Two–Stage Procedures Under Nonnormality

Makoto Aoshima
University of Tsukuba, Institute of Mathematics
Ibaraki 305-8571, Japan
aoshima@math.tsukuba.ac.jp

Hirofumi Wakaki
Hiroshima University, Department of Mathematics
Hiroshima 739–8526, Japan
wakaki@math.sci.hiroshima-u.ac.jp

1. Introduction

Let $\boldsymbol{X}$, $\boldsymbol{X}_1$, $\boldsymbol{X}_2$,... be a sequence of i.i.d. random vectors with values in R^p. Let $\boldsymbol{\mu} = E(\boldsymbol{X})$, $\boldsymbol{\Sigma} = Cov(\boldsymbol{X})$ and $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ (> 0) be the eigen values of $\boldsymbol{\Sigma}$. Let X_j denote the j-th element of $\boldsymbol{X}$ and $\mu_{i_1 \cdots i_r}$ be the moment of $\boldsymbol{X}$ defined by $\mu_{i_1 \cdots i_r} = E(X_{i_1} \cdots X_{i_r})$. Similarly the corresponding cumulant of $\boldsymbol{X}$ is denoted by $\kappa_{i_1 \cdots i_r}$. Let K_r be the r-th order tensor whose $(i_1, ..., i_r)$-element is $\kappa_{i_1 \cdots i_r}$. Having recorded $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$, we consider the problem of constructing an ellipsoidal confidence region R_n for $\boldsymbol{\mu}$ such that (i) the maximum diameter of $R_n \leq 2d$ and (ii) $P(\boldsymbol{\mu} \in R_n) = 1 - \alpha$ for given d (> 0) and α $(0 < \alpha < 1)$.

Let us write $\overline{\boldsymbol{X}}_n = \sum_{i=1}^n \boldsymbol{X}_i / n$. When the population distribution is supposed to be $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if $\boldsymbol{\Sigma}$ were known, such confidence region is given by $R_n = \{\boldsymbol{\omega} \in R^p : n(\overline{\boldsymbol{X}}_n - \boldsymbol{\omega})' \boldsymbol{\Sigma}^{-1} (\overline{\boldsymbol{X}}_n - \boldsymbol{\omega}) \leq a_0\}$ along with the optimal fixed–sample size $n = a_0 \lambda_1 / d^2$. Here, a_0 denotes the upper α point of the chi–square distribution function with p d.f. When $\boldsymbol{\Sigma}$ is unknown, it is known that no confidence region to meet (i)–(ii) can be constructed on the basis of a fixed number of observations. Healy (1956) gave a solution to it by proposing a Stein–type (1945) two–stage procedure as follows: Take a pilot sample $\boldsymbol{X}_1, ..., \boldsymbol{X}_m$ of size m $(> p)$ fixed and compute $\boldsymbol{S}_m = (m-1)^{-1} \sum_{i=1}^m (\boldsymbol{X}_i - \overline{\boldsymbol{X}}_m)(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_m)'$. Let ℓ_m be the largest eigen value of $\boldsymbol{S}_m$. Then, the sample size n is estimated by

$$N = \max \left\{ m, \left[\frac{a_m \ell_m}{d^2} \right] + 1 \right\}, \tag{1}$$

where $a_m = p(m-1)(m-p)^{-1} F_{p,m-p}(\alpha)$ with $F_{a,b}(\alpha)$ denoting the upper α point of F distribution with d.f.s (a, b). Here, $[x]$ stands for the largest integer less than a real number x. According as (1), take the additional sample of size $N - m$ and compute $\overline{\boldsymbol{X}}_N$ by combining the initial sample and the additional sample. Finally, the required confidence region to meet (i)–(ii) is constructed as $R_N = \{\boldsymbol{\omega} \in R^p : N(\overline{\boldsymbol{X}}_N - \boldsymbol{\omega})' \boldsymbol{S}_m^{-1} (\overline{\boldsymbol{X}}_N - \boldsymbol{\omega}) \leq a_m\}$.

However, once the assumption about normality is broken, the constant a_m defined above would not work as a design constant any more. The purpose of this article is to study how robust Healy's solution is against general distributions and to reform it under general distributions so as to meet (i)–(ii), even though asymptotically, by giving a suitable modification to a_m.

2. Main results

We may assume that $\boldsymbol{\mu} = \boldsymbol{0}$, $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. Accordingly, the largest eigen value ℓ_m should be defined for $\boldsymbol{\Lambda}^{1/2} \boldsymbol{S}_m \boldsymbol{\Lambda}^{1/2}$ where $\boldsymbol{\Lambda} = diag(\lambda_1, ..., \lambda_p)$. We formally put the design constant as

$a_m = a_0\,(1 + m^{-1}\widehat{a}_1)$ where $\widehat{a}_1$ is an appropriate estimator of certain smooth function of $\boldsymbol{\Sigma}$, K_3 and K_4. The following assumptions are required for technical reasons.

(A0) Cramér's condition for the joint distribution of $(\boldsymbol{X},\ \boldsymbol{X}\boldsymbol{X}')$ holds;

$$\sup_{||\boldsymbol{t}_1||+\left(tr\left(\boldsymbol{T}_2'\boldsymbol{T}_2\right)\right)^{1/2}>b} E\left\{\exp\left(i(\boldsymbol{t}_1'\boldsymbol{X} + \boldsymbol{X}'\boldsymbol{T}_2\boldsymbol{X})\right)\right\} < 1 \qquad \text{for any } b > 0.$$

(A1) $\lambda_1 > \lambda_2$.

(A2) $E(||\boldsymbol{X}||^8) < \infty$, where $||\cdot||$ denotes the Euclidean norm.

(A3) $\lim_{m\to\infty} md^2 = c$ for some constant c such that $0 < c < a_0\lambda_1$.

Let $\boldsymbol{Z}_1 = m^{-1/2}\sum_{i=1}^m \boldsymbol{X}_i$, $\boldsymbol{V}_1 = m^{-1/2}\sum_{i=1}^m(\boldsymbol{X}_i\boldsymbol{X}_i' - \boldsymbol{I}_p)$ and $U = [a_m\ell_m/d^2] + 1 - a_m\ell_m/d^2$. Then, under the assumptions, we can expand $r = N/m$ as

$$r = \rho\left\{1 + \frac{1}{\sqrt{m}}\boldsymbol{e}_1'\boldsymbol{V}_1\boldsymbol{e}_1 + \frac{1}{m}\left(1 - (\boldsymbol{e}_1'\boldsymbol{Z}_1)^2 + \boldsymbol{e}_1'\boldsymbol{V}_1\boldsymbol{\Xi}\boldsymbol{V}_1\boldsymbol{e}_1 + \rho^{-1}U + \widehat{a}_1\right)\right\} + o_p(m^{-1}), \qquad (2)$$

where $\rho = n/m$, $\boldsymbol{e}_1 = (1, 0, ..., 0)'$ and $\boldsymbol{\Xi} = diag\,(0,\ \lambda_2(\lambda_1 - \lambda_2)^{-1}, ..., \lambda_p(\lambda_1 - \lambda_p)^{-1})$. Let us write that $\kappa_3^{<1>} = \sum_{i=1}^p\sum_{j=1}^p\sum_{k=1}^p \kappa_{ijk}^2$, $\kappa_3^{<2>} = \sum_{i=1}^p\sum_{j=1}^p\sum_{k=1}^p \kappa_{iij}\kappa_{jkk}$, $\eta_{31}^{<2>} = \sum_{j=1}^p \kappa_{11j}\kappa_{j11}$, $\eta_{32}^{<2>} = \sum_{j=1}^p\sum_{k=1}^p \kappa_{11j}\kappa_{jkk}$, $\xi_3^{<2>} = \sum_{i=1}^p\sum_{j=1}^p \lambda_j(\lambda_1 - \lambda_i)^{-1}\kappa_{1ij}^2$, $\kappa_4^{<1>} = \sum_{i=1}^p\sum_{j=1}^p \kappa_{iijj}$ and $\eta_{41}^{<1>} = \sum_{j=1}^p \kappa_{11jj}$. After several complicated calculations, we finally obtain an asymptotic expansion for the distibution of $T_N^2 = N(\overline{\boldsymbol{X}}_N - \boldsymbol{\mu})'\boldsymbol{S}_m^{-1}(\overline{\boldsymbol{X}}_N - \boldsymbol{\mu})$:

$$P\left(T_N^2 \le x\right) = G_p(x) + \frac{1}{m}\sum_{j=0}^3 \beta_j G_{p+2j}(x) + o(m^{-1}), \qquad (3)$$

where $G_p(\cdot)$ denotes the chi–square distribution function with p d.f. and

$$\begin{aligned}
\beta_0 =& \left\{2(2\rho^2 + 2\rho - 1)\kappa_3^{<1>} + 3(2\rho - 1)\kappa_3^{<2>} - 24\rho^2\eta_{31}^{<2>} + 24\rho^2\xi_3^{<2>}\right.\\
&\left. -3\rho(\rho^2 + \rho + 1)\kappa_4^{<1>} + 12\rho^2\eta_{41}^{<1>} - 6\rho^3p^2\right\}/(24\rho^3),\\
\beta_1 =& \left\{-2(2\rho^2 + 2\rho - 1)\kappa_3^{<1>} - 3(2\rho - 1)\kappa_3^{<2>} + 8\rho^2\eta_{31}^{<2>} - 8\rho^2\eta_{32}^{<2>}\right.\\
&\left. -8\rho^2\xi_3^{<2>} - 2\rho(\rho^2 - 3\rho - 1)\kappa_4^{<1>} - 4\rho^2\eta_{41}^{<1>} - 4\rho^3p\right\}/(8\rho^3),\\
\beta_2 =& \left\{2(2\rho - 1)\kappa_3^{<1>} - (4\rho^2 - 6\rho + 3)\kappa_3^{<2>} + 8\rho^2\eta_{32}^{<2>} + \rho(3\rho^2 - 5\rho - 1)\kappa_4^{<1>}\right.\\
&\left. +2\rho^3p(p + 2)\right\}/(8\rho^3),\\
\beta_3 =& \left\{2(4\rho^2 - 2\rho + 1)\kappa_3^{<1>} + 3(4\rho^2 - 2\rho + 1)\kappa_3^{<2>}\right\}/(24\rho^3).
\end{aligned}$$

By using this expansion formula, we observe that Healy's solution could be robust against the nonnormal distributions with the fourth–order cumulant positive and large. Cornish–Fisher expansion produces a suitable modification to the constant a_m and then a super efficiency phenomena are visible when the fourth–order cumulant is positive and large.

REFERENCES

Fujikoshi, Y. (1980). Asymptotic expansions for the distributions of the sample roots under nonnormality. *Biometrika,* **67**, 45–51.

Healy, W. C., Jr. (1956). Two–sample procedures in simultaneous estimation. *Ann. Math. Statist.,* **27**, 687–702.

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