김일성종합대학학보
수 학

주체109(2020)년
제66권 제2호

상관을 가지는 고차원평균벡토르들의 안정한 동일성검정

김 성 국

경애하는 최고령도자 **김정은**동지께서는 다음과 같이 말씀하시였다.

《우리는 과학기술을 발전시켜도 남들이 걸은 길을 따라만 갈것이 아니라 우리 과학자들의 애국충정과 우리 인민의 슬기와 민족적자존심을 폭발시켜 년대와 년대를 뛰여넘으며 비약해나가야 합니다.》

선행연구[1]에서는 상관을 가지는 고차원벡토르들의 동일성문제에 대하여 통계량을 제기하고 그것의 점근적성질을 연구하였다.

론문에서는 이상값들이 많은 경우 이상값들의 영향을 받지 않으면서도 안정한 통계량을 제기하고 점근분포를 밝혔으며 수치실험을 통하여 안정성을 확인하였다. 다변량해석에서 평균에 대한 동일성검정의 령가설과 대립가설이 $H_0:\mu_1=\mu_2$, $H_1:\mu_1\neq\mu_2$ 로 주어졌을 때 호텔링의 T^2 통계량은

$$T^2=(\bar{x}_1-\bar{x}_2)^{\mathrm{T}}S^{-1}(\bar{x}_1-\bar{x}_2)$$

이다. 만일 우연량들이 서로 독립이 아니고 표본공분산행렬 S 가 퇴화인 경우에는 우의 검정통계량을 리용할수 없다.

표본공분산행렬 S 가 퇴화로 되는 경우는 표본의 개수가 우연량의 수보다 작을 때이다. 표본의 개수가 우연량의 수보다 작거나 같은 경우를 고차원우연량 혹은 고차원설정이라고 표현한다. 일반적으로 고차원이라는것은 다변량해석에서 변수의 개수가 표본의 수보다 큰 경우이다. 고차원평균의 동일성검정에서는 고전적인 호텔링의 T^2 통계량을 리용할수 없다.

선행연구[1]에서는 우연벡토르의 공분산행렬이

$$\Sigma=V\Lambda V^{\mathrm{T}}+\sigma^2 I_p,\ \Lambda=\mathrm{diag}(\lambda_1,\cdots,\lambda_r),\ r\geq 1$$

인 경우에 평균의 동일성검정을 위한 통계량을 제안하고 그것의 분포를 밝혔다. 또한 선행연구[2]에서는 꼴모고로브거리에 기초한 검정통계량

$$\mathrm{M}_\Omega=\frac{n_1n_2}{n_1+n_2}\max\left\{\frac{\overline{Z}_1^2}{v_{11}},\ \cdots,\ \frac{\overline{Z}_p^2}{v_{pp}}\right\}$$

$$\Omega=\Sigma^{-1}=(v_{ij}),\ \overline{Z}=\Omega(\overline{X}_1-\overline{X}_2)$$

를 제안하여 평균의 동일성검정을 진행하였다.

선행연구들에서는 고차원적인 설정에서 공분산행렬의 퇴화를 피하고 고유값들을 리용한 통계량들을 제안함으로써 평균의 동일성검정을 진행할수 있게 하였다. 그러나 이러한 방법들은 이상값이 많은 자료에 대하여서는 평균과 분산의 이지러짐으로 하여 옳은 결과를 얻을수 없었다. 우리는 이상값들이 많은 경우 평균의 동일성검정을 위한 검정통계량을 제안하고 그것의 점근분포를 밝혔으며 수치실험을 통하여 개선된 성능을 확인하였다.

보조정리 1 $X_1,\cdots,X_n$: $N(\mu,\Sigma)$로부터 취한 표본들로서 독립이라고 하자.

$\exists C_0>0:C_0^{-1}\leq\lambda_{\min}(\Sigma)\leq\lambda_{\max}(\Sigma)\leq C_0$ 이라고 하자.

① $0<p<\infty,\ n\to\infty \Rightarrow \sqrt{n}(\overline{X}_{\varepsilon}-\mu)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}G(X_i)\xrightarrow{p}0$

여기서

$$\overline{X}_{\varepsilon}=\frac{\sum_{i=[n\varepsilon]+1}^{n-[n\varepsilon]}X_{(i)}}{n-2[n\varepsilon]},\ G(X_i)=\begin{cases}[F^{-1}(\varepsilon)-\theta]/(1-2\varepsilon), & X_i<F^{-1}(\varepsilon)\\ (X_i-\theta)/(1-2\varepsilon), & F^{-1}(\varepsilon)\le X_i\le F^{-1}(1-\varepsilon)\\ [F^{-1}(1-\varepsilon)-\theta]/(1-2\varepsilon), & X_i>F^{-1}(1-\varepsilon)\end{cases}$$

② $\varepsilon_i\in[b_1,\ b_2],\ 0<b_1<b_2<1/2,\ \varepsilon_i\to b_1;\ n,\ p\to\infty$ 일 때

$$\frac{\{\ln(pn)\}^7}{n}\to 0\Rightarrow\sqrt{n}(\overline{X}_{\varepsilon}-\mu)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}G(X_i)\xrightarrow{p}0$$

증명 선행연구[3]의 보조정리 2에 따르면 $X_1,\ \cdots,\ X_n$이 μ에 대하여 대칭인 밀도 F로부터 취한 우연표본, $0<\varepsilon<0.5$일 때

$$\overline{X}_{\varepsilon}=\mu+\frac{1}{n}\sum_{i=1}^{n}G(X_i)+o_p(n^{-1/2})$$

이다. 따라서 $\forall i\in\{1,\ \cdots,\ p\}$, $n\to\infty$일 때

$$\sqrt{n}(\overline{X}_{\varepsilon,i}-\mu_i)-\frac{1}{\sqrt{n}}\sum_{j=1}^{n}G(X_{ji})\xrightarrow{p}0 \tag{1}$$

이다. 따라서 $0<p<\infty,\ n\to\infty$일 때

$$\sqrt{n}(\overline{X}_{\varepsilon}-\mu)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}G(X_i)\xrightarrow{p}0$$

이다. $p\to\infty$일 때도 만족된다는것을 보자. $\forall i\in\{1,\ \cdots,\ p\}$에 대하여 식 (1)로부터 $n\to\infty$일 때

$$\sqrt{n}\left[\frac{1}{n-2[n\varepsilon_i]}\sum_{k=[n\varepsilon_i]+1}^{n-[n\varepsilon_i]}\frac{X_{(k)i}-\mu_i}{\sqrt{\sigma_{ii}}}\right]-\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\frac{G(X_{ki})}{\sqrt{\sigma_{ii}}}\xrightarrow{p}0$$

으로 된다. 이제 $\Phi^{-1}(\varepsilon)=\inf_x\{x:\Phi(x)\ge\varepsilon\}$, $\Phi^{-1}(1-\varepsilon)=\sup_x\{x:\Phi(x)<1-\varepsilon\}$으로 놓자. 그러면

$$\frac{G(X_{ki})}{\sqrt{\sigma_{ii}}}=\begin{cases}\dfrac{\Phi^{-1}(\varepsilon_i)}{(1-2\varepsilon_i)}, & (X_{ki}-\mu_i)<\Phi^{-1}(\varepsilon_i)\\ \dfrac{(X_{ki}-\mu_i)}{\{\sqrt{\sigma_{ii}}(1-2\varepsilon_i)\}}, & \Phi^{-1}(\varepsilon_i)\le(X_{ki}-\mu_i)\le\Phi^{-1}(1-\varepsilon_i)\\ \dfrac{\Phi^{-1}(1-\varepsilon_i)}{(1-2\varepsilon_i)}, & \dfrac{(X_{ki}-\mu_i)}{\sqrt{\sigma_{ii}}}\ge\Phi^{-1}(1-\varepsilon_i)\end{cases}$$

이다. 이때 $W_{ki}=(X_{ki}-\mu_i)/\sqrt{\sigma_{ii}}\sim N(0,\ 1)$이다. $n\to\infty$일 때

$$\sqrt{n}\left[\frac{1}{n-2[n\varepsilon_i]}\sum_{k=[n\varepsilon_i]+1}^{n-[n\varepsilon_i]}W_{(k)i}-\frac{1}{n}\sum_{k=1}^{n}G(W_{ki})\right]\xrightarrow{p}0 \tag{2}$$

이다. 여기서

$$G(W_{ki})=\begin{cases}\dfrac{\Phi^{-1}(\varepsilon_i)}{(1-2\varepsilon_i)}, & W_{ki}<\Phi^{-1}(\varepsilon_i)\\ \dfrac{W_{ki}}{(1-2\varepsilon_i)}, & \Phi^{-1}(\varepsilon_i)\le W_{ki}<\Phi^{-1}(1-\varepsilon_i)\\ \dfrac{\Phi^{-1}(1-\varepsilon_i)}{(1-2\varepsilon_i)}, & W_{ki}\ge\Phi^{-1}(1-\varepsilon_i)\end{cases}$$

이다. 따라서 식 (1)과 (2)는 동등하다.

이제

$$T_i=\sqrt{n}(\overline{X}_{\varepsilon_i,i}-\mu_i)-\frac{1}{\sqrt{n}}G(X_{ki}),\ Q_i=Q_n(\eta_i)=\sqrt{n}\left[\frac{1}{n-2[n\varepsilon_i]}\sum_{k=[n\varepsilon_i]+1}^{n-[n\varepsilon_i]}W_{ki}-\frac{1}{n}\sum_{k=1}^{n}G(W_{ki})\right]$$

로 놓자. $T=(T_1,\ \cdots,\ T_p)^{\mathrm{T}}=\sqrt{n}(\overline{X}_{\varepsilon}-\mu)-\{G(X_1)+\cdots+G(X_n)\}/\sqrt{n}$ 이라고 하자. 그러면 $\varepsilon\in[b_1,\ b_2]$에 대하여 과정 $Q_n(\varepsilon)$은 $[b_1,\ b_2]$우에서 오른쪽련속이며 왼쪽극한을 가지는 실함수공간에 놓이게 된다.

$n,\ p\to\infty$일 때 $\{\ln(pn)\}^7/n\to 0$으로부터 $n,\ p\to\infty$일 때 $\{\ln(p)\}^7/n\to 0$이여야 한다. 그러므로 여기서 모든 $p>M_1$과 $n\ge N$에 대하여 $p/e^{1/7}<1$인 $M_1,\ N_1>0$이 존재한다. 또한 $\varepsilon_i\to b_i$라는데로부터 모든 $\varepsilon_0>0$에 대하여 $M_2>M_1$이 존재하여 모든 $i>M_2$에 대하여 $|Q_n(\varepsilon_i)-Q(b_1)|<\varepsilon/2e^{N_1^{1/7}}$인 과정 $Q_n(\varepsilon)$의 오른쪽련속성으로부터 나온다. 그러면 모든 $n\ge N_1$과 $p>M_2$에 대하여

$$\sum_{i=M_2+1}^{p}|Q_n(\varepsilon_i)-Q_n(b_1)|<\frac{p\varepsilon_0}{(2e^{N_1^{1/7}})}<\frac{\varepsilon_0}{2}$$

이다. 그러므로 $n,\ p\to\infty$일 때

$$P\left\{\sum_{i=M_2+1}^{p}|Q_n(\varepsilon_i)-Q_n(b_1)|>\frac{\varepsilon_0}{2}\right\}\to 0 \tag{3}$$

이다. 임의의 i에 대하여 $n\to\infty$일 때 $Q_n(\varepsilon_i)\xrightarrow{p}0$이므로

$$P\left\{\sum_{i=1}^{M_2}|Q_n(\varepsilon_i)|>\frac{\varepsilon_0}{4}\right\}\to 0 \tag{4}$$

이다. 또한 $N_2>N_1$인 모든 $n>N_2$에 대하여

$$P\left\{|Q_n(b_1)|>\frac{\varepsilon_0}{(4e^{N_1^{1/7}})}\right\}<\frac{\varepsilon_0}{(4e^{N_1^{1/7}})}$$

이 성립한다. 따라서

$$P\left\{\sum_{i=M_2+1}^{p}|Q_n(b_1)|>\frac{\varepsilon_0}{4}\right\}\le p\times P\left\{|Q_n(b_1)|>\frac{\varepsilon_0}{(4e^{N_1^{1/7}})}\right\}<\frac{p\varepsilon_0}{(4e^{N_1^{1/7}})}<\frac{\varepsilon_0}{4}$$

$$\sigma_{\max} = \max(\sqrt{\sigma_{11}},\ \cdots,\ \sqrt{\sigma_{pp}}) < \infty$$

로 놓자. 식 (3)－(5)를 결합하면 $\sqrt{T_1^2 + \cdots + T_p^2} \le |T_1| + \cdots + |T_p|$ 이고 $|Q_i| \ge |T_i| / \sigma_{\max}$ 이므로

$$P\{\|T\| > \varepsilon_0 \sigma_{\max}\} \le P\left\{\sum_{i=1}^{p} |T_i| > \varepsilon_0 \sigma_{\max}\right\} \le P\left\{\sum_{i=1}^{p} |Q_i| > \varepsilon_0\right\} \le$$

$$\le P\left\{\sum_{i=1}^{M_2} |Q_i(\varepsilon_i)| > \frac{\varepsilon_0}{2}\right\} + P\left\{\sum_{i=M_2+1}^{p} |Q_n(\varepsilon_i) - Q_n(b_1)| > \frac{\varepsilon_0}{4}\right\} + P\left\{\sum_{i=M_2+1}^{p} |Q_n(b_1)| > \frac{\varepsilon_0}{4}\right\} \to 0$$

이다. 따라서 보조정리는 성립한다.(증명끝)

보조정리 2[4] $b > 0$ 인 상수, $B_n \ge 1$ 로서 $n \to \infty$ 일 때 무한히 커지는 수렬이라고 하자.

① $\forall i \in \{1,\ \cdots,\ p\},\ \sum_{k=1}^{n} \frac{E(U_{ki}^2)}{n} \ge b$

② $\forall i \in \{1,\ \cdots,\ p\},\ \gamma \in \{1,\ 2\},\ \sum_{k=1}^{n} \frac{E(|U_{kl}|^{2+\gamma})}{n} \le B_n^{\gamma}$

③ $\forall k \in \{1,\ \cdots,\ n\},\ \forall i \in \{1,\ \cdots,\ p\},\ E\left\{\exp\left(\frac{|U_{ki}|}{B_n}\right)\right\} \le 2$

조건 ①－③이 성립하면

$$\rho_n(A) = \sup_{A \in A^p} |P(S_n^U \in A) - P(S_n^V \in A)| \le C\left[\frac{B_n^2 (\ln(pn))^7}{n}\right]^{1/6}$$

이다. 여기서 C는 b에만 관계된다.

정리 1 $X_1,\ \cdots,\ X_n$: $N(\mu,\ \Sigma)$로부터 취한 표본들로서 독립이라고 하자.

$\exists C_0 > 0$: $C_0^{-1} \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le C_0$ 이라고 하자.

$\varepsilon_i \in [b_1,\ b_2],\ 0 < b_1 < b_2 < \frac{1}{2},\ \varepsilon_i \to b_1,\ n,\ p \to \infty$ 일 때

$$\frac{\{\ln(pn)\}^7}{n} \to 0 \Rightarrow \sqrt{n}(\overline{X}_{\varepsilon} - \mu) \xrightarrow{p} N(0,\ \Sigma_{\varepsilon})$$

증명 우선 보조정리 2의 조건 ①－③이 성립한다는것을 알수 있다.

A^{re} 를 R^p 의 부분모임 $A = \{w \in R^p : a_i \le w_i \le b_i,\ \forall i \in \{1,\ \cdots,\ p\}\}$ 들전부의 모임족이라고 하자. 여기서 $a_i = \frac{\{F_i^{-1}(\varepsilon_i) - \mu_i\}}{(1 - 2\varepsilon_i)},\ b_i = \frac{\{F_i^{-1}(1 - \varepsilon_i) - \mu_i\}}{(1 - 2\varepsilon_i)}$ 이다.

이제 $V_1,\ \cdots,\ V_n$ 을 $N_p[0,\ E\{G(X)\}\{G(X)\}^{\mathrm{T}}]$ 에 따르는 독립이고 동일분포하는 관측렬이라고 하자. 보조정리 2로부터

$$\sup_{A \in A^{re}} |P(S_n^{G(X)} \in AV) - P(S_n^V \in A)| \le C\left[\frac{B_n^2 \{\ln(pn)\}^7}{n}\right]^{1/6}$$

이다. $n,\ p \to \infty$ 일 때 $\{\ln(pn)\}^7 / n \to 0$ 이므로

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n} G(X_k) \to N[0,\ E_X\{G(X)\}\{G(X)\}^{\mathrm{T}}]$$

이다. 그러므로 $\sqrt{n}(\overline{X}_{\varepsilon}-\mu) \to N[E_X\{G(X)\}\{G(X)\}^{\mathrm{T}}]$이다.

이제 점근공분산행렬 $E_X\{G(X)\}\{G(X)\}^{\mathrm{T}}$를 계산하면 정리의 결과가 나온다.(증명끝)

정리 2 다음의 조건들이 성립한다고 하자.

① $\exists C_0 > 0 : C_0^{-1} \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le C_0$

② $\exists r_1 \in (0,\ 1) : \max_{1\le i<j\le p} |r_{ij}| \le r_1 < 1$

③ $\forall i \in \{1,\ \cdots,\ p\},\ \varepsilon_i \in [0,\ b_2],\ \varepsilon_i \xrightarrow{p\to\infty} 0 (b_2 \in (0,\ 0.5))$

④ $n,\ p \to \infty : \dfrac{(\ln(pn))^7}{n} \to 0$

⑤ $\sup_{1\le i\le p} \dfrac{1}{|F_i^{-1}(\varepsilon_i)|} = o\left(\dfrac{1}{\sqrt{\ln p}}\right)$

$n,\ p \to \infty$일 때

$$P_{H_0}\{M_{\Omega_\varepsilon} - 2\ln(p) + \ln(\ln(p)) \le x\} \to \exp\left[-\exp\left\{-\frac{[x+\ln(\pi)]}{2}\right\}\right]$$

이다. 정리 2는 정리 1과 선행연구[5]의 보조정리 1로부터 나온다.

정리 1에 의하여 제안검정량의 점근정규성이 나오며 이것을 리용하여 정리 2에서는 검정량의 점근분포에 대하여 주었다.

수치실험을 다음과 같이 진행한다.

$X_{1k} \sim N(\mu_1,\ \Sigma),\ X_{2k} \sim N(\mu_2,\ \Sigma)$ 이며 표본의 개수는 $n=100$ 이며 벡토르의 차원수는 $p=50$ 또는 100이다. 그리고 이상값들은 $N(5,\ 1):5\%$ 로서 상계쪽으로, $N(-5,\ 1):5\%$ 로서 하계쪽으로 주어 전체 자료의 10%를 이상값으로 주었다. 이상값은 표본별로 주었다.

실험결과에 대한 평가는 우의 과정을 1 000번 반복하며 유의수준을 5%로 주었을 때 검정능력을 평가하는 방법으로 진행한다.

실험을 진행한 결과는 다음과 같다.

비교결과	이상값이 없을 때		이상값이 10%일 때	
방법	선행연구[2]	제안	선행연구[2]	제안
능력	0.973	0.969	0.102	0.887

이와 같이 론문에서 제안한 방법은 이상값이 없을 때는 선행한 방법과 류사한 성능을 보여주었으며 이상값이 많을 때도 이상값이 없을 때와 류사한 안정한 성능을 보여주었다.

참 고 문 헌

[1] R. Wang et al.; J. Multivariage Anal., 167, 225, 2018.

[2] M. Avella-Medina et al.; Biometrika, 103, 1, 2016.

[3] A. DasGupta; Asymptotic Theory of Statistics and Probability, Springer, 1~208, 2008.

[4] V. Chernozhukov et al.; Ann. Probab., 45, 2309, 2017.
[5] T. Cai et al.; J. R. Stat. Soc. Ser. B Stat. Methodol, 76, 349, 2014.

주체108(2019)년 12월 15일 원고접수

Robust Identification Test of High Dimensional Mean Vectors with Correlation

Kim Song Guk

In case that there are many outliers in dataset, we propose robust identification test statistics for mean vector. For the independent vectors, test statistics is well-known, but if vectors have correlation, Hotelling statistics cannot be used. We propose robust test statistics of mean vector with correlation, prove the asymptotic distribution and show the numerical experiment results.

Keywords: high-dimensional vector, identification test of mean, asymptotic distribution