 和 η_{AOk} 也有相同的分布。

引理 3　设 $\{X_t\}$ 是零均值方差为 1，自相关系数为 $\rho(h)$ 的 Gauss 平稳时间序列，如果 $\{X_t\}$ 对 Berman 条件成立，令 $M_n=\max\{X_1^2,X_2^2,\cdots,x_n^2\}$，则当 $n\to\infty$ 时有

$$P(M_n\leqslant e_n+2x)\to\Lambda(x)$$

其中 $e_n=2\log n-\log(\log n)-\log\pi$

定理 4　在定理 2 的条件下，分别定义 $\max_{t=1}^{n}\{(\lambda_t^{IO})^2\}\triangleq\vartheta_{IO}$ 和 $\max_{t=1}^{n}\{(\lambda_t^{AO})^2\}\triangleq\vartheta_{AO}$，则当 $n\to\infty$ 时有

$$P(\vartheta_{IO}\leqslant e_n+2x)\to\Lambda(x)\tag{11}$$

$$P(\vartheta_{AO}\leqslant e_n+2x)\to\Lambda(x)\tag{12}$$

证明　由定理 3 和引理 3 易知结论成立。

注 2：对于成片异常点，定义 $\max_{t=1}^{n}(\lambda_t^{IOk})^2\triangleq\vartheta_{IOk}$ 和 $\max_{t=1}^{n}(\lambda_t^{AOk})^2\triangleq\vartheta_{AOk}$，则定理 4 对 ϑ_{IOk} 和 ϑ_{AOk} 也有相同的渐近分布。

2.3　异常点的诊断步骤

前面已经得到了绝对值统计量和平方统计量的渐近分布，接下来用分布的渐近临界值来获得检验 $H_0:\omega_t^{IO}=0$ 或 $H_0:\omega_t^{AO}=0$ 的拒绝域和渐近 p 值，仅以绝对值统计量对一个 IO 异常点的诊断说明，平方统计量及 AO 的情形同样处理，诊断步骤如下：

1）首先获得参数估计和模型残差，并令诊断统计量为 $T=\eta_{IO}$；

2）在每一个观测值计算似然比统计量 λ_q^{IO} 的大小，接着计算 $T_0=\max_{t=1}^{n}\{|\lambda_t^{IO}|\}$；

3）应用诊断统计量 T 在 $H_0:\omega_t^{IO}=0$ 下的渐近分布获得一定水平 α 下的临界值，即由式（9）得渐近临界值 $CV=c_{2n}+d_{2n}\cdot\Lambda^{-1}(1-\alpha)=c_{2n}+d_{2n}\cdot\{-\log[-\log(1-\alpha)]\}$，模拟中 α 取 0.05；

4）对于式（2）计算的值 T_0，若 $T_0>CV$，则表明第 q 个观测值为一个 IO；

5）p 值的计算，$p-value=P_{H_0}(T>T_0)=1-\Lambda(T_0)$，$\Lambda(x)$ 为标准的 Gumbel 分布函数，如果 $p-value<\alpha$，则拒绝原假设 $H_0:\omega_t^{IO}=0$，确认存在异常点 IO。

上面步骤仅给出应用绝对值统计量 η_{IO} 来诊断 IO 的步骤，对于 AO 的诊断只需替换为诊断统计量 η_{AO} 即可。当然哪种形式的统计量诊断效果好，可以通过比较它们的渐近 p 值得知，这里 p 值与临界值的计算只需要样本量 n，而不再进一步的仿真，这也是极值渐近分布诊断异常点的最大优点。对于多个异常点的诊断可以结合 CHEN 等[4]的迭代挖掘程序逐个挖掘。

3 模拟例子

例 1：对下面 AO 异常点模型 $\begin{cases} y_t = x_t + \omega_{n/2} \cdot \delta(t-n/2) \\ x_t = 0.8x_{t-1} + \varepsilon_t, \varepsilon_t \sim N(0,1), \omega_{n/2} = 4.8 \end{cases}$

用 Matlab 软件分别随机产生样本量 $n=50$、120、400、1 000 四种情况来比较对不同样本量的诊断效果，各随机模拟 50 次，表中 $\hat{a}$，$\hat{\sigma}$，$\hat{\omega}$ 分别是自回归系数，残差标准差和异常点大小的平均估计。通过表 1 的比较看到绝对值统计量和平方统计量对诊断 AO 异常点也很有效，且计算的平均渐近 P 值均小于 0.05，但是 CHEN 等[4]（统计量是 η_{AO}，临界值取 3.5）的方法无法计算 P 值。从平均的 P 值来看，平方统计量的平均 P 值比绝对值统计量稍小，即平方统计量比绝对值统计量诊断效果要好。而且发现随着样本量的增大，模型的参数估计受异常点的影响逐渐变小，两种检验统计量的渐近 P 值随样本量的增大有先减小后变大的趋势，对于中小样本量我们的方法是有效的。

表 1 例 1 中不同统计量对应的诊断次数、平均 P 值及平均参数估计表

Tab. 1 Diagnostic time, average p-value of different test statistics and parameters estimation of Example 1

		$\eta_{AO}>3.5$	绝对值统计量	平方统计量	$\hat{a}$	$\hat{\sigma}$	$\hat{\omega}$
n=50	诊断次数	50	50	50	0.636 8	1.289 2	4.634 7
	平均 P 值	——	0.004 4	0.003 3			
n=120	诊断次数	50	50	50	0.711 7	1.175 0	4.833 6
	平均 P 值	——	0.001 6	0.001 2			
n=400	诊断次数	50	50	50	0.784 9	1.053 0	4.897 6
	平均 P 值	——	0.000 8	0.000 6			
n=1 000	诊断次数	50	49	49	0.794 5	1.022 0	4.739 8
	平均 P 值	——	0.007 7	0.007 4			

例 2：对下面 IO 异常点模型 $\begin{cases} y_t = 0.8y_t + \eta_t \\ \eta_t = \varepsilon_t + \omega_{n/2} \cdot \delta(t-n/2), \varepsilon_t \sim N(0,1), \omega_{n/2} = 5 \end{cases}$

用 Matlab 软件分别随机产生样本量 $n=50$、120、400、1 000 四种情况来比较对不同样本量的诊断效果，各随机模拟 50 次。通过表 2 的比较可以看到此方法对诊断 IO 异常点也很有效，不同样本量下的平均渐近 P 值均小于 0.05，诊断结果是有效的。而且从平均的渐近 P 值来看，平方统计量比绝对值统计量对诊断异常点效果要好。而且发现随着样本量的变大，模型的参数估计受异常点的影响也越小，但是两种检验统计量的渐近 P 值有逐渐变大的趋势，这一点与 AO 情形不太相同，这说明渐近临界值对于中小样本量有更好的效果。

表 2 例 2 的各诊断统计量对应的诊断次数、平均 P 值及平均参数估计表

Tab. 2 Diagnostic time, average p-value of different test statistics and parameters estimation of Example 2

		$\eta_{IO}>3.5$	绝对值统计量	平方统计量	$\hat{a}$	$\hat{\sigma}$	$\hat{\omega}$
n=50	诊断次数	48	48	48	0.777 2	1.207 2	5.028 2
	平均 P 值	——	0.002 7	0.002 3			
n=120	诊断次数	47	45	45	0.770 8	1.091 2	4.795 7
	平均 P 值	——	0.018 4	0.018 1			
n=400	诊断次数	47	45	45	0.793 5	1.037 0	5.076 7
	平均 P 值	——	0.022 3	0.022 2			
n=1 000	诊断次数	50	42	41	0.796 7	1.029 6	5.153 8
	平均 P 值	——	0.033 7	0.033 5			

4 结论

本文主要通过极值理论给出 AR（p）模型的几个异常点诊断检验统计量的渐近分布，并依据渐近分布给出了诊断异常点的临界值取法与诊断步骤，这种取法相对于大多数文献中利用仿真给出的临界值，其最大的优点就是保证可控制在一定的显著性水平之下，并可以计算检验的 p 值，而且从模拟例子来看，平方统计量相对于绝对值统计量对诊断异常点有稍好的效果。

[参考文献] (References)

[1] BOX E P, JENKINS G M, REINSEL G C. Time series analysis: forecasting and control[M]. New Jersey: Prentice Hall, Englewood Cliffs, 1994.

[2] FOX A J. Outliers in time series[J]. Journal of the Royal Statistics Society Series, 1972(48): 39～47.

[3] CHANG I, TIAO G C, CHEN C. Estimation of time series parameters in the presence of outliers[J]. Technometrics, 1988(30): 193～204.

[4] CHEN C, LIU L M. Jiont estimation of model parameters and outlier essects in time series[J]. Journal of the American Statistical Association, 1993(88): 284～297.

[5] ALEJANDRO. Extreme value theory-based *P* values in time series outlier detection[D]. USA: University of Wisconsin-Madison, 2005.

[6] CHAREKA P, MATARISE F, TURNER R. A test for additive outliers applicable to long-memory time series[J]. Journal of Economic Dynamics & Control, 2006(30): 595～621.