第31卷 第2期
1992年

中山大学学报(自然科学版)
ACTA SCIENTIARUM NATURALIUM
UNIVERSITATIS SUNYATSENI

Vol.31 No.2
1992

不完整的多元正态数据的附加信息检验*

余锦华**
(数学系)

摘 要 给出了检验不完整观测的多元正态数据中所含附加信息是否显著影响总体参数估计的一种方法，从而提供了一种决定是否应该利用这些不完整数据而采用更为复杂的估计方法的途径。

关键词 附加信息，不完整观测数据，似然比检验

不完整的多元正态数据的参数估计，多年来被广泛研讨，许多学者先后找出了在各种情形下参数的MLE，如Wilks (1932)首次给出了有缺失数据时二元正态总体的参数估计；Edgett[1]给出了仅在一个变元中出现缺失数据时三元正态总体的参数估计；Anderson[2]研讨了具有某种特定形式的缺失数据时P元正态总体的参数估计；Orchard与Woodbury[3]提出了一种更为一般的估计方法并称之为缺失信息原理(MIP)，Dempster，Laird与Rubin (1977)进一步归纳出求MLE的EM算法等，然而，给定的一组不完整观测的多元正态数据中，究竟是否含有显著影响总体参数估计的附加信息，则是应用这些方法之前所必须回答的首要问题。本文将首先推出利用包括残缺数据在内的全部数据时，总体参数的极大似然估计，进而给出检验不完整数据中所含附加信息是否显著的一种似然比检验，从而为决定是否应该利用这些不完整数据，以及如何利用这些数据提供了一种途径。

1 参数(μ,Σ)的混合估计

设$x\sim N_p(\mu,\Sigma)$，$x^{(1)},x^{(2)},\cdots,x^{(N)}$为来自总体$x$的一组观测值，其中有$n_1$个为完整数据，即$x^{(1)},\cdots,x^{(n_1)}$ iid $N_p(\mu,\Sigma)$，其余$N-n_1\triangleq n_2$个为不完整数据，即$x^{(n_1+1)},\cdots x^{(N)}$ iid $N_q(\mu_{(2)},\Sigma_{(2)})$，其中$q<p$，$\mu_{(2)}=D\mu$，$\Sigma_{(2)}=D\Sigma D'$，而$D$是一个以0，1为元素的反映不完整数据与完整数据之间的结构关系（0对应缺失分量，1对应保留的分量）的$q\times p$矩阵。

记仅用完整观测数据求得的μ,Σ的MLE为$\hat{\mu}_1,\hat{\Sigma}_1$；而仅用不完整数据求得的$\mu_{(2)}$，$\Sigma_{(2)}$的MLE为$\hat{\mu}_2,\hat{\Sigma}_2$，则

本文1991年5月29日收到
* 国家自然科学基金资助项目
** 数学系86级本科生高宽贤提供有关数据

$$\hat{\mu}_1=\frac{1}{n_1}\sum_{i=1}^{n_1}x^{(i)}\ ,\qquad \hat{\Sigma}_1=\frac{1}{n_1}\sum_{i=1}^{n_1}(x^{(i)}-\hat{\mu}_1)(x^{(i)}-\hat{\mu}_1)'$$

$$\hat{\mu}_2=\frac{1}{n_2}\sum_{i=1}^{n_2}x^{(n_1+i)}\ ,\qquad \hat{\Sigma}_2=\frac{1}{n_2}\sum_{i=1}^{n_2}(x^{(n_1+i)}-\hat{\mu}_2)(x^{(n_1+i)}-\hat{\mu}_2)'$$

参数(μ,Σ)的似然函数可表为

$$L=L_1\cdot L_2$$

取对数后即

$$\ln L=\ln L_1+\ln L_2$$

其中，
$$L_1=(2\pi)^{-n_1p/2}\,|\Sigma|^{-n_1/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\Sigma^{-1}\left[\sum_{i=1}^{N_1}(x^{(i)}-\mu)(x^{(i)}-\mu)'\right]\right\}$$

令若

$$M_1=\sum_{i=1}^{n_1}(x^{(i)}-\mu)(x^{(i)}-\mu)'=n_1[\hat{\Sigma}_1+(\hat{\mu}_1-\mu)(\hat{\mu}_1-\mu)']$$

则

$$\ln L_1=-\frac{pn_1}{2}\ln(2\pi)-\frac{n_1}{2}\ln|\Sigma|-\frac{1}{2}\mathrm{tr}\,\Sigma^{-1}M_1$$

同理

$$\ln L_2=-\frac{qn_2}{2}\ln(2\pi)-\frac{n_2}{2}\ln|D\Sigma D'|-\frac{1}{2}\mathrm{tr}[(D\Sigma D')^{-1}\cdot M_2]$$

其中，
$$M_2=\sum_{i=1}^{2}(x^{(n_1+i)}-D\mu)(x^{(n_1+i)}-\quad\mu)'=n_2[\hat{\Sigma}_2+(\hat{\mu}_2-D\mu)(\hat{\mu}_2-D\mu)']$$

从而$\ln L$最终可表为

$$\ln L=(c_1+c_2)-\frac{1}{2}[n_1\ln|\Sigma|+n_2\ln|D\Sigma D'|]-\frac{1}{2}\mathrm{tr}[\Sigma^{-1}M_1+(D\Sigma D')^{-1}M_2]\qquad(1)$$

定理 1　令$Z=D\hat{\mu}_1-\hat{\mu}_2$，$B(\Sigma)=n_2\Sigma D'(D\Sigma D')^{-1}/N$，则$\ln L$的极大值点为

$$\tilde{\mu}=\hat{\mu}_1-B(\Sigma_1)Z,\qquad \tilde{\Sigma}=\hat{\Sigma}_1-\frac{N}{n_2}B(\Sigma_1)(D\hat{\Sigma}_1D'-\hat{\Sigma}_2)B(\Sigma_1)'+\frac{n_1}{n_2}B(\Sigma_1)ZZ'B(\Sigma_1)'$$

称$\tilde{\mu},\tilde{\Sigma}$为使用全部数据时$\mu,\Sigma$的混合极大似然估计。

证明　令$\dfrac{\partial\ln L}{\partial\mu}=0$　可推出

$$n_1\Sigma^{-1}(\hat{\mu}_1-\mu)+n_2D'(D\Sigma D')^{-1}(\hat{\mu}_2-D\mu)=0\qquad(2)$$

由此不难求得

$$\tilde{\mu}=\hat{\mu}_1-B(\Sigma)Z\qquad(3)$$

将$\tilde{\mu}$代入M_1，M_2后可得

$$\mathrm{tr}(\Sigma^{-1}M_1)+\mathrm{tr}[(D\Sigma D')^{-1}M_2]=n_1\,\mathrm{tr}(\Sigma^{-1}\hat{\Sigma}_1)+n_2\mathrm{tr}[(D\Sigma D')^{-1}\hat{\Sigma}_2]$$
$$+\frac{n_1n_2}{N}Z'(D\Sigma D')^{-1}Z$$

将上式代入(1)式后令$\frac{\partial \ln L}{\partial \Sigma}=\mathbf{0}$ 可推出

$$\frac{\partial}{\partial \Sigma}\{n_1 \ln|\Sigma| + n_2 \ln|D\Sigma D'| + n_1 \operatorname{tr}(\Sigma^{-1}\hat{\Sigma}_1) + n_2 \operatorname{tr}[(D\Sigma D')^{-1}\hat{\Sigma}_2]$$

$$+\frac{n_1 n_2}{N}Z'(D\Sigma D')^{-1}Z\} = \mathbf{0} \tag{4}$$

记矩阵A的对角线元素所组成的对角阵为$\operatorname{diag} A$，则

$$\frac{\partial}{\partial \Sigma}\ln|\Sigma| = 2\Sigma^{-1} - \operatorname{diag}\Sigma^{-1}$$

$$\frac{\partial}{\partial \Sigma}\ln|D\Sigma D'| = D'[2(D\Sigma D')^{-1} - \operatorname{diag}(D\Sigma D')^{-1}]D$$

$$\frac{\partial}{\partial \Sigma}\operatorname{tr}(\Sigma^{-1}\hat{\Sigma}_1) = -2\Sigma^{-1}\hat{\Sigma}_1\Sigma^{-1} + \operatorname{diag}(\Sigma^{-1}\hat{\Sigma}_1\Sigma^{-1})$$

$$\frac{\partial}{\partial \Sigma}\operatorname{tr}(D\Sigma D')^{-1}ZZ' = D'\{-2(D\Sigma D')^{-1}ZZ'(D\Sigma D')^{-1}$$

$$+\operatorname{diag}[(D\Sigma D')^{-1}ZZ'(D\Sigma D')^{-1}]\}D$$

将上述几式代入(4)式并化简后可得

$$\left(I+\frac{N}{n_1}BD\right)\Sigma = \hat{\Sigma}_1 + \frac{N^2}{n_1 n_2}B\hat{\Sigma}_2 B' + \frac{N}{n_2}BZZ'B' \tag{5}$$

两边左乘$(I-BD)$并化简后可得

$$\Sigma = \hat{\Sigma}_1 - (N/n_2)B(D\hat{\Sigma}_1 D' - \hat{\Sigma}_2)B' + (n_1/n_2)BZZ'B' \tag{6}$$

从而

$$\tilde{\Sigma}_1 = \hat{\Sigma}_1 - (N/n_2)B(\tilde{\Sigma})(D\hat{\Sigma}_1 D' - \hat{\Sigma}_2)B(\tilde{\Sigma})' + (n_1/n_2)B(\tilde{\Sigma})ZZ')B(\tilde{\Sigma})' \tag{7}$$

不难验证 $DB(\tilde{\Sigma}) = DB(\hat{\Sigma}_1)$

进而可推知 $B(\tilde{\Sigma}) = B(\hat{\Sigma}_1)$

故(3)式和(7)式中的$B(\tilde{\Sigma})$可代以$B(\hat{\Sigma}_1)$，定理得证。

事实上，只要适当安排各分量次序，总可将D表为

$$D = (I_q \vdots \mathbf{0})_{q\times p}$$

这样一来，可明显看出

$$Z = D\hat{\mu}_1 - \hat{\mu}_2.$$

正好表示分别用完整和不完整数据对μ的前q个分量作出的两种估计之差；而$D\hat{\Sigma}_1 D' - \hat{\Sigma}_2$则表示两组数据对$\mu$的前$q$个分量的协方差阵所作的不同估计之差。

从而，μ,Σ的混合估计$\tilde{\mu},\tilde{\Sigma}$可视为利用完全数据作出估计之后，再根据上述对公共部份所作出的两种估计之差异信息加以适当修正而获得的。显而易见，此差异信息的大小反映了$\tilde{\mu},\tilde{\Sigma}$与$\hat{\mu}_1,\hat{\Sigma}_1$的差异大小，反映了采用混合估计的必要性大小，因而可以用来对

不完整观测作附加信息检验。

2　不完整观测的附加信息检验

为了检验不完整观测中的附加信息是否显著，有无必要采用混合估计，令

$$Q=\ln L(\tilde{\mu},\tilde{\Sigma})-\ln L(\hat{\mu}_1,\hat{\Sigma}_1) \tag{8}$$

显然$Q\geqslant 0$,且Q越小，越接近 0，表明不完整观测组所含的附加信息越不显著[5]。

下面往求Q的表达式，为推导简便，记

$$A_t=\mathrm{tr}(\Sigma_{(t)}^{-1}M_t),\quad t=1,2,\quad \Sigma_{(1)}=\Sigma$$

$A_t(\sim)$,$A_t(\Lambda)$分别表示以$\tilde{\mu},\tilde{\Sigma}$及$\hat{\mu}_1,\hat{\Sigma}_1$代入$\mu$，$\Sigma$后的函数值，则由(1)式可推知

$$Q=-\frac{1}{2}[n_1(\ln|\tilde{\Sigma}|-\ln|\hat{\Sigma}_1|)+n_2(\ln|D\tilde{\Sigma}D'|-\ln|D\hat{\Sigma}_1D'|)]$$

$$-\frac{1}{2}[A_1(\sim)-A_1(\Lambda)+A_2(\sim)-A_2(\Lambda)] \tag{9}$$

引理 1　$A_1(\Lambda)=n_1P$

证明　$A_1(\Lambda)=\mathrm{tr}\,\hat{\Sigma}_1^{-1}\cdot n_1[\hat{\Sigma}_1+(\hat{\mu}_1-\hat{\mu}_1)(\hat{\mu}_1-\hat{\mu}_1)']=n_1\,\mathrm{tr}\,I_p=n_1P$

引理 2　$A_1(\sim)=n_1p+n_2\,\mathrm{tr}[I_q-(D\tilde{\Sigma}D')^{-1}\hat{\Sigma}_2-\left(\frac{n_1}{N}\right)^2(D\tilde{\Sigma}D')^{-1}ZZ']$

证明　$A_1(\sim)=\mathrm{tr}\,\tilde{\Sigma}^{-1}\cdot n_1[\hat{\Sigma}_1+(\hat{\mu}_1-\tilde{\mu})(\hat{\mu}_1-\tilde{\mu})']=n_1\,\mathrm{tr}\,\tilde{\Sigma}^{-1}[\hat{\Sigma}_1+BZZ'B']$

而由(6)式可知

$$\hat{\Sigma}_1+(n_1/n_2)BZZ'B'=\tilde{\Sigma}+(N/n_2)B(D\hat{\Sigma}_1D'-\hat{\Sigma}_2)B'$$

从而$A_1(\sim)=n_1\,\mathrm{tr}\,\tilde{\Sigma}^{-1}[\tilde{\Sigma}+(N/n_2)B(D\hat{\Sigma}_1D'-\hat{\Sigma}_2)B'+[(n_2-n_1)/n_2]BZZ'B']$

$$=n_1\{P+(N/n_2)\mathrm{tr}(B'\tilde{\Sigma}^{-1}B)(D\hat{\Sigma}_1D'-\hat{\Sigma}_2)+[(n_2-n_1)/n_2]\mathrm{tr}[(B'\tilde{\Sigma}^{-1}B)ZZ']\}$$

将　$B'\tilde{\Sigma}^{-1}B=(n_2/N)^2(D\tilde{\Sigma}D')^{-1}$代入上式即得

$$A_1(\sim)=n_1P+[(n_1\cdot n_2)/N]\mathrm{tr}[(D\tilde{\Sigma}D')^{-1}(D\hat{\Sigma}_1D')]-[(n_1\cdot n_2)/N]\mathrm{tr}(D\tilde{\Sigma}D')^{-1}\hat{\Sigma}_2$$

$$+[(n_2-n_1)n_1n_2/N^2]\cdot\mathrm{tr}[(D\tilde{\Sigma}D')^{-1}ZZ']$$

再将　$D\hat{\Sigma}_1D'=(N/n_1)(D\tilde{\Sigma}D')-(n_2/n_1)\hat{\Sigma}_2-(n_1/N)ZZ'$

代入上式化简整理后得

$$A_1(\sim)=n_1P+n_2\mathrm{tr}[I_q-(D\tilde{\Sigma}D')^{-1}\hat{\Sigma}_2-(n_1/N)^2(D\tilde{\Sigma}D')^{-1}ZZ']$$

引理 3　$A_2(\sim)=n_2\mathrm{tr}[(D\tilde{\Sigma}D')^{-1}\hat{\Sigma}_2+(n_1/N)^2(D\tilde{\Sigma}D')^{-1}ZZ']$

证明　$A_2(\sim)=\mathrm{tr}(\tilde{\Sigma}_2^{-1}M_2)=n_2\,\mathrm{tr}(D\tilde{\Sigma}D')^{-1}[\hat{\Sigma}_2+(\hat{\mu}_2-D\tilde{\mu})(\hat{\mu}_2-D\tilde{\mu})']$

而　$\hat{\mu}_2-D\tilde{\mu}=\hat{\mu}_2-D[\hat{\mu}_1-B(D\hat{\mu}_1-\hat{\mu}_2)]$

$$=\hat{\mu}_2-D\hat{\mu}_1+(n_2/N)D\hat{\mu}_1-(n_2/N)\hat{\mu}_2=-(n_2/N)Z$$

从而

$$A_2(\sim)=n_2\,\mathrm{tr}[(D\tilde{\Sigma}D')^{-1}\hat{\Sigma}_2+(n_1/N)^2(D\tilde{\Sigma}D')^{-1}ZZ']$$

类似地可推知

$$A_2(A)=n_2\,\mathrm{tr}[(D\hat{\Sigma}_1D')^{-1}\hat{\Sigma}_2+(D\hat{\Sigma}_1D')^{-1}ZZ'] \tag{10}$$

引理 4 $|\hat{\Sigma}_1^{-1}\tilde{\Sigma}|=|(n_1/N)I_q+(n_2/N)(D\hat{\Sigma}_1D')^{-1}\hat{\Sigma}_2+(n_1n_2/N^2)(D\hat{\Sigma}_1D')^{-1}ZZ'|$

证明 $$|\hat{\Sigma}_1^{-1}\tilde{\Sigma}|=|\hat{\Sigma}_1^{-1}[\hat{\Sigma}_1-(N/n_2)B(\hat{\Sigma}_1)(D\hat{\Sigma}_1D'-\hat{\Sigma}_2)B'(\hat{\Sigma}_1)$$

$$+(n_1/n_2)B(\hat{\Sigma}_1)ZZ'B'(\hat{\Sigma}_1)]|$$

$$=|I_q-(n_2/N)[I_q-(D\hat{\Sigma}_1D')^{-1}\hat{\Sigma}_2-(n_1/N)(D\hat{\Sigma}_1D')^{-1}ZZ']|$$

$$=|(n_1/N)I_q+(n_2/N)(D\hat{\Sigma}_1D')^{-1}\hat{\Sigma}_2+(n_1n_2/N^2)(D\hat{\Sigma}_1D')^{-1}ZZ'|.$$

引理 5 $|(D\hat{\Sigma}_1D')^{-1}(D\tilde{\Sigma}D')=|\hat{\Sigma}_1^{-1}\tilde{\Sigma}|$

证明 由(12)式及引理 4 即知

$$|(D\hat{\Sigma}_1D')^{-1}(D\tilde{\Sigma}D')|=|(n_1/N)I_q+(n_2/N)(D\hat{\Sigma}_1D')^{-1}\hat{\Sigma}_2$$

$$+(n_1n_2/N)(D\hat{\Sigma}_1D')^{-1}ZZ'|=|\hat{\Sigma}_1^{-1}\tilde{\Sigma}|$$

综上可得

定理 2 $Q=-\frac{1}{2}W-\frac{1}{2}V=-\frac{1}{2}(W+V)$

其中，

$$W=A_1(\sim)-A_1(A)+A_2(\sim)-A_2(A)$$

$$=n_2q-n_2Z'(D\hat{\Sigma}_1D')^{-1}Z-n_2\mathrm{tr}(D\hat{\Sigma}_1D')^{-1}\hat{\Sigma}_2$$

$$=n_2\,\mathrm{tr}[I_q-(D\hat{\Sigma}_1D')^{-1}ZZ'-(D\hat{\Sigma}_1D')^{-1}\hat{\Sigma}_2]$$

与

$$V=n_1\ln|\tilde{\Sigma}_1\tilde{\Sigma}|+n_2\ln|(D\hat{\Sigma}_1D')^{-1}(D\tilde{\Sigma}D')|=(n_1+n_2)\ln|\hat{\Sigma}_1^{-1}\tilde{\Sigma}|$$

$$=N\ln|(n_1/N)I_q+(n_2/N)(D\hat{\Sigma}_1D')^{-1}\hat{\Sigma}_2+(n_1n_2/N^2)(D\hat{\Sigma}_1D')^{-1}ZZ'|$$

分别为Q的迹部分及行列式部分。不难看出，计算W，V之值时无须先求出混合估计$\tilde{\mu}$及$\tilde{\Sigma}$，因而Q确实可以作为不完整观测数据所含附加信息的显著性检验的检验统计量，而且是似然比类型的统计量。遗憾的是，Q的精确分布不易确定，为此，构造新的统计量

$$T_1^2=Z'(D\hat{\Sigma}_1D')^{-1}Z, \qquad T_2^2=\mathrm{tr}(D\hat{\Sigma}_1D_1')^{-1}\hat{\Sigma}_2$$

分别表述完整数据组与不完整观测组在均值与协方差阵方面的差异程度。两组观测数据的均值与协方差阵之间差异越大，则T_1^2与T_2^2越大，Q也将越大，表明不完整数据组所含附加信息越显著，可见，T_1^2与T_2^2可作为新的检验统计量。

定理3　当原假设（Ho:不完全观测所含附加信息不显著）成立时，统计量 T_1^2 与 T_2^2 分别服从Hotelling T^2 分布与广义Hotelling T^2 分布。

证明　显然　$\hat{\mu}_1 \sim N_p(\mu, \Sigma/n_1)$，$\hat{\mu}_2 \sim N_q(D\mu, D\Sigma D'/n_2)$

$\therefore$　$D\hat{\mu}_1 \sim N_q(D\mu, D\Sigma D'/n_1)$ 且与 $\hat{\mu}_2$ 独立

从而　　$Z \sim N_q(\mathbf{0}, (N/n_1n_2)\cdot D\Sigma D')$

另外　　$n_1\hat{\Sigma}_1 \sim W_p(n_1-1, \Sigma)$

$\therefore$　$n_1(D\hat{\Sigma}_1D') \sim W_q(n_1-1, D\Sigma D')$

且　$(D\hat{\Sigma}_1D')$ 与 Z 相互独立，从而

$$\frac{n_1n_2}{N}Z'\left(\frac{n_1D\hat{\Sigma}_1D'}{n_1-1}\right)^{-1}Z \sim T^2(q, n_1-1)$$

即

$$T_1^2 \sim \frac{N}{n_2(n_1-1)}T^2(q, n_1-1)$$

或者说　$$T_1^2 \sim \frac{Nq}{n_2(n_1-q)}F_{q,n_1-q}.$$

又由　　$n_1(D\hat{\Sigma}_1D') \sim W_q(n_1-1, D\Sigma D')$

$n_2\hat{\Sigma}_2 \sim W_q(n_2-1, D\Sigma D')$

且两者相互独立，及广义Hotelling T^2 分布定义[4]，即知

$$\mathrm{tr}(n_1D\hat{\Sigma}_1D')^{-1}(n_2\cdot\hat{\Sigma}_2) \sim \text{广义}T^2\text{分布}$$

即　　$$\frac{n_2}{n_1}T_2^2 \sim \text{广义}T^2\text{分布}.$$

当 Σ 已知时，Q 变为[5] $(n_1n_2/N)Z'(D\Sigma D')^{-1}Z \sim \chi_q^2$.

3　结束语

运用上述检验统计量可以对不完整观测组作附加信息的显著性检验：给定显著水平 α，由Hotelling T^2 分布及广义Hotelling T^2 分布可确定其临界值 $C(\alpha)$，若 $T_1^2+T_2^2 \geqslant C(\alpha)$，则表明所考察的不完整观测所含附加信息显著，应加以利用，且与完整观测共同使用求得 μ,Σ 的混合估计 $\tilde{\mu}$,$\tilde{\Sigma}$，进而作其它统计推断；若 $T_1^2+T_2^2 < C(\alpha)$，则说明不完整观测组所含附加信息不显著，可舍去不用。

值得一提的是：对于一组具有相同模式的不完整观测，必要时也可以逐一对各次观测运用上述方法作附加信息检验，利用其中附加信息显著的各次观测而舍弃不显著的各次观测，避免或者全部采用，或者全部舍弃的粗糙处理方法，使更科学、更合理，更充分有效地利用现有数据。

参 考 文 献

1 Edgett G L. Journal of the American Statistical Association, 1956, 51:122~131

2 Anderson T W. Journal of the American Statistical Association, 1957, 52:200~203

3 Orchard T, Woodbury M A. A missing information principle: theory and applications, Proceedings of the Sixth Anunal Berkeley Symposiuml, 1972. 697~715

4 Pillai K C S, Sampson P. Biometrika, 1959, 46:160~168

5 Smith W B, Riggs M W. Statistics & Probability Letters, 1984, 2:337~343

A Method for Testing the Added Information in Partial Multinormal Data

*Yu Jinhua** *Gao Luanxian*

Abstract A test procedure is developed for determining whether information contained in partial multinormal vectors is sufficient to significantly alter estimates of multivariate normal population parameters. The test will help the user in deciding whether a rather complicated estimation procedure should be undertaken.

Keywords added information, partial data, likelihood ratio testing

* Department of