国防科技大学学报
JOURNAL OF NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY
第 19 卷第 2 期　1997 年 4 月　Vol. 19 No. 2

先验信息下的改进主成分估计

胡庆军

(国防科技大学系统工程与数学系　长沙　410073)

摘　要　对于线性模型 $Y = XC + \epsilon, E(\epsilon) = 0, Cov(\epsilon, \epsilon) = \sigma^2 I$, 在先验信息 $(E(C) = 0, Cov(C, C) = \sigma_0^2 I)$ 下, 本文给出主成分个数的一种更合理取法, 并提出了一种改进的主成分估计。文中证明了该估计是可容许估计, 且在"平均"均方误差意义下, 优于最小二乘估计, 亦优于主成分估计。

关键词　线性模型, 先验信息, 改进主成分估计, 可容许性, "平均"均方误差
分类号　0212. 1

Improved Principal Components Estimate under Prior Information

Hu Qing Jun
(Department of Systems Engineering and Maths, NUDT, Changsha, 410073)

Abstract With regard to the linear model

$$Y = XC + \epsilon, E(\epsilon) = 0, Cov(\epsilon, \epsilon) = \sigma^2 I$$

possessing the prior information: $E(C) = 0, Cov(C, C) = \sigma_0^2 I$, This article gives a best method of selecting the principal components´ numbers in the principal components estimate and presents an improved principal components estimate. Theoretically, we prove that the improved estimate is admissible amd that it is marked much letter than Least Square Estimate and is better than the principal components estimate in a sense of "average" error of mean square. Simulation verifies the inference above.

Key words linear model, prior information, improved principal components estimate, admissible, "average" error of mean square.

考虑线性模型

* 1996 年 10 月 31 日修订

$$Y_{m\times 1} = XC_{n\times 1} + \epsilon_{m\times 1} \tag{1}$$

$$E(\epsilon) = 0, \mathrm{Cov}(\epsilon, \epsilon) = \sigma^2 I \tag{2}$$

其中 Y 为观测数据，ϵ 为不可观测的随机误差向量，σ^2 为误差方差且未知，X 为已知的设计矩阵且列满秩，C 为待估参数向量且具有先验信息（其中 σ_0^2 已知）

$$E(C) = 0, \mathrm{Cov}(C, C) = \sigma_0^2 I \tag{3}$$

最小二乘估计 $\hat{C}^{LS}$ 是参数 C 的线性无偏估计类中的最优估计，但当 X “病态”严重时，由于 $\hat{C}^{LS}$ 的方差太大，$\hat{C}^{LS}$ 不再是 C 的良好估计。于是相继提出了多种有偏估计（诸如主成分估计、岭估计等）[1,2]，旨在减少估计的均方误差以达到估准参数 C 的目的。主成分估计是 1965 年文[1]提出的一种有偏估计方法，在处理实际问题中颇受推崇。但主成分估计有两方面是值得改进的：其一，主成分个数的选取[2]问题；其二，主成分估计是过分地压小随机误差 ϵ 的方差而增大估计的偏差。

本文在先验信息（3）下，首先解决了主成分个数的选取问题，并提出一种改进的主成分估计，在“平均”均方误差意义下，此估计优于主成分估计。

1 记号和基本引理

在模型（1）～（3）下，记

$$\hat{C}^{LS} = (X'X)^{-1}X'Y \tag{4}$$

$$\hat{\sigma}^2 = Y'(I - P_X)Y/(m - n) \tag{5}$$

$$P_X = X(X'X)^{-1}X', P'P = I$$

$$P'X'XP = \Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_p, \lambda_{p+1}, \cdots, \lambda_n)$$

$$\alpha = (\alpha_1, \cdots, \alpha_n)' = P'C \tag{6}$$

$$\Lambda_1 = \mathrm{diag}(\lambda_1, \cdots, \lambda_p), \Lambda_2 = \mathrm{diag}(\lambda_{p+1}, \cdots, \lambda_n)$$

$$\lambda_1 \geq \cdots \geq \lambda_p > \lambda_{p+1} \geq \cdots \geq \lambda_n > 0 \tag{7}$$

$$\tilde{C}^{PC} = P\begin{pmatrix} \Lambda_1^{-1} & \\ & 0 \end{pmatrix} P'X'Y \tag{8}$$

$$\tilde{C}^{MPC}(r) = P\begin{pmatrix} \Lambda_1^{-1} & \\ & (\Lambda_2 + rI)^{-1} \end{pmatrix} P'X'Y \tag{9}$$

其中 $r > 0$ 为待定常数。注意到 $\hat{C}^{LS}$ 即为 C 的最小二乘估计，$\tilde{C}^{PC}$ 为主成分估计。

引理 1　在模型（1）、（2）下，则

1)
$$\hat{C}^{LS} = P\Lambda^{-1}P'X'Y = P\begin{pmatrix} \Lambda_1^{-1} & \\ & \Lambda_2^{-1} \end{pmatrix} P'X'Y \tag{10}$$

$$E\,\|\hat{C}^{LS} - C\|^2 = \sum_{i=1}^{n}(\sigma^2/\lambda) \tag{11}$$

2)
$$\tilde{C}^{PC} = P\begin{pmatrix} I_p & \\ & 0 \end{pmatrix} P'C + P\begin{pmatrix} \Lambda_1^{-1} & \\ & 0 \end{pmatrix} P'X'\epsilon \tag{12}$$

$$\|E(\tilde{C}^{PC}) - C\|^2 = \sum_{i=p+1}^{n}\alpha_i^2 \tag{13}$$

$$\mathrm{tr}[\mathrm{Cov}(\tilde{C}^{PC},\tilde{C}^{PC})] = \sum_{i=1}^{p}(\sigma^2/\lambda_i) \tag{14}$$

$$\mathrm{E}\,\|\tilde{C}^{pc} - C\|^2 = \sum_{i=1}^{p}(\sigma^2/\lambda) + \sum_{i=p+1}^{n}\alpha_i^2 \tag{15}$$

3) $$\mathrm{E}\,\|\tilde{C}^{MPC}(r) - C\|^2 = \sum_{i=1}^{p}(\sigma^2/\lambda) + \sum_{i=p+1}^{n}[(\sigma^2\lambda_i + r^2\alpha_i^2)/(\lambda + r)^2] \tag{16}$$

引理 2　对于模型 (1)、(2)，A 是 n 阶常数矩阵，在均方损失函数下，则 $A\tilde{C}^{LS}$ 为 C 的可容许估计的充要条件是下式成立：

$$A(X'X)^{-1}A' \leq A(X'X)^{-1}.$$

引理 3　在模型 (1) ～ (3) 下，则

$$\mathrm{E}_c(\mathrm{E}\,\|\tilde{C}^{PC} - C\|^2) = \sum_{i=1}^{p}(\sigma^2/\lambda_i) + (n-p)\sigma_0^2 \tag{17}$$

此处 "E_c" 表示关于先验信息 C 对 $\mathrm{E}\,\|\tilde{C}^{PC} - C\|^2$ 求期望，即求 "平均" 均方误差，以下类同。

引理 1、3 易直接验证，引理 2 参见 [3]。

从 (12) ～ (5) 式可见，若主成分个数 p 取得少，这时虽然 C 的估计 $\tilde{C}^{PC}$ 的方差小，但其偏差平方和 $\sum_{i=p+1}^{n}\alpha_i^2$ 将偏大；反之，偏差平方和虽小，但当小特征值 λ_i 较多时，方差将迅速增大；再者，均方误差还与 σ^2、α_i^2 密切相关。这就是说，在主成分估计中，主成分个数的选取不仅要考虑特征根 λ_i，而且要顾及到 σ^2、α_i^2 的大小对估计的影响。结合 (8)、(10) 式，可见，偏差平方和 $\sum_{i=p+1}^{n}\alpha_i^2$ 是由 $\begin{pmatrix}\Lambda_1^{-1} & \\ & \Lambda_2^{-1}\end{pmatrix}$ 换成 $\begin{pmatrix}\Lambda_1^{-1} & \\ & 0\end{pmatrix}$ 所造成的，即由 Λ_2^{-1} 换成零矩阵造成的。这显然不够合理，特别是 Λ_2 中的前几个对角元 λ_{p+1}，…。那么，一个自然的想法是考虑形如 (9) 式的估计 $\tilde{C}^{MPC}(r)$。下面讨论主成分个数选取问题及形如 (9) 式估计的 r 确定问题。

2　主成分个数的合理选取

在 "平均" 均方误差最小意义下，主成分个数的最优选取准则就是：确定个数 p 使 (17) 式关于 p 达到最小。注意到 (7) 式特征根 λ_i 的大小顺序，使 (17) 式达最小又等价于确定 p，使 (7) 式满足

$$\lambda_{p+1} \leq \frac{\sigma^2}{\sigma_0^2} < \lambda_p \tag{18}$$

在实用中用 (5) 式 $\hat{\sigma}^2$ 估计 σ^2，因此 (18) 式改为

$$\lambda_{p+1} \leq \hat{\sigma}^2/\sigma_0^2 < \lambda_p \tag{19}$$

若上式成立，则取 $\Lambda_1 = \mathrm{diag}(\lambda_1, \cdots, \lambda_p)$，就得到主成分估计 (8) 式。这就是在先验信息 (3) 下，主成个个数的最优选择准则。

3　改进主成分估计

定理 1　在模型 (1) ～ (2) 下，对任给定的常数 $r > 0$，则 (9) 式 $\tilde{C}^{MPC}(r)$ 是 C 的可容许估计。

证明　由(9)、(10)式知，$\tilde{C}^{MPC}(r) = A \cdot \hat{C}^{LS}$，其中

$$A = P\begin{pmatrix}\Lambda_1^{-1} & \\ & (\Lambda_2 + rI)^{-1}\end{pmatrix}\begin{pmatrix}\Lambda_1 & \\ & \Lambda_2\end{pmatrix}P' = P\begin{pmatrix}I_p & \\ & (\Lambda_2 + rI)^{-1}\Lambda_2\end{pmatrix}P'$$

易直接验证有

$$A(X'X)^{-1}A' = P\begin{pmatrix}\Lambda_1^{-1} & \\ & (\Lambda_2 + rI)^{-2}\Lambda_2\end{pmatrix}P'$$

$$A(X'X)^{-1} = P\begin{pmatrix}\Lambda_1^{-1} & \\ & (\Lambda_2 + rI)^{-1}\end{pmatrix}P'$$

故

$$A(X'X)^{-1} - A(X'X)^{-1}A' = r\ P\begin{pmatrix}0 & \\ & (\Lambda_2 + rI)^{-2}\end{pmatrix}P' \geq 0$$

即有$A(X'X)^{-1}A' \leq A(X'X)^{-1}$。由引理2知，$\tilde{C}^{MPC}(r)$是$C$的可容许估计。(证毕)。

引理4　在模型(1)～(3)下，则

$$f(r) \equiv E_C(E\|\tilde{C}^{MPC}(r) - C\|^2)$$

$$= \sum_{i=1}^{p}(\sigma^2/\lambda_i) + \sum_{i=p+1}^{n}[(\sigma^2\lambda_i + r^2\sigma_0^2)/(\lambda_i + r)^2] \tag{20}$$

定理2　条件同引理4，则$f(r)$的最小值点为$\hat{r} = \sigma^2/\sigma_0^2$，即$f(\hat{r}) = \min_{r\geq 0} f(r)$。且

$$f(\hat{r}) = \sum_{i=1}^{p}(\sigma^2/\lambda_i) + \sum_{i=p+1}^{n}[\sigma^2/(\lambda_i + \sigma^2/\sigma_0^2)] \tag{21}$$

由引理1及(3)式得引理4。对$f(r)$关于r求导可得定理2(略)。

这样就得到了形如(9)式估计的最优r选取方法。记

$$\tilde{C}^{MPC} = P\begin{pmatrix}\Lambda_1^{-1} & \\ & (\Lambda_2 + \frac{\sigma^2}{\sigma_0^2}I)^{-1}\end{pmatrix}P'(X'X)X' \tag{22}$$

本文称$\tilde{C}^{MPC}$为在先验信息(3)下参数C的改进主成分估计。其中$\lambda_1, \cdots, \lambda_p, \lambda_{p+1}, \cdots, \lambda_n$满足(7)和(18)式，由(16)及(21)式，则有

$$E\|\tilde{C}^{MPC} - C\|^2 = \sum_{i=1}^{p}\frac{\sigma^2}{\lambda_i} + \sum_{i=p+1}^{n}\frac{\sigma^2\lambda_i + \frac{\sigma^4}{\sigma^4}\alpha^2}{(\lambda_i + \frac{\sigma^2}{\sigma^2})^2} \tag{23}$$

$$E_C(E\|\tilde{C}^{MPC} - C\|^2) = \sum_{i=1}^{p}\frac{\sigma^2}{\lambda_i} + \sum_{i=p+1}^{n}\frac{\sigma^2}{\lambda_i + \frac{\sigma^2}{\sigma_0^2}} \tag{24}$$

定理3　在模型(1)～(3)下，则

$$E_C(E\|\tilde{C}^{MPC} - C\|^2) - E\|\hat{C}^{LS} - C\|^2 = -\sigma^4\sum_{i=p+1}^{n}\frac{1}{\lambda_i(\lambda_i\sigma_0^2 + \sigma^2)} \tag{25}$$

$$E_C(E\|\tilde{C}^{MPC} - C\|^2) - E_C(E\|\tilde{C}^{PC} - C\|^2) = -\sigma_0^4\sum_{i=p+1}^{n}\frac{\lambda_i}{\lambda_i\sigma_0^2 + \sigma^2} \tag{26}$$

在“平均”均方误差越小越好的意义下，由(25)、(26)式知，当X⑦X“病态”严重，或者说X⑦X有很小的特征根λ(即$\lambda_i<<\sigma^2/\sigma_0^2$)时，有

$$E_c(E\|\tilde{C}^{MPC}-C\|^2)<<E\|\hat{C}^{LS}-C\|^2$$

$$E_c(E\|\tilde{C}^{MPC}-C\|^2)<E_c(E\|\tilde{C}^{PC}-C\|^2)$$

以上说明，改进主成分估计$\tilde{C}^{MPC}$与$\hat{C}^{LS}$相比，大大提高了参数估计的精度，且优于主成分估计$\tilde{C}^{PC}$。实用中用$\hat{\sigma}^2$估计σ^2，由(22)式得到在先验信息(3)下的改进主成分估计，仍记为

$$\tilde{C}^{MPC}=P\begin{pmatrix}\Lambda_1^{-1} & \\ & (\Lambda_2+\frac{\hat{\sigma}^2}{\sigma_0^2}I)^{-1}\end{pmatrix}P\text{⑦}X\text{⑦}Y \tag{27}$$

其中λ_1，…，λ_r满足(19)式，P，Λ_1，Λ_2记号同前。

4 模拟计算

针对某问题的工具误差系数模型符合(1)～(3)的假设，X为$m\times n$(其中$m=6000$，$n=36$)的已知设计矩阵，“病态”严重，$X'X$的最小特征根约为0.5×10^{-6}，而$\hat{\sigma}^2\approx39$，$\sigma_0^2=(\frac{1}{2.7})^2$。下面对$C$取仿真值，用模拟随机数组$\epsilon$及已知的$X$，形成模型(1)～(3)，作有偏估计的模拟计算，用两种准则来衡量改进主成分估计$\tilde{C}^{MPC}$的模拟结果。

准则1：估计值$\tilde{C}$与仿真值C的误差平方和$\|\tilde{C}-C\|^2$及相对误差平方和$\|\tilde{C}-C\|^2/\|C\|^2$越小越好。

准则2：仿真值C的分量C_i与估计值$\tilde{C}_i$($i=1,2,\cdots,n$)满足不等式

$$0.5<\tilde{C}_i/C_i<2 \tag{28}$$

的个数越多表明该估计方法越好[4]。

具体做法：对于给定的σ_0^2，取$C\sim N(0,\sigma_0^2I)$为C的仿真值，取$\epsilon\sim N(0,\sigma^2I)$为随机误差数组(对给定的$\sigma^2$)，形成模型(1)～(3)，再求各种有偏估计。今做了大量的模拟计算，仅给出部分模拟结果。表1是一组仿真值及对应的估计值，表2列出了多次模拟计算的各种估计的误差平方和、相对误差平方和及满足不等式(28)式的个数。

从模拟结果看，改进主成分估计明显优于最小二乘估计，且优于主成分估计。

表 1　一组仿真值 C 的有偏估计（取 $C \sim N(0, 10^2 \cdot I)$，$\epsilon \sim N(0, I)$）

序号	$\hat{C}^{LS}$	真值 C	$\tilde{C}^{MPC}$	$\tilde{C}^{PC}$	序号	$\hat{C}^{LS}$	真值 C	$\tilde{C}^{MPC}$	$\tilde{C}^{PC}$
1	- 3.28	- 1.690	- 2.002	- 1.723	19	18.2	22.343	3.527	2.940
2	- 8.52	- 9.253	- 8.933	- 9.092	20	- 272.	- 4.480	- 3.904	- 2.856
3	- 3.40	- 4.974	- 4.856	- 4.959	21	0.369	- 5.187	- 8.636	- 9.503
4	20.9	17.485	17.071	14.849	22	- 3.72	- 6.430	- 6.497	- 6.642
5	- 11.1	- 10.748	- 10.905	- 10.799	23	6.19	3.940	4.308	4.428
6	- 22.4	- 22.361	- 22.258	- 22.247	24	29.2	28.401	28.641	28.474
7*	- 37.4	- 14.771	- 14.492	0.339	25	- 21.2	- 21.093	- 21.105	- 21.077
8	- 10.3	- 10.717	- 10.510	- 10.638	26	- 5.10	0.649	- 2.929	- 1.755
9	- 6.71	- 7.979	8.455	11.121	27	- 11.3	- 8.150	- 1.072	1.053
10	- 7.43	- 4.757	- 4.522	- 4.385	28	- 72.8	1.184	1.008	1.451
11	- 43.3	- 9.022	- 6.401	- 5.285	29	2.12	3.494	3.197	3.726
12	12.3	13.392	12.955	13.286	30	7.62	7.587	7.828	7.975
13	66.0	14.264	16.839	18.834	31	- 13.4	- 14.704	- 12.412	- 12.90
14*	7.20	- 1.227	- 2.192	- 2.649	32	17.5	17.460	17.461	17.461
15	976.	- 4.784	- 1.845	- 1.798	33	1.80	1.587	1.222	0.836
16	6.70	- 5.407	- 6.546	- 7.017	34	8.98	9.046	8.976	8.966
17	10.2	9.699	9.982	9.823	35	21.7	21.746	21.737	21.737
18*	- 50.7	- 18.115	- 14.308	- 6.907	36	- 4.67	- 5.474	- 5.481	- 5.505
误差平方和 $\|\tilde{C} - C\|^2$						1.04×10^6		744.7	1266.9
相对误差平方和 $\|\tilde{C} - C\|^2 / \|C\|^2$						190.		0.136	0.231
$\frac{1}{2} < C_i / C_i < 2$ 的个数						25		31	28

表 2　不同 C 及 ϵ 时各种估计的误差模拟结果（$C\sim N\ (0,\ \sigma_0^2 I)$，$\epsilon\sim N\ (0,\ \sigma^2 I)$

		$\|\tilde{C}-C\|^2$			$\frac{\|\tilde{C}-C\|^2}{\|C\|^2}$			$\frac{1}{2}<\frac{\tilde{c}_i}{c_i}<2$ 的个数			仿真值 $\|C\|^2$
σ_0	σ	$\hat{C}^{LS}$	$\tilde{C}^{MPC}$	$\tilde{C}^{PC}$	$\hat{C}^{LS}$	$\tilde{C}^{MPC}$	$\tilde{C}^{PC}$	$\hat{C}^{LS}$	$\tilde{C}^{MPC}$	$\tilde{C}^{PC}$	
$\frac{1}{2.7}$	0.01	496.	0.858	0.874	135.	0.233	0.238	26	30	30	3.678
$\frac{1}{2.7}$	1.	9.1×10^5	2.041	2.054	1.94×10^5	0.434	0.437	9	22	20	4.703
$\frac{1}{2.7}$	1.7	9.72×10^6	2.112	2.137	1.86×10^6	0.404	0.409	8	19	16	5.230
$\frac{1}{2.7}$	3.1	2.27×10^5	2.799	3.117	5.35×10^4	0.660	0.735	7	15	12	4.240
$\frac{1}{2.7}$	3.9	1.38×10^7	3.943	4.111	2.15×10^6	0.611	0.637	2	15	13	6.453
$\frac{1}{2.7}$	4.7	2.31×10^7	1.839	2.248	7.24×10^6	0.577	0.705	2	13	11	3.189
$\frac{1}{2.7}$	5.6	1.56×10^8	3.877	4.123	2.73×10^7	0.680	0.723	4	10	11	5.703
$\frac{1}{2.7}$	6.2	1.27×10^8	2.428	2.555	2.88×10^7	0.549	0.578	3	15	12	4.422
$\frac{1}{2.7}$	6.8	1.55×10^8	2.405	2.303	3.18×10^7	0.492	0.472	3	14	11	4.883
$\frac{1}{2.7}$	6.8	1.09×10^6	2.199	2.272	3.08×10^5	0.619	0.610	6	14	11	3.551
1.	1.	3.31×10^6	11.755	12.022	1.02×10^5	0.361	0.372	15	23	23	32.288
$\frac{10}{2.7}$	6.	1.98×10^6	151.01	133.32	5.36×10^3	0.410	0.362	9	19	18	368.733
5.	5.	5.20×10^8	263.26	283.55	3.32×10^5	0.168	0.181	13	23	22	1563.6
5.	10.	1.51×10^9	243.99	357.57	2.21×10^6	0.358	0.525	13	20	16	681.29
10.	1.	1.04×10^6	744.73	1266.92	190.	0.136	0.231	25	31	28	5494.7

参考文献

1　Massy, W. F. Principal Components Regression in Exploratory Statistical Research. J. Amer. Statist. Assoc., (1965) 60: 234～266.

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

3　王松桂。线性模型的理论及其应用。安徽教育出版社，1987.

4　胡庆军，程光显，罗建书。新的有偏估计的研究——主扫描估计。导弹与航天运载技术，1995（3）：36～44.

（责任编辑　潘生）