DOI:10.13203/j.whugis1995.02.010
第20卷第2期 武汉测绘科技大学学报 Vol. 20 No. 2
1995年6月 Journal of Wuhan Technical University of Surveying and Mapping June 1995

粗差数学期望平移模型的理论研究*

王金岭 陈永奇
(武汉测绘科技大学地球科学与测量工程学院,武汉市珞喻路39号,430070)

摘 要 现有文献对数学期望平移模型的理论分析仅考虑了观测值统计独立的特殊情况。基于观测值统计相关的一般情况,导出了数学期望平移参数估值($\hat{\nabla S_i}$)的简明表达式。在此基础上,采用统计预测理论对$\hat{\nabla S_i}$进行了直观的理论解释,扩展了统计学文献中的有关结论。借助于实例,分析了$\hat{\nabla S_i}$与最小二乘残差的本质区别。

关键词 相关观测值;粗差;数学期望平移模型;预测残差;
分类号 P207;O211.67

1 数学期望平移模型及其参数估计

数学期望平移模型是处理观测值粗差的主要模型之一[1,2],由统计学者 Dixon 和 Paulson 在50年代最早提出并应用于样本观测值的粗差检测[3]。数学期望平移模型的基本思想是,作为粗差检验的原假设,设有几个样本观测值 $l_i(i=1,2,\cdots,n)$来源于正态分布 $N(\mu,\sigma^2)$,即 $H_0:l_i\sim N(\mu,\sigma^2),i=1,2,\cdots,n$。其中 μ 为观测值的数学期望,σ^2 为方差。如果某一观测值 l_j 含有粗差 ∇S_j,则将 l_j 视为与其它观测值具有相同的方差 σ^2,但具有不同数学期望的一个子样。因此,粗差检验的备选假设为 $H_a:l_j\sim N(\mu+\nabla S_j,\sigma^2),l_i\sim N(\mu,\sigma^2)(i\neq j)$。这里$\nabla S_j$ 是观测值 l_j 的数学期望平移参数。根据这一模型可以构造出样本观测值粗差检验的统计量。

处理样本观测值粗差的数学期望平移模型在线性模型的统计诊断领域(如文献[4,5]等)和测量系统观测值粗差定位研究(如文献[1,2,6,8]等)中得到了广泛应用。设 l 是 n 维观测值向量,其方差阵为 $D_l(|D_l|>0)$,l 服从多维正态分布,即 $l\sim N_n(Ax,D_l)$。粗差检验的原假设为:

$$H_0:\quad E(l)=Ax \tag{1}$$

$$D_l=\sigma_0^2Q=\sigma_0^2P^{-1} \tag{2}$$

这里,A 为 $n\times t$ 设计矩阵,其秩 $rk(A)=t$;x 为 $t\times 1$ 模型参数。考虑将观测值 l 分成 l_1 和 l_2 两组;第一组观测值 l_1 不含粗差,其维数为 n_1;第二组观测值 l_2 可能含有粗差∇S,其维数为 n_2。∇S 视为观测值 l_2 的数学期望平移向量,即

$$l=\begin{bmatrix}l_1\\l_2\end{bmatrix}\sim N_n\left[\begin{pmatrix}A_1x\\A_2x+\nabla S\end{pmatrix},\quad \sigma_0^2\begin{pmatrix}Q_{11}&Q_{12}\\Q_{21}&Q_{22}\end{pmatrix}\right]$$

粗差检验的备选假设为: H_a:

$$E(l_1)=A_1x \tag{3}$$

$$E(l_2)=A_2x+\nabla S \tag{4}$$

$$D_{l_1}=\sigma_0^2Q_{11},\ D_{l_2}=\sigma_0^2Q_{22},\ D_{l_1l_2}=\sigma_0^2Q_{12} \tag{5}$$

或者进一步表示为: H_a:

$$E(l)=Ax+H\nabla S \tag{6}$$

收稿日期:1994-12-05. 王金岭,男,33岁,副教授,博士生,现从事测量数据处理、质量控制及软件技术的研究.
* 国家自然科学基金和香港理工大学科研基金资助项目.

$$D_l=\sigma_0^2Q=\sigma_0^2P^{-1} \tag{7}$$

其中，$H=\begin{bmatrix}0\\E\end{bmatrix}$，$E$ 为 $n_2\times n_2$ 单位阵。在上述数学期望平移模型中，数学期望平移参数∇S(即粗差)可以依据最小二乘原理加以估计，由文献[2]，

$$\hat{\nabla S}=P_{ss}^{-1}[H^{T}pl-H^{T}PA(A^{T}PA)^{-1}A^{T}Pl]=-P_{ss}^{-1}H^{T}PV \tag{8}$$

$$Q_{\hat{\nabla S}}=P_{ss}^{-1}=[H^{T}PQ_{V}PH]^{-1} \tag{9}$$

公式(8)，(9)中，$Q_V=P^{-1}-A(A^{T}PA)^{-1}A^{T}$；$V$ 是模型(1)，(2)的最小二乘残差。

正如(8)式所示，数学期望平移参数的估值$\hat{\nabla S}$通常表示为原始平差模型(1)，(2)最小二乘残差 V 的线性函数，而最小二乘估计对观测值粗差有“均摊”作用，残差并不能真实地反映粗差值，因此，$\hat{\nabla S}$能否充分地估计出观侧值 l_2 中的粗差从形式上看并不是很直观。对$\hat{\nabla S}$的表达式进一步简化并对其进行理论分析是十分必要的。文献[4,5]等仅在观测值之间统计独立的特殊情况下给出了$\hat{\nabla S}$的直观解释。本文针对观测值统计相关的一般情况，推导了$\hat{\nabla S}$的简化表达式。在此基础上，借助于统计预测理论对$\hat{\nabla S}$进行了直观的理论解释，扩展了现有文献的有关结论。

2　数学期望平移参数估值的理论分析

2.1　$\hat{\nabla S}$的进一步推导

为了对数学期望平移参数估值$\hat{\nabla S}$进行理论分析，首先对(8)式进行简化。为讨论问题方便起见，推证结果以定理的形式表述如下。

定理　在数学期望平移模型(6)，(8)中，平移参数的最小二乘估值$\hat{\nabla S}$可由下式表示：

$$\hat{\nabla S}=l_2-\hat{l}_2-Q_{21}Q_{11}^{-1}(l_1-\hat{l}_1) \tag{10}$$

其中，　$\hat{l}_1=A_1\hat{x}$，　$\hat{l}_2=A_2\hat{x}$，　$\hat{x}=(A_1^{T}Q_{11}^{-1}A_1)^{-1}A_1^{T}Q_{11}^{-1}l_1$

证明　将观测值 l 的权阵和协因数阵按 l_1 和 l_2 的分组方式分块，则

$$\begin{bmatrix}Q_{11}&Q_{12}\\Q_{21}&Q_{22}\end{bmatrix}\begin{bmatrix}P_{11}&P_{12}\\P_{21}&P_{22}\end{bmatrix}=\begin{bmatrix}E_1&0\\0&E_2\end{bmatrix} \tag{11}$$

由(11)式不难推证：

$$Q_{11}^{-1}=P_{11}-P_{12}P_{22}^{-1}P_{21} \tag{12}$$

$$P_{22}^{-1}P_{21}=-Q_{21}Q_{11}^{-1} \tag{13}$$

公式(12)，(13)将在下面的推导过程中被多次采用。

为证明(10)式成立，首先利用分块矩阵对(8)式右边的逐项分别简化。

$$\begin{aligned}P_{ss}^{-1}&=(H^{T}PQ_{V}PH)^{-1}=[H^{T}PH-H^{T}PA(A^{T}PA)^{-1}A^{T}PH]^{-1}\\&=[P_{22}-(P_{21}A_1+P_{22}A_2)(A^{T}PA)^{-1}(A_1^{T}P_{12}+A_2^{T}P_{22})]^{-1}\end{aligned}$$

依据矩阵反演公式，则

$$\begin{aligned}P_{ss}^{-1}=&P_{22}^{-1}+(P_{22}^{-1}P_{21}A_1+A_2)[A^{T}PA-(A_1^{T}P_{12}P_{22}^{-1}+A_2^{T})\\&\times P_{22}(P_{22}^{-1}P_{21}A_1+A_2)]^{-1}(A_1^{T}P_{12}P_{22}^{-1}+A_2^{T})\end{aligned} \tag{14}$$

令，$A_2+P_{22}^{-1}P_{21}A_1=C$，并顾及(12)式，可证明：

$$A^{T}PA=A_1^{T}Q_{11}^{-1}A_1+C^{T}P_{22}C \tag{15}$$

将(15)式代入(14)式，得：

$$P_{ss}^{-1}=P_{22}^{-1}+C(A_1^{T}Q_{11}^{-1}A_1)^{-1}C^{T} \tag{16}$$

在(16)式两边右乘 $H^TPA(A^TPA)^{-1}$,可进一步导出：

$$\begin{aligned}P_{ss}^{-1}H^TPA(A^TPA)^{-1} &= P_{22}^{-1}(P_{21}A_1 + P_{22}A_2)(A^TPA)^{-1} + C(A_1^TQ_{11}^{-1}A_1)^{-1} \\ &\quad \times C^T(P_{21}A_1 + P_{22}A_2)(A^TPA)^{-1} \\ &= C(A^TPA)^{-1} + C(A_1^TQ_{11}^{-1}A_1)^{-1}(A_2^TP_{21}A_1 \\ &\quad + A_2^TP_{22}A_2 + A_1^TP_{12}A_2 + A_1^TP_{12}P_{22}^{-1}P_{21}A_1)(A^TPA)^{-1}\end{aligned} \tag{17}$$

顾及(12),(15)式,上式可简化为：

$$\begin{aligned}P_{ss}^{-1}H^TPA(A^TPA)^{-1} &= C(A^TPA)^{-1} + C(A_1^TQ_{11}^{-1}A_1)^{-1}(A^TPA - A_1^TQ_{11}^{-1}A_1)(A^TPA)^{-1} \\ &= C(A_1^TQ_{11}^{-1}A_1)^{-1}\end{aligned} \tag{18}$$

此外,容易导出：

$$H^TPl = P_{21}l_1 + P_{22}l_2 \tag{19}$$

$$A^TPl = A_1^TP_{11}l_1 + A_2^TP_{21}l_1 + A_2^TP_{22}l_2 + A_1^TP_{12}l_2 \tag{20}$$

将(16,18～20)式代入(8)式,得

$$\begin{aligned}\hat{\nabla S} &= P_{22}^{-1}P_{21}l_2 + l_2 + C(A_1^TQ_{11}^{-1}A_1)^{-1}(A_2^TP_{21}l_1 + A_2^TP_{22}l_2 + A_2^TP_{12}P_{22}^{-1}P_{21}l_1 \\ &\quad + A_1^TP_{12}l_2) - C(A_1^TQ_{11}^{-1}A_1)^{-1}(A_1^TP_{11}l_1 + A_2^TP_{21}l_1 + A_2^TP_{22}l_2 + A_1^TP_{12}l_2) \\ &= l_2 + P_{22}^{-1}P_{21}l_1 - C(A_1^TQ_{11}^{-1}A_1)^{-1}A_1^T(P_{11} - P_{12}P_{22}^{-1}P_{21})l_1\end{aligned} \tag{21}$$

顾及(12),(13)式,上式可进一步简化为：

$$\begin{aligned}\hat{\nabla S} &= l_2 - Q_{21}Q_{11}^{-1}l_1 - C(A_1^TQ_{11}^{-1}A_1)^{-1}A_1^TQ_{11}^{-1}l_1 \\ &= l_2 - A_2(A_1^TQ_{11}^{-1}A_1)^{-1}A_1^TQ_{11}^{-1}l_1 - Q_{21}Q_{11}^{-1}[l_1 - A_1(A_1^TQ_{11}A_1)^{-1}A_1^TQ_{11}^{-1}l_1]\end{aligned} \tag{22}$$

比较(22)式和(10)式,定理证毕。

2.2 $\hat{\nabla S}$的理论分析

为了对$\hat{\nabla S}$进行理论分析,引入统计预测的概念。因为观测值 l_1 和 l_2 服从联合正态分布,由概率论可知,给定 l_1 之值时,l_2 的条件分布为正态。其条件数学期望为：

$$E(l_2|l_1) = \tilde{\mu}_2 = \mu_2 + Q_{21}Q_{11}^{-1}(l_1 - \mu_1) = f(l_1) \tag{23}$$

(23)式中,μ_1,μ_2 分别为观测值 l_1,l_2 的数学期望值。根据统计预测理论,$f(l_1)$是在均方误差最小的意义下(用 l_1 的函数)对 l_2 的最佳逼近[7]。一般来说,观测值的数学期望值(μ_1,μ_2)是未知的。一个合理的方案是采用它们的估值(如 $\hat{l}_1$,$\hat{l}_2$)代替。那么,$\hat{l}_2+Q_{21}Q_{11}^{-1}(l_1-\hat{l}_1)=\tilde{l}_2$ 可视为观测值 l_2 条件数学期望的预测值。观测值 l_2 与其条件期望的预测值 $\tilde{l}_2$ 之差

$$\hat{\nabla l_2} = l_2 - \hat{l}_2 = l_2 - \tilde{l}_2 - Q_{21}Q_{11}^{-1}(l_1 - \hat{l}_1) \tag{24}$$

定义为观测值 l_2 的预测残差。比较(10)式和(24)式,可以看出,观测值 l_2 数学期望平移参数的最小二乘估值$\hat{\nabla S}$实质上就是观测值 l_2 的预测残差。

分析(10)式不难发现,按最小二乘原理求解观测值 l_2 数学期望平移参数实质上等价于,先由第一组不含粗差的观测值 l_1 求解测量系统的模型参数的最小二乘估值 $\hat{x}=(A_1^TQ_{11}^{-1}A_1)^{-1}\times A_1^TQ_{11}^{-1}l_1$;然后,由 $\hat{x}$ 求解观测值 l_1、l_2 的数学期望估值(平差值)$\hat{l}_1$ 和 $\hat{l}_2$;最后,依据(24)式和(10)式求得观测值 l_2 的预测残差(即$\hat{\nabla S}$)。根据最小二乘估计理论,$\hat{x}$ 是最优无偏估值。因此,$\hat{l}_1$ 和 $\hat{l}_2$ 是数学期望 μ_1 和 μ_2 的最优无偏估值。显然,若观测值 l_2 中含有粗差,则会在数学期望平移参数估值$\hat{\nabla S}$(即预测残差)中得到充分的反映,$\hat{\nabla S}$是粗差值∇S 的最优无偏估计。

总之,借助于预测残差的概念对$\hat{\nabla S}$ 加以理论解释,其本质含义非常直观。$\hat{\nabla S}$虽然可以表示为最小二乘残差的线性函数[如(8)式],但却与最小二乘残差 V 有着根本区别。下面的实例分析,可以进一步证实这一结论。

3 算例分析

图 1 为某城市三等加密三角网，网点 1，2，3 为城市二等控制点。方向观测值精度为 $m_r=\pm 1.273''$。经检验，方向观测值中未含粗差。为了检验这三个二等点坐标值是否相容，即坐标值中是否含有粗差，将其坐标值视为有误差的观测值与三等网的方向观测值一起平差，基于数学期望平移模型进行粗差检测，有关公式详见文献[8]。由二等网平差结果中得到坐标观测值的协方差阵(cm^2)为：

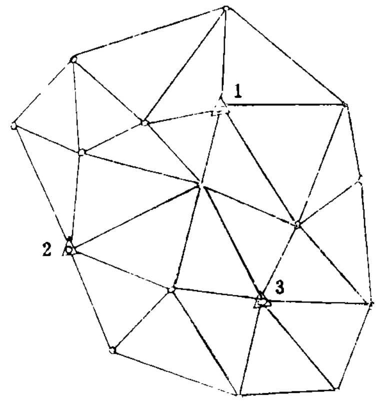

图 1

$$D_c=\begin{bmatrix} 14.48 & 7.30 & 8.20 & 2.56 & 6.42 & 6.46 \\ & 37.51 & 12.38 & 21.77 & 0.27 & 25.54 \\ & & 11.77 & 8.72 & 5.11 & 9.22 \\ & & & 16.52 & 0.59 & 15.79 \\ & (\text{对称}) & & & 5.99 & 1.20 \\ & & & & & 20.28 \end{bmatrix}$$

为了分析坐标观测值最小二乘残差 V 与数学期望平移参数估值 ∇S 的差别，分别在每个坐标观测值中模拟了同样的粗差 +50cm。表 1 列出了这些模拟方案中，含粗差的坐标观测值的残差 V_i 和数学期望平移参数估值 $\hat{\nabla S_i}$，由 $\hat{\nabla S_i}$ 和 V_i 构造的粗差检验统计量 $W_i=|\hat{\nabla S_i}|/\sigma_{\hat{\nabla S_i}}$，$\overline{W}_i=|V_i|/\sigma_{V_i}$ 的最大值 $W_{i\max}$ 和 $\overline{W}_{i\max}$ 及其对应的观测值位置。

表 1　$\hat{\nabla S_i}$ 与最小二乘残差 V_i 的比较

粗差(+50cm)位置	x_1	y_1	x_2	y_2	x_3	y_3
V_i/cm	−12.65	−11.43	−8.74	−0.38	0.27	−7.62
$\overline{W}_{i\max}$	12.75	13.33	13.59	12.28	15.64	14.49
$\overline{W}_{i\max}$ 位置	y_3	x_2	y_2	y_1	y_1	x_3
$\hat{\nabla S_i}$/cm	46.62	49.29	52.36	47.53	51.0	53.47
$W_{i\max}$	12.75	13.51	13.57	13.66	15.65	14.54
$W_{i\max}$ 位置	x_1	y_1	x_2	y_2	x_3	y_3

分析表 1 中的数据可以看出，含粗差的观测值其残差与相应的模拟粗差值相差甚远。根据残差理论，粗差 ∇S_i 在其对应的观测值残差 V_i 中的反映与多余观测分量 $r_i=(Q_vP)_{ii}$ 成正比，即 $V_i=-r_i\nabla S_i$，见文献[2]。进一步分析发现，坐标观测值 y_2 和 x_3 的多余观测分量都接近于 0，因此，当 50cm 的粗差分别加到这两个观测值上时其残差 V_i 几乎没有变化。若采用基于残差的统计量 $\overline{W}_i$ 定位粗差，在模拟的所有方案中均未能得到正确结果。另一方面，数学期望平移估值 ∇S_i 都接近相应的模拟粗差值，基于 $\hat{\nabla S_i}$ 构造的粗差检验统计量 W_i 可以有效地定位粗差。

4 结束语

当观测值统计独立时，$Q_{21}=Q_{12}=0$，由(10)式可导出统计学文献(如文献[4,5])中的有关结论。因此，本文的研究结果具有更一般的理论意义。

近年来，一些文献提出采用预测残差构造粗差检验统计量。根据上述分析结果，这种比较直观的粗差检验统计量构造方法在理论上与数学期望平移模型等价。为了保证数学期望平移

参数估值(或预测残差)用于粗差检验的有效性,关键问题是对观测值正确分组。

文献[2]明确指出,稳健估计的一个致命不足是将影响函数及权函数选择成观测值最小二乘残差的函数,而稳健准则是针对粗差规定的。根据上述分析结果,数学期望平移参数估值 $\hat{\nabla S}$(或预测残差)对粗差的估计远比最小二乘残差可靠。因此,借助于 $\hat{\nabla S}$ 构造权函数可能更为合理。这一问题尚需进一步探讨。

参 考 文 献

1 陈永奇. 变形观测数据处理. 北京:测绘出版社,1988.

2 李德仁. 误差处理和可靠性理论. 北京:测绘出版社,1988.

3 Barnett V, Lewis T. Outliers in Statistical Data (Second Edition). New York: John Wiley and Sons, 1984.

4 Cook R D, Weisberg S. Residuals and Influence in Regression. New York: Chapman and Hall, 1982.

5 韦博成,鲁国斌,史建清. 统计诊断引论. 南京:东南大学出版社,1991.

6 Forstner W. Reliability and Discernability of Extended Gauss-Markov Models. DGK, Reihe A, 1983, 98: 79～103

7 陈希孺,王松桂. 近代回归分析——原理方法及应用. 合肥:安徽教育出版社,1987.

8 王金岭,陈永奇. 控制网连接点坐标值粗差的可定位性研究. 武汉测绘科技大学学报,1993,18(4):40～47

A Research on the Mean Shift Outlier Models

Wang Jinling Chen Yongqi

(School of Earth Science and Surveying Engineering, WTUSM, 39 Luoyu Road, Wuhan, China, 430070)

Abstract Based on the mean shift outlier model, outliers are considered as the mean shift parameters of the corresponding observations, which can be estimated by using the least squares method. In literature, the estimation of the mean shift parameters is often formulated as the complicated linear function of the least squares residuals biased by the presence of outliers, and its statistical performance is not lucid. In this paper, the estimated mean shift parameters have been further derived and represented as a concise formula in the general case of correlated observations, from which the estimated mean shift parameters are interpreted with the concept of statistical prediction. Finally, the essential distinction between the estimated mean shift parameters and residuals is demonstrated by a practical example.

Key words correlated observations; outliers; mean shift outlier model; statistical prediction