EDF 통계량을 이용한 다변량 정규성 검정[1)]

김 남 현[2)]

요 약

EDF에 근거한 Cramer-von Mises 형태의 통계량을 합교원리를 이용하여 다변량으로 일반화한다. 그리고 제안된 통계량의 귀무가설에서의 극한분포를 적절한 공분산 함수를 가진 가우스 과정의 적분의 형태로 표현하고 통계량의 근사적인 계산방법을 고려한다.

주요용어 : 다변량 정규분포, EDF, Cramer-von Mises 통계량, 가우스 과정

1. 서론

$X_1, \ldots, X_n$을 다변량 확률변수 X의 분포에서 관측한 확률표본이라고 하자. 또한 $N_d(\mu, \Sigma)$를 평균이 μ이고 공분산행렬이 Σ인 d-변량 정규분포라고 하자. 대부분의 다변량 통계기법은 자료가 다변량 정규분포에 따른다는 가정

$$H_0 : X\text{의 분포가 } N_d(\mu, \Sigma) \text{ 를 따른다 (} \mu\text{와 } \Sigma\text{는 미지).}$$

을 기반으로 한다. 이와 같은 이유로 많은 다변량 정규분포의 검정방법이 제안되어왔다.

위의 복합귀무가설 H_0의 검정을 위한 다변량 정규분포의 검정통계량은 기본적으로 affine invariance의 성질을 갖는 것이 바람직하다. 왜냐하면 X가 정규분포일 때 $AX+b$ ($b \in R^d$, $A \in R^{d \times d}$, A는 정칙행렬)도 역시 정규분포를 따르기 때문이다. 즉, $T_n = T_n(X_1, \ldots, X_n)$이 H_0의 검정통계량일 때

$$T_n(AX_1 + b, \ldots, AX_n + b) = T_n(X_1, \ldots, X_n)$$

을 만족해야 한다.

일변량 정규성 검정통계량을 위의 affine invariance의 성질을 갖는 다변량 정규성 검정 통계량으로 확장하는 한가지 일반적인 방법은 Roy(1953)의 합교 원리(Roy's union-intersection principle)를 이용하는 것이다. 이것은 X가 d-변량 정규분포를 따를 때, 모든 $c \in R^d$, $c \neq 0$에 대해서 $c'X$는 일변량 정규분포를 따른다는 사실에 기반을 두고 있다. 여기서 $'$는 전치(transpose)를 의미한다. $U_n(Z_1, \ldots, Z_n)$을 일변량 정규분포의 검정통계량이라고 하자. 그리고 U_n의 값이 클 때 귀무가설을 기각하고 U_n은 affine transformation ($aZ+b$)에 대해서 불변(invariant)이라고 하자. 그러면

---

1) 이 논문은 2004년도 한국학술진흥재단의 지원에 의하여 연구되었음. (KRF-2004-041-C00073)

2) (121-791) 서울시 마포구 상수동 72-1 홍익대학교 기초과학과 부교수
E-mail : nhkim@hongik.ac.kr

$$T_n(X_1, \ldots, X_n) = \max_{c \in R^d,\, c \neq 0} U_n(c'X_1, \ldots, c'X_n)$$

은 affine invariance의 성질을 갖는 '사영추적형태(projection pursuit type)'의 합리적인 통계량이고 이론적인 연구의 가치가 충분하다. 예를 들어 Malkovich & Afifi(1973), Fattorini(1986), Kim & Bickel(2003) 등에서 제안한 통계량이 이와 유사한 형태이다.

많은 다변량 정규성 검정통계량은 또한 X가 $N_d(\mu, \Sigma)$를 따르면 $(X-\mu)'\Sigma^{-1}(X-\mu)$는 χ^2_d 분포를 따른다는 사실을 이용한다. 이와 같은 통계량은 μ와 Σ를 추정량으로 대치한 제곱 반지름(squared radii)

$$D_{n,j} = \| Y_{n,j} \|^2 = (X_j - \overline{X}_n)' S_n^{-1} (X_j - \overline{X}_n)$$

의 함수이다. 여기서 $\overline{X}_n$는 $X_1, \ldots, X_n$의 평균벡터 $\overline{X}_n = n^{-1}\sum_{j=1}^{n} X_j$, S_n은 표본 공분산 행렬

$$S_n = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)(X_j - \overline{X}_n)'$$

이고 $Y_{n,j}$는 척도화된 잔차(scaled residuals)

$$Y_{n,j} = S_n^{-1/2}(X_j - \overline{X}_n)$$

이다. $D_{n,j}$는 같은 분포를 가지나 독립은 아니며, H_0에서는 근사적으로 χ^2_d 분포를 따른다.

$D_{n,1}, \ldots, D_{n,n}$의 확률 plot에 대해서 Gnanadesikan(1977)과 Singh(1993)가 언급하였다. $\widehat{G}_n(t)$,

$$\widehat{G}_n(t) = \frac{1}{n}\sum_{j=1}^{n} I(D_{n,j} \leq t)$$

를 $D_{n,1}, \ldots, D_{n,n}$의 EDF(경험분포함수, empirical distribution function)라고 하고 G_d를 χ^2_d의 분포함수라고 할 때 Malkovich & Afifi(1973), Koziol(1982)는 경험과정(empirical process) $g_n(t)$,

$$g_n(t) = \sqrt{n}(\widehat{G}_n(t) - G_d(t)), \quad 0 \leq t < \infty$$

를 고려하여 Cramér-von Mises 통계량

$$J_n = \int_0^{\infty} (g_n(t))^2 dG_d(t)$$

를 제안하였다. Malkovich & Afifi(1973)는 또한 Kolmogorov-Smirnov 통계량

$$KS = \max_t |\widehat{G}_n(t) - G_d(t)|$$

도 제안하였다. Malkovich & Afifi(1973)에 따르면 J_n이나 KS는 다른 통계량에 비해서 그리 좋은 검정력을 갖지 못하며 이는 $D_{n,j}$의 근사분포로 χ^2_d 분포보다 더 좋은 분포가 필요하다는 것을 시사한다.

본 논문에서는 일변량 Cramér-von Mises 형태의 통계량을 다변량으로 확장하기 위해서 $D_{n,1}, \ldots, D_{n,n}$의 EDF를 고려하는 대신 X의 일차결합 $c'X$의 EDF와 합교원리를 이

용하는 방법을 제안한다. Kim & Bickel(2003)에서 언급한 바와 같이 일변량 정규성 검정통계량을 합교원리를 이용하여 다변량으로 확장했을 때, 일변량 통계량에 나타나는 검정력의 양상이 다변량에도 유사하게 나타나는 것을 볼 수 있다. 따라서 일변량에서 비교적 우수한 검정력을 보여주는 Cramér-von Mises 형태의 통계량을 다변량으로 확장하는 것은 의미 있는 연구라고 생각된다.

우선 $F_n(x;\ c)$를 일차결합 $c'X$의 EDF,

$$F_n(x;\ c)=\frac{1}{n}\sum_{i=1}^{n} I(\ c'\ X_i \le x) \tag{1.1}$$

라고 하자. 복합귀무가설 H_0에서는 모평균벡터 μ와 공분산행렬 Σ가 미지이므로 이를 각각 $\overline{X_n}$와 S_n으로 추정한 분포함수(distribution function)를

$$\hat{F}(x;\ c)=\Phi\left(\frac{x-c'\overline{X_n}}{(\ c'\ S_n\ c)^{1/2}}\right) \tag{1.2}$$

라고 하면, 일차결합 $c'X$에 대한 Cramer-von Mises 통계량은

$$W_1^2(\ c)=n\int_{-\infty}^{\infty}(\ F_n(x;\ c)-\hat{F}(x;\ c))^2 d\hat{F}$$

이다. 이로부터 다변량 정규성 검정통계량으로

$$W_d^2=\max_{c,\ c\neq 0} W_1^2(\ c) \tag{1.3}$$

을 고려할 수 있다. 물론 W_d^2은 적절한 치환을 통하여

$$W_d^2=\max_{c,\ c\neq 0} n\int_0^1(\ \hat{G}_n(t;\ c)-t)^2 dt$$

으로 나타낼 수 있다. 여기서 $\hat{G}_n(t;\ c)$는

$$\hat{G}_n(t;\ c)=\frac{1}{n}\sum_{i=1}^{n} I\left(\frac{c'(\ X_i-\overline{X}_n)}{(\ c'\ S_n\ c)^{1/2}}\le \Phi^{-1}(t)\right)$$

을 나타낸다. 또한 W_d^2은 $Z(c;\ \overline{X}_n,\ S_n)=\Phi\left(\frac{c'(X-\overline{X}_n)}{(\ c'\ S_n\ c)^{1/2}}\right)$라고 할 때

$$W_d^2=\max_{\|c\|=1}\sum_{j=1}^{n}\left((Z(c;\overline{X_n},\ S_n))_{(j)}-\frac{2j-1}{2n}\right)^2+\frac{1}{12n}$$

로 계산할 수 있다. 여기서 $\|\cdot\|$은 L^2-norm을 의미하고, $()_{(j)}$는 ()안의 통계량의 순서통계량을 의미한다. 위의 통계량은 affine invariance의 성질을 만족하므로 c는 $\|c\|=1$로 제한해도 무방하다. 본 논문에서는 W_d^2의 근사분포와 구체적인 계산방법에 대해서 언급한다.

2. 제안된 통계량의 극한분포

정리 식(1.3)의 W_d^2은 H_0에서

$$W_d^2 \xrightarrow{d} \sup_{c,\ c\neq 0}\int_0^1(\ \hat{U}(t;\ c))^2 dt$$

을 만족한다. 여기서 $\widehat{U}(t;\ c)$는

$$\widehat{U}(t;\ c)=\ U(t;c)+\ c'\ Z\phi(\Phi^{-1}(t))+\frac{1}{2}\ c'\ V\ c\phi(\Phi^{-1}(t))\Phi^{-1}(t)\ ,$$

$Z\sim\ N_d(\mathbf{0},\ I)$, $V=(Z_{ij})_{d\times d}$, $Z_{ii}\ \underset{\sim}{i.i.d.}\ N_1(0,2)$, $Z_{ij}\ \underset{\sim}{i.i.d.}\ N_1(0,1)$, $i\neq j$, Z_{ij}도 독립이고 Z와 V도 역시 독립이다. 또한 $U(t;\ c)$는 공분산 함수

$$Cov(\ U(y_1,\ c_1),\ U(y_2,\ c_2))=P(\Phi(\ c_1X)\leq y_1\ \text{and}\ \Phi(\ c_2X)\leq y_2)-y_1y_2$$

를 갖는 가우스 과정(Gaussian process)이다.

증명 W_d^2은 affine invariant를 만족하므로 귀무가설에서의 분포를 고려할 때 $\mu=\mathbf{0}$, $\Sigma=I$, $\|\ c\ \|=1$이라고 가정해도 무방하다. 우선 식(1.1)과 (1.2)에 정의 된 $F_n(x;\ c)$과 $\widehat{F}(x;\ c)$에 대하여, 추정된 경험과정(estimated empirical process)을

$$\widehat{\beta}_n(x;\ c)=\sqrt{n}(\ F_n(x;\ c)-\widehat{F}(x;\ c))$$

라고 하자. 그러면

$$\widehat{\beta}_n(x;\ c)=\sqrt{n}(\ F_n(x;\ c)-\widehat{F}(x;\ c))$$

$$=\sqrt{n}\left(\ F_n(x;\ c)-\Phi\left(\frac{x-\ c'\mu}{(\ c'\Sigma c)^{1/2}}\right)\right)-\sqrt{n}\left(\widehat{F}(x;\ c)-\Phi\left(\frac{x-\ c'\mu}{(\ c'\Sigma c)^{1/2}}\right)\right)$$

이고, 경험과정(empirical process)

$$\sqrt{n}\left(\ F_n(x;\ c)-\Phi\left(\frac{x-\ c'\mu}{(\ c'\Sigma c)^{1/2}}\right)\right)=\sqrt{n}(\ F_n(x;\ c)-\Phi(x))$$

는 공분산 함수

$$Cov(\ U(y_1,\ c_1),\ U(y_2,\ c_2))=P(\Phi(\ c_1X)\leq y_1\ \text{and}\ \Phi(\ c_2X)\leq y_2)-y_1y_2$$

를 갖는 Gaussian process $U(y,\ c)$로 수렴함을 보일 수 있다. 여기서 X는 $X\sim N_d(\mathbf{0},\ I)$이다. 즉,

$$\sqrt{n}(\ F_n(x;\ c)-\Phi(x))\xrightarrow{d}\ U(\Phi(x);\ c)$$

임이 성립한다. 이는 Dudley(1978) 또는 Massart(1989)를 이용하여 보일 수 있다. 또한

$$\sqrt{n}(\widehat{F}(x;\ c)-\Phi(x))=\sqrt{n}\left(\Phi\left(\frac{x-\ c'\overline{X_n}}{(\ c'\ S_n\ c)^{1/2}}\right)-\Phi(x)\right)$$

$$=\sqrt{n}(\ c'\overline{X_n})(-\phi(x))+\sqrt{n}\left[(\ c'\ S_n\ c)^{1/2}-(\ c'\ I\ c)^{1/2}\right](-x\phi(x))+o_p(1)$$

이고

$$\sqrt{n}(\ c'\overline{X_n})\xrightarrow{d}\ c'\ Z,\ \ Z\sim\ N(\mathbf{0},\ I)$$

$$\sqrt{n}((\ c'\ S_n\ c)^{1/2}-(\ c'\ I\ c)^{1/2})\xrightarrow{d}\frac{1}{2}\ c'\ Vc$$

임이 성립한다. 여기서 V는

$$V=(Z_{ij})_{d\times d},\ Z_{ii}\ \underset{\sim}{i.i.d.}\ N(0,2),\ Z_{ij}\ \underset{\sim}{i.i.d.}\ N(0,1),\ i\neq j$$

이다. 따라서

$$\sqrt{n}(\hat{F}(x;c)-\Phi(x)) \xrightarrow{d} c' Z(-\phi(x))+\frac{1}{2} c' V c(-x\phi(x))$$

이다. 이를 통하여 추정된 경험과정 $\hat{\beta}_n$은

$$\hat{\beta}_n(x;c) \xrightarrow{d} \hat{U}(\Phi(x);c) = U(\Phi(x);c)+ c' Z\phi(x)+\frac{1}{2} c' V c(x\phi(x))$$

이고, 이로부터

$$W_d^2 \xrightarrow{d} \max_{c,\, c\neq 0} \int_0^1 (\hat{U}(t;c))^2 dt$$

가 성립한다.

□

3. 제안된 통계량의 계산방법

식(1.3)과 같이 사영추적형태(projection pursuit type)로 정의되어 있는 통계량의 단점은 해석적인 최대값을 갖는 벡터 c를 찾아내는 것이 간단하지 않다는 것이다. 이를 위해서는 Fang & Wang(1993)이 연구한 NTM(number theoretic methods)을 이용하는 것도 한 가지 방법이다. 이는 충분히 큰 k에 대해서 단위 d-차원 구면(unit d-sphere)

$$U^d=\{c\in R^d: \|c\|=1\}$$

에서 균일하게 퍼져 있는 $c_1, \ldots, c_k$를 이용하여

$$W_d^2 \approx \max_{1\le j\le k} W_d^2(c_j)$$

로 W_d^2을 근사적으로 계산하는 방법이다. NTM에 대한 자세한 내용은 Fang & Wong(1993)을 참고로 한다.

또한 c를 자료 $X_1, \ldots, X_n$에 의존하도록 택하는 방법, 예를 들면 $c_l = S_n^{-1/2}(X_l - \overline{X_n})/\| S_n^{-1/2}(X_l - \overline{X_n})\|$, $l=1,\ldots,n$으로 택하는 방법도 고려할 수 있다. 다시 말해서 단위 d-차원 구면 $U^d=\{c\in R^d: \|c\|=1\}$의 모든 벡터 c에 대해서 최대를 고려하는 대신에 EDF가 근사적으로 U^d에서의 균일분포를 따르는 표준화된 척도화 잔차(normalized scaled residuals)에 대해서 최대를 고려하자는 것이다. Fattorini(1986), Kim(2004a), Kim(2004b)는 이러한 방법 또는 이와 유사한 방법으로 합교원리를 이용하여 다변량으로 확장한 통계량을 근사적으로 구하였다.

참고문헌

Dudley, R. M. (1978). Central limit theorems for empirical measures. *The Annals of Statistics*, 6, 899-929.

Fang, K. T., and Wang, Y. (1993). *Number-theoretic methods in statistics*. Monographs on statistics and applied probability. Chapman and Hall, London.

Fattorini, L. (1986). Remarks on the use of the Shapiro-Wilk statistic for testing multivariate normality, *Statistica* 46, 209-217.

Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations.* Wiley, New York.

Kim, N. (2004a). An approximate Shapiro-Wilk statistic for testing multivariate normality. *The Korean Journal of Applied statistics*, 17, 35-37.

Kim, N. (2004b). Remarks on the use of multivariate skewness and kurtosisi for testing multivarite normality. *The Korean Journal of Applied statistics*, 17, 507-518.

Kim, N. and Bickel, P. J. (2003). The limit distribution of a test statistic for bivariate normality. *Statistica Sinica*, 13, 327-349.

Koziol, J. A. (1982). A class of invariant procedures for assessing multivariate normality. *Biometrika*, 69, 423-427.

Malkovich, J. F., and Afifi, A. A. (1973). On tests for multivariate normality. *Journal of the American Statistical Association*, 68, 176-179.

Massart, P. (1989). Strong approximation for multivarite empirical and related processes, via KMT constructions. *The Annals of probability*, 17, 266-291.

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Annals of Mathematical Statistics*, 24, 220-238.

Singh, A. (1993). Omnibus robust procedures for assessment of multivariate normality and detection of multivariate outliers. In : Multivariate Environmental Statistics (G. P. Patil and C. R. Rao, eds.) North-Holland, Amsterdam, 445-488.