Ann. Inst. Statist. Math.
38 (1986), Part A, 523-528

A CONSTRUCTION METHOD OF CERTAIN MATRICES REQUIRED IN THE MULTIVARIATE HETEROSCEDASTIC METHOD

HIROTO HYAKUTAKE

(Received Mar. 4, 1985; revised Apr. 25, 1985)

Summary

For statistical inference about several normal means, the heteroscedastic method was proposed by Dudewicz and Bishop (1979, *Optimizing Methods in Statistics*, Academic Press, 183–203). However, the practical application in the multivariate case was not possible because it had not been known how to construct the certain matrices required in the method. In this paper, a construction method of the matrices is given.

1. Introduction

We consider a statistical inference on means $\mu_1, \cdots, \mu_k$ of k p-variate normal populations $N_p(\mu_i, \Sigma_i)$, $\Sigma_i > 0$, $i=1, \cdots, k$, where μ_i's and Σ_i's are unknown, and Σ_i's are different. When $k=1$, Stein [7] gave the two-stage sampling scheme for the univariate case in which the power for a test, a confidence coefficient for a confidence interval, and etc. are completely free from the population variance. This sampling scheme was generalized to the multivariate case by Chatterjee [1]. Using this procedure, it is possible to give the heteroscedastic method proposed by Dudewicz and Bishop [2] to overcome the complexity arisen in many cases of comparison of several (k) normal populations with different variances or covariance matrices. This is available to multiple comparisons, construction of simultaneous confidence intervals, ranking and selection problems, and etc. However, in the multivariate case, it was not possible to use the heteroscedastic method in practice up to the present because although the existence of certain matrices was known, an algorithm for calculation of the matrices was not known. In this paper, we give the matrices required in the multivariate sampling procedure.

Keywords and phrases: Multivariate normal populations, heteroscedastic method, two-stage sampling scheme, matrix quadratic equation.

In the sampling procedure of the multivariate heteroscedastic method, we first take samples of size N_0 from each of k populations and compute

$$(1.1)\qquad \bar{\boldsymbol{x}}^{(i)}=\frac{1}{N_0}\sum_{r=1}^{N_0}\boldsymbol{x}_r^{(i)}\,,\qquad S_i=\frac{1}{\nu}\sum_{r=1}^{N_0}(\boldsymbol{x}_r^{(i)}-\bar{\boldsymbol{x}}^{(i)})(\boldsymbol{x}_r^{(i)}-\bar{\boldsymbol{x}}^{(i)})'\,,$$

for $i=1,\cdots,k$, where $\nu=N_0-1\geqq p$. Then for each i, we define N_i by

$$(1.2)\qquad N_i=\max\{N_0+p^2,\,[c\cdot\mathrm{tr}\,(TS_i)]+1\}\,,$$

where c is a positive constant, T is a $p\times p$ given positive definite (p.d.) matrix, and $[y]$ denotes the greatest integer not greater than y. Note that we use the same c and T for all $i=1,\cdots,k$. Next we take N_i-N_0 additional observations $\boldsymbol{x}_{N_0+1}^{(i)},\cdots,\boldsymbol{x}_{N_i}^{(i)}$ and construct the basic random vector $\boldsymbol{z}_i$ $(i=1,\cdots,k)$ in the following way:

We choose p matrices $A_l^{(i)}$: $p\times N_i=[\boldsymbol{a}_{l1}^{(i)},\cdots,\boldsymbol{a}_{lN_0}^{(i)},\boldsymbol{a}_{l,N_0+1}^{(i)},\cdots,\boldsymbol{a}_{lN_i}^{(i)}]$, $l=1,\cdots,p$, in such a way that

(I) $\boldsymbol{a}_{l1}^{(i)}=\cdots=\boldsymbol{a}_{lN_0}^{(i)}$,

(II) $A_l^{(i)}\boldsymbol{j}_{N_i}=\boldsymbol{e}_l$, where $\boldsymbol{j}_{N_i}$: $N_i\times 1=(1,\cdots,1)'$ and $\boldsymbol{e}_l$: $p\times 1=(0,\cdots,0,\overset{(l)}{1},0,\cdots,0)'$,

(III) $A^{(i)}A^{(i)\prime}=(1/c)T^{-1}\otimes S_i^{-1}$, where $A^{(i)}$: $p^2\times N_i=[A_1^{(i)\prime},A_2^{(i)\prime},\cdots,A_p^{(i)\prime}]'$.

Then we define

$$(1.3)\qquad \boldsymbol{z}_i\colon\ p\times 1=[\mathrm{tr}\,(A_1^{(i)}X^{(i)\prime}),\,\mathrm{tr}\,(A_2^{(i)}X^{(i)\prime}),\cdots,\mathrm{tr}\,(A_p^{(i)}X^{(i)\prime})]'\,,$$

where $X^{(i)}$: $p\times N_i=[\boldsymbol{x}_1^{(i)},\cdots,\boldsymbol{x}_{N_0}^{(i)},\boldsymbol{x}_{N_0+1}^{(i)},\cdots,\boldsymbol{x}_{N_i}^{(i)}]$.

The heteroscedastic inference method is based on the vectors $\boldsymbol{z}_1,\cdots,\boldsymbol{z}_k$. The distributions of $\boldsymbol{z}_i$ and the statistics based on these random vectors are fully independent of the population covariance matrices. Hyakutake and Siotani [4] and Hyakutake, Siotani, Li and Mustafid [5] gave the asymptotic approximations for their distributions.

Our purpose is to find the matrices $A^{(i)}=[A_1^{(i)\prime},\cdots,A_p^{(i)\prime}]'$, $i=1,\cdots,k$, which satisfy the conditions (I), (II) and (III), for practical application.

2. Main result

In this section, we give a method of constructing the matrices $A^{(i)}=[A_1^{(i)\prime},\cdots,A_p^{(i)\prime}]'$ which satisfy the conditions (I), (II) and (III). This construction is done independently for each population in the heteroscedastic method. Since a result from one population does not affect results from other populations, we can drop the suffix i without loss of generality.

First of all, we put $A=[A_1',\cdots,A_p']'=[A_0;B]$, where A_0: $p^2\times(N-m)=[\boldsymbol{a}_0,\cdots,\boldsymbol{a}_0]=\boldsymbol{a}_0\boldsymbol{j}_{N-m}'$, B is a $p^2\times m$ matrix, and $p^2\leqq m\leqq N-N_0$ by (I). From (II), $A\boldsymbol{j}_N=[A_0;B]\boldsymbol{j}_N=(N-m)\boldsymbol{a}_0+B\boldsymbol{j}_m=\boldsymbol{e}$, hence

$$a_0=\frac{1}{N-m}(e-Bj_m)\,, \tag{2.1}$$

where e: $p^2\times 1=(e_1',\cdots,e_p')'$. Substituting (2.1) into (III), the condition (III) is written in the term of B as

$$\frac{1}{c}T^{-1}\otimes S^{-1}=B\Big[I_m+\frac{1}{N-m}j_mj_m'\Big]B' -\frac{1}{N-m}ej_m'B'-\frac{1}{N-m}Bj_me'+\frac{1}{N-m}ee'\,, \tag{2.2}$$

which is a matrix quadratic equation. Let G be an $m\times m$ matrix satisfying

$$GG'=I_m+\frac{1}{N-m}j_mj_m'\,. \tag{2.3}$$

In general, the existence of such a decomposition of a p.d. or positive semi-definite (p.s.d.) matrix and its algorithm are well known (see e.g. appendix in Muirhead [6]). The matrix G may be a lower triangular matrix or a symmetric matrix. A solution for symmetric matrix is given by $G=I_m-(1/m)j_mj_m'\pm(\sqrt{N}/m\sqrt{N-m})j_mj_m'$. Letting $D=BG$, the equation (2.2) can be written in term of D as

$$(D-E)(D-E)'=\frac{1}{c}T^{-1}\otimes S^{-1}-\frac{1}{N}ee'\,, \tag{2.4}$$

where

$$E=\frac{1}{N-m}ej_m'(GG')^{-1}G=\frac{1}{N-m}ej_m'\Big(I_m-\frac{1}{N}j_mj_m'\Big)G=\frac{1}{N}ej_m'G\,.$$

To show the existence of the solution in the equation (2.4), it is necessary that $(1/c)T^{-1}\otimes S^{-1}-(1/N)ee'$ is a p.d. or p.s.d. matrix. This is shown by considering the inverse matrix of $(1/c)T^{-1}\otimes S^{-1}-(1/N)ee'$, which is given by

$$cT\otimes S+(cT\otimes S)ee'(cT\otimes S)/(N-e'(cT\otimes S)e)\,. \tag{2.5}$$

Since $c>0$, T is a p.d. matrix, and S is a p.d. matrix with probability one, the first term of (2.5) is a p.d. matrix. The second term of (2.5) is a p.s.d. matrix clearly because of $e'(cT\otimes S)e=c\cdot\mathrm{tr}\,(TS)<N$ by (1.2). Therefore the matrix in (2.5) is a p.d. matrix, and hence $(1/c)T^{-1}\otimes S^{-1}-(1/N)ee'$ is a p.d. matrix.

Now we can choose a matrix K: $p^2\times m$ such that

$$KK'=\frac{1}{c}T^{-1}\otimes S^{-1}-\frac{1}{N}ee'\,. \tag{2.6}$$

Hence

$$B=EG^{-1}+KG^{-1}=\frac{1}{N}ej'_m+KG^{-1} \tag{2.7}$$

and from (2.1),

$$\begin{aligned} A_0=a_0j'_{N-m}&=\frac{1}{N-m}\left\{e-\left(\frac{1}{N}ej'_m+KG^{-1}\right)j_m\right\}j'_{N-m} \\ &=\frac{1}{N}ej'_{N-m}-\frac{1}{N-m}KG^{-1}j_mj'_{N-m}\,. \end{aligned} \tag{2.8}$$

Thus we obtain

$$A=\frac{1}{N}ej'_N+KG^{-1}\left[-\frac{1}{N-m}j_mj'_{N-m};\ I_m\right]\,. \tag{2.9}$$

We summarize the above argument in the following.

THEOREM. *A matrix A satisfying the conditions* (I), (II) *and* (III) *is given by* (2.9), *where $G: m\times m$ and $K: p^2\times m$ are satisfying* (2.3) *and* (2.6), *respectively, and $p^2\leqq m\leqq N-N_0$.*

If we take $G=I_m-(1/m)j_mj'_m+(\sqrt{N}/m\sqrt{N-m})j_mj'_m$, the matrix A in (2.9) is written as

$$\begin{aligned} A=\frac{1}{N}ej'_N+K\Big[&-\frac{1}{\sqrt{N(N-m)}}j_mj'_{N-m};\ I_m-\frac{1}{m}j_mj'_m \\ &+\frac{\sqrt{N-m}}{m\sqrt{N}}j_mj'_m\Big]\,. \end{aligned} \tag{2.10}$$

It is noted that the matrix A is not unique since K and G are not unique.

3. Numerical examples

To give numerical examples of the previous result in the bivariate case, i.e. $p=2$, we use a part of the Iris Data discussed by Fisher [3]. We take the first-stage samples of size $N_0=10$, then the sample covariance matrix is

$$S=\begin{pmatrix}.5289 & .1933\\ .1933 & .1157\end{pmatrix}\,.$$

Now, if we take $c=1$ and $T=I_2$, the total sample size N defined in (1.2) is 14. Hence the additional (second-stage) sample size $N-N_0$ is 4 and $m=4$. Then the total observation matrix is given by

$$X=\begin{pmatrix}7.0 & 6.4 & 6.9 & 5.5 & 6.5 & 5.7 & 6.3 & 4.9 & 6.6 & 5.2 & 5.0 & 5.9 & 6.0 & 6.1\\ 3.2 & 3.2 & 3.1 & 2.3 & 2.8 & 2.8 & 3.3 & 2.4 & 2.9 & 2.7 & 2.0 & 3.0 & 2.2 & 2.9\end{pmatrix}.$$

Next if we take the lower triangular matrices K and G with positive diagonal satisfying (2.6) with $e=(1, 0, 0, 1)'$ and (2.3), respectively, the matrices satisfying (I), (II) and (III) are

$$A_1=\begin{pmatrix}-.13723\cdots-.13723 & 2.1580 & .07143 & .07143 & .07143\\ .10105\cdots .10105 & -3.7925 & 2.7820 & 0 & 0\end{pmatrix},$$

$$A_2=\begin{pmatrix}-.17652\cdots-.17652 & -.17651 & -.17651 & 2.1182 & 0\\ .15619\cdots .15619 & .12196 & .10953 & -3.6860 & 2.8926\end{pmatrix}.$$

Then the basic random vector is

$$z=(7.366, 5.352)'.$$

On the other hand, if we use the formula (2.10) with the same lower triangular matrix K as before, we obtain

$$A_1=\begin{pmatrix}-.11276\cdots-.11276 & 2.1751 & -.03129 & -.03129 & -.03129\\ .06790\cdots .06790 & -3.6812 & 2.9369 & .03123 & .03123\end{pmatrix},$$

$$A_2=\begin{pmatrix}-.18556\cdots-.18556 & -0.8535 & -.08535 & 2.1194 & -.08535\\ .14143\cdots .14143 & .07098 & .06192 & -3.5814 & 3.0313\end{pmatrix}.$$

Then

$$z=(7.314, 5.244)'.$$

From the above numerical results, we can see that the choice of a lower triangular or symmetric one as a solution of G will affect the basic random vectors, but little affect.

Acknowledgements

The author wishes to thank Professor M. Siotani and Professor Y. Fujikoshi, Hiroshima University, for their valuable comments and encouragement.

HIROSHIMA UNIVERSITY

REFERENCES

[1] Chatterjee, S. K. (1959). On an extension of Stein's two-sample procedure to the multi-normal problem, *Calcutta Statist. Assoc. Bull.*, 8, 121-148.

[2] Dudewicz, E. J. and Bishop, T. A. (1979). The heteroscedastic method, *Optimizing Methods in Statistics*, (ed. J. S. Rustagi), Academic Press, New York, 183-203.

[3] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, *Ann.*

Eugen., **7**, 179–188.

[4] Hyakutake, H. and Siotani, M. (1984). Distributions of some statistics in heteroscedastic method, *Technical Report*, No. 108, Statistical Research Group, Hiroshima University, Japan.

[5] Hyakutake, H., Siotani, M., Li, C. and Mustafid (1984). Distributions of some statistics in heteroscedastic inference method II: Tables of percentage points and power functions, *Technical Report*, No. 142, Statistical Research Group, Hiroshima University, Japan.

[6] Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*, John Wiley, New York.

[7] Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance, *Ann. Math. Statist.*, **16**, 243–258.