

河段汇流参数抗差估计研究

赵超,包为民,王叶琴,王浩,张阳

(河海大学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)

摘要 :当河段实测出流流量值被不服从正态分布的异常值污染时,最小二乘法不能获得参数最优估计.为此,采用崩溃率为 20%、附有条件的抗差最小二乘法估计马斯京根汇流参数最优值.通过对闽江永安河段理想模型参数的计算,比较了参数抗差最小二乘法与最小二乘法抵抗 2 类误差的效果.结果表明,抗差最小二乘法能排除异常值对参数估计的影响而获得稳健性的参数.

关键词 汇流参数 抗差估计 异常值 崩溃率

中图分类号 P333 **文献标识码** A **文章编号** 1000-1980(2006)01-0014-04

在流域水文模型中,绝大多数模型需要用水文资料来分析和确定模型参数.一般情况下,若假设误差服从零均值的正态分布或对数正态分布,通常的最小二乘法等方法可以取得比较好的效果.当水文资料被异常值污染时,一般的参数估计方法很难达到要求.为了排除异常值对参数估计的影响,抗差估计理论提出了污染分布概念,并利用极值函数、导函数和权函数来描述误差分布特征^[1-3].本文应用抗差理论与方法对河段马斯京根模型参数进行了估计,同时分析了河段实测出流流量值被异常值污染时抗差估计法保持参数估计稳健性的有效性.

1 参数估计方法

1.1 最小二乘法

马斯京根模型要求河道汇流参数满足 $C^T = [c_0 \ c_1 \ c_2]$, $c_0 + c_1 + c_2 = 1$ 的条件.本文首先采用拉格朗日乘子法^[4],通过乘子的引入,将约束问题化为无约束最优化问题,然后将约束与乘子相结合并添加到目标函数中,从而构成了新的目标函数.

1.2 附有条件的抗差最小二乘法

抗差估计具有排除异常值的能力,基本上能消除具有污染分布的观测值对估值的影响.为了满足 $c_0 + c_1 + c_2 = 1$ 的条件,本文采用附有条件的抗差最小二乘法,即

$$\begin{cases} XC + V = Y \\ BC + W = 0 \end{cases} \tag{1}$$
$$X = \begin{bmatrix} Q_{i_2} & Q_{i_1} & Q_{o_1} \\ Q_{i_3} & Q_{i_2} & Q_{o_2} \\ \vdots & \vdots & \vdots \\ Q_{i_{m+1}} & Q_{i_m} & Q_{o_m} \end{bmatrix} \quad Y^T = [Q_{o_2} \ Q_{o_3} \ \cdots \ Q_{o_{m+1}}] \quad B = [1 \ 1 \ 1] \quad W = -1$$

式中: V ——残差矩阵; Q_{i_j} , Q_{o_j} ——上断面入流流量和下断面出流流量; m ——时间序列编号.根据等价权原理,组成了准则函数

$$\min \Omega = \min \left\{ \sum_{i=1}^m p_i \rho(v_i) + \lambda (BX + W) \right\} \tag{2}$$

并得到参数和 λ 的抗差最小二乘解

收稿日期 2005-04-26
基金项目 浙江省重点科技项目(2002CZ3010)
作者简介 赵超(1977—),女,安徽铜陵人,博士研究生,主要从事水文水资源方面的研究.

$$\begin{bmatrix} X^T \bar{P} X & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \hat{C} \\ \hat{\lambda} \end{bmatrix} + \begin{bmatrix} -X^T \bar{P} Y \\ W \end{bmatrix} = 0 \tag{3}$$

迭代解为

$$\begin{bmatrix} \hat{C}^{(k+1)} \\ \hat{\lambda}^{(k+1)} \end{bmatrix} = \begin{bmatrix} X^T \bar{P}^{(k)} X & B^T \\ B & 0 \end{bmatrix}^{-1} \begin{bmatrix} X^T \bar{P}^{(k)} Y \\ -W \end{bmatrix} \tag{4}$$
$$\bar{P} = \begin{bmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_m \end{bmatrix}$$

式中： p_i ——等价权函数； γ_0 ——极值函数； λ ——拉格朗日乘子； $\bar{P}$ ——等价权函数矩阵。

对于不同问题,等价权函数形式和自变量是不同的,其确定的结果是否反映问题的特点,是抗差最小二乘法应用效果好坏的关键^[5-6]。本文采用的等价权函数为 IGG 权函数^[1]

$$p_i = \begin{cases} p_i & (\sigma_i \leq k_1) \\ p_i k_1 \sigma_i & (k_1 < \sigma_i \leq k_2) \\ 0 & (\sigma_i > k_2) \end{cases} \tag{5}$$

$$\sigma_i = |Q_i - Q_i| \left[\frac{\sum_{i=1}^m p_i |Q_i - Q_i|^2}{m - 3 - t} \right]^{-1} \tag{6}$$

式中： Q_i ——计算出流流量； t ——剔除异常值的个数； m ——资料样本数； k_1 ——分位参数， $k_1 = 1.5$ ； k_2 ——淘汰参数， $k_2 = 2.5$ 。

1.3 抗差估计的初值选取

抗差估计需要迭代求解。如果极值函数严凸,目标函数有唯一解且迭代过程能收敛到这个唯一解,则可以用最小二乘法结果作为迭代初值。如果极值函数不严凸,可能含有局部极值点,只有当迭代初值在真值附近时,才能保证迭代收敛到全局极小解。

本文所选的 IGG 极值函数不严凸。为了提供比较好的迭代初值,假定观测资料中包含的异常值少于 $m/5$ 个。如果把观测资料分为 $m/5$ 组,每组 5 个点,那么至少有 1 组的 5 个点全无异常值。采用 1.1 中的最小二乘法对各组的 5 个点分别进行计算,可给出 $m/5$ 个解。用式(2)计算出的最小解作为迭代初值。该初值位于真值附近,一般由 5 个正常值得得。显然,该抗差估计的崩溃率为 20%,即出现异常值的个数不能超过观测值的 20%。

2 应用实例

2.1 理想系统设计

为了分析抗差估计效果,采用参数、输入和输出都是已知且不存在任何误差的理想系统^[7],并将已知汇流参数、入流流量和出流流量过程作为系统已知真值。对于河段汇流模型,如果已知汇流参数,则根据上断面入流流量过程可计算出下断面出流流量过程。

本文采用的汇流参数 $c_0 = 0.28, c_1 = 0.52, c_3 = 0.20$,上断面入流流量为闽江永安断面 20 场不同量级和不同特点的实际洪水流量。下断面出流流量过程根据闽江永安断面 20 场不同量级和不同特点的实际洪水流量求出。理想系统中参数、上断面入流流量和下断面出流流量真值都是已知的,且系统无任何误差。

2.2 误差生成

误差随机生成时随机误差服从零均值正态分布,而异常值误差不具有此统计特征,故异常值误差采用如下生成模式^[6]：

$$e_i = \begin{cases} (r - 0.5) \bar{Q}_p & i = \text{int}(i/L) L \\ 0 & i \neq \text{int}(i/L) L \end{cases} \tag{7}$$

式中： e_i ——生成的误差； r ——分布随机数； p ——影响误差大小的常数； L ——异常值产生的频率,随机产生； $\bar{Q}$ ——入流流量均值。

2.3 计算结果

在理想系统中添加随机误差后,分别采用最小二乘法和抗差最小二乘法估计了汇流参数,结果见表 1. 在理想系统中添加异常值误差后,分别采用最小二乘法和抗差最小二乘法估计了汇流参数,结果见表 2. 2 种方法抵抗 2 类误差的效果见表 3.

表 1 随机误差的参数率定结果

Table 1 Result of parameter calibration with random error taken into account

洪号	c_{ls0}	c_{ls1}	c_{ls2}	V_{ls}	c_{r0}	c_{r1}	c_{r2}	V_r
33880520	0.2799	0.5201	0.1999	0.032	0.2798	0.5204	0.1998	0.032
33880620	0.2797	0.5206	0.1997	0.031	0.2798	0.5204	0.1997	0.048
33890511	0.2800	0.5200	0.2000	0.031	0.2800	0.5200	0.2000	0.031
33900601	0.2799	0.5202	0.1999	0.033	0.2800	0.5200	0.2000	0.033
33900608	0.2800	0.5199	0.2001	0.037	0.2798	0.5204	0.1998	0.037
33920514	0.2800	0.5199	0.2001	0.041	0.2801	0.5197	0.2002	0.041
33920616	0.2800	0.5200	0.2000	0.031	0.2800	0.5200	0.2000	0.031
33920704	0.2800	0.5201	0.2000	0.035	0.2799	0.5203	0.1998	0.035
33930615	0.2801	0.5199	0.2000	0.038	0.2801	0.5200	0.1999	0.038
33940614	0.2800	0.5200	0.2000	0.031	0.2800	0.5200	0.2000	0.031
33950603	0.2800	0.5201	0.2000	0.032	0.2800	0.5200	0.2000	0.032
33950614	0.2799	0.5203	0.1999	0.016	0.2798	0.5205	0.1998	0.135
33960801	0.2796	0.5208	0.1996	0.033	0.2787	0.5228	0.1985	0.032
33970605	0.2801	0.5198	0.2001	0.040	0.2802	0.5196	0.2002	0.040
33970618	0.2800	0.5201	0.2000	0.039	0.2797	0.5208	0.1995	0.039
33970808	0.2800	0.5200	0.2000	0.031	0.2800	0.5200	0.2000	0.031
33980215	0.2800	0.5199	0.2000	0.030	0.2800	0.5200	0.2000	0.030
33980301	0.2800	0.5199	0.2001	0.031	0.2800	0.5200	0.2001	0.163
33980509	0.2802	0.5196	0.2002	0.039	0.2803	0.5193	0.2004	0.039
33980608	0.2800	0.5201	0.2000	0.030	0.2799	0.5202	0.1999	0.030

注:下标 ls 和 r 分别表示用最小二乘法和抗差最小二乘法估计的结果; V 为抗差均方差, $V = \left[\frac{1}{m-1} \sum_{i=1}^m p_i (Q_i - Q_{ii})^2 \right]^{0.5}$.

表 2 异常值的参数率定结果

Table 2 Result of parameters calibration with outlier taken into account

洪号	c_{ls0}	c_{ls1}	c_{ls2}	V_{ls}	c_{r0}	c_{r1}	c_{r2}	V_r
33880520	0.2473	0.5802	0.1726	1.1934	0.2800	0.5200	0.2000	0.0000
33880620	0.3701	0.4004	0.2296	2.5413	0.2800	0.5200	0.2000	0.0000
33890511	0.2209	0.6755	0.1036	1.9608	0.2800	0.5200	0.2000	0.0033
33900601	0.2648	0.5764	0.1588	1.6019	0.2800	0.5200	0.2000	0.0000
33900608	0.2299	0.6781	0.0920	2.7620	0.2800	0.5200	0.2000	0.0000
33920514	0.2289	0.6633	0.1079	2.9641	0.2800	0.5200	0.2000	0.0000
33920616	0.2636	0.5650	0.1714	1.5421	0.2800	0.5200	0.2000	0.0000
33920704	0.2767	0.5493	0.1740	1.5321	0.2800	0.5200	0.2000	0.0046
33930615	0.2333	0.6332	0.1335	1.7208	0.2800	0.5200	0.2000	0.0000
33940614	0.2776	0.5306	0.1918	0.3536	0.2800	0.5200	0.2000	0.0000
33950603	0.2604	0.5771	0.1625	1.4956	0.2800	0.5200	0.2000	0.0000
33950614	0.2516	0.6028	0.1457	1.5152	0.2800	0.5200	0.2000	0.0000
33960801	0.2766	0.5631	0.1604	1.0667	0.2800	0.5200	0.2000	0.0000
33970605	0.2056	0.6836	0.1108	3.0518	0.2800	0.5199	0.2000	0.0782
33970618	0.2048	0.6922	0.1030	1.6032	0.2800	0.5200	0.2000	0.0000
33970808	0.2376	0.6098	0.1527	2.1113	0.2800	0.5200	0.2000	0.0000
33980215	0.1795	0.7164	0.1041	2.8955	0.2800	0.5200	0.2000	0.0000
33980301	0.1825	0.7378	0.0797	2.6659	0.2800	0.5200	0.2000	0.0000
33980509	0.1259	0.8424	0.0317	3.0101	0.2800	0.5200	0.2000	0.0000
33980608	0.2401	0.6156	0.1443	1.6779	0.2800	0.5200	0.2000	0.0256

由表 1~表 3 可知:

a. 当误差服从零均值正态分布时,最小二乘法和抗差最小二乘法估计的参数基本相同,并且与真值相差不大.比较参数均值和抗差均方差,最小二乘法略优于抗差最小二乘法.这与当误差严格服从正态分布,最小二乘法具有最优一致且方差最小的特性相符合.

表 3 2 种方法抵抗 2 类误差的效果对比

Table 3 Effects of two methods resisting two kinds of errors

误差类型	效 果	c_{b0}	c_{b1}	c_{b2}	V_{ls}	c_{d0}	c_{d1}	c_{d2}	V_r
随机误差	平 均	0.2800	0.5201	0.2000	0.033	0.2799	0.5202	0.1999	0.0460
	均方差	0.0001	0.0003	0.0001		0.0003	0.0007	0.0004	
异常值误差	平 均	0.2389	0.6246	0.1365	1.963	0.2800	0.5200	0.2000	0.0056
	均方差	0.0491	0.0927	0.0452		0.0000	0.0000	0.0000	

b. 当存在异常值时,最小二乘法估计的参数偏离真值,抗差最小二乘法估计的参数与真值基本相同.这是因为抗差估计具有排除异常值的能力,基本上能消除具有污染分布的观测值对参数估值的影响.抗差最小二乘法估计的抗差均方差小于最小二乘法估计的抗差均方差,说明抗差估计有抗御异常值的效果.

c. 参数估计的均方差,在统计意义上反映了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,不同的估计方法,通过均方差可以评判 2 种方法的好坏.当误差服从零均值正态分布时,2 种方法参数均方差都较小,说明 2 种方法都是比较稳定的.最小二乘法估计的参数均方差略小于抗差最小二乘法.当有异常值出现时,抗差最小二乘法的估计结果明显小于最小二乘法的估计结果,说明被异常值污染时,抗差最小二乘法估计的参数比较稳定.异常值误差越大,抗差最小二乘法估计的参数越稳定,而最小二乘法估计的参数越不稳定.

d. 根据本文初值的选择方法,抗差估计算法的崩溃率为 20%.当异常值个数大于或等于观测值 20% 的个数时,抗差估计的结果不能接受.

3 结 论

当河段实测出流流量被异常值污染时,通常的最小二乘法会造成马斯京根汇流参数的崩溃.采用附有条件的抗差最小二乘法可基本上消除异常值对参数的影响而获得稳健性的参数.

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] HUBER P J. Robust statistics[M]. New York: John Wiley & Sons, 1981:22-34.
[3] HUBER P J. Robust estimation of location parameter[J]. Ann Math Statist, 1964, 35:73-101.
[4] 薛履中.工程最优化技术[M].天津:天津大学出版社,1988:146-161.
[5] 包为民,嵇海祥,胡其美,等.抗差理论及在水文学中的应用[J].水科学进展,2003,14(4):528-532.
[6] 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.
[7] 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 reach flow concentration parameters

ZHAO Chao, BAO Wei-min, WANG Ye-qin, WANG Hao, ZHANG Yang

(State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering,
College of Water Resources and Environment, Hohai University, Nanjing 210098, China)

Abstract Owing to the fact that the measured river outflow is seriously influenced by outlier, the least square method is ineffective for optimal parameter estimation. Therefore, the robust least square method with some additional conditions at the breakdown ratio being 20% was employed for optimal estimation of Muskingum parameters. The study of the effect of robust estimation relative to least square estimation through parameter calculation for the ideal model of Yong'an Reach of the Minjiang River shows that the robust estimation can avoid the influence of outlier and obtain stable parameters.

Key words flow concentration parameter; robust estimation; outlier; breakdown ratio