ZHANG Qianqian, GUI Qingming, WANG Yanting. Bayesian Methods for Outliers Detection in Autoregressive Model Based on Different Types of Classification Variables[J]. Acta Geodaetica et Cartographica Sinica, 2012, 41(3): 378-384. (张倩倩, 归庆明, 王延停. 不同类型识别变量的自回归模型异常值探测的 Bayes 方法[J]. 测绘学报, 2012, 41(3): 378-384.)

不同类型识别变量的自回归模型异常值探测的 Bayes 方法

张倩倩[1], 归庆明[1,2], 王延停[1]

1. 信息工程大学 理学院, 河南 郑州 450001; 2. 信息工程大学 测绘学院, 河南 郑州 450052

Bayesian Methods for Outliers Detection in Autoregressive Model Based on Different Types of Classification Variables

ZHANG Qianqian[1], GUI Qingming[1,2], WANG Yanting[1]

1. Institute of Science, Information Engineering University, Zhengzhou 450001, China; 2. Institute of Surveying and Mapping, Information Engineering University, Zhengzhou 450052, China

Abstract: A Bayesian procedure for outlier detection in time series is discussed. The main idea of this method is introducing different types of classification variables into autoregressive model. Then outliers can be detected by comparing the posterior probabilities of these classification variables with a given threshold. Besides, a procedure for computing the posterior probabilities of classification variables and obtaining the estimates of outliers is designed based on Gibbs sampling. A large number of simulation experiments and an experiment of real clock error data are carried out. It is shown that the new procedure is applicable to detect additive and innovational outliers occurring at the same time or not in time series.

Key words: AR model; additive outlier; innovation outlier; classification variable; Bayesian method; Gibbs sampling; satellite clock error

摘 要: 讨论基于自回归模型(AR 模型)的时间序列数据中异常值探测的 Bayes 方法。该方法针对自回归模型引入不同类型的识别变量, 通过比较这些识别变量的后验概率值与事先给定的阈值来进行异常值定位; 基于 Gibbs 抽样算法, 提出识别变量后验概率值的计算方法和异常值的估算方法; 进行了大量的模拟试验并把该方法应用于卫星钟差实测数据的异常值探测, 结果表明, 该方法对于解决时间序列数据中在同一时刻或不同时刻出现加性异常值或革新异常值的探测问题是可行的和有效的。

关键词: 自回归模型; 加性异常值; 革新异常值; 识别变量; Bayes 方法; Gibbs 抽样; 卫星钟差

中图分类号: P207 **文献标识码**: A **文章编号**: 1001-1595(2012)03-0378-07

基金项目: 国家自然科学基金(40974009; 41174005); 中国卫星导航学术年会青年优秀论文获奖者资助课题; 郑州市科技计划攻关项目(0910SGYG21198)

1 引 言

时间序列分析是测绘导航数据处理的基本技术手段[1-2], 而异常值探测是时间序列分析中的一个重要环节[3-10]。通常, 时间序列中的异常值分为加性异常值和革新异常值两种类型[3-4]。加性异常值, 又称 AO 类异常值, 是指只影响异常干扰发生的那一个时刻的观测值, 而不影响该时刻以后的观测值; 革新异常值, 又称 IO 类异常值, 是指造成这种异常值的干扰不仅作用于该时刻的观测值而且影响该时刻以后的所有观测值。目前, 关于自回归模型(自回归模型是时间序列分析中最具代表性的一类线性模型[1-2])中这两类异常值的探测主要有非 Bayesian 方法和 Bayesian 方法。前者主要有迭代似然比法[5], 基于稳健估计和影响分析思想的探测法[6]和序贯检验法等[7]。后者主要有文献[3]提出的探测方法, 该方法是先探测出 AO 类异常值, 对数据进行修正后, 再继续探测 IO 类异常值, 但是该方法未给出明确的异常值探测规则, 文献[8]用 Gibbs 抽样算法讨论了该方法涉及的有关后验概率值的计算问题; 文献[9—10]借鉴文献[11—12]的思想, 在一定限制条件下, 将 AR 模型的异常值探测问题转化为线性模型的异常值探测问题, 从而提出了异常值定位和定值的 Bayes 方法。然而, 这些 Bayesian 方法并没有对定位的异常值进行区分, 更未讨论如何解决时间序

列中在同一时刻同时出现 AO 类和 IO 类异常值这一问题。为此本文通过进一步发展文献[13—14]的思想，提出一种基于不同类型识别变量，同时探测 AR 模型中 AO 类异常值和 IO 类异常值的 Bayes 方法。为了验证该方法的正确性，本文进行了大量的模拟试验，并把该方法应用于钟差实测数据的优化处理中，通过比较异常值消除前后各数据点的均方误差，论证了该方法对于解决时间序列数据中在同一时刻或不同时刻出现加性异常值或革新异常值的探测问题是可行和有效的。

2　基于不同类型识别变量的自回归模型异常值定位的 Bayes 方法

设有一组时间序列数据$\{x_1,\cdots,x_n\}$符合如下的$AR(p)$模型

$$\left.\begin{aligned} x_t&=\phi_1x_{t-1}+\phi_2x_{t-2}+\cdots+\phi_px_{t-p}+a_t\\ a_t\ &i.i.d.\ N(0,\sigma^2),t=1,\cdots,n\end{aligned}\right\}\quad(1)$$

式中，$i.i.d.$为相互独立同分布；$\boldsymbol{\Phi}=[\phi_1\quad\phi_2\quad\cdots\quad\phi_p]^{\mathrm{T}}$和$\sigma^2$为未知参数。

对每个观测值x_t分别引入 AO 类异常值识别变量

$$\delta_t^{\mathrm{AO}}=\begin{cases}1 & x_t\text{ 受到 AO 类异常扰动}\\ 0 & x_t\text{ 不受到 AO 类异常扰动}\end{cases}\quad(2)$$

和 IO 类异常值识别变量

$$\delta_t^{\mathrm{IO}}=\begin{cases}1 & x_t\text{ 受到 IO 类异常扰动}\\ 0 & x_t\text{ 不受到 IO 类异常扰动}\end{cases}\quad(3)$$

记$w_1^{\mathrm{AO}},\cdots,w_n^{\mathrm{AO}}$和$w_1^{\mathrm{IO}},\cdots,w_n^{\mathrm{IO}}$分别代表每个时刻观测值的 AO 类异常值大小和 IO 类异常值大小，并假设：

（1）每个观测值x_t受到 AO 类异常扰动或 IO 类异常扰动的先验概率都为α[3]，即$P(\delta_t^{\mathrm{AO}}=1)=\alpha$，$P(\delta_t^{\mathrm{IO}}=1)=\alpha$。

（2）根据共轭先验分布的选取准则[12-13,16]和实际应用的需要，取参数的先验分布为

$w_t^{\mathrm{AO}}\ i.i.d.\ N(\mu_1,\xi^2)$、$w_t^{\mathrm{IO}}\ i.i.d.\ N(\mu_2,\xi^2)$、$\delta_t^{\mathrm{AO}}\ i.i.d.\ b(1,\alpha)$、$\delta_t^{\mathrm{IO}}\ i.i.d.\ b(1,\alpha)$、

$\boldsymbol{\Phi}\sim N_p(\boldsymbol{\Phi}_0,\boldsymbol{V}^{-1})$、$\sigma^2\sim IG(\frac{\upsilon}{2},\frac{\upsilon\lambda}{2})$。

其中，μ_1、μ_2、ξ、α、$\boldsymbol{\Phi}_0$、$\boldsymbol{V}$、υ和λ为超参数。

根据以上假设，观测值x_t可表示为

$$x_t=z_t+w_t^{\mathrm{AO}}\delta_t^{\mathrm{AO}}+\phi^{-1}(B)w_t^{\mathrm{IO}}\delta_t^{\mathrm{IO}}\quad(4)$$

式中，z_t代表未受到异常扰动的干净数据；$\phi(B)=I-\phi_1B-\phi_2B^2-\cdots-\phi_pB^p$，$B$为后推算子，即$B^kx_t=x_{t-k}$。式(4)右边的第 3 项之所以含有因子$\phi^{-1}(B)$是考虑到 IO 类异常值的定义[3]。由此，可得基于 AR 模型的 AO 类和 IO 类异常值同时探测的模型

$$\left.\begin{aligned} z_t&=\phi_1z_{t-1}+\phi_2z_{t-2}+\cdots+\phi_pz_{t-p}+a_t\\ x_t&=z_t+w_t^{\mathrm{AO}}\delta_t^{\mathrm{AO}}+\phi^{-1}(B)w_t^{\mathrm{IO}}\delta_t^{\mathrm{IO}}\end{aligned}\right\}\quad(5)$$

为了计算简便，对式(5)进行等价转换，令$y_t=z_t+\phi^{-1}(B)w_t^{\mathrm{IO}}\delta_t^{\mathrm{IO}}$，则可得与式(5)等价的模型

$$\left.\begin{aligned} y_t&=\phi_1y_{t-1}+\phi_2y_{t-2}+\cdots+\phi_py_{t-p}+a_t+w_t^{\mathrm{IO}}\delta_t^{\mathrm{IO}}\\ x_t&=y_t+w_t^{\mathrm{AO}}\delta_t^{\mathrm{AO}}\end{aligned}\right\}\quad(6)$$

式中，y_t代表仅受到 IO 类异常扰动的数据。不妨设前p个观测值$x_1,\cdots,x_p$为正常值[8]，为判定观测值中是否含有 AO 类或者 IO 类异常值以及确定它们的判别阈值，构造如下两个 Bayes 假设检验问题。

原假设$H_{01}:x_t$是正常观测值，即$\delta_t^{\mathrm{AO}}=0$；

备选假设$H_{11}:x_t$为 AO 类异常值，即$\delta_t^{\mathrm{AO}}=1$。

原假设$H_{02}:x_t$是正常观测值，即$\delta_t^{\mathrm{IO}}=0$；

备选假设$H_{12}:x_t$为 AO 类异常值，即$\delta_t^{\mathrm{IO}}=1$。

根据 Bayes 假设检验的思想和原理[16]，当备选假设H_{11}对应的后验概率$P(\delta_t^{\mathrm{AO}}=1|\boldsymbol{X})$大于原假设$H_{01}$对应的后验概率$P(\delta_t^{\mathrm{AO}}=0|\boldsymbol{X})$，即$P(\delta_t^{\mathrm{AO}}=1|\boldsymbol{X})>0.5$时，认为备选假设成立，从而认为观测值$x_t$为 AO 类异常值，反之，认为观测值$x_t$是正常观测值。同理，若$P(\delta_t^{\mathrm{IO}}=1|\boldsymbol{X})$大于$P(\delta_t^{\mathrm{IO}}=0|\boldsymbol{X})$，即$P(\delta_t^{\mathrm{IO}}=1|\boldsymbol{X})>0.5$时，认为备选假设成立，从而认为观测值$x_t$为 IO 类异常值，反之，认为观测值$x_t$是正常观测值。若$P(\delta_t^{\mathrm{AO}}=1|\boldsymbol{X})>0.5$且$P(\delta_t^{\mathrm{IO}}=1|\boldsymbol{X})>0.5$，则判断$x_t$同时受到 AO 和 IO 两种异常扰动。其中，$\boldsymbol{X}=[x_{p+1}\quad x_{p+2}\quad\cdots\quad x_n]^{\mathrm{T}}$。这样，问题归结为计算每个观测值$x_t$含有 AO 类异常值的后验概率$P(\delta_t^{\mathrm{AO}}=1|\boldsymbol{X})$和含有 IO 类异常值的后验概率$P(\delta_t^{\mathrm{IO}}=1|\boldsymbol{X})$。

3　基于 Gibbs 抽样的后验概率值的计算和异常值的估计

3.1　参数的完全条件分布

由于后验概率$P(\delta_t^{\mathrm{AO}}=1|\boldsymbol{X})$和$P(\delta_t^{\mathrm{IO}}=1|\boldsymbol{X})$涉及的分布比较复杂，下面引入 Gibbs 抽样算法[15]来解决这些后验概率值的计算问题。为此，根据 Bayes 定理[16]可得下列未知参数的完全条件分布。

（1）$\boldsymbol{X}$、σ^2、$\boldsymbol{\delta}^{\mathrm{AO}}$、$\boldsymbol{\delta}^{\mathrm{IO}}$、$\boldsymbol{W}^{\mathrm{AO}}$、$\boldsymbol{W}^{\mathrm{IO}}$给定时，$\boldsymbol{\Phi}$的完

全条件分布为

$$\boldsymbol{\Phi}|\boldsymbol{X},\sigma^2,\boldsymbol{\delta}^{\mathrm{AO}},\boldsymbol{\delta}^{\mathrm{IO}},\boldsymbol{W}^{\mathrm{AO}},\boldsymbol{W}^{\mathrm{IO}}\sim N_p(\boldsymbol{\Phi}_0^*,\hat{\boldsymbol{V}}^{-1}) \quad (7)$$

式中

$$\boldsymbol{\Phi}_0^*=\hat{\boldsymbol{V}}^{-1}\left(\frac{1}{\sigma^2}\sum_{t=p+1}^{n}\boldsymbol{Y}_{t-1}(x_t-w_t^{\mathrm{AO}}\boldsymbol{\delta}_t^{\mathrm{AO}}-w_t^{\mathrm{IO}}\boldsymbol{\delta}_t^{\mathrm{IO}})+\boldsymbol{V}\boldsymbol{\Phi}_0\right)$$

$$\hat{\boldsymbol{V}}=\boldsymbol{V}+\frac{1}{\sigma^2}\sum_{t=p+1}^{n}\boldsymbol{Y}_{t-1}\boldsymbol{Y}_{t-1}^{\mathrm{T}},\boldsymbol{Y}_{t-1}=[y_{t-1}\ \cdots\ y_{t-p}]^{\mathrm{T}}$$

$$\boldsymbol{W}^{\mathrm{AO}}=[w_1^{\mathrm{AO}}\ \cdots\ w_n^{\mathrm{AO}}]^{\mathrm{T}},\ \boldsymbol{W}^{\mathrm{IO}}=[w_1^{\mathrm{IO}}\ \cdots\ w_n^{\mathrm{IO}}]^{\mathrm{T}}$$

$$\boldsymbol{\delta}^{\mathrm{AO}}=[\delta_1^{\mathrm{AO}}\ \cdots\ \delta_n^{\mathrm{AO}}]^{\mathrm{T}},\quad \boldsymbol{\delta}^{\mathrm{IO}}=[\delta_1^{\mathrm{IO}}\ \cdots\ \delta_n^{\mathrm{IO}}]^{\mathrm{T}}$$

(2) $\boldsymbol{X}$、$\boldsymbol{\Phi}$、$\boldsymbol{\delta}^{\mathrm{AO}}$、$\boldsymbol{\delta}^{\mathrm{IO}}$、$\boldsymbol{W}^{\mathrm{AO}}$、$\boldsymbol{W}^{\mathrm{IO}}$给定时，$\sigma^2$ 的完全条件分布为

$$\sigma^2|\boldsymbol{X},\boldsymbol{\Phi},\boldsymbol{\delta}^{\mathrm{AO}},\boldsymbol{\delta}^{\mathrm{IO}},\boldsymbol{W}^{\mathrm{AO}},\boldsymbol{W}^{\mathrm{IO}}\sim IG\left(\frac{\upsilon_1}{2},\frac{\upsilon_1\lambda_1}{2}\right) \quad (8)$$

式中

$$\upsilon_1=n-p+\upsilon$$

$$\lambda_1=\frac{1}{n-p+\upsilon}\times\Big[\sum_{t=p+1}^{n}(x_t-\sum_{i=1}^{p}\phi_i y_{t-i}-w_t^{\mathrm{AO}}\delta_t^{\mathrm{AO}}-w_t^{\mathrm{IO}}\delta_t^{\mathrm{IO}})^2+\upsilon\lambda\Big]$$

(3) $\boldsymbol{X}$、$\boldsymbol{\Phi}$、σ^2、$\boldsymbol{\delta}_{(-j)}^{\mathrm{AO}}$、$\boldsymbol{\delta}^{\mathrm{IO}}$、$\boldsymbol{W}^{\mathrm{AO}}$、$\boldsymbol{W}^{\mathrm{IO}}$给定时 δ_j^{AO} 的完全条件分布为

$$\delta_j^{\mathrm{AO}}|\boldsymbol{X},\boldsymbol{\Phi},\sigma^2,\boldsymbol{\delta}_{(-j)}^{\mathrm{AO}},\boldsymbol{\delta}^{\mathrm{IO}},\boldsymbol{W}^{\mathrm{AO}},\boldsymbol{W}^{\mathrm{IO}}\sim b(1,p_j^{\mathrm{AO}}) \quad (9)$$

式中

$$\delta_{(-j)}^{\mathrm{AO}}=(\delta_1^{\mathrm{AO}},\cdots,\delta_{j-1}^{\mathrm{AO}},\delta_{j+1}^{\mathrm{AO}},\cdots,\delta_n^{\mathrm{AO}})^{\mathrm{T}}$$

$$p_j^{\mathrm{AO}}=P(\delta_j^{\mathrm{AO}}=1|\boldsymbol{X},\boldsymbol{\Phi},\sigma^2,\boldsymbol{\delta}_{(-j)}^{\mathrm{AO}},\boldsymbol{\delta}^{\mathrm{IO}},\boldsymbol{W}^{\mathrm{AO}},\boldsymbol{W}^{\mathrm{IO}})=\frac{q_{j1}^{\mathrm{AO}}}{q_{j1}^{\mathrm{AO}}+q_{j2}^{\mathrm{AO}}}$$

$$q_{j1}^{\mathrm{AO}}=\alpha\exp\{-\frac{1}{2\sigma^2}\sum_{t=j}^{T}(x_t^*-\sum_{i=1}^{p}\phi_i x_{t-i}^*-w_t^{\mathrm{IO}}\delta_t^{\mathrm{IO}}+C_{t-j}w_j^{\mathrm{AO}})^2\}$$

$$q_{j2}^{\mathrm{AO}}=(1-\alpha)\exp\{-\frac{1}{2\sigma^2}\sum_{t=j}^{T}(x_t^*-\sum_{i=1}^{p}\phi_i x_{t-i}^*-w_t^{\mathrm{IO}}\delta_t^{\mathrm{IO}})^2\}$$

$$x_t^*=\begin{cases}x_t, & t=j\\ y_t, & t\neq j\end{cases},\ C_{t-j}=\begin{cases}-1, & t=j\\ \phi_i, & i=1,\cdots,p\\ 0, & i>p\end{cases},$$

$$T=\min(n,p+j)$$

(4) $\boldsymbol{X}$、$\boldsymbol{\Phi}$、σ^2、$\boldsymbol{\delta}^{\mathrm{AO}}$、$\boldsymbol{\delta}_{(-j)}^{\mathrm{IO}}$、$\boldsymbol{W}^{\mathrm{AO}}$、$\boldsymbol{W}^{\mathrm{IO}}$给定时，$\delta_j^{\mathrm{IO}}$ 的完全条件分布为

$$\delta_j^{\mathrm{IO}}|\boldsymbol{X},\boldsymbol{\Phi},\sigma^2,\boldsymbol{\delta}^{\mathrm{AO}},\boldsymbol{\delta}_{(-j)}^{\mathrm{IO}},\boldsymbol{W}^{\mathrm{AO}},\boldsymbol{W}^{\mathrm{IO}}\sim b(1,p_j^{\mathrm{IO}}) \quad (10)$$

式中

$$\boldsymbol{\delta}_{(-j)}^{\mathrm{IO}}=[\delta_1^{\mathrm{IO}}\ \cdots\ \delta_{j-1}^{\mathrm{IO}}\ \delta_{j+1}^{\mathrm{IO}}\ \cdots\ \delta_n^{\mathrm{IO}}]^{\mathrm{T}}$$

$$p_j^{\mathrm{IO}}=P(\delta_j^{\mathrm{IO}}=1|\boldsymbol{X},\boldsymbol{\Phi},\sigma^2,\boldsymbol{\delta}^{\mathrm{AO}},\boldsymbol{\delta}_{(-j)}^{\mathrm{IO}},\boldsymbol{W}^{\mathrm{AO}},\boldsymbol{W}^{\mathrm{IO}})=\frac{q_{j1}^{\mathrm{IO}}}{q_{j1}^{\mathrm{IO}}+q_{j2}^{\mathrm{IO}}}$$

$$q_{j1}^{\mathrm{IO}}=\alpha\exp\{-\frac{1}{2\sigma^2}(y_j-\sum_{i=1}^{p}\phi_i y_{j-i}-w_j^{\mathrm{IO}})^2\}$$

$$q_{j2}^{\mathrm{IO}}=(1-\alpha)\exp\{-\frac{1}{2\sigma^2}(y_j-\sum_{i=1}^{p}\phi_i y_{j-i})^2\}$$

(5) $\boldsymbol{X}$、$\boldsymbol{\Phi}$、σ^2、$\boldsymbol{\delta}^{\mathrm{AO}}$、$\boldsymbol{\delta}^{\mathrm{IO}}$、$\boldsymbol{W}_{(-j)}^{\mathrm{AO}}$、$\boldsymbol{W}^{\mathrm{IO}}$给定时，$w_j^{\mathrm{AO}}$ 的完全条件分布为

$$w_j^{\mathrm{AO}}|\boldsymbol{X},\boldsymbol{\Phi},\sigma^2,\boldsymbol{\delta}^{\mathrm{AO}},\boldsymbol{\delta}^{\mathrm{IO}},\boldsymbol{W}_{(-j)}^{\mathrm{AO}},\boldsymbol{W}^{\mathrm{IO}}\sim N(\hat{w}_j^{\mathrm{AO}},(\xi_j^2)^{\mathrm{AO}}) \quad (11)$$

式中

$$\hat{w}_j^{\mathrm{AO}}=(\xi_j^2)^{\mathrm{AO}}\{\delta_j^{\mathrm{AO}}[(x_j-\sum_{i=1}^{p}\phi_i y_{j-i}-w_j^{\mathrm{IO}}\delta_j^{\mathrm{IO}})+\sum_{t=j+1}^{T}\phi_{t-j}(\sum_{i=1}^{p}\phi_i x_{t-i}^*+w_t^{\mathrm{IO}}\delta_t^{\mathrm{IO}}-x_t^*)]+\mu_1\}$$

$$(\xi_j^2)^{\mathrm{AO}}=[\frac{(\delta_j^{\mathrm{AO}})^2}{\sigma^2}(1+\sum_{i=1}^{p}\phi_i^2)+\frac{1}{\xi^2}]^{-1}$$

(6) $\boldsymbol{X}$、$\boldsymbol{\Phi}$、σ^2、$\boldsymbol{\delta}^{\mathrm{AO}}$、$\boldsymbol{\delta}^{\mathrm{IO}}$、$\boldsymbol{W}^{\mathrm{AO}}$、$\boldsymbol{W}_{(-j)}^{\mathrm{IO}}$给定时，$w_j^{\mathrm{IO}}$ 的完全条件分布为

$$w_j^{\mathrm{IO}}|\boldsymbol{X},\boldsymbol{\Phi},\sigma^2,\boldsymbol{\delta}^{\mathrm{AO}},\boldsymbol{\delta}^{\mathrm{IO}},\boldsymbol{W}^{\mathrm{AO}},\boldsymbol{W}_{(-j)}^{\mathrm{IO}}\sim N(\hat{w}_j^{\mathrm{IO}},(\xi_j^2)^{\mathrm{IO}}) \quad (12)$$

式中

$$\hat{w}_j^{\mathrm{IO}}=(\xi_j^2)^{\mathrm{IO}}[\delta_j^{\mathrm{IO}}(y_j-\sum_{i=1}^{p}\phi_i y_{j-i})+\mu_2]$$

$$(\xi_j^2)^{\mathrm{IO}}=[\frac{(\delta_j^{\mathrm{IO}})^2}{\sigma^2}+\frac{1}{\xi^2}]^{-1}$$

3.2 识别变量后验概率值的计算

设 $\boldsymbol{\Phi}^{(r)}$，$(\sigma^2)^{(r)}$，$(\boldsymbol{\delta}^{\mathrm{AO}})^{(r)}$，$(\boldsymbol{\delta}^{\mathrm{IO}})^{(r)}$，$(\boldsymbol{W}^{\mathrm{AO}})^{(r)}$，$(\boldsymbol{W}^{\mathrm{IO}})^{(r)}$ $r=1,\cdots,R$ 为用 Gibbs 抽样算法从上述完全条件分布中抽取的样本，则计算 AO 类异常值和 IO 类异常值的识别变量后验概率值的公式分别为

$$P(\delta_j^{\mathrm{AO}}=1\mid X)\approx\frac{1}{R}\sum_{r=1}^{R}\frac{(q_{j1}^{\mathrm{AO}})^{(r)}}{(q_{j1}^{\mathrm{AO}})^{(r)}+(q_{j2}^{\mathrm{AO}})^{(r)}} \quad (13)$$

和

$$P(\delta_j^{\mathrm{IO}}=1\mid X)\approx\frac{1}{R}\sum_{r=1}^{R}\frac{(q_{j1}^{\mathrm{IO}})^{(r)}}{(q_{j1}^{\mathrm{IO}})^{(r)}+(q_{j2}^{\mathrm{IO}})^{(r)}} \quad (14)$$

式中

$$(q_{j1}^{\mathrm{AO}})^{(r)}=\alpha\exp\{-\frac{1}{2(\sigma^2)^{(r)}}\sum_{t=j}^{T}(x_t^*-\sum_{i=1}^{p}\phi_i^{(r)}x_{t-i}^*-(w_t^{\mathrm{IO}})^{(r)}(\delta_t^{\mathrm{IO}})^{(r)}+C_{t-j}(w_j^{\mathrm{AO}})^{(r)})^2\}$$

$$(q_{j2}^{\mathrm{AO}})^{(r)}=(1-\alpha)\exp\{-\frac{1}{2(\sigma^2)^{(r)}}\sum_{t=j}^{T}(x_t^*-\sum_{i=1}^{p}\phi_i^{(r)}x_{t-i}^*-(w_t^{\mathrm{IO}})^{(r)}(\delta_t^{\mathrm{IO}})^{(r)})^2\}$$

$$(q_{j1}^{\mathrm{IO}})^{(r)}=\alpha\exp\{-\frac{1}{2(\sigma^2)^{(r)}}(y_j-\sum_{i=1}^{p}\phi_i^{(r)}y_{j-i}-(w_j^{\mathrm{IO}})^{(r)})^2\}$$

$$(q_{j2}^{\mathrm{IO}})^{(r)}=(1-\alpha)\exp\{-\frac{1}{2(\sigma^2)^{(r)}}(y_j-\sum_{i=1}^{p}\phi_i^{(r)}y_{j-i})^2\}$$

3.3　AO 类和 IO 类异常值的估计

由第 2 节中的异常值探测模型知，$w_1^{\mathrm{AO}},\cdots,w_n^{\mathrm{AO}}$ 和 $w_1^{\mathrm{IO}},\cdots,w_n^{\mathrm{IO}}$ 分别为 AO 类和 IO 类异常值的大小。根据 Bayes 估计原理[16]，取它们的后验均值作为 AO 类和 IO 类异常值的估计值，即

$$\hat{w}_j^{\mathrm{AO}}=[\frac{(\delta_j^{\mathrm{AO}})^2}{\sigma^2}(1+\sum_{i=1}^{p}\phi_i^2)+\frac{1}{\xi^2}]^{-1}\cdot\{\delta_j^{\mathrm{AO}}[(x_j-\sum_{i=1}^{p}\phi_i y_{j-i}-w_j^{\mathrm{IO}}\delta_j^{\mathrm{IO}})+\sum_{t=j+1}^{T}\phi_{t-j}(\sum_{i=1}^{p}\phi_i x_{t-i}^{*}+w_t^{\mathrm{IO}}\delta_t^{\mathrm{IO}}-x_t^{*})]+\mu_1\}\tag{15}$$

$$\hat{w}_j^{\mathrm{IO}}=[\frac{(\delta_j^{\mathrm{IO}})^2}{\sigma^2}+\frac{1}{\xi^2}]^{-1}[\delta_j^{\mathrm{IO}}(y_j-\sum_{i=1}^{p}\phi_i y_{j-i})+\mu_2]\tag{16}$$

4　自回归模型异常值探测的 Bayes 方法的实施过程

第 1 步，确定先验分布中的超参数。例如，本文在算例 1 中给出这些超参数的一组具体取值如下

$$\boldsymbol{\Phi}_0=\begin{pmatrix}0\\ \vdots\\ 0\end{pmatrix}_{p\times 1},\quad \boldsymbol{V}=\begin{pmatrix}10 & & \\ & \ddots & \\ & & 10\end{pmatrix}_{p\times p},$$

$\upsilon=3,\lambda=0.5,\alpha=0.05,$

$\xi^2=2,\mu_1=0,\mu_2=0$

第 2 步，根据 Bayes 估计方法[16]以及超参数的取值，确定 Gibbs 抽样的初值。

第 3 步，假定第 $s\geqslant 1$ 次抽样开始时样本值向量为

$(\boldsymbol{\Phi}^{(s-1)},(\sigma^2)^{(s-1)},(\boldsymbol{\delta}^{\mathrm{AO}})^{(s-1)},(\boldsymbol{\delta}^{\mathrm{IO}})^{(s-1)},(\boldsymbol{W}^{\mathrm{AO}})^{(s-1)},(\boldsymbol{W}^{\mathrm{IO}})^{(s-1)})$

则第 s 次抽样按下列方式产生样本值向量

$(\boldsymbol{\Phi}^{(s)},(\sigma^2)^{(s)},(\boldsymbol{\delta}^{\mathrm{AO}})^{(s)},(\boldsymbol{\delta}^{\mathrm{IO}})^{(s)},(\boldsymbol{W}^{\mathrm{AO}})^{(s)},(\boldsymbol{W}^{\mathrm{IO}})^{(s)})$

$\boldsymbol{\Phi}^{(s)}$ 从以下条件分布中抽取

$p(\boldsymbol{\Phi}\mid\boldsymbol{X},(\sigma^2)^{(s-1)},(\boldsymbol{\delta}^{\mathrm{AO}})^{(s-1)},(\boldsymbol{\delta}^{\mathrm{IO}})^{(s-1)},(\boldsymbol{W}^{\mathrm{AO}})^{(s-1)},(\boldsymbol{W}^{\mathrm{IO}})^{(s-1)})$

$(\sigma^2)^{(s)}$ 从以下条件分布中抽取

$p(\sigma^2\mid\boldsymbol{X},\boldsymbol{\Phi}^{(s)},(\boldsymbol{\delta}^{\mathrm{AO}})^{(s-1)},(\boldsymbol{\delta}^{\mathrm{IO}})^{(s-1)},(\boldsymbol{W}^{\mathrm{AO}})^{(s-1)},(\boldsymbol{W}^{\mathrm{IO}})^{(s-1)})$

$(\delta_j^{\mathrm{AO}})^{(s)}$ 从以下分布中抽取

$p(\delta_j^{\mathrm{AO}}=1\mid\boldsymbol{X},\boldsymbol{\Phi}^{(s)},(\sigma^2)^{(s)},(\boldsymbol{\delta}_{(-j)}^{\mathrm{AO}})^{(s-1,j+1)},(\boldsymbol{\delta}^{\mathrm{IO}})^{(s-1,j)},(\boldsymbol{W}^{\mathrm{AO}})^{(s-1,j+1)},(\boldsymbol{W}^{\mathrm{IO}})^{(s-1,j)})$

$(\boldsymbol{\delta}^{\mathrm{IO}})^{(s-1,j)},(\boldsymbol{W}^{\mathrm{AO}})^{(s-1,j)},(\boldsymbol{W}^{\mathrm{IO}})^{(s-1,j)})$

$(w_j^{\mathrm{AO}})^{(s)}$ 从以下分布中抽取

$p(w_j^{\mathrm{AO}}\mid\boldsymbol{X},\boldsymbol{\Phi}^{(s)},(\sigma^2)^{(s)},(\boldsymbol{\delta}^{\mathrm{AO}})^{(s-1,j+1)},(\boldsymbol{\delta}^{\mathrm{IO}})^{(s-1,j)},(\boldsymbol{W}_{(-j)}^{\mathrm{AO}})^{(s-1,j+1)},(\boldsymbol{W}^{\mathrm{IO}})^{(s-1,j)})$

$(\delta_j^{\mathrm{IO}})^{(s)}$ 从以下分布中抽取

$p(\delta_j^{\mathrm{IO}}=1\mid\boldsymbol{X},\boldsymbol{\Phi}^{(s)},(\sigma^2)^{(s)},(\boldsymbol{\delta}^{\mathrm{AO}})^{(s-1,j+1)},(\boldsymbol{\delta}_{(-j)}^{\mathrm{IO}})^{(s-1,j+1)},(\boldsymbol{W}^{\mathrm{AO}})^{(s-1,j+1)},(\boldsymbol{W}^{\mathrm{IO}})^{(s-1,j)})$

$(w_j^{\mathrm{IO}})^{(s)}$ 从以下分布中抽取

$p(w_j^{\mathrm{IO}}\mid\boldsymbol{X},\boldsymbol{\Phi}^{(s)},(\sigma^2)^{(s)},(\boldsymbol{\delta}^{\mathrm{AO}})^{(s-1,j+1)},(\boldsymbol{\delta}^{\mathrm{IO}})^{(s-1,j+1)},(\boldsymbol{W}^{\mathrm{AO}})^{(s-1,j+1)},(\boldsymbol{W}_{(-j)}^{\mathrm{IO}})^{(s-1,j+1)})$

其中，向量的上角标 (i,k) 的含义为该向量的第 1 个分量至第 $k-1$ 个分量是第 $i+1$ 次抽样抽取的样本，第 k 个分量至最后一个分量为第 i 次抽样抽取的样本。例如

$$(\boldsymbol{\delta}^{\mathrm{AO}})^{(s-1,j+1)}=((\delta_1^{\mathrm{AO}})^{(s)},\cdots,(\delta_j^{\mathrm{AO}})^{(s)},(\delta_{j+1}^{\mathrm{AO}})^{(s-1)},\cdots,(\delta_n^{\mathrm{AO}})^{(s-1)})^{\mathrm{T}}$$

重复上述抽样过程 R 次，每次均直到 Markov 链达到稳定，就得到 R 个 Gibbs 样本

$(\boldsymbol{\Phi}^{(1)},(\sigma^2)^{(1)},(\boldsymbol{\delta}^{\mathrm{AO}})^{(1)},(\boldsymbol{\delta}^{\mathrm{IO}})^{(1)},(\boldsymbol{W}^{\mathrm{AO}})^{(1)},(\boldsymbol{W}^{\mathrm{IO}})^{(1)})$

…

$(\boldsymbol{\Phi}^{(R)},(\sigma^2)^{(R)},(\boldsymbol{\delta}^{\mathrm{AO}})^{(R)},(\boldsymbol{\delta}^{\mathrm{IO}})^{(R)},(\boldsymbol{W}^{\mathrm{AO}})^{(R)},(\boldsymbol{W}^{\mathrm{IO}})^{(R)})$

第 4 步，按公式(13)和(14)计算识别变量的后验概率值，并按照第 2 节中的判别规则进行异常值定位。

第 5 步，按 3.3 节中的方法估计异常值的大小。

5　算例与分析

5.1　算例 1

考虑 $AR(2)$ 模型

$$\left.\begin{aligned}&z_t=0.8z_{t-1}+0.1z_{t-2}+a_t\\&a_t\ i.i.d.\ N(0,1)\end{aligned}\right\}\tag{17}$$

取 $z_1=0.3,z_2=0.2$，经模拟产生 100 个数据。采用如下 3 种方案进行试验和计算。

方案 1：在第 20 个观测值上加大小为 -15 的 IO 类异常值。

方案 2：在第 50 个、第 80 个观测值上分别加大小为 10、-6 的 AO 类异常值。

方案 3：在第 30 个观测值上加大小为 12 的 AO 类异常值和大小为 5 的 IO 类异常值，在第 78 个观测值上加大小为 -9 的 IO 类异常值。

3 种方案中异常值识别变量的后验概率值分别如图 1～图 6 所示。经判断，方案 1 中的第 20

个观测值为 IO 类异常值，方案 2 中的第 50 个、第 80 个观测值都为 AO 类异常值，方案 3 中的第 30 个观测值同时 AO 类异常值和 IO 类异常值，第 78 个观测值为 IO 类异常值。又经估计，异常值的大小分别为 −14.454 7、9.654 8、−5.235 2、12.908 6、5.906 9、−8.398 7。

由此可见，无论是同一种类型的异常值还是不同类型的异常值，无论它们在同一时刻出现还是在不同时刻出现，利用本文的方法进行定位和定值均能取得很好的效果。

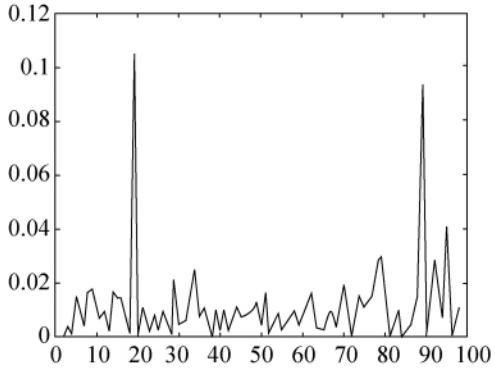


图 1 AO 类异常值识别变量的后验概率值(方案 1)

Fig. 1 Posterior probability of additive outlier classification variable of scheme 1

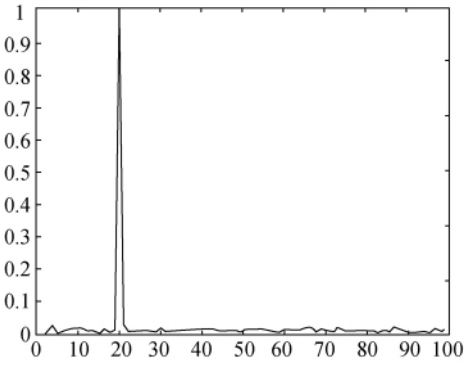


图 2 IO 类异常值识别变量的后验概率值(方案 1)

Fig. 2 Posterior probability of innovational outlier classification variable of scheme 1

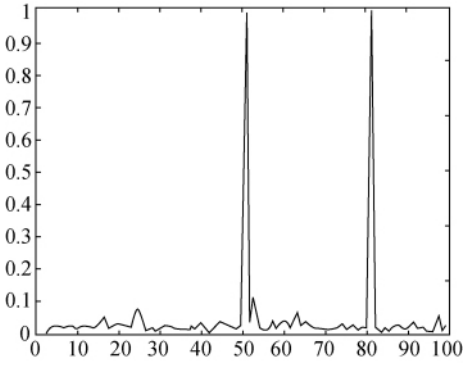


图 3 AO 类异常值识别变量后验概率值(方案 2)

Fig. 3 Posterior probability of additive outlier classification variable of scheme 2

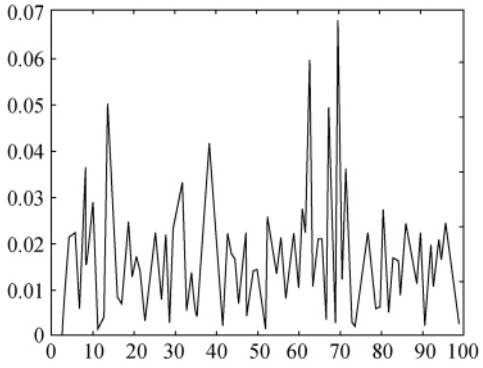


图 4 IO 类异常值识别变量后验概率值(方案 2)

Fig. 4 Posterior probability of innovational outlier classification variable of scheme 2

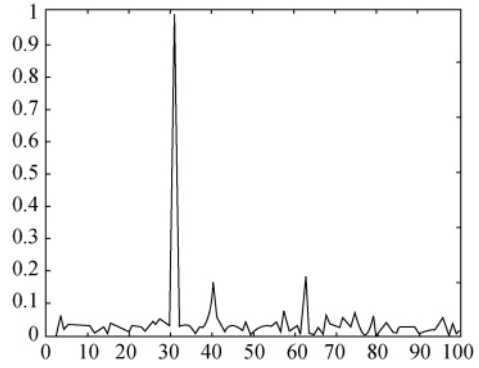


图 5 AO 类异常值识别变量的后验概率值(方案 3)

Fig. 5 Posterior probability of additive outlier classification variable of scheme 3

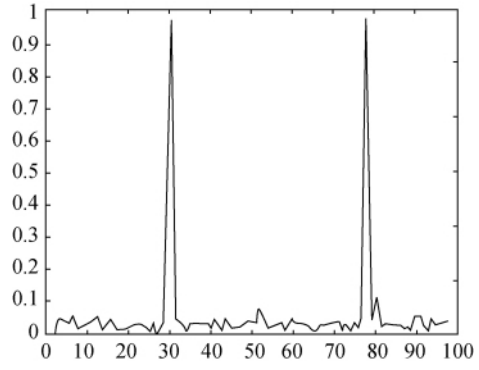


图 6 IO 类异常值识别变量的后验概率值(方案 3)

Fig. 6 Posterior probability of innovational outlier classification variable of scheme 3

5.2　算例 2

取 IGS 发布的 2007-07-31 全天的一组卫星钟差数据[11]，每 5 min 一个测值，故有 288 个观测值，如图 7 所示。本文对这组观测序列进行异常值探测，并且比较修正前后观测序列的均方误差。

采用参数检验法[17]对钟差序列进行平稳性检验，发现该序列不平稳，为此对该序列进行差分，结果如图 8 所示。从图 8 可以看出一次差分后的序列为平稳序列。用 AIC 准则对一次差分后的序列进行模型识别和定阶，发现该序列符合如下的$AR(1)$模型

$$\left.\begin{aligned} x_t &= 0.1178x_{t-1} + a_t \\ a_t\ &i.i.d.\ \ N(0, 1.00^2) \end{aligned}\right\} \tag{18}$$

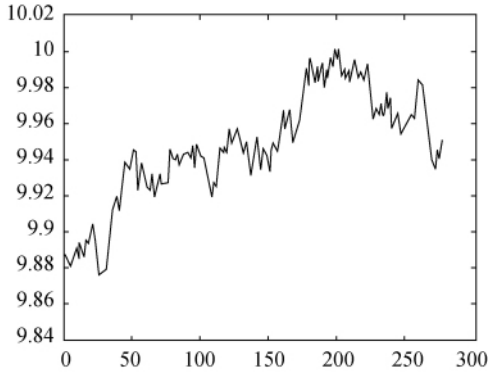


图 7　原始钟差序列

Fig. 7　Primal observations of satellite clock error

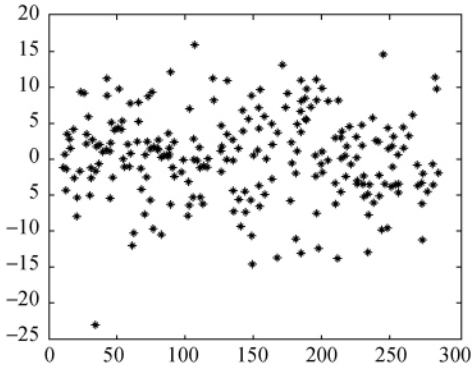


图 8　一次差分后的钟差序列

Fig. 8　Differential observations of satellite clock error

异常值识别变量后验概率值的计算结果如图 9和图 10 所示。经判断，第 25 个观测值为 AO 类异常值，进一步，其异常扰动的大小估计为 $wa(25) = -0.022\,561\,1$ ns。

根据 AO 异常值的定义对第 25 个数据点进行修正，并比较修正前后各个数据点的均方误差，如图 11 所示。可以看出，消除异常值后各个数据点的均方误差的大部分都小于消除异常值前的均方误差。事实上，经统计发现，288 个数据点的均方误差中，有 215 个点的均方误差有所减小，占总数据的 70% 左右。而且观测值序列总体的均方误差在消除异常值之前为 6.493 0e-005，而在消除异常值之后为 5.723 0e-005，亦有所减小。

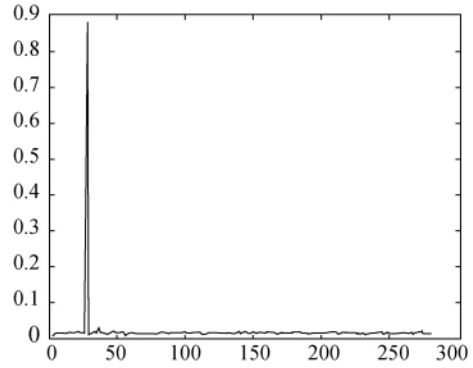


图 9　AO 类异常值识别变量后验概率值

Fig. 9　Posterior probability of additive outlier classification variable

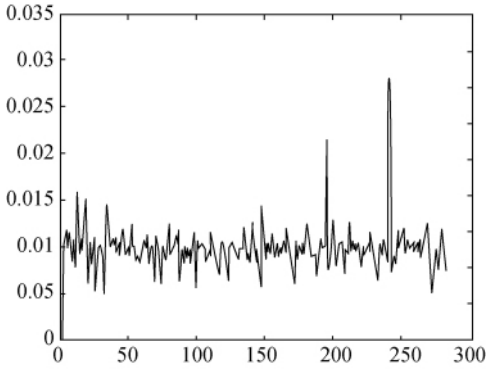
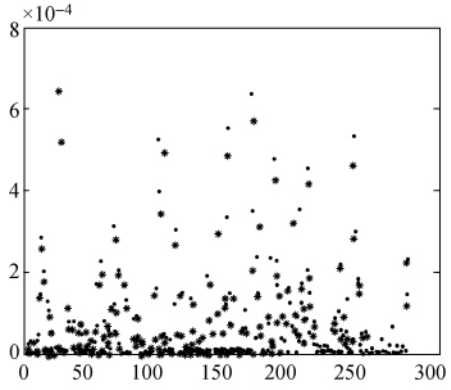


图 10　IO 类异常值识别变量后验概率值

Fig. 10　Posterior probability of innovational outlier classification variable



注：为消除前；* 为消除后。

图 11　修正前后各个序列值的均方误差的比较

Fig. 11　Comparison of mean square errors of observations modified and unmodified

6 结 论

(1) 针对每个时刻观测值中可能含有的不同类型的异常值分别引入不同类型的识别变量,通过比较这些识别变量后验概率值与事先给定的阈值来进行异常值探测,有效地克服了以往探测方法的模糊性及探测标准选择困难的问题,并且较好地解决了时间序列数据中在同一时刻或不同时刻出现不同类型异常值的探测问题。

(2) 在正态—Gamma 先验分布下,基于 Gibbs 抽样算法,提出了识别变量后验概率值的计算方法和异常值的估算方法。

(3) 基于识别变量的自回归模型异常值探测的 Bayes 方法,较为成功地应用于卫星钟差实测数据的优化处理中,验证了新方法的可行性和有效性。

(4) 根据 Bayes 假设检验的思想和原理,确定出了异常值判别阈值。

(5) 将 Gibbs 抽样算法等 MCMC 现代统计计算方法引入动态测量数据处理中,使得以往认为不可能实施计算的一些方法变得可行,特别对 Bayes 统计推断理论与方法在动态测量数据处理中的应用开辟了广阔的前景。

参考文献:

[1] HUANG Shengxiang, YIN Hui, JIANG Zheng. The Data Processing of Deformation Monitoring[M]. Wuhan: Wuhan University Press, 2003. (黄声享,尹晖,蒋征. 变形监测数据处理[M]. 武汉:武汉大学出版社,2003)

[2] WU Fumei, YANG Yuanxi. Gyroscope Random Drift Model Based on the Higher-order AR Model[J]. Acta Geodaetica et Cartographic Sinica, 2007, 36(4): 389-394. (吴富梅,杨元喜. 基于高阶 AR 模型的陀螺随机漂移模型[J]. 测绘学报,2007, 36(4): 389-394.)

[3] ABRAHAM B, BOX G E P. Bayesian Analysis of Some Outlier Problems in Time Series[J]. Biometrika, 1979, 66: 229-236.

[4] WEI Bocheng, LU Guobin, SHI Jianqing. Introduction to Statistics Diagnose [M]. Nanjing: Southeast University Press, 1991. (韦博成,鲁国斌,史建清. 统计诊断引论[M]. 南京:东南大学出版社,1991.)

[5] TSAY R S. Time Series Model Specification in the Presence of Outliers [J]. Journal of American Statistical Association, 1986, 393(81): 132-141.

[6] MARTIN D, YOHAI V J. Influence Functionals in Time Series[J]. The Annals of Statistics, 1986, 3(14): 781-818.

[7] LOUNI H. Outlier Detection in ARMA Models [J]. Journal of Time Series Analysis, 2005, 6(29): 1057-1065.

[8] MCCULLOCH R E, TSAY R S. Bayesian Analysis of Autoregressive Time Series via the Gibbs Sample[J]. Journal of Time Series Analysis, 1992, 2(15): 235-250.

[9] GUI Qingming, LI Tao, HENG Guanghui. Bayesian Method for Time Series Outliers Detection and Applications in Ionospheric VTEC Data Processing[J]. Geomatics and Information Science of Wuhan University, 2011, 36(7): 802-806. (归庆明,李涛,衡广辉. 时间序列异常值探测的 Bayes 方法及其在电离层 VTEC 数据处理中的应用[J]. 武汉大学学报:信息科学版, 2011, 36(7): 802-806.)

[10] LI Tao, HENG Guanghui, GUI Qingming. The Application of Bayesian Method in Autoregressive Model Outliers Detecting of Satellite Clock Error Prediction[J]. CNSS World of China, 2010, 35(4): 15-20. (李涛,衡广辉,归庆明. AR 序列异常值探测的 Bayes 方法在卫星钟差预报中的应用[J]. 全球定位系统, 2010, 35(4): 15-20.)

[11] GUI Qingming, GONG Yisong, LI Guozhong, et al. Bayesian Method for Detection of Gross Errors [J]. Acta Geodaetica et Cartographic Sinica, 2006, 35(4): 303-307. (归庆明,宫轶松,李国重,等. 粗差探测的 Bayes 方法[J]. 测绘学报, 2006, 35(4): 303-307.)

[12] GUI Q, GONG Y, LI G, et al. A Bayesian Approach to the Detection of Gross Errors Based on Posterior Probability[J]. Journal of Geodesy, 2007, 81: 651-659.

[13] LI Xinna, GUI Qingming, XU Apei. Bayesian Method for Detection of Gross Errors Based on Classification Variables [J]. Acta Geodaetica et Cartographica Sinica, 2008, 37(3): 355-360. (李新娜,归庆明,许阿裴. 基于识别变量的粗差探测 Bayes 方法[J]. 测绘学报, 2008, 37(3): 355-360.)

[14] GUI Q, LI X, GONG Y, et al. A Bayesian Unmasking Method for Locating Multiple Gross Errors Based on Posterior Probabilities of Classification Variables [J]. Journal of Geodesy, 2011, 85: 191-203.

[15] CHRISTIAN P R. Monte Carlo Statistical Methods [M]. Berlin: Springer , 2004.

[16] KOCH K R. Bayesian Inference with Geodetic Applications [M]. Berlin: Springer , 1990.

[17] WANG Zhenlong, HU Yonghong. The Application of Time Series Analysis[M] . Beijing: Science Press, 2007. (王振龙,胡永宏. 应用时间序列分析[M]. 北京:科学出版社,2007.)

(责任编辑:宋启凡)

收稿日期: 2011-05-23
修回日期: 2011-09-17

第一作者简介: 张倩倩(1987—),女,硕士生,研究方向为动态测量数据处理。
First author: ZHANG Qianqian(1987—), female, postgraduate, majors in data processing of dynamic surveying.
E-mail: zhangqianqian0216@163.com