[文章编号] 1004- 0609(2000)03- 0455- 04

# 附加参数与基本参数空间关系的度量①

卢秀山，宋淑丽，韩晓冬
(山东科技大学 地球科学系，泰安 271019)

[摘 要] 合理地选择参数，避免参数间的强复共线性，是参数有效估计的关键之一。通过将参数 $\boldsymbol{X}$ 分为两类，即基本参数 $\boldsymbol{X}_1$ 和附加参数 $\boldsymbol{X}_2$，并假定附加参数的选取不当是造成参数间强复共线性的主要原因，提出了用点($\boldsymbol{X}_2$ 中的一列向量)到空间(由 $\boldsymbol{X}_1$ 所对应的列向量张成 Hilbert 空间的子空间 $\boldsymbol{H}_0$)夹角的正弦值作为度量标准来优选附加参数，给出了对该方法的理论论证，并用两个算例对度量方法进行了验证和说明。

[关键词] 参数估计；复共线性；度量；Hilbert 空间
[中图分类号] O 177.1 [文献标识码] A

用参数来描述某一物理量时，要求参数间线性独立；为了客观地刻画该物理量，需要选择足够多的参数。存在的问题是，参数之间存在强复共线性时，会严重地影响参数的最小二乘估值质量，并且所选择的参数愈多，这种复共线性愈强。因此，需要剔除那些作用不大的参数，以保证参数估值的质量。在所选择的参数中，一些是描述物理量的基本参数，一些是描述物理量的附加参数(备选参数)。由于对这些附加参数之间、附加参数与基本参数之间的关系缺乏足够的认识，附加参数的选取不当是造成参数间强复共线性的主要原因[1~5]。

文献[6]就附加参数的统计检验和优选作了较为详细的论述，检验的方法有：基于残差平方和的检验方法、基于附加参数的检验方法；在附加参数的比较和优选方面，有残差平方和准则及赤池信息量(AIC)准则。同时，文献[6]也指出其相应的缺点，即统计检验带有主观的倾向、AIC 准则要求模型的分布类型必须是已知的。文献[7]给出了一种比较附加参数优劣的方法，但是该方法没有避免基本参数可能具有某种复共线性的干扰。本文提出的优选附加参数的方法，是通过度量向量($\boldsymbol{x}_i$ 所对应的列向量，$\boldsymbol{x}_i \in \boldsymbol{X}_2$)与空间(由 $\boldsymbol{X}_1$ 对应的列向量所张成的空间)夹角的正弦值而决定附加参数去留的。

本文根据 Hilbert 空间(H 空间)中的最佳逼近理论[8, 9]，首先分析了附加参数所对应的向量(H 空间中的点)与基本参数所对应的列向量张成的空间(H 空间的子空间)之间的关系。分析表明，点到子空间的距离大小反映了附加参数与基本参数间复共线性强弱。该距离与相应点的范数之比，即为 H 空间中的点(向量)与 H 的子空间(由矩阵的列向量张成)夹角的正弦值。用这一正弦值作为附加参数与基本参数间复共线性强弱的度量指标，几何概念明确，附加参数间排队方便，避免了上述的一些缺点，且具有较强的可操作性。

## 1 向量 α 到空间 $\boldsymbol{H}_0$ 的距离

### 1.1 向量 α 与空间 $\boldsymbol{H}_A$

将参数 $\boldsymbol{X}$ 分为两类，基本参数 $\boldsymbol{X}_1$ 和附加参数 $\boldsymbol{X}_2$。为了研究 $\boldsymbol{X}_2$ 与 $\boldsymbol{X}_1$ 之间复共线性强弱，需要分析参数所对应的向量及其向量空间。

$\boldsymbol{X}$ 所对应的系数阵为 $\boldsymbol{A}$，且 $\boldsymbol{A}=(\alpha_1, \alpha_2, \cdots \alpha_n)$，其中 $\alpha_i$ 为 $m$ 维列向量，与 $\boldsymbol{x}_i(\boldsymbol{x}_i \in \boldsymbol{X})$ 对应。由 $\boldsymbol{A}$ 的列向量 α 张成向量空间 $\boldsymbol{H}_A$，定义内积为

$$\forall \alpha_i, \alpha_j \in \boldsymbol{H}_A:\ (\alpha_i, \alpha_j)=\alpha_i^{\mathrm{T}}\alpha_j \in \mathbf{R}$$

由内积导出的范数为

$$\forall \alpha, \in \boldsymbol{H}_A:\ \|\alpha\|=\sqrt{(\alpha, \alpha)} \in \mathbf{R}$$

定义距离 $d$ 为

$$\forall \alpha_i, \alpha_j \in \boldsymbol{H}_A:\ d(\alpha_i, \alpha_j)=\|\alpha_i-\alpha_j\|=\sqrt{(\alpha_i-\alpha_j)^{\mathrm{T}}(\alpha_i-\alpha_j)} \in \mathbf{R}$$

式中 $\boldsymbol{H}_A$ 是有限维内积空间，因而完备，故 $\boldsymbol{H}_A$ 是 $n$ 维 H 空间。$\alpha_i$ 是 H 空间中的一点。

将 $\boldsymbol{A}$ 分组，$\boldsymbol{A}=(\boldsymbol{A}_1 \vdots \boldsymbol{A}_2)$，其中，$\boldsymbol{A}_1=(\alpha_1, \alpha_2, \cdots, \alpha_q)$，$\boldsymbol{A}_2=(\alpha_{q+1}, \alpha_{q+2}, \cdots, \alpha_n)$。由 $\boldsymbol{A}_1$ 的列向量张成 H 空间的子空间 $\boldsymbol{H}_0(\boldsymbol{H}_0 \subset \boldsymbol{H}_A)$。由

① [收稿日期] 1999- 06- 02；[修订日期] 2000- 03- 20 [作者简介] 卢秀山(1961- )，男，博士，教授.

于 $\boldsymbol{A}_2$ 中的任一 α 是 H 空间中的一点，故可以度量 α 到 $\boldsymbol{H}_0$ 的距离。

**1.2 α 到 $\boldsymbol{H}_0$ 的距离**

对于任一附加参数 $\boldsymbol{x}_i(\boldsymbol{x}_i \in \boldsymbol{X}_2)$，其对应的列向量 $\alpha_i$ 为 $\boldsymbol{H}_0$ 外的一点，根据文献[8]的定理 2.2.4，可以度量 $\alpha_i$ 到 $\boldsymbol{H}_0$ 的距离 $D$，且

$$D^2=\frac{|\boldsymbol{G}(\boldsymbol{e}_1,\ \boldsymbol{e}_2,\ \cdots,\ \boldsymbol{e}_q,\ \alpha_i)|}{|\boldsymbol{G}(\boldsymbol{e}_1,\ \boldsymbol{e}_2,\ \cdots,\ \boldsymbol{e}_q)|} \tag{1}$$

式中

$$|\boldsymbol{G}(\boldsymbol{e}_1,\ \boldsymbol{e}_2,\ \cdots,\ \boldsymbol{e}_q)|=1$$

$$|\boldsymbol{G}(\boldsymbol{e}_1,\ \boldsymbol{e}_2,\ \cdots,\ \boldsymbol{e}_q,\ \alpha_i)|=\begin{vmatrix} 1 & \cdots & & (\alpha_i,\ \boldsymbol{e}_1) \\ \vdots & \ddots & & \vdots \\ & & 1 & (\alpha_i,\ \boldsymbol{e}_q) \\ (\boldsymbol{e}_1,\ \alpha_i) & \cdots & (\boldsymbol{e}_q,\ \alpha_i) & (\alpha_i,\ \alpha_i) \end{vmatrix}=$$

$$\|\alpha_i\|^2-\sum_{j=1}^{q}|\boldsymbol{c}_j|^2$$

由于 $\boldsymbol{c}_j=(\alpha_i,\ \boldsymbol{e}_j)$，因此有

$$\begin{aligned} D^2 &= \|\alpha_i\|^2-\sum_{j=1}^{q}|\boldsymbol{c}_j|^2 \\ &= \|\alpha_i\|^2-\sum_{j=1}^{q}|(\alpha_i,\ \boldsymbol{e}_j)|^2 \end{aligned} \tag{2}$$

$\boldsymbol{H}_0$ 的标准正交基的确定过程为：按照 Gram-Schmidt 正交化方法将 $\boldsymbol{A}_1$ 按列正交化，得与 $\boldsymbol{A}_1$ 相对应的正交化矩阵 $\boldsymbol{B}$，$\boldsymbol{B}=(\beta_1,\ \beta_2,\ \cdots,\ \beta_q)$，其中 $\beta_j$ 为 $m$ 维列向量。以 $\boldsymbol{B}$ 的列向量为 $\boldsymbol{H}_0$ 的一组正交基，则其标准正交基为 $(\boldsymbol{e}_1,\ \boldsymbol{e}_2,\ \cdots,\ \boldsymbol{e}_q)$，其中 $\boldsymbol{e}_j=\beta_j/\|\beta_j\|$。

H 空间内的一点，到不包含这一点的 H 子空间的距离，度量了该点与相应子空间的空间关系。因此，可以基于 α 到 $\boldsymbol{H}_0$ 的距离，进行参数间复共线性度量指标的讨论。

## 2 参数的空间关系度量

**2.1 参数间的复共线性与距离的关系**

式(2)右侧的第一项，是点(向量) $\alpha_i$ 范数的平方；求和号中的任一项，是 $\alpha_i$ 在对应正交基上投影的平方。投影值的大小取决于 $\alpha_i$ 的范数及其与对应正交基的夹角。当 $\alpha_i$ 与对应正交基正交时，投影值为零；当 $\alpha_i$ 与对应的正交基平行时，投影值等于 $\alpha_i$ 的范数。一般地，$\alpha_i$ 与对应的正交基既不平行也不正交，投影值小于 $\alpha_i$ 的范数，并且投影值的大小与夹角的大小作反向变化。因此，求和号中的每一项，反映了 $\alpha_i$ 与正交基的关系。求和项的整体，反映了 $\alpha_i$ 与子空间的关系。当求和项为零时，表明 $\alpha_i$ 与子空间正交；当求和项等于 $\alpha_i$ 的范数时，表明 $\alpha_i$ 与子空间中的某一正交基平行，或者 $\alpha_i$ 与子空间平行。

换句话说，求和项为零表明附加参数与基本参数独立；求和项等于 $\alpha_i$ 的范数，表明附加参数与某一基本参数共线，或者与基本参数复共线。一般地，求和项既不等于 $\alpha_i$ 的范数，也不为零。但是，求和项的值愈接近于 $\alpha_i$ 的范数，表明附加参数与基本参数间的复共线性愈强。

因此，点 $\alpha_i$ 到子空间的距离大小反映了附加参数与基本参数间复共线性强弱，距离的大小与复共线性强弱呈反向变化。

**2.2 参数空间关系的度量**

点 $\alpha_i$ 到子空间的距离，既与参数间的复共线性有关，也与 $\alpha_i$ 范数的大小成正比。然而，距离与对应的范数之比则是一个相对量。从几何意义上看，若 $\tilde{\alpha}_i$ 为 $\alpha_i$ 在 $\boldsymbol{H}_0$ 上的投影，则

$$D=\|\alpha_i-\tilde{\alpha}_i\|$$

且，当 $\tilde{\alpha}_i=0$ 时，$D=\|\alpha_i\|$；当 $\tilde{\alpha}_i=\alpha_i$ 时，$D=0$。因此，定义

$$\sin\theta_i=\frac{D}{\|\alpha_i\|} \tag{3}$$

即 H 空间中的点 $\alpha_i$(向量)与 H 的子空间 $\boldsymbol{H}_0$(由矩阵的列向量张成)夹角的正弦值作为衡量附加参数与基本参数间复共线性强弱的指标。这一指标具有明确的几何意义。

式(3)右侧的分子、分母，分别是附加参数所对应向量到基本参数向量空间的距离和向量的范数。正弦值的大小，反映了附加参数与基本参数间复共线性的强弱。当附加参数 $\boldsymbol{x}_i$ 与基本参数 $\boldsymbol{X}_1$ 完全独立时，向量 $\alpha_i$ 与空间 $\boldsymbol{H}_0$ 正交，$\sin\theta_i=1$；当附加参数 $\boldsymbol{x}_i$ 与基本参数 $\boldsymbol{X}_1$ 共线时，向量 $\alpha_i$ 与空间 $\boldsymbol{H}_0$ 平行，$\sin\theta_i=0$；因此有

$$0\leqslant\sin\theta_i\leqslant 1 \tag{4}$$

附加参数与基本参数复共线性强弱，与式(3)正弦值成反比。以正弦值为指标，比较附加参数相应正弦值的大小，进行附加参数排队，可以作为附加参数去留的主要依据。

## 3 算例

**3.1 算例 1**

本例原始数据引自文献[7]。其中，$\boldsymbol{X}_1(\boldsymbol{x}_1,\ \boldsymbol{x}_2,\ \boldsymbol{x}_3)$ 为基本参数，$\boldsymbol{x}_4$，$\boldsymbol{x}_5$，$\boldsymbol{x}_6$ 为附加参数，各参

数对应的向量如表 1 所示。为了避免系统病态，保证参数的估值质量，需要判断每一个附加参数与基本参数复共线性情况。

表 1　算例 1 参数对应的向量

**Table 1**　Vectors corresponding to parameters in Example 1[7]

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ |
|---|---|---|---|---|---|
| -1.0 | 2.0 | 5.0 | 0.5 | 0.8 | 0.6 |
| 2.0 | 2.0 | 4.0 | 0.2 | 0.4 | 1.2 |
| 3.0 | 1.0 | 4.0 | 0.8 | 1.4 | 1.4 |
| -5.0 | -1.0 | 2.0 | 0 | -0.5 | -1.0 |
| 4.0 | -1.0 | -2.0 | 0 | 0.4 | 0.8 |
| -1.0 | -1.0 | -2.0 | 0.4 | 0.3 | 0.2 |
| 1.0 | 3.0 | 1.0 | 0 | 0.1 | 0.2 |
| 5.0 | 2.0 | 3.0 | 0.1 | 0.6 | 1.1 |
| 2.0 | -3.0 | -1.0 | 0.5 | 0.7 | 0.9 |
| 3.0 | -3.0 | -2.0 | 0 | 0.3 | 0.6 |

应用本文提出的方法，分别计算附加参数与基本参数的空间夹角正弦值。从计算结果看，$x_4$ 与基本参数的空间夹角最大，$x_5$ 次之，$x_6$ 最小。如果需要去掉一个附加参数，应该去掉 $x_6$。计算数据如表 2 所示。

表 2　算例 1 中附加参数与基本参数的空间关系度量值

**Table 2**　Measure of space relationships between basic and additional parameters in Example 1

| | $X_1$ | | $X_2$ | $\sin\theta_4$ |
|---|---|---|---|---|
| $x_1$ | $x_2$ | $x_3$ | $x_4$ | 0.698 |
| | $X_1$ | | $X_2$ | $\sin\theta_5$ |
| $x_1$ | $x_2$ | $x_3$ | $x_5$ | 0.515 |
| | $X_1$ | | $X_2$ | $\sin\theta_6$ |
| $x_1$ | $x_2$ | $x_3$ | $x_6$ | 0.335 |

### 3.2　算例 2

本例原始数据引自文献[10]。在该参数估计问题中，有 5 个待估参数，系数矩阵严重“病态”，各参数对应的向量如表 3 所示。相关分析[10]表明(表 4)，$x_3$ 与 $x_1$ 具有很强的相关性，$x_4$ 与 $x_1$，$x_3$ 也具有较强的相关性。因此，是否可将 $x_3$，$x_4$ 全去掉就成了一个需要仔细考虑的问题。

事实上，$x_4$ 是不应去掉的。因为相关分析不能有效地说明一个参数与其它所有参数的复相关性(复共线性)。应用本文所提出的参数优选方法，可有效地指明需剔除的参数。将 $x_1$，$x_2$，$x_5$ 作为基本参数，应用本文提出的方法，分别计算得 $x_3$，$x_4$ 与基本参数的空间夹角的正弦值为 0.010 和 0.232。如果将 $x_3$ 作为附加参数，其余四个参数作为基本参数，则附加参数与基本参数的空间夹角正弦值为 0.008。显然，参数 $x_3$ 是必须要去掉的。计算数据见表 5。

表 3　算例 2 参数对应的向量

**Table 3**　Vectors corresponding to parameters in Example 2[10]

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|
| 15.57 | 2463 | 472.92 | 18.0 | 4.45 |
| 44.02 | 2048 | 1339.75 | 9.5 | 6.92 |
| 20.42 | 3940 | 620.25 | 12.8 | 4.28 |
| 18.74 | 6505 | 568.33 | 36.7 | 3.90 |
| 49.20 | 5723 | 1497.60 | 35.7 | 5.50 |
| 44.92 | 11520 | 1365.83 | 24.0 | 4.60 |
| 55.48 | 5779 | 1687.00 | 43.3 | 5.62 |
| 59.28 | 5969 | 1639.92 | 46.7 | 5.15 |
| 94.39 | 8461 | 2872.33 | 76.7 | 6.18 |
| 128.02 | 20106 | 3655.08 | 180.5 | 6.15 |
| 96.00 | 13313 | 2912.00 | 60.9 | 5.88 |
| 131.42 | 10771 | 3921.00 | 103.7 | 4.88 |
| 127.21 | 15543 | 3865.67 | 126.8 | 5.50 |
| 252.90 | 36194 | 7684.10 | 157.7 | 7.00 |
| 409.20 | 34703 | 12446.33 | 169.4 | 10.78 |
| 463.70 | 39204 | 14098.40 | 331.4 | 7.05 |
| 510.22 | 86533 | 15524.00 | 371.6 | 6.35 |

表 4　参数间相关系数表

**Table 4**　Correlation coefficients among parameters in Example 2

| Parameter | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|
| $x_1$ | 1.0000 | 0.9074 | 0.9999 | 0.9357 | 0.6712 |
| $x_2$ | | 1.0000 | 0.9072 | 0.9015 | 0.4466 |
| $x_3$ | | | 1.0000 | 0.9332 | 0.6711 |
| $x_4$ | | | | 1.0000 | 0.4629 |
| $x_5$ | | | | | 1.0000 |

表 5　算例 2 的附加参数与基本参数的空间关系度量值

**Table 5**　Measure of space relationships between basic and additional parameters in Example 2

| | $X_1$ | | | $X_2$ | $\sin\theta_3$ |
|---|---|---|---|---|---|
| $x_1$ | $x_2$ | $x_5$ | | $x_3$ | 0.010 |
| | $X_1$ | | | $X_2$ | $\sin\theta_4$ |
| $x_1$ | $x_2$ | $x_5$ | | $x_4$ | 0.335 |
| | $X_1$ | | | $X_2$ | $\sin\theta_3$ |
| $x_1$ | $x_2$ | $x_4$ | $x_5$ | $x_3$ | 0.008 |

## 4　结束语

(1) 本文提出的附加参数与基本参数空间关系

的度量方法，其宗旨是通过度量参数间的空间关系，优选附加参数，避免附加参数与基本参数存在强复共线性。

(2) 与其他优选附加参数方法相比，参数空间关系度量方法的显著优点在于可以度量附加参数向量与基本参数向量张成的空间的关系，即度量了一个参数与多个参数的关系。

(3) 本文所提出的度量指标具有明晰的几何意义。

(4) 参数空间关系度量指标是有界的，这是此指标的又一个显著特点。

## [REFERENCES]


[1] LU Xiu-shan(卢秀山). 病态系统分析理论及其在测量中的应用 [D]. Wuhan: Surveying and Geophysics Institute of The Chinese Academy of Sciences, 1999.

[2] ZHOU Jiang-wen(周江文), YANG. Yuan-xi(杨元喜), HUANG You-cai(黄幼才), et al. Robustified Least Squares Method(抗差最小二乘法) [M]. Wuhan: Huazhong University of Science and Technology Press, 1997: 272- 290.

[3] HU Qing-jun(胡庆军). 诊断复共线性的新方法 [J]. Mathematics in Practice and Theory(数学的实践与认识), 1995, (4): 55- 62.

[4] WANG Zhi-zhong(王志忠) and ZHU Jian-jun(朱建军). 平差模型的影响分析 [J]. The Chinese Journal of Nonferrous Metals(中国有色金属学报), 1999, 9(2): 418- 424.

[5] JIN Feng-xiang(靳奉祥), WANG Tong-xiao(王同孝), DU Zhi-xing(独知行), et al. 测量数据信息的模式识别方法 [J]. The Chinese Journal of Nonferrous Metals (中国有色金属学报), 1999, 9(2): 425- 430.

[6] HUANG Wei-bin(黄维彬). Modern Theory of Survey Adjustment of Observation with Applications(近代测量平差理论及其应用) [M]. Beijing: People's Liberation Army Press, 1992: 402- 412.

[7] LU Xiu-shan(卢秀山) and OU Ji-kun(欧吉坤). 值域空间正交分解及参数分组估计 [J]. The Chinese Journal of Nonferrous Metals(中国有色金属学报), 1998, 8(3): 541- 546.

[8] LIU Zhong-kan(柳重堪). Applied Functional Analysis (应用泛函分析) [M]. Beijing: National Defence Industry Press, 1986: 36- 55.

[9] Erwin Kreyszig (University of Windsor). Introductory Functional Analysis with Applications [M]. Beijing: Beijing Aviation Institute Press, 1987: 80- 100.

[10] ZHANG Jin-huai(张金槐). Estimation of Parameters with Improvements in Linear Models(线性模型参数估计及其改进) [M]. Changsha: National University of Defence Technology Press, 1992. 109- 118.


# Measure of space relationships among basic and additional parameters


LU Xiu-shan, SONG Shu-li, HAN Xiao-dong

*(Department of Geoscience, Shandong University of Science and Technology, Tai'an 271019, P. R. China)*



**[Abstract]** Properly selecting parameters to avoid stronger multicollinearity is a key to efficient estimation of parameters. Through separating parameters $\boldsymbol{X}$ into basic parameter $\boldsymbol{X}_1$ and additional parameter $\boldsymbol{X}_2$, and assuming that non-properly choosing additional parameters is the main reason causing stronger multicollinearity among parameters, the measure criterion which applies sine of the angle from a point (a vector in $\boldsymbol{X}_2$) in Hilbert space to space $\boldsymbol{H}_0$(a subspace of Hilbert space, from columns of $\boldsymbol{X}_1$) was proposed to optimumize the additional parameters. Theoretical reasoning and two demonstration spanded examples were also given.



**[Key words]** parameter estimation; multicollinearity; measures; Hilbert space


(编辑　何学锋)