

AR 模型参数的抗差估计研究

包为民¹,王 浩¹,赵 超¹,闻 璐²

(1. 水文水资源与水利工程科学国家重点实验室、河海大学水资源环境学院,江苏 南京 210098;
2. 河海大学水利水电工程学院,江苏 南京 210098)

摘要 实时校正一般以实测洪水流量为校正依据.研究实测洪水流量过程出现异常值时,采用抗差递推最小二乘法代替传统递推最小二乘法估计 AR 模型参数,能获得更稳健的参数结果.将闽江七里街流域的洪水资料人工生成异常值,对采用抗差递推最小二乘法和传统递推最小二乘法所得的校正结果进行比较,结果表明抗差递推最小二乘法具有更强的容差能力,是一种稳健的参数估计方法.

关键词 AR 模型 实时洪水校正 异常值 抗差递推最小二乘估计

中图分类号 P333 **文献标识码** A **文章编号** 1000-1980(2006)03-0258-04

误差自回归模型(AR 模型)是一种重要的实时校正模型.一般实测资料误差服从正态分布,采用传统递推最小二乘法可以得到满意的 AR 模型参数,从而获得更接近实测资料的预报结果.但当实测流量资料遭到污染出现异常值,即存在不满足正态分布的误差甚至错误时,则传统递推最小二乘估计的结果有偏,致使校正结果不可信,或者更差.因此,需要寻找一种有较强容错能力的、参数稳健的估计方法.

针对这一问题,抗差理论提出了污染分布概念,利用极值函数、导函数和权函数来描述误差分布特征,抵御异常值对参数估计的影响^[1-5].

本文把抗差理论与方法运用于 AR 模型参数在线估计中,分析在实测流量受异常值污染时,抗差估计的有效性.

1 AR 模型参数的抗差估计方法

1.1 递推最小二乘法

实时洪水校正需要利用所有历史流量观测资料,数据量大,需占用大量内存.本文采用的递推最小二乘法估计模型参数,克服了上述不足.

一个 N 阶的 AR 模型

$$e(t) = \psi_t^T \theta \quad (t = 1, 2, \dots) \tag{1}$$

式中 $e(t)$ —— t 时刻实测流量与计算流量的误差; $\psi_t = [e(t-1), e(t-2), \dots, e(t-N)]^T$; θ ——模型参数矩阵.

用矩阵表示 t 时刻的参数为

$$\hat{\theta}(t) = (\Phi_t^T \Phi_t)^{-1} \Phi_t^T Y_t \tag{2}$$

式中 $\Phi_t = [\psi_1, \psi_2, \dots, \psi_t]^T$; $Y_t = [e(1), e(2), \dots, e(t)]^T$.

设 $P(t) = (\Phi_t^T \Phi_t)^{-1}$, 式(2)可以写为

$$\hat{\theta}(t) = P(t) \Phi_t^T Y_t \tag{3}$$

根据矩阵分块乘法,可得到递推算式为

$$\hat{\theta}(t+1) = \hat{\theta}(t) + \frac{P(t) \psi_{t+1}}{I + \psi_{t+1}^T P(t) \psi_{t+1}} (e(t+1) - \psi_{t+1}^T \hat{\theta}(t)) \tag{4}$$

$$P(t+1) = P(t) - \frac{P(t) \psi_{t+1} \psi_{t+1}^T P(t)}{I + \psi_{t+1}^T P(t) \psi_{t+1}} \tag{5}$$

收稿日期 2005-10-18
基金项目 基金项目 浙江省重点科技项目(2002CZ3010)
作者简介 包为民(1956—),男,浙江东阳人,教授,主要从事水文水资源研究.

式中 I 为单位矩阵.

1.2 抗差递推最小二乘法

给 $e(t)$ 系列中的每个值加上初始权重,初始权重都为单位权. 式(2)变为

$$\hat{\theta}(t) = (\Phi_t^T W_t \Phi_t)^{-1} \Phi_t^T W_t Y_t \tag{6}$$

式中: $W_t = \text{diag}[w(1), w(2), \cdots, w(t)]$, $w(t)$ 为 t 时刻的权重.

设 $P(t) = (\Phi_t^T W_t \Phi_t)^{-1}$, 推导出加权递推最小二乘估计公式为

$$\hat{\theta}(t+1) = \hat{\theta}(t) + \frac{w(t+1)P(t)\psi_{t+1}}{I + w(t+1)\psi_{t+1}^T P(t)\psi_{t+1}}(e(t+1) - \psi_{t+1}^T \hat{\theta}(t)) \tag{7}$$

$$P(t+1) = P(t) - \frac{w(t+1)P(t)\psi_{t+1}\psi_{t+1}^T P(t)}{I + w(t+1)\psi_{t+1}^T P(t)\psi_{t+1}} \tag{8}$$

本文选用的抗差权函数为 IGG I 方案^[1]形式:

$$w(t) = \begin{cases} w(t) & |v(t)| \leq 1.5\sigma \\ w(t)1.5\sigma/|v(t)| & 1.5\sigma < |v(t)| \leq 2.5\sigma \\ 0 & |v(t)| > 2.5\sigma \end{cases} \tag{9}$$

式中: v ——残差; σ ——加权残差均方差.

参数抗差估计方法的计算步骤如下:

- a. 根据式(7)计算 $\hat{\theta}^{(1)}(t+1)$, 由式(8)计算得到 $P(t+1)$, 计算出 $e'^{(1)}(t+1)$, 计算 $|e(t+1) - e'^{(1)}(t+1)|$ 的值. 其中 $e'^{(1)}(t+1)$ 为式(1)根据 $\hat{\theta}^{(1)}(t+1)$ 计算出的误差.
- b. 按式(9)计算权重.
- c. 重复步骤 a 和 b, 迭代权重, 直到 $|\theta^{(k)}(t+1) - \theta^{(k+1)}(t+1)| < \epsilon$, 迭代停止, 此时 $\hat{\theta}(t+1) = \theta^{(k+1)}(t+1)$. 如果判断出 $t+1$ 时刻为异常值, 那么 $t+1$ 时刻的权重为零, 根据式(7)计算出的 $\hat{\theta}(t+1)$ 与 $\hat{\theta}(t)$ 相同, 抵御了异常值对参数的影响.

2 应用实例

本文选用了闽江七里街流域 1988~1998 年 43 场洪水资料, 其中 1988~1996 年的 36 场洪水作为模拟洪水, 1997~1998 年的 7 场洪水作为预报洪水.

首先根据资料给本流域 AR 模型定阶, 采用 F 检验定阶法^[6], 将其定为一阶自回归模型. 然后比较有常数项和无常数项的 AR 模型洪水校正结果, 发现无常数项的结果更好. 所以本文采用的是一阶无常数项的自回归模型. 对 1988~1996 年的 36 场洪水采用整批最小二乘估计计算 AR 模型参数, 并将此作为实时校正的初始参数 $\theta(0) = 0.9848$.

本文实测的洪水流量过程误差系列满足正态分布, 利用传统递推最小二乘法和抗差递推最小二乘法计算各时段的参数并进行实时校正. 表 1 为应用 2 种方法对预报洪水在无异常值污染情况下的实时校正效果, 预见期选为 1 h.

表 1 无异常值时 2 种方法对预报洪水的实时校正效果

Table 1 Statistics of rectified results of forecast floods without outliers by two methods									
洪 号	洪量误差			洪峰误差			确定性系数		
	e	e_{ls}	e_r	f	f_{ls}	f_r	dc	dc_{ls}	dc_r
45970605	-0.10	0.00	-0.01	-0.02	0.02	0.02	0.72	0.99	0.98
45970618	0.08	0.00	0.00	-0.10	0.04	0.04	0.89	1.00	1.00
45970701	-0.10	0.00	0.00	-0.01	0.03	0.03	0.88	1.00	1.00
45970808	0.07	0.00	0.00	0.07	0.03	0.03	0.25	0.98	0.98
45980215	-0.02	0.00	0.00	0.13	0.01	0.01	0.81	1.00	1.00
45980301	0.02	0.00	0.00	-0.01	0.02	0.02	0.89	1.00	1.00
45980509	0.08	0.00	-0.01	0.12	0.01	0.01	0.91	0.99	0.99
平均	0.00	0.00	0.00	0.03	0.03	0.03	0.78	1.00	0.99

注: 下标 ls 和 r 分别表示递推最小二乘法和抗差递推最小二乘法估计的结果; 没有下标的表示预报模型计算流量与实测流量之间的比较结果.

由表 1 可知,2 种方法的效果非常一致,表示当实测的洪水流量过程误差服从正态分布时,抗差估计与最小二乘法估计效果基本相同,并且把这个参数估计作为真值 θ .

采用 Monte Carlo 方法生成异常值系列,加入实测洪水流量资料中,这样实测洪水流量资料中的误差分布不再符合正态分布.

采用递推最小二乘法和抗差递推最小二乘法这 2 种方法进行模拟洪水流量实时校正,并对这 2 种方法的校正效果进行比较,表 2 为模拟洪水的流量校正效果,表 3 为预报洪水的流量校正效果.预见期均为 1 h.

表 2 有异常值时 2 种方法对模拟洪水的校正效果

Table 2 Statistics of rectified results of simulated floods with outliers by two methods

洪 号	洪量误差			洪峰误差			确定性系数			参数均方差		<i>r</i>
	<i>e</i>	<i>e_{ls}</i>	<i>e_r</i>	<i>f</i>	<i>f_{ls}</i>	<i>f_r</i>	<i>dc</i>	<i>dc_{ls}</i>	<i>dc_r</i>	<i>V_{ls}</i>	<i>V_r</i>	
45880228	-0.02	-0.11	0	0.02	-0.18	0.04	0.96	0.78	1.00	0.013 8	0.000 1	0.91
45880509	0.01	-0.15	0	-0.26	-1.13	0.03	0.17	-0.42	0.98	0.015 9	0	0.81
45880520	-0.02	-0.06	0	0.10	-0.13	0.01	0.95	0.94	1.00	0.016 2	0.000 5	0.85
45880620	-0.02	-0.08	0	0.12	-0.39	0	0.89	0.57	0.98	0.012 4	0.000 2	0.90
45890419	-0.11	-0.17	0	0.07	-0.45	0.02	0.85	0.49	0.99	0.011 0	0	0.88
45890511	0.03	-0.10	0	-0.38	-0.57	0.02	0.57	0.73	0.99	0.029 0	0.000 3	0.89
45890618	0	-0.13	0	0.01	-0.29	0.03	0.75	0.74	0.97	0.014 1	0.000 1	0.95
45890629	-0.07	-0.09	0	0.04	-0.18	0.02	0.92	0.66	0.99	0.012 5	0.000 1	0.90
45890723	-0.21	-0.12	0	0.11	-0.21	0.01	0.87	0.82	0.99	0.003 7	0.000 1	0.88
45900608	0.04	-0.10	0	-0.15	-0.72	0.04	0.67	0.67	0.99	0.015 0	0.000 3	0.84
45900819	0.10	-0.34	0	-0.02	-1.82	0.02	0.70	-2.70	0.99	0.010 9	0	0.94
45900908	0.05	-0.28	0	0.28	-1.18	0.03	0.83	-1.77	0.99	0.008 9	0	0.93
45910326	-0.14	-0.14	-0.01	-0.20	-1.02	0.04	0.67	0.26	0.99	0.014 9	0.000 1	0.83
45910426	-0.06	-0.21	0	-0.07	-1.78	0.02	0.16	-4.68	0.97	0.012 6	0	0.95
45910617	-0.12	-0.44	0	-0.24	-3.28	-0.04	0.24	-9.00	0.94	0.005 8	0	0.92
45920321	-0.10	-0.10	0	-0.11	-0.65	0	0.71	0.50	0.99	0.020 9	0	0.83
45920501	0.02	-0.22	-0.01	0.07	-1.20	0.01	0.57	-1.15	0.99	0.011 4	0	0.94
45920514	0.02	-0.11	0	-0.10	-0.27	0.01	0.82	0.79	1.00	0.012 2	0.000 2	0.95
45920616	0.12	-0.11	0	-0.01	-0.29	0.02	0.82	0.76	0.99	0.011 6	0	0.95
45920704	0	-0.07	0	-0.04	-0.10	0.01	0.96	0.94	1.00	0.017 2	0	0.90
45920831	0.04	-0.20	0	0.02	-1.34	-0.01	0.96	-0.16	1.00	0.004 5	0	1.00
45930502	0.06	-0.17	0	-0.25	-0.56	0.03	0.65	0.15	0.99	0.020 3	0.000 3	0.91
45930523	-0.08	-0.13	0	-0.17	-1.20	-0.02	0.43	-0.26	0.99	0.017 4	0.000 1	0.83
45930615	-0.09	-0.08	0	0.05	-0.05	0.01	0.89	0.88	1.00	0.020 5	0.000 3	0.86
45930630	0.07	-0.08	0	0.18	-0.23	0.03	0.88	0.72	0.99	0.007 7	0	0.86
45940425	0.14	-0.12	0	0.11	-0.29	0.04	0.88	0.80	0.99	0.016 3	0.000 1	0.86
45940521	-0.04	-0.17	-0.01	0.04	-0.70	0.01	0.84	0.37	0.99	0.016 0	0.000 1	0.78
45940614	0.02	-0.07	0	-0.01	-0.33	0.04	0.68	0.86	0.99	0.018 4	0.000 5	0.91
45950424	-0.07	-0.12	0	-0.03	-0.35	0.03	0.80	0.56	0.99	0.025 0	0.000 3	0.89
45950603	-0.05	-0.09	0	-0.03	-0.42	0.03	0.75	0.78	0.99	0.012 1	0.000 1	0.90
45950614	-0.06	-0.06	0	-0.07	-0.11	0.03	0.87	0.90	0.99	0.024 6	0.000 1	0.91
45950703	-0.08	-0.05	0	0.13	-0.12	-0.04	0.61	0.71	0.98	0.004 5	0	0.82
45950813	-0.10	-0.12	0	0.05	-0.60	-0.01	0.93	0.66	1.00	0.005 0	0	1.00
45960317	-0.13	-0.13	0	-0.01	-0.51	-0.02	0.50	0.47	1.00	0.013 9	0.000 1	0.91
45960328	0.15	-0.25	0	0.03	-1.70	0.02	0.67	-1.76	0.99	0.012 6	0	0.90
45960530	-0.12	-0.08	0	-0.06	-0.50	0.01	0.90	0.77	1.00	0.006 2	0	0.92
45960801	0.27	-0.46	0	0.48	-1.53	0.05	0.24	-4.17	0.97	0.007 5	0	0.93
平均值	-0.01	-0.15	0	-0.01	-0.71	0.02	0.72	-0.21	0.99	0.013 6	0.000 1	0.90

注： r 为抗差估计正确定位异常值的概率； V 为估计参数的均方差， $V = [\frac{1}{m-1} \sum_{i=1}^m (\theta(i) - \hat{\theta}(i))^2]^{0.5}$.

3 结果分析

a. 误差服从正态分布时,采用传统递推最小二乘法和采用抗差递推最小二乘法估计 AR 模型参数,效果基本相同,都能满足实时预报的要求.

b. 误差服从正态分布时,不进行实时修正,洪量误差、洪峰误差、确定性系数平均值分别为 -0.012, -0.003,0.725 有异常值污染后,采用传统的最小二乘法估计预报洪水的洪量误差、洪峰误差和确定性系数,其平均值分别为 -0.14,-0.60,-0.50 采用抗差递推最小二乘法的估计结果分别为0,0.01,0.99.当有

表 3 有异常值时 2 种方法对预报洪水流量的校正效果

洪 号	洪量误差			洪峰误差			确定性系数			参数均方差		<i>r</i>
	<i>e</i>	<i>e_{ls}</i>	<i>e_r</i>	<i>f</i>	<i>f_{ls}</i>	<i>f_r</i>	<i>dc</i>	<i>dc_{ls}</i>	<i>dc_r</i>	<i>V_{ls}</i>	<i>V_r</i>	
45970605	-0.10	-0.13	-0.01	-0.02	-0.35	0.02	0.72	0.70	0.98	0.011 3	0.000 2	1.00
45970618	0.08	-0.20	0	-0.10	-0.55	-0.04	0.89	0.26	0.99	0.019 2	0	0.88
45970701	-0.10	-0.08	0	-0.01	-0.24	0.03	0.88	0.84	1.00	0.018 9	0.000 1	0.88
45970808	0.07	-0.24	0	0.07	-1.91	0.03	0.25	-7.17	0.97	0.011 9	0	0.95
45980215	-0.02	-0.13	0	0.13	-0.53	0.01	0.81	0.12	0.99	0.016 9	0	0.82
45980301	0.02	-0.09	0	-0.01	-0.38	0.02	0.89	0.80	1.00	0.018 0	0	0.90
45980509	0.08	-0.09	-0.01	0.12	-0.27	0.01	0.91	0.92	0.99	0.013 0	0.000 1	0.86
平均值	0	-0.14	0	0.03	-0.60	0.01	0.76	-0.50	0.99	0.015 6	0.000 1	0.90

异常值出现,传统递推最小二乘法的校正方法会出现越校正越差的情况,而抗差递推最小二乘法能很好地解决这个问题,这说明抗差递推最小二乘法是一种具有较强容差能力的、参数稳健的估计方法。

c. 参数估计的均方差在统计意义上反映参数估计值的稳定程度。一般均方差越大,估计参数波动越大。对于相同资料,采用 2 种不同的估计方法,通过均方差可以评判该 2 种方法的好坏。所有洪水抗差采用递推最小二乘法估计的参数均方差均较传统递推最小二乘法的小。对于预报洪水,抗差递推最小二乘法估计的参数均方差为 0.0001,传统递推最小二乘法为 0.015 6。

d. 传统递推最小二乘法随异常值偏离程度的增加,参数偏离的越厉害,而抗差估计的参数趋近于稳定。其根本原因就是,抗差估计排除了异常值对参数估计的影响,从而具有稳健的估计结果。可以看到抗差估计可以抵御粗差对参数估计的影响,按本文的误差生成模式,准确排除粗差的平均概率为 90%。正是由于抗差估计能高效率地排除粗差对参数的影响,从而使参数的均方差较小。但也要注意,抗差估计对异常值的数目是有要求的,当数据中的大部分被异常值污染时,抗差估计结果也会偏向于异常值,估计出有偏估值。对于这种比例的研究将有待进一步探讨。

参考文献：

[1] 周江文,黄幼才,杨元喜,等. 抗差最小二乘法[M]. 武汉:华中理工大学出版社,1997:104-116.
[2] TUKEY J W. Study of robustness by simulation :Particularly in improvement by djustment and combination[J]. Robustness in Statistics, 1979(2):75-106.
[3] HUBER P J. Robust estimation of location parameter[J]. Ann Math Statist,1964,35:73-101.
[4] HUBER P J. Robust Statistics[M]. New York:John Wiley & Sons,1981:22-34.
[5] 周江文. 经典误差理论与抗差估计[J]. 测绘学报,1989,18(2):115-120.
[6] 许国辉,张新长. 用 AR 模型建立变形预测模型的研究[J]. 中山大学学报:自然科学版,2004,43(1):107-109.
[7] 包为民,瞿思敏,黄贤庆,等. 水文系统抗差权函数分析与检验[J]. 清华大学学报:自然科学版,2003,43(8):1127-1129.
[8] 瞿思敏,包为民,石朋,等. AR 模式误差修正方程参数抗差估计[J]. 河海大学学报:自然科学版,2003,31(5):497-500.

Robust estimation of AR model parameters

BAO Wei-min¹, WANG Hao¹, ZHAO Chao¹, WEN Jun²

(1. State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering,
College of Water Resources and Environment, Hohai University, Nanjing 210098, China ;
2. College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing 210098, China)

Abstract The real-time rectification is generally based on the measured data of flood discharge. The parameters of AR model obtained by the robustified recursive least square method are more robust than those obtained by the standard recursive least square method when outliers occur in the measured data of flood discharge. By artificial generation of outliers among the flood data of the Qilijie Basin of the Minjiang River, the rectified results of parameters by the two approaches were compared. The result shows that the robustified recursive least square method has potential to reduce estimation bias in the presence of noise, and that it is superior to the standard recursive least square method.

Key words AR model ; real-time flood rectification ; outlier ; estimation with robustified recursive least square method