DOI:10.13203/j.whugis1987.04.008

第12卷 第4期 武汉测绘科技大学学报 Vol.12 No.4
1987年 12月 Journal of Wuhan Technical University of Surveying and Mapping Dec. 1987

岭估计及其应用

黄 幼 才

摘 要

最小二乘估计是最小无偏估计。平差后，观测值的残差平方和为最小。在法方程系数矩阵A^TA接近于单位矩阵（矩阵状态良好）的情况下，未知参数的最小二乘估值是可靠的。但是，在法方程系数阵处于病态的情况下，观测值残差平方和最小并不能保证未知参数估值的方差也小。岭估计能缩短未知参数估值与其真值之间的距离，即减少未知参数估值的方差。

【关键词】 岭估计；偏度；方差；可靠性

一、前 言

当法方程系数矩阵A^TA偏离单位矩阵很大时，未知参数的最小二乘估值会对很多误差产生敏感。在实际工作中，如物理、化学、测量等，人们事先可能对未知数有一定的了解。当法方程系数矩阵状态不佳的情况下，用最小二乘法求得平差参数的估值往往很不理想。HOERL于1962年首次提出了用岭估计的方法来解决这类问题。岭估计的基本思想是利用原来最小二乘估计的数学模型，在其法方程系数矩阵A^TA的对角线上加上一个很小的正数。这个很小的正数能使法方程系数矩阵的正交性大大地加强，减少了其列向量之间的相关程度，使未知参数的估值的精度得到很大的改善。由于法方程系数矩阵的对角线上加了一个小正数，所以平差后观测值的残差平方和不是最小。因此，岭估计又被称之为有偏估计。

二、岭估计数学模型

设观测方程的数学模型为

$$L = AX + \varepsilon \tag{1}$$

L是观测值向量，A是观测方程系数矩阵，X是未知参数向量，ε是服从于正态分布的误差向量，它的期望值为零，方差为$\sigma^2 I$。

本文1987年4月收到.

未知参数X的最小二乘估计为

$$\hat{X}=(A^TA)^{-1}A^TL \tag{2}$$

平差后，观测值残差的平方和为

$$\phi(\hat{X})=(L-A\hat{X})^T(L-A\hat{X}) \tag{3}$$

未知参数估值$\hat{X}$的方差是

$$Var(\hat{X})=\sigma^2(A^TA)^{-1} \tag{4}$$

$\hat{X}$与其真值X之间的欧氏距离可以写成下式

$$S^2=(\hat{X}-X)^T(\hat{X}-X)$$

S^2的期望值为

$$E(S^2)=\sigma^2 trace(A^TA)^{-1} \tag{5}$$

未知参数最小二乘估计的欧氏长度为

$$E(\hat{X}^T\hat{X})=X^TX+\sigma^2 trace(A^TA)^{-1} \tag{6}$$

其中 trace 是迹的意思。（6）式表明未知参数的最小二乘估值的长度与其真值的长度的差别是与其协方差矩阵 $(A^TA)^{-1}$ 的迹有关。迹越大，未知参数的估值偏离其真值的程度越大。距离S^2的方差为

$$Var(S^2)=2\sigma^4 trace(A^TA)^{-2} \tag{7}$$

由于法方程系数矩阵A^TA是实对称矩阵，对矩阵A^TA进行正交变换不会改变其迹的大小。由此，距离S^2可以表示成矩阵A^TA的特征值的函数

$$E(S^2)=\sigma^2\sum_{i=1}^{p}(1/\lambda_i) \tag{8}$$

p 是未知参数的个数，λ_i 是矩阵 A^TA 的特征值。比较（5）式与（8）式可以看出，矩阵 $(A^TA)^{-1}$的迹的大小是与特征值 λ_i 大小成反比。同样，距离S^2的方差可以表示为

$$Var(S^2)=2\sigma^4\sum(1/\lambda_i)^2 \tag{9}$$

（8）式表明未知参数估值$\hat{X}$偏离其真值的程度取决于法方程系数矩阵的特征值大小。当法方程系数矩阵处于病态，也就是说某些特征值很小的情况下，用最小二乘估计的方法是不可能得到未知参数的精确估值。加大法方程系数矩阵的特征值或减小协方差矩阵的迹是改善平差参数估值的精度的关键。HOERL于1962年首次提出用岭估计来解决这类问题。岭估计一般数学表达式可以写成

$$\hat{X}^*=[A^TA+KI]^{-1}A^TL=WA^TL \tag{10}$$

K是大于零的小数，$W=[A^TA+KI]^{-1}$。

下面推导未知参数X的岭估计$\hat{X}^*$与最小二乘估计$\hat{X}$的关系式

$$\begin{aligned}\hat{X}^*&=[A^TA+KI]^{-1}A^TL=[A^TA+KI]^{-1}[(A^TA)^{-1}]^{-1}(A^TA)^{-1}A^TL\\&=[I_p+K(A^TA)^{-1}]^{-1}\hat{X}=Z\hat{X}\end{aligned} \tag{11}$$

（11）式表明未知参数的岭估计是其最小二乘估计的函数，其系数矩阵为Z。由于矩阵W和Z是基于法方程系数矩阵A^TA，故W和Z也是正定实对称矩阵。通过正交变换很容易求出矩阵W和Z的特征值

$$\lambda_i(W)=1/(\lambda_i+K) \tag{12}$$

$$\lambda_i(Z)=\lambda_i/(\lambda_i+K) \tag{13}$$

当$K>0$时，$\lambda_{i(W)}<1/\lambda_i$。因此，岭估计法方程系数矩阵的特征值大于原来最小二乘估计法方程系数阵的特征值。下面讨论未知参数的岭估计的欧氏长度。由（11）式知

$$Z=(A^TA+KI)^{-1}A^TA \tag{14}$$

设法方程系数矩阵A^TA的全部特征值都不等于零，用一正交矩阵G对（11）式中$Z\hat{X}$进行正交变换，使矩阵Z^TZ变为对角矩阵

$$\hat{X}^*=(ZG)(G^T\hat{X}) \tag{15}$$

此时，未知参数的岭估计的长度为

$$(\hat{X}^*)^T(\hat{X}^*)=[ZG)(G^T\hat{X})]^T[(ZG)(G^T\hat{X})]=\hat{X}^TG[G^TZ^TZG]G^T\hat{X}$$

由于$GG^T=I$，则得

$$\sum_{i=1}^{P}(\hat{X}_i^*)^2=\sum_{i=1}^{P}\left(\frac{\lambda_i}{\lambda_i+K}\right)^2(\hat{X}_i)^2 \tag{16}$$

其中$(\hat{X}^*)^T=(\hat{x}_1^*,\ \hat{x}_2^*,\ \cdots,\ \hat{x}_P^*)$，$(\hat{X})^T=(\hat{x}_1,\ \hat{x}_2,\ \cdots,\ \hat{x}_P)$。

当$K>0$时，$\lambda_i/(\lambda_i+K)<1$，所以

$$(\hat{X}^*)^T(\hat{X}^*)<(\hat{X})^T(\hat{X}) \tag{17}$$

未知参数的岭估计与其真值之间的欧氏距离可以写成为

$$E(S_K^2)=\sigma^2\,\mathrm{trace}\,[A^TA+K]^{-1} \tag{18}$$

顾及（5）式，因$\mathrm{trace}[A^TA+KI]^{-1}<\mathrm{trace}[A^TA]^{-1}$，所以

$$E(S_K^2)<E(S^2) \tag{19}$$

（17）式和（19）式给出了一个很重要的概念，即当法方程系数矩阵A^TA偏离单位矩阵时，未知参数的岭估计长度小于其最小二乘估计的长度。由（6）式可知，$\sigma^2\,\mathrm{trace}(A^TA)^{-1}$总是大于零的正数，所以未知参数最小二乘估计的长度平均大于其真值的长度。（17）式和（19）式表明未知参数的岭估计优于其最小二乘估计。

三、岭估计数理統计上的含义

上面讨论了岭估计的数学模型并与传统的最小二乘估计作了比较。下面对岭估计进行更深入的分析。现设x_a为未知参数X的任意一估计。则平差后观测值残差的平方和为

$$\phi=(L-AXa)^T(L-AXa)=(L+A\hat{X}-A\hat{X}-AX_a)^T(L+A\hat{X}-A\hat{X}-AXa)$$

$$=[(L-A\hat{X})-A(Xa-\hat{X})]^T[(L-A\hat{X})-A(Xa-\hat{X})]$$

$$= (L - A\hat{X})^T (L - A\hat{X}) + (Xa - \hat{X})^T A^T A (Xa - \hat{X})$$

$$- 2(Xa - \hat{X})^T A^T (L - A\hat{X})$$

顾及（2）式，则上式中第三项中的 $A^T(L - A\hat{X})$ 为

$$A^T(L - A\hat{X}) = A^T L - A^T A\hat{X} = A^T L - A^T A (A^T A)^{-1} A^T L = 0$$

所以

$$\phi = (L - A\hat{X})^T (L - A\hat{X}) + (Xa - \hat{X})^T A^T A (Xa - \hat{X})$$

$$= \phi_{min} + \phi(Xa) \tag{20}$$

（20）式中，ϕ_{min} 是由最小二乘估计而得观测值残差的平方和。式中第二部分是由于用非最小二乘估计所带来的偏差。如果以未知数的最小二乘估计 $\hat{X}$ 为中心，则第二部分描述了一个多维的等值椭球面。在这个椭球面上所得任何未知参数的估值偏离 $\hat{X}$ 大小都相等。如果在这个椭球面上找到一个长度为最小的未知参数估值，则既可以提高未知参数估值的精度又可以把偏差限制在一定范围。这实际上是在满足条件

$$(Xa - \hat{X})^T A^T A (Xa - \hat{X}) = C$$

C 为一常数，求下列函数的极值

$$\min F = X_a^T Xa + (1/K)[(Xa - \hat{X})^T A^T A (Xa - \hat{X}) - C] \tag{21}$$

K 是拉格朗日乘常数。对（21）式微分

$$\frac{\partial F}{\partial Xa} = 2Xa + (1/K)[2(A^T A)Xa - 2(A^T A)\hat{X}] = 0 \tag{22}$$

解方程（22），得

$$Xa = \hat{X}^* = [A^T A + KI]^{-1} A^T L \tag{23}$$

（23）式就是岭估计的数学模型。用图 1 可以更清楚地描述岭估计的几何上的含义。

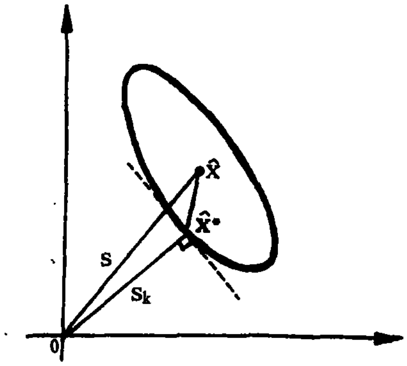

图 1　岭估计几何上的含义

S —— 最小二乘估计 $\hat{X}$ 的欧氏长度

S_K —— 岭估计 $\hat{X}^*$ 的欧氏长度

四、岭估计的方差和偏度

未知参数岭估计与其真值之间距离的期望值可以表示为

$$E(S_K^2) = E[(\hat{X}^* - X)^T (\hat{X}^* - X)]$$

因为 $\hat{X}^* = Z\hat{X}$，则

$$E(S_K^2) = E[(Z\hat{X} - ZX + ZX - X)^T (Z\hat{X} - ZX + ZX - X)]$$

$$= E\{[(Z(\hat{X} - X) + ZX - X)]^T [Z(\hat{X} - X) + ZX - X]\}$$

$$=E\{(\hat{X}-X)^T Z^T Z(\hat{X}-X)+(\hat{X}-X)^T Z^T(ZX-X)$$

$$+(ZX-X)^T[Z(\hat{X}-X)]+(ZX-X)^T(ZX-X)\}$$

因为 $E(\hat{X}-X)=0$，并顾及（5）式，则得

$$E(S_K^2)=\sigma^2\ \mathrm{trace}[(A^TA)^{-1}Z^TZ]+(ZX-X)^T(ZX-X)$$

$$=\sigma^2\mathrm{trace}[(A^TA)^{-1}Z^TZ]+X^T(Z-I)^T(Z-I)X \tag{24}$$

由（11）式知

$$Z=[I_p+K(A^TA)^{-1}]^{-1}$$

又因为

$$Z=[A^TA+KI]^{-1}(A^TA)=[A^TA+KI]^{-1}[(A^TA+KI)-KI]$$

$$=I-K(A^TA+KI)^{-1}$$

将上面Z的两个表达式按需分别代入（24）式的第一项和第二项，则得

$$E(S_K^2)=\sigma^2\ \mathrm{trace}\{(A^TA)^{-1}[I+K(A^TA)^{-1}]^{-1}\}^T$$

$$[I-K(A^TA+KI)^{-1}]+K^2X^T(A^TA+KI)^{-2}X$$

$$=\sigma^2\mathrm{trace}\{[(A^TA+KI)^{-1}]^T[I-K(A^TA+KI)^{-1}]\}$$

$$+K^2X^T(A^TA+KI)^{-2}X$$

$$=\sigma^2\mathrm{trace}(A^TA+KI)^{-1}-\sigma^2\mathrm{trace}K(A^TA+KI)^{-2}$$

$$+K^2X^T(A^TA+KI)^{-2}X$$

$$=\sigma^2\Sigma\, 1/(\lambda_i+K)-\sigma^2K\Sigma\, 1/(\lambda_i+K)^2$$

$$+K^2X^T(A^TA+KI)^{-2}X$$

$$=\sigma^2\sum_{i=1}^{p}\lambda_i/(\lambda_i+K)^2+K^2X^T(A^TA+KI)^{-2}X$$

$$=r_1(K)+r_2(K) \tag{25}$$

参照方程式（24），r_2（K）可以被看成用岭估计代替最小二乘估计所引起的未知参数的偏离度。很显然，当 $K=0$ 时，$Z=I$，则 $r_2(0)=0$，这种情况就是最小无偏的最小二乘估计。（25）式中第一项就是未知参数估值的方差总和，又称之为全方差。因为

$$\hat{X}^*=Z\hat{X}=Z(A^TA)^{-1}A^TL$$

则$\hat{X}^*$的方差为

$$\mathrm{Var}(\hat{X}^*)=Z(A^TA)^{-1}A^T\mathrm{Var}(L)A(A^TA)^{-1}Z^T$$

$$=\sigma^2Z(A^TA)^{-1}Z^T \tag{26}$$

故$\hat{X}^*$的方差总和也就是（25）式中$Z(A^TA)^{-1}Z^T$对角线上元素的总和。因此，未知参数岭估计与其真值之间的距离由两部分组成；1、岭估计值的全方差，2、引进岭估计所产生的偏度。

五、K 值的选择

从（25）式中可以看出，当K 增加时，未知参数岭估计方差减少，而其偏度增加。为了选择较好的K，需要研究 $r_1(K)$ 和 $r_2(K)$ 随K变化的速率。现以K为变量对 $r_1(K)$ 和$r_2(K)$ 求导并取极限

$$\lim_{K\to 0^+}(dr_1(K)/dK) = -2\sigma^2\Sigma(1/\lambda_i^2) \tag{27}$$

$$\lim_{K\to 0^+}(dr_2(K)/dK) = 0 \tag{28}$$

由（25）式可知，$r_1(K)$ 是K的递减函数。$r_2(K)$ 是K的递增函数。分析（27）和(28）式，当法方程系数矩阵A^TA 处于病态的情况下，矩阵 A^TA 某些特征值很小，从（27）式可知，$r_1(K)$ 的变化很快，也就是全方差变化很快。在相同的情况下，当K值很小时，$r_2(K)$ 的变化是很缓慢的，因为曲线$r_2(K)$ 是平坦的，它的斜率趋近于零。由于全方差和偏度随K（当K很小时）变化的速率有明显的不同，这就为我们选择一个最佳K 提供了条件。我们可以选择一个很小的K使其偏度没有显著地增加，但又大大地减少未知参数估计的全方差。这个最佳K 值一般可以用作图的方法来获得。这部分内容将在算例中详细讨论。

六、算例分析

在摄影测量中，控制点的分布、几何结构、数学模型以及未知参数的选择都会影响平差中法方程系数矩阵的结构。例如，当被摄地区是平坦的，如果用直接线性变换的方法就不太可能获得较为理想的精度。因为控制点高程上相差不大会导致直接线性变换数学模型中某些未知参数不同程度上的线性相关，因而法方程式系数矩阵的状态不会很好。同样，这个问题会出现在空间后方交会之中。如果象平面平行于摄影地区，而选择摄影机主距为未知参数，这时未知参数主距与摄影站坐标中高程有较强的相关。在求未知参数的最小二乘估值时就会出现主距和摄影站的高程的平差值对控制点的分布和误差很敏感。现以空间后交为例用计算机模拟一套实验数据。这套数据设计的基本思想如下，根据控制点布点情况分成二组，在第一组中，控制点之间高程最大差别为摄影距离的60%。在第二组中，控制点之间最大高差为摄影距离的 0.1% 。然后用严格的共线方程算出各个控制点所对应的象点坐标。假定象点坐标的中误差，利用现有的计算机软件算出象点坐标误差的随机量并分别加到各点坐标之中。这样组成了两种不同情况的观测数据。在空间后交中，选择七个未知参数，即摄影站坐标的改正数Δx，Δy，Δz，旋转角的改正数$\Delta\omega$，$\Delta\varphi$，$\Delta\kappa$，主距改正数Δf。由于象点坐标的误差很小，并且选定这七个待求值的原设计值为近似值，所以平差后这七个未知参数的最小二乘估值应是接近于零的很小数值。这也是衡量平差结果是否可靠的一个标准。下面列出二组观测数据以及用最小二乘法算得七个参数的估值。

表 1　第一组　控制点坐标

点　号	控制点坐标（米）			象点坐标（毫米）	
	X	Y	Z	x	y
1	-50	50	70	-89.2857	89.2857
2	50	50	130	48.0769	48.0769
3	-50	-50	100	-62.5	-62.5
4	50	-50	90	69.4444	-69.4444

表 2　第二组　控制点坐标

点　号	控制点坐标（米）			象点坐标（毫米）	
	X	Y	Z	x	y
1	-50	50	99.9	-62.5626	62.5626
2	50	50	100.1	62.4376	62.4376
3	-50	-50	100.	-62.5	-62.5
4	50	-50	99.95	62.5313	-62.5313

初始值：主距 $f=125$ 毫米

摄影站坐标　$X=0$，$Y=0$，$Z=0$

旋转角　$\omega=0$，$\varphi=0$，$\kappa=0$

表 3　未知参数的最小二乘估值

未知参数	Δx(米)	Δy(米)	Δz(米)	Δφ(度)	Δω(度)	Δκ(度)	Δf(毫米)
第一组	4.1×10^{-7}	-6×10^{-8}	-2.9×10^{-6}	0	0	0	0.0998
第二组	0.026	-0.027	-16.032	-0.0003	0.0002	0.0001	20.065

表 4　第二组法方程系数矩阵特征值

特征值	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
	0.2418	0.2432	5.1350	31256.5	101536.7	101653.2	0.000001

表 3 列出了第一组和第二组未知参数的最小二乘估值。由于第一组的法方程系数矩阵状态良好，所以未知参数的估值都很小。也就是说接近其真值。第二组的法方程系数矩阵趋于病态。因此其未知参数的估值都比较大，特别是未知参数Δz 和Δf 。这就说明了Δz 和Δf 有较强的相关性。法方程系数矩阵状态的好坏可由表 4 中列出的特征值大小来判定 。在 第 二 组中，最小特征值为0.000001。由此可以断定第二组的法方程系数矩阵处于病态。下面讨论如何用岭估计来解决第二组的病态问题。式（25）表明未知参数的岭估计与其真值之间的距离由两部分组成。即全方差和偏度。同时考虑这两部分因素就存在一个优选K 值的问题。选择K 值比较简单而有效的方法是图解法。下面以第二组数据为例，根据方 程（25）分 别 算 出 $r_1(K)$ 、$r_2(K)$ 、均方差、偏度以及迹。

表 5　岭估计有关参数表

K	均　方　差	r_1 (K)	r_2 (K)	迹
0	0.00281169	2811.7136	0	234424
0.001	0.00312563	0.02635215	0.00606721	1008
0.002	0.00312579	0.0246034	0.00607010	508.323
0.003	0.00312586	0.02348466	0.00607107	341.688
0.004	0.00312592	0.02316469	0.00607155	258.144
0.005	0.00312597	0.02292028	0.00607184	208.322
0.006	0.00312603	0.02270692	0.00607204	174.961
0.007	0.00312609	0.02250934	0.00607218	151.122
0.008	0.00312616	0.02232114	0.00607228	133.235
0.009	0.00312624	0.02213934	0.00607236	119.315
0.01	0.00312632	0.02196201	0.00607243	108.173
0.02	0.00312749	0.02034029	0.00607276	57.8701

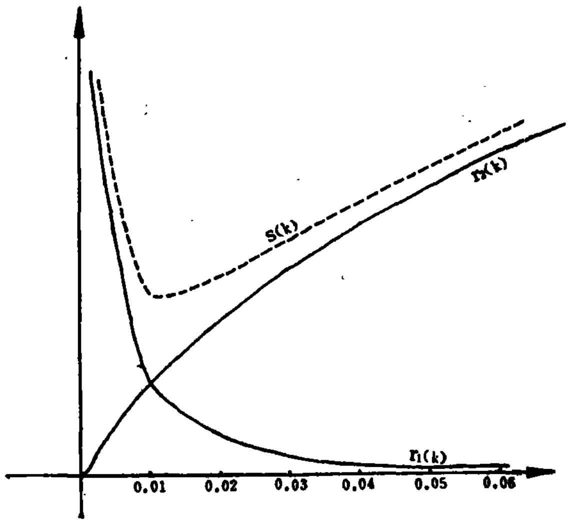

图 2　岭估计示意图

图 2 表明 $r_1(K)$ 是K值递降函数，$r_2(K)$ 是K值递增函数。图中虚线是 $r_1(K)$ 和 $r_2(K)$ 之和。很明显 K 值应选在 $r_1(K)$ 和 $r_2(K)$ 的交点处（$K=0.008$）。这样既可大大地减少全方差又可以使偏度不致于很大。表 5 表明在 $K=0.008$ 处，$r_1(K)$ 从 2811.7136 减少到 0.02232114。而 $r_2(K)$ 仅仅增加了0.00607228。另外，可从表 5 中看出很重要的一点是观测值的均方差仅增加0.00031447。

七、小　　結

岭估计自从1962 年 Hoerl 提出以来在应用数学领域里得到较快的发展并逐步应用到其他学科中去。岭估计是基于最小二乘估计，解决由于法方程系数矩阵状态不佳而导致未知参数估值不可靠的问题。本文所讨论的岭估计的数学模型较为简单。它是在法方程系数矩阵对角线上都加上同样大小的K 值。这种方法虽然解决了由于法方程系数矩阵状态不佳导致估值不可靠的主要因素，但或多或少对相关性不强的其他未知参数有所影响（有利或不利的影响）。但这影响是很小的，一般能得到满意的结果。因此这种方法在实际中得以广泛应用。人们现在探索用解析的方法求对不同的未知参数采取不同 K 的方法。现举一例简单地说明这种方法。

设有一正交矩阵G，通过对法方程系数矩阵 A^TA 进行正交变换可得

$$A^TA = G^T \wedge G \tag{29}$$

其中 $\wedge = (\delta_{ij}\lambda_i)$ 是 A^TA 矩阵的特征矩阵。由此设

$$A = A^*G \tag{30}$$

故原来的观测方程可以写成

$$L = A^*\alpha + \varepsilon \tag{31}$$

式中 $\alpha = GX$，$(A^*)^T(A^*) = \wedge$，$\alpha^T\alpha = X^TX$。

求 α 的岭估计

$$\alpha^* = [(A^*)^T(A^*) + KI]^{-1}(A^*)^TL \tag{32}$$

其中 $K = (\delta_{ij}K_i)$，$K_i \geqslant 0$。

求欧氏距离 $S_\alpha^2 = (\hat{\alpha}^* - \alpha)^T(\hat{\alpha}^* - \alpha)$ 极小可得最优K值（$K_i = \sigma^2/\alpha_i^2$）。因为 σ 和 α_i 一般未知，所以先取 $\hat{K}_i = \hat{\sigma}^2/\hat{\alpha}_i^2$，然后用迭代方法求得 K_i 最后值。

参 考 文 献

〔1〕Baerjer, K.S. and Carr, R.N., A Comment on Ridge Regression. Biased Estimation for Non-orthogonal Problems, Technometrics, 13, 895-898, 1971.

〔2〕Hoerl, A.E. and Kennard, R.W., Ridge Regression, Biased Esimation for Non-orthogonal Problem, Techometrics, 12, 55-67, 69-82, 1970.

〔3〕Marquardt, D. W., Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation, Technomertics, 12, 591-611, 1970.

〔4〕Mayer, L.S. and Willke, T. A., On Biased Estimation in Linear Models, Technometrics, 15, 497-508, 1973.

[5] Theobald, C.M., Generalization of Mean Square Error Applied to Ridge Regression, J.R. Stat. Soc., B, 36, 103-106, 1974.
[6] 王之卓，摄影测量原理，测绘出版社，1979.
[7] 杨簏引、马正午、孙宇，电子计算机应用数学，冶金工业出版社，1979.

The Ridge Estimation and Its Application

Huang Youcai

Abstract

The least squares estimator is a minimum and unbiased estimator. The parameter estimates are based on the minimum sum of residual spuares. The parameter estimates may be reliable if the coefficient matrix A^TA is nearly a unit matrix. However, minimizing the sum of residual squares may not guarrantee that the variances of the parameter estimates are small if the matrix A^TA is ill-conditioned. The distances between parameter estimates and their true values may be reduced by using the ridge estimator. In other words, the ridge estimator may be used to improve the accuracies with regard to the parameters of imprecised estimates.

[Key words] Ridge estimation; bias; variance; reliability