DOI: 10.13203/j.whugis1998.01.002

第 23卷第 1期 武汉测绘科技大学学报 Vol. 23 No. 1
1998年 3月 Journal of Wuhan Technical University of Surveying and Mapping March 1998

基于相关分析的粗差理论*

施 闯 刘经南

(武汉测绘科技大学地学测量工程学院,武汉市珞喻路 129号, 430079)

摘 要 提出了基于相关分析的可靠性研究和粗差分析理论,并应用于相关观测量的多粗差分析。

关键词 相关分析法;相关观测量;粗差分析;误差处理

分类号 P207

现代测量手段趋于向数据采集的自动化和快速化发展。然而由于各种因素的影响,由仪器和计算机自动采集的数据中,常带有一定数量的粗差观测量。这些粗差观测量如果直接参与数据处理,将大大降低成果应有的精度,甚至导致错误的估计。现有的粗差分析方法在实用上多限定于独立观测量或等权独立观测量的单个或多个粗差的条件,而对于相关观测量的多粗差问题,被认为是一个难题。对此,本文提出了用相关分析的方法,解决相关观测量的多粗差问题。

1 原 理

在相关观测量的最小二乘平差中,一个观测量的误差,通过观测量之间随机特性的相关性和图形几何条件的关联性,反映于其它观测量的平差改正数之中。当一个观测量含有粗差时,它必将或多或少地影响到其它观测量的改正数。那么粗差观测量是否与改正数向量 V 之间存在着某种必然的联系?

1.1 观测量误差对改正数向量的影响

由可靠性矩阵[1]:

$$R = Q_{VV}P_{ll} = E - A(A^{\mathrm{T}}P_{ll}A)A^{\mathrm{T}}P_{ll} \quad (1)$$

且有:

$$R\Delta = -V \quad (2)$$

其中,A 是平差系统的图形设计矩阵;P_{ll} 是观测量的权矩阵;Δ 是观测值误差;V 是观测值改正数向量;Q_{VV} 是观测值改正数的协因数阵。将(2)式展开整理得:

$$R\Delta = \begin{bmatrix} r_{11} \\ r_{21} \\ \vdots \\ r_{n1} \end{bmatrix}\Delta_1 + \begin{bmatrix} r_{12} \\ r_{22} \\ \vdots \\ r_{n2} \end{bmatrix}\Delta_2 + \cdots + \begin{bmatrix} r_{1n} \\ r_{2n} \\ \vdots \\ r_{nn} \end{bmatrix}\Delta_n = -[v_1, v_2, \cdots, v_n]^{\mathrm{T}} \quad (3)$$

式中,Δ_i 为数值变量 $(i = 1, 2, \cdots, n)$。令

$$F_i = [r_{1i}, r_{2i}, \cdots, r_{ni}]^{\mathrm{T}}, (i = 1, 2, \cdots, n)$$

则有:

$$\Delta_1F_1 + \Delta_2F_2 + \cdots + \Delta_nF_n = -V \quad (4)$$

称 F_i 为观测量 i 的误差 Δ_i 对改正数向量 V 的影响向量。

F_i 由平差系统的图形设计矩阵 A 和观测量的权矩阵 P_{ll} 所决定,它反映了观测量 i 的误差对改正数向量 V 的内在的影响关系和作用程度。

由(4)式可以看出:观测量的改正数向量 V 可以表示为 Δ_iF_i $(i = 1, 2, \cdots, n)$ 的向量和;各观测量的误差 Δ_i 在其影响向量 F_i 的作用下反映于改正数向量 V 之中;Δ_i 起到了对向量 F_i 的缩放作用。

1.2 粗差与改正数向量的相关关系

当某观测量出现粗差时,无论是将其归于函数模型还是随机模型的误差[1],都将表现为该观测量的 $\| \Delta_iF_i \|_2$ 值,即向量 Δ_iF_i 的模较大,且在(4)式左半部分各项中占有优势。这时,改正数向量 V 将突出地表现为与该观测量的影响向量 F_i 有较强的相关性。

图 1显示出观测量中含有一个粗差时,粗差观测量 $(i = 18)$ 的影响向量 F_{18} 的各分量 $r_{1i}, r_{2i}, \cdots, r_{ni}$ $(i = 18, n = 25)$ 与 V 对应分量 $v_1, v_2, \cdots, v_n$ $(n = 25)$ 的关系。

当观测量中含有多个粗差时,就表现为这种相关关系的叠加。图 2中的数据是实测数据(相关观测量)中,加入了 3倍观测量中误差的粗差所得到的结果。

下面对向量 F_i 与 V 的相关性进行定量的分析。

收稿日期: 1996-12-02。施闯,男, 28岁,博士生,现从事高精度 GPS网数据分析理论和软件研究。

* 国家自然科学基金资助项目,编号 49674206

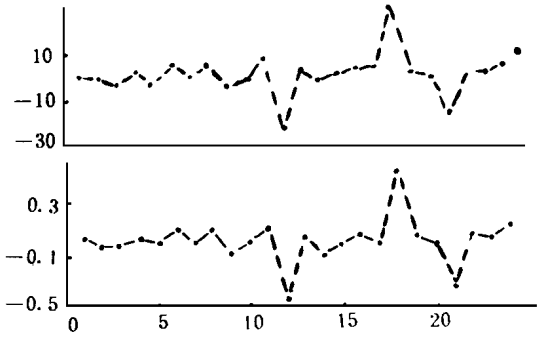


图 1 含单个粗差时,粗差的影响向量 F_{18}与 V的关系

Fig. 1 Relation Between the Outlier Influence Vector F_{18} and the Residual Vector V(Single Outlier)

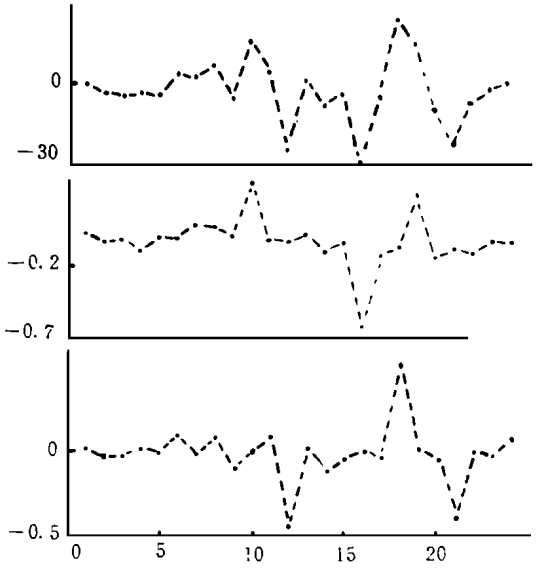


图 2 含两粗差时,粗差的影响向量 F_{18} F_{16}与 V的关系

Fig. 2 Relation Between the Outlier Influence Vector F_{18}, F_{16} and the Residual Vector V(Two Outliers)

如图 1,图 2所示,若要定量地描述两者的相关程度,可以考虑用它们对应分量 F_{ji}($F_{ji}=r_{ji}$)与 v_j($j=1,2,\cdots,n$)差的平方和的最小值 Q_0来表示:

$$Q_0=\min_{a,b}\frac{1}{n}\sum_{j=1}^{n}(v_j-a-bF_{ji})^2 \qquad (5)$$

如果存在某个 a和 b,使得 $Q_0=0$,则可以说 F_i与 V完全相关,否则就用 Q_0的大小来描述两者的相关程度,为了求 Q_0,对函数

$$Q(a,b)=\frac{1}{n}\sum_{j=1}^{n}(v_j-a-bF_{ji})^2 \qquad (6)$$

求关于 a、b的偏导数,并令其等于 0,即

$$\frac{\partial Q}{\partial a}=-\frac{2}{n}\sum_{j=1}^{n}(v_j-a-bF_{ji})=0$$

$$\frac{\partial Q}{\partial b}=-\frac{2}{n}\sum_{j=1}^{n}((v_j-a)F_{ji}-bF_{ji}^2)=0$$

解得:

$$b=\sum_{j=1}^{n}(F_{ji}-\bar{F}_i)(v_j-\bar{v})\Big/\sum_{j=1}^{n}(F_{ji}-\bar{F}_i)^2$$

$$a=\bar{v}-b\bar{F}_i \qquad (7)$$

将式(7)代入式(6)得:

$$Q_0=\frac{1}{n}\sum_{j=1}^{n}(v_j-\bar{v})^2(1-\mathrm{d}^2_{F_i,V}) \qquad (8)$$

其中记:

$$\mathrm{d}_{F_i,V}=\frac{\sum_{j=1}^{n}(F_{ji}-\bar{F}_i)(v_j-\bar{v})}{\left(\sum_{j=1}^{n}(F_{ji}-\bar{F}_i)^2\sum_{j=1}^{n}(v_j-\bar{v})^2\right)^{1/2}} \qquad (9)$$

称 $\mathrm{d}_{F_i,V}$为 F_i与 V的相关系数。$\mathrm{d}_{F_i,V}$是一个无量纲的量,它定量地反映了 F_i与 V之间的相关程度。

相关系数 $\mathrm{d}_{F_i,V}$具有下列性质:

$|\mathrm{d}_{F_i,V}|\leqslant 1$;$|\mathrm{d}_{F_i,V}|$越大,则 F_i与 V越相关;$|\mathrm{d}_{F_i,V}|=0$时,F_i与 V不相关。

一个粗差的出现,会对整个平差系统产生较大的影响,这种影响可以通过 F_i的作用,反映于改正数向量 V之中。要想从 V中发现粗差的踪迹,必须从分析 F_i与 V的关系入手,而 $\mathrm{d}_{F_i,V}$正是定量地反映了两者之间的相关程度,这也就是对F_i与 V的相关程度。当$|\mathrm{d}_{F_i,V}|$越接近于 1时,相关性越强,说明改正数向量 V的变化受来自观测量 i的误差的影响越显著。如果存在粗差,则该观测量为粗差的可能性也就越大。反之,$|\mathrm{d}_{F_i,V}|$越接近于 0时,对F_i对 V的影响越不显著,则它为粗差的可能性就越小。

当观测量中含有多个粗差时,改正数向量 V的变化规律,将表现为来自这些粗差对其影响的叠加(如图 2所示)。各粗差观测量的 F_i都将显示出与 V有着相对显著的相关性,相关性的大小取决于粗差个数、分布和粗差值的大小。所以当相关观测量中含有多个粗差时,只要粗差是可测的,通过 $\mathrm{d}_{F_i,V}$仍能分析出粗差所在的位置。

2 相关系数 $\mathrm{d}_{F_i,V}$的检验

如果 F_i、V的分量由随机变量 f_i、v组成,则其相关系数可由下式给出:

$$\tilde{\mathrm{d}}_{F_i,V}=\mathrm{cov}(f_i,v)\Big/\sqrt{D(f_i)D(v)} \qquad (10)$$

其中, $\mathrm{cov}(f_i, v) = E\{[f_i - E(f_i)][v - E(v)]\}$

$D(f_i) = E(f_i^2) - [E(f_i)]^2$

$D(v) = E(v^2) - [E(v)]^2$

在实际问题中,如果有 n 个观测量,计算 F_i、V 的相关系数是用式(9)求 $\rho_{F_i,V}$ 来估计 $\tilde{\rho}_{F_i,V}$。

如果 f_i、V 服从二维正态分布,则当 $\tilde{\rho}_{F_i,V} = 0$ 时, $\rho_{F_i,V}$ 的概率密度函数为:

$$\tilde{\rho}_{F_i,V}(\rho_{F_i,V}) = \frac{1}{\pi}\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)}(1-\rho_{F_i,V}^2)^{\frac{n-4}{2}} \tag{11}$$

其中, $\Gamma\left(\frac{n-1}{2}\right)$、$\Gamma\left(\frac{n-2}{2}\right)$ 是 Gamma函数。

检验时,取零假设 H_0: $\tilde{\rho}_{F_i,V} = 0$,即 F_i、V 不相关,则有统计量:

$$t = \rho_{F_i,V}(n-2)^{1/2}/(1-\rho_{F_i,V}^2)^{1/2} \tag{12}$$

t 是遵从 $n-2$ 自由度的 t 分布。

给定置信度 $\alpha = 0.001$,则由(12)式可得:

若 $$\rho_{F_i,V} > \frac{t_\alpha}{\sqrt{t_\alpha^2 + (n-2)}} \tag{13}$$

成立,则拒绝零假设,认为 F_i 与 V 相关是显著的;否则接受零假设。

3 相关分析法的几何解释

由式(9)定义的相关系数可以表达为下列形式:

$$\rho_{F_i,V} = \frac{F_i^{\mathrm{T}} HV}{\|HF_i\| \cdot \|HV\|} = \cos\theta \tag{14}$$

式中, H 为中心化矩阵, θ 为向量 HF_i 和 HV 的夹角。这就是说, F_i 与 V 的相关系数的几何意义是其经过中心化变换后的向量 HF_i 与 HV 的夹角的余弦。 F_i 与 V 的相关性越强, HF_i 与 HV 的夹角越接近于 0°或 180°。

根据式(4),若 i 出现粗差,即 $\|\nabla F_i\|_2$ 显著大于其它值,这时 ∇F_j ($j = 1, 2, \cdots, n$)的向量之和必然使 HV 的方向接近于 H_iF_i,即 F_i 与 V 具有较强的相关性。

同样,观测量中存在多个粗差时,由于观测量的个数远远大于粗差的个数, F 向量空间的维数也远远大于粗差个数。这时, ∇F_i ($i = 1, 2, \cdots, n$)的向量和,同样会使 HV 在 n 维空间中的方向接近于各粗差测量的 HF_i 的方向。

当然有可能会出现多个粗差共同作用的结果,使 HV 的方向更接近于某个非粗差观测量的 HF_i,但这种情况仅限于某些特殊的图形结构和粗差组合。在粗差分析中,可以通过反向搜索的方法,去伪存真。

4 平差系统中是否含有粗差

以 GPS网为例,对验后单位权中误差 $\hat{\sigma}_0^2$ 进行 χ^2 检验。 $\hat{\sigma}_0^2$ 的理论值为 1,则有 χ^2 检验量:

$$\chi^2 = V^{\mathrm{T}}PV \sim \chi^2(f)$$

取零假设: $H_0 = \hat{\sigma}_0^2 = 1$;选置信水平 $1-\alpha$, $\alpha = 0.005$,如果:

$$\chi^2_{p_1} \leqslant V^{\mathrm{T}}PV \leqslant \chi^2_{p_2} \text{ 或 } \frac{V^{\mathrm{T}}PV}{\chi^2_{p_1}} \leqslant \hat{\sigma}_0^2 \leqslant \frac{V^{\mathrm{T}}PV}{\chi^2_{p_2}}$$

其中, $p_1 = (1-\alpha)/2$, $p_2 = (1+\alpha)/2$ 成立,则接受零假设。否则,拒绝零假设,认为平差系统中可能存在粗差。

5 粗差的可测性和可区分性

在一些情况下,某些观测量的粗差是不可能被平差系统自身所发现的;而有些粗差即使被发现,也不可能区分它存在于哪个观测量之中。这就是粗差的可测性和可区分性问题。

基于本文中的理论,不难得出下列结论:

1)观测量对改正数向量的影响向量 F_i 的模 $\|F_i\|_2$ 越接近于 0,则平差系统对该观测量的误差越不敏感,该观测量的粗差越不易被发现。当 $F_i = 0$ 时,该观测量的粗差是不可测的,称为粗差不可测观测量。

2)对于两个观测量,其影响向量 F_i、F_j 的相关系数为:

$$\rho_{F_i,F_j} = \frac{\sum_{k=1}^{n}(F_{ki} - \bar{F}_i)(F_{kj} - \bar{F}_j)}{\left(\sum_{k=1}^{n}(F_{ki} - \bar{F}_i)^2 \sum_{k=1}^{n}(F_{kj} - \bar{F}_j)^2\right)^{1/2}} \tag{15}$$

当 $|\rho_{F_i,F_j}|$ 越接近于 1时,这两个观测量的粗差的可区分性越差;若 $|\rho_{F_i,F_j}| = 1$,则这两个观测量的粗差是不可区分的,称为粗差不可区分观测量。

6 粗差分析实现的方法和步骤(以 GPS网为例)

6.1 同时探测多个粗差

1)对观测量进行最小二乘平差,求得观测量改正数向量 V 和验后单位权中误差 $\hat{\sigma}_0$。

2)对 $\hat{\sigma}_0^2$ 进行 χ^2 检验,若不通过,则系统中可能含有粗差。

3)计算可靠性矩阵 R 及各观测量对改正数的影响向量 F_i。

4)计算各观测量的 F_i 与 V 的相关系数 $d_{F_i,V}$。

5)对各 $d_{F_i,V}$ 的显著性进行 $t_{(n-2)}$ 检验,若显著,则对该观测量做粗差标记。

6)对所有粗差标记的观测量进行剔除或降权处理后,重复 1)~ 6)。若 2)中 $\hat{\sigma}_0^2$ 检验通过,则进行 7)。

7)将有粗差标记的观测量,逐一恢复到平差系统中,重复 1)、2),若 $\hat{\sigma}_0^2$ 检验仍能通过,则该观测量为误判,应确认恢复;否则确认为粗差。在这一步中,应注意不可区分粗差的特殊情况。

8)最后确认所有被标记的观测量中含有粗差,进行适当的处理后,完成粗差分析,重新进行平差。

6.2 逐个探测多粗差

1)、2)、3)、4)步同 § 6.1;

5)对 $|d_{F_i,V}|$ 最大的一个观测量作粗差标记;

6)对该观测量进行剔除或降权处理后,重复 1)~ 6),若其中 2) $\hat{\sigma}_0^2$ 检验通过,则进行 7);

7)、8)同 § 6.1

比较这两种方法,前者适用于观测量间相关性较弱的情况下的多粗差探测,而后者适用于观测量之间相关性较强时的粗差探测。

7 实践和算例

高精度 GPS网的基线处理,大都采用网解方式,一个同步网的基线解的协方差阵是满阵,即在一个同步网中所有基线向量之间及基线向量各分量之间都是相关的,而整个 GPS网可由若干个同步网组成。

7.1 模拟粗差

图 3是一实测高精度 GPS网,由 4个同步网组成,每个同步网含两条基线向量。在同步网 B9502601.LCX 中基线 TN08-TN09 的 ΔX 和 ΔY 分量上,加入 3倍观测量平差值中误差的粗差,则粗差分析中得出表 1。

若简单地从观测量改正数 V 及其精度入手,则难以发现粗差所在位置。根据(13)式,对 $d_{F_i,V}$ 进行 t_{n-2} 检验,这时 $n=22$,取 $\alpha=0.001$ 时,检验量为 0.628。不难发现:$d_{F_4,V}=0.849>0.628$,$d_{F_5,V}=0.691>0.628$,正是粗差所在的观测量。

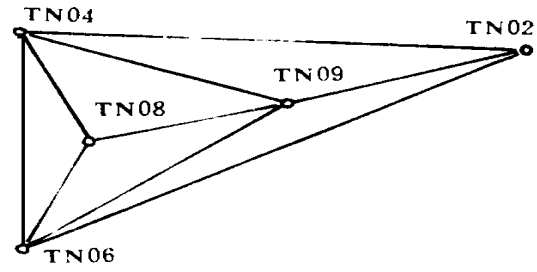


图 3 GPS网图

Fig. 3 GPS Network Plot

表 1 粗差分析结果

Tab. 1 Results of Outlier Analysis

序号	测站	观测量	V/m	精度/m	d	同步网
1	TN08	DX	-0.004 4	±.010 7	-.378	
	TN04	DY	0.003 1	±.015 4	-.273	
		DZ	0.001 6	±.009 1	-.208	B9502601.LCX
2	TN08#	DX	0.017 3	±.012 1	.849	
	TN09#	DY	0.019 0	±.030 0	.691	
		DZ	-0.000 4	±.009 8	.190	B9502601.LCX
3	TN02	DX	0.012 1	±.010 6	.378	
	TN04	DY	-0.001 7	±.021 8	.273	
		DZ	-0.005 2	±.012 0	.208	B9502701.LCX
4	TN02	DX	-0.002 6	±.014 9	-.104	
	TN06	DY	0.001 8	±.028 7	-.080	
		DZ	-0.000 4	±.015 2	-.199	B9502701.LCX
5	TN09	DX	0.019 1	±.013 8	.266	
	TN02	DY	0.036 1	±.034 3	.155	
		DZ	0.008 6	±.017 6	-.004	B9502602.LCX
6	TN09	DX	0.019 0	±.015 2	.332	
	TN06	DY	0.036 8	±.036 6	.365	
		DZ	0.009 6	±.018 5	-.043	B9502702.LCX
7	TN08	DX	-0.016 5	±.021 8	-.254	
	TN06	DY	0.003 8	±.035 7	-.363	
		DZ	0.010 1	±.032 3	.230	B9502703.LCX
8	TN08	DX	-0.035 0	±.023 7	-.529	
	TN09	DY	-0.028 8	±.058 0	-.145	
		DZ	-0.002 7	±.046 9	-.257	B9502703.LCX

7.2 实践算例

图 4是国家高精度 GPS B 级网某测区中,任意提取的一个子块。

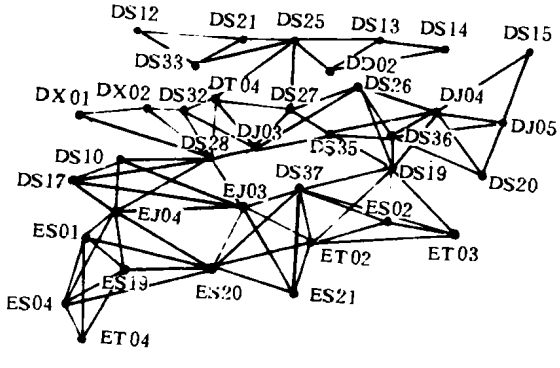
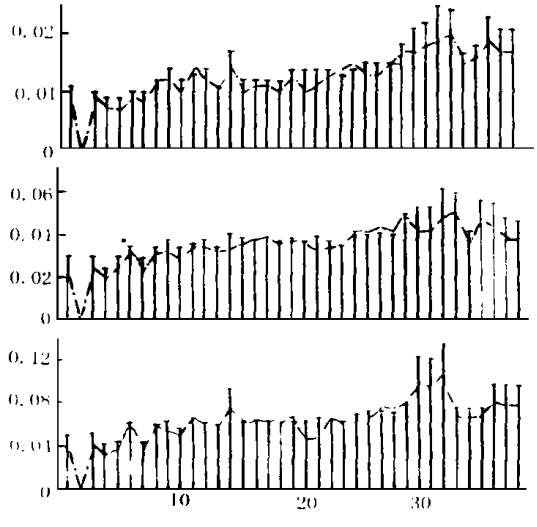


图 4 GPS网图

Fig. 4 GPS Network Plot

对原始基线观测量进行最小二乘无约束平差,其验后单位权中误差 $\hat{\sigma}_0^2$= 1.374, χ^2 检验不通过,说明平差系统中含有小的粗差。按§ 6.2节逐个探测的方法进行粗差分析,最后确认 3个同步网中含有粗差基线。对其进行简单的剔除后,重新平差,则 $\hat{\sigma}_0^2$= 1.076, χ^2 检验通过。

图 5是粗差处理前后平差结果的坐标分量精度比较。



T和曲线分别表示粗差处理前后的坐标分量精度

图 5 粗差处理前后坐标分量 rms比较

Fig. 5 Comparisons of Component rms

8 结 论

1)观测量误差对改正数向量的影响向量 F_i 由平差系统的图形设计矩阵和观测量权矩阵共同决定,它反映了观测量的误差对改正数向量的内在的影响关系和作用程度。

2)无论将粗差归于函数模型的误差还是随机模型的误差,都将在‖ ∇F_i ‖ $_2$ 上有显著的反映。

3)‖ F_i ‖ $_2$ 的大小反映了该观测量粗差的可测性程度。

4)观测量之间的相关系数 ρ_{F_i,F_j} 反映了两个观测量粗差的可区分性程度。

5)若观测量中含有粗差,则其对应的 F_i 与 V 将表现为较强的相关性,这时 $\rho_{F_i,V}$ 的 $t_{(n-2)}$ 检验显著。

本文所涉及的理论、算法和实践,由武汉测绘科技大学研制的 GPSNAS软件实现。

参 考 文 献

1 李德仁.误差处理和可靠性理论.北京: 测绘出版社, 1988.
2 崔希璋,刘经南.国家高精度 GPS网数据处理个别问题探讨.武汉测绘科技大学学报, 1994, 19(增刊)
3 胡定国,张润楚.多元数据分析方法.天津: 南开大学出版社, 1990.
4 方开泰,张尧庭.广义多元分析.北京: 科学出版社, 1993.
5 熊西文等译.数值分析的理论及其应用.上海: 上海科技出版社, 1980.
6 孙继广.矩阵扰动分析.北京: 科学出版社, 1987.
7 施　闯.国家高精度 GPS网数据处理方案、模型和软件研究: [学位论文].武汉: 武汉测绘科技大学, 1995
8 程云鹏.矩阵论.西安: 西北工业大学出版社, 1989.

Correspondence Based Outlier Analysis

Shi Chuang　Liu Jingnan

(School of Geo-science and Surveying Engineering, WTUSM, 129 Luoyu Road, Wuhan, China, 430079)

Abstract The specialty of reliability for correlative observables is studied. The residual vector can be expressed as a linearization equation of observable's errors and column components of the reliablity matrix (the influence vectors). The correspondence coefficient of the influence vector and the residual vector is used to study on the condition that the observable's error is outlier. In this paper, the theory of correspondence analysis is proposed and discussed. This theory is used to solve the problem of mult-outlier analysis in correlative observables.

Key words correspondence analysis; correlative observable; outlier analysis; error discussing