Proceedings of the Autumn Conference in 2001
and the 30th Anniversary of the Korean Statistical Society

Projection Pursuit을 이용한 이변량 정규분포의 검정

김 남현[1)]

요약

projection pursuit을 이용하여 이변량 정규분포의 적합도 검정을 위한 통계량을 제안한다. 기본적인 생각은 이변량 정규분포의 가정하에 표준정규분포를 갖는 모든 선형조합을 고려하여 이들의 순서통계량과 이론적인 분위수를 비교하는 것이다. 이와 같이 제안된 통계량은 선형변환에 대해서 불변(invariant)이다. 본 논문에서는 제안된 통계량의 극한분포를 적절한 Gaussian process의 적분으로 표현한다.

주요용어 : 이변량 정규분포, 적합도검정, Gaussian process, projection pursuit

1. 서론

다변량 정규분포의 검정에 대해서는 여러 가지 방법이 제안되어왔다. 일반적인 방법에 대해서는 Gnanadesikan(1977), Mardia(1980), Cox와 Small(1978)과 D'Agostino와 Stephens(1987, 9.7절)에 잘 설명되어있다. Mardia(1970, 1974, 1975), Mardia와 Foster(1983), Malkovich와 Afifi(1973), Marchado(1983)는 다변량 왜도와 첨도를 제안하고 이들을 이용하여 다변량 정규분포를 검정하는 방법을 고려하였다. 그리고 Horsewell과 Looney(1992)는 다변량 왜도와 첨도를 이용한 여러 가지 검정법을 비교하였다.

$\boldsymbol{X}_1=(X_{11},X_{21})^T,\ldots,\boldsymbol{X}_n=(X_{1n},X_{2n})^T$을 이변량 확률변수 $\boldsymbol{X}=(X_1,X_2)^T$의 분포에서 관측한 확률표본이라고 하자. 여기서 T는 전치행렬을 의미한다. 또한 평균이 μ_1,μ_2, 분산이 σ_1^2, σ_2^2, 상관계수가 ρ인 이변량 정규분포를 $BVN(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$라고 하자.

Kim(1997)에서는 이변량 정규분포의 단순귀무가설을 검정하기 위한 통계량을 생각해 보고 이의 극한분포를 구하였다. 즉, 단순귀무가설

$$H_0^s:\ \boldsymbol{X}\text{의 분포가 } BVN(0,0,1,1,\rho),\ \rho\text{는 기지, 를 따른다.}$$

를 검정하기 위하여 통계량

$$P_n^{\,0}=\sup_{c_1,\,c_2\ni c_1^2+c_2^2+2\rho c_1c_2=1}\sum_{i=1}^{n}\left\{(c_1X_1+c_2X_2)_{(i)}-\Phi^{-1}\left(\frac{i}{n+1}\right)\right\}^2 \tag{1.1}$$

를 고려하였다. 여기서 $(\cdot)_{(i)}$는 괄호 안의 확률변수의 i번째 순서통계량을 의미한다. Projection pursuit은 projection index라고 불리는 어떤 목적함수를 최대로 하는 저차원의 projection을 찾고자 한다. 식(1.1)의 $P_n^{\,0}$는 일종의 projection index로 이해될 수 있다. projection pursuit에 대한 일반적인 개념은 Huber(1985)에 잘 설명되어있다. 이변량 분포 (X_1,X_2)가 이변량 정규분포를 따른다는 것은 X_1,X_2의 모든 선형조합 $c_1X_1+c_2X_2$이 정규분포를 따른다는 것과 필요충분조건이므로, 하나의 선형조합에서라도 비정규성이 발견되면 이변량 정규분포의 가정도 기각될 수 밖에 없다. 식(1.1)의 $P_n^{\,0}$-통계량은 단순귀무가설 H_0^s

1) 121-791 서울시 마포구 상수동 72-1 홍익대학교 기초과학과 조교수.

하에서 표준정규분포를 갖는 모든 가능한 선형조합을 조사하는 것이다. 만일 단순귀무가설 H_0^s가 사실이 아니라면, P_n^0의 최대치는 가장 정규분포와 거리가 먼 선형조합에서 나타날 것이고 이 선형조합이 관심의 대상이 될 것이다.

그러나 실제적으로는 복합귀무가설

H_0: X의 분포가 $BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho$는 미지, 를 따른다.

를 검정하는 경우가 좀 더 일반적이다. 따라서 식(1.1)의 P_n^0-통계량을 복합귀무가설 H_0에서의 검정통계량으로 일반화하면 자연스럽게

$$P_n = \sup_{c_1, c_2} \sum_{i=1}^{n} \left\{ \frac{(c_1X_1 + c_2X_2)_{(i)} - (c_1\overline{X_1} + c_2\overline{X_2})}{sd(c_1X_1 + c_2X_2)} - \Phi^{-1}\left(\frac{i}{n+1}\right) \right\}^2 \quad (1.2)$$

과 같은 통계량을 고려할 수 있다. 여기서 $\overline{X_k}$는 표본평균, $\widehat{\sigma_k^2}$은 표본분산이고, $\hat{\rho}$은 표본상관계수이다. 그리고

$$sd^2(c_1X_1 + c_2X_2) = c_1^2\widehat{\sigma_1^2} + c_2^2\widehat{\sigma_2^2} + 2c_1c_2\hat{\rho}\widehat{\sigma_1}\widehat{\sigma_2}$$

이다. P_n은 벡터 합과 정칙 행렬 곱에 대해서 불변(invariant)이다. 따라서 귀무가설 H_0하에서 P_n의 분포는 모수 μ_1, μ_2, σ_1^2, σ_2^2, ρ에 의존하지 않는다.

2. P_n-통계량의 극한분포

이 절에서는 복합귀무가설 H_0하에서 P_n의 근사통계량의 극한분포를 Gaussian process의 적분의 형태로 표현하고자 한다.

정리 1. P_n의 근사통계량 P_n^T를

$$P_n^T := \sup_{c_1, c_2} \sum_{i=I_n}^{n-I_n} \left\{ \frac{(c_1X_1 + c_2X_2)_{(i)} - (c_1\overline{X_1} + c_2\overline{X_2})}{sd(c_1X_1 + c_2X_2)} - \Phi^{-1}\left(\frac{i}{n+1}\right) \right\}^2 \quad (2.1)$$

라고 하고

$$a_n^T := \frac{1}{n} \sum_{i=I_n}^{n-I_n} \left(\frac{i}{n+1}\right)\left(1 - \frac{i}{n+1}\right) / \phi^2\left(\Phi^{-1}\left(\frac{i}{n+1}\right)\right)$$

라고 하자. $I_n/n = n^{-\delta}$, $0 < \delta < 1/8$일 때 복합귀무가설 H_0하에서 통계량 P_n^T는 다음과 같은 극한분포를 갖는다 :

$$P_n^T - a_n^T \xrightarrow{d} \sup_{\theta \in [0, 2\pi)} \left[\int_0^1 \frac{B^2(y, \theta) - y(1-y)}{\phi^2(\Phi^{-1}(y))} dy - \left(\int_0^1 \frac{B(y, \theta)}{\phi(\Phi^{-1}(y))} dy \right)^2 - \left(\int_0^1 \frac{B(y, \theta)}{\phi(\Phi^{-1}(y))} \Phi^{-1}(y) dy \right)^2 \right]. \quad (2.2)$$

여기에서 $B(y, \theta)$는 공분산 함수

$$E(B(y_1, \theta_1)B(y_2, \theta_2))$$
$$= \Pr\left(Z_1 \sin\theta_1 + Z_2 \cos\theta_1 \leq \Phi^{-1}(y_1) \text{ and } Z_1 \sin\theta_2 + Z_2 \cos\theta_2 \leq \Phi^{-1}(y_2)\right) - y_1 y_2$$

를 가진 Gaussian process이고 Z_1, Z_2는 표준정규분포를 따르는 확률변수이다.

증명. 식(1.2)의 P_n은 벡터합과 정칙행렬곱에 대해서 불변이므로, 귀무가설 H_0를

$$H_0 : \boldsymbol{X}\text{의 분포는 } BVN(0,0,1,1,0)\text{이다.}$$

로 가정해도 무방하다. $Q_n(y, \boldsymbol{c})$를 $c_1X_{1i}+c_2X_{2i}$의 표본분위함수라고하고 $c_1X_{1i}+c_2X_{2i}$에 대해서

$$Q_n(y,\ \boldsymbol{c}) := \begin{cases} (c_1X_1+c_2X_2)_{(k)}, & \frac{k-1}{n+1} < y \le \frac{k}{n+1},\ k=1,\ldots,n \\ (c_1X_1+c_2X_2)_{(n)}, & \frac{n}{n+1} < y \le 1 \end{cases}$$

과 같이 정의하자. $\boldsymbol{c}=(c_1, c_2)^T$일 때

$$\hat{\sigma}^2(\boldsymbol{c}) := c_1^2\,\hat{\sigma}_1^{\,2} + c_2^2\,\hat{\sigma}_2^{\,2} + 2\hat{\rho}c_1c_2\hat{\sigma}_1\hat{\sigma}_2 = sd^2(c_1X_1+c_2X_2),$$

$$\widetilde{Q_n}(y,\ \boldsymbol{c}) := (Q_n(y,\ \boldsymbol{c}) - (c_1\overline{X_1} + c_2\overline{X_2}))/\hat{\sigma}(\boldsymbol{c})$$

라고 하자. 그러면 양끝이 절단된(truncated) 식(2.1)의 P_n^T는

$$P_n^T = \sup_{c,\ c\neq 0} \int_{n^{-\delta}}^{1-n^{-\delta}} n(\widetilde{Q_n}(y,\ \boldsymbol{c}) - \Phi^{-1}(y))^2 dy,\quad 0<\delta<1/8,$$

과 같이 쓸 수 있다.

$$\cos\theta = c_1/(c_1^2+c_2^2)^{1/2},\quad \sin\theta = c_2/(c_1^2+c_2^2)^{1/2}$$

라고 하고

$$Q_n(y,\theta) := Q_n(y, (\cos\theta, \sin\theta)),$$

$$\widetilde{Q_n}(y,\theta) := \widetilde{Q_n}(y, (\cos\theta, \sin\theta)),$$

$$\hat{\sigma}^2(\theta) := \hat{\sigma}^2((\cos\theta, \sin\theta))$$

라고 하면 $\widetilde{Q_n}(y,\ \boldsymbol{c}) - \Phi^{-1}(y)$는 다음과 같이 쓸 수 있다 :

$$\begin{aligned} &\widetilde{Q_n}(y,\ \boldsymbol{c}) - \Phi^{-1}(y) \\ &= Q_n(y,\theta) - \Phi^{-1}(y) + \Phi^{-1}(y)(1-\hat{\sigma}(\theta)) - (\overline{X_1}\cos\theta + \overline{X_2}\sin\theta) \\ &\quad + (Q_n(y,\theta) - \Phi^{-1}(y))\frac{1-\hat{\sigma}(\theta)}{\hat{\sigma}(\theta)} + \Phi^{-1}(y)\frac{(1-\hat{\sigma}(\theta))^2}{\hat{\sigma}(\theta)} \\ &\quad - (\overline{X_1}\cos\theta + \overline{X_2}\sin\theta)\frac{1-\hat{\sigma}(\theta)}{\hat{\sigma}(\theta)}\ . \end{aligned}$$

따라서 식(2.1)의 P_n^T는

$$\begin{aligned} &|P_n^T - M_n^T| \\ &:= \Big| \sup_{\theta\in[0,\,2\pi)} \int_{n^{-\delta}}^{1-n^{-\delta}} n(\widetilde{Q_n}(y,\theta) - \Phi^{-1}(y))^2 dy \\ &- \sup_{\theta\in[0,\,2\pi)} \int_{n^{-\delta}}^{1-n^{-\delta}} n(Q_n(y,\theta) - \Phi^{-1}(y) + \Phi^{-1}(y)(1-\hat{\sigma}(\theta)) - (\overline{X_1}\cos\theta + \overline{X_2}\sin\theta))^2 dy \Big| \end{aligned}$$

$$\leq C \sup_{\theta\in[0,2\pi)} \int_{n^{-\delta}}^{1-n^{-\delta}} n(Q_n(y,\theta)-\Phi^{-1}(y))^2 dy \sup_{\theta\in[0,2\pi)} \frac{(\hat{\sigma}(\theta)-1)^2}{\hat{\sigma}^2(\theta)}$$
$$+ \int_0^1 (\Phi^{-1}(y))^2 dy \sup_{\theta\in[0,2\pi)} \frac{n(\hat{\sigma}(\theta)-1)^4}{\hat{\sigma}^2(\theta)}$$
$$+ \sup_{\theta\in[0,2\pi)} n(\overline{X_1}\cos\theta+\overline{X_2}\sin\theta)^2 \sup_{\theta\in[0,2\pi)} \frac{(\hat{\sigma}(\theta)-1)^2}{\hat{\sigma}^2(\theta)}\Bigg) \tag{2.3}$$

이 된다. 한편

$$\sqrt{n}\,(\overline{X_1}\cos\theta+\overline{X_2}\sin\theta) = \int_0^1 \sqrt{n}(Q_n(y,\theta)-\Phi^{-1}(y))\,dy \tag{2.4}$$

이고

$$\hat{\sigma}^2(\theta) = \int_0^1 (Q_n(y,\theta)^2 dy - \left(\int_0^1 Q_n(y,\theta)dy\right)^2$$
$$= 1 + 2\int_0^1 (Q_n(y,\theta)-\Phi^{-1}(y))\Phi^{-1}(y)\,dy$$
$$+ \int_0^1 (Q_n(y,\theta)-\Phi^{-1}(y))^2 dy - \left(\int_0^1 Q_n(y,\theta)-\Phi^{-1}(y)\,dy\right)^2$$

이므로

$$\sqrt{n}\,(\hat{\sigma}(\theta)-1) \approx \sqrt{n}\,(\hat{\sigma}^2(\theta)-1)/2$$
$$\approx \sqrt{n}\int_0^1 (Q_n(y,\theta)-\Phi^{-1}(y))\Phi^{-1}(y)\,dy \tag{2.5}$$

이다. M_n^T의 적분 안의 제곱을 전개하고 식(2.4), (2.5)를 이용하면 어렵지 않게

$$\Bigg| M_n^T - \sup_{\theta\in[0,2\pi)} \Bigg\{ \int_{n^{-\delta}}^{1-n^{-\delta}} n(Q_n(y,\theta)-\Phi^{-1}(y))^2 dy$$
$$-\left(\int_{n^{-\delta}}^{1-n^{-\delta}} \sqrt{n}(Q_n(y,\theta)-\Phi^{-1}(y))dy\right)^2$$
$$-\left(\int_{n^{-\delta}}^{1-n^{-\delta}} \sqrt{n}(Q_n(y,\theta)-\Phi^{-1}(y))\Phi^{-1}(y)\,dy\right)^2\Bigg\}\Bigg|$$
$$= \Bigg| M_n^T - \sup_{\theta\in[0,2\pi)} \Bigg\{ \int_{n^{-\delta}}^{1-n^{-\delta}} \frac{\rho_n^2(y,\theta)}{\phi^2(\Phi^{-1}(y))}\,dy$$
$$-\left(\int_{n^{-\delta}}^{1-n^{-\delta}} \frac{\rho_n(y,\theta)}{\phi(\Phi^{-1}(y))}\,dy\right)^2 - \left(\int_{n^{-\delta}}^{1-n^{-\delta}} \frac{\rho_n(y,\theta)}{\phi(\Phi^{-1}(y))}\Phi^{-1}(y)\,dy\right)^2\Bigg\}\Bigg|$$
$$\xrightarrow{p} 0$$

임을 보일 수 있다. 여기서 $\rho_n(y,\theta)$는

$$\rho_n(y,\theta) := \phi(\Phi^{-1}(y))\sqrt{n}(Q_n(y,\theta)-\Phi^{-1}(y))$$

이다. 따라서 주어진 결과를 보이기 위해서 (적절한 공간에서)

$$\Big| \sup_{\theta\in[0,2\pi)} \left\{ \int_{n^{-\delta}}^{1-n^{-\delta}} \frac{\rho_n^2(y,\theta)}{\phi^2(\Phi^{-1}(y))}\,dy - \left(\int_{n^{-\delta}}^{1-n^{-\delta}} \frac{\rho_n(y,\theta)}{\phi(\Phi^{-1}(y))}\,dy \right)^2 - \left(\int_{n^{-\delta}}^{1-n^{-\delta}} \frac{\rho_n(y,\theta)}{\phi(\Phi^{-1}(y))}\,\Phi^{-1}(y)\,dy \right)^2 \right\}$$

$$- \sup_{\theta\in[0,2\pi)} \left\{ \int_{n^{-\delta}}^{1-n^{-\delta}} \frac{B_n^2(y,\theta)}{\phi^2(\Phi^{-1}(y))}\,dy - \left(\int_{n^{-\delta}}^{1-n^{-\delta}} \frac{B_n(y,\theta)}{\phi(\Phi^{-1}(y))}\,dy \right)^2 - \left(\int_{n^{-\delta}}^{1-n^{-\delta}} \frac{B_n(y,\theta)}{\phi(\Phi^{-1}(y))}\,\Phi^{-1}(y)\,dy \right)^2 \right\} \Big|$$

$$:= |S\rho_n - SB_n|$$

$$\xrightarrow{p} 0$$

임을 보이면 될 것이고 이것은 Kim(1997)의 정리 10과 정리 11을 이용하면 쉽게 보일 수 있다. 식(2.3)의 우변의 각 항은 모두 $o_p(1)$ 이므로 P_n^T도 역시

$$|P_n^T - SB_n| \xrightarrow{p} 0$$

이 성립한다.

마지막으로 식(2.2)의 우변의 존재성에 대해서 고려해야한다.

$$\sup_{\theta\in[0,2\pi)} \int_0^1 \frac{B^2(y,\theta) - y(1-y)}{\phi^2(\Phi^{-1}(y))}\,dy$$

의 존재성은 Kim(1997)에서 고려하였다. 또한

$$\sup_{\theta\in[0,2\pi)} \sqrt{n}\,(\overline{X_1}\cos\theta + \overline{X_2}\sin\theta) \le \sqrt{n}\sqrt{\overline{X_1}^2 + \overline{X_2}^2} = O_p(1)$$

이므로

$$\sup_{\theta\in[0,2\pi)} \left(\int_0^1 \frac{B(y,\theta)}{\phi(\Phi^{-1}(y))}\,dy \right)^2 < \infty$$

이 성립한다. 또한

$$\sup_{\theta\in[0,2\pi)} n(\widehat{\sigma}^2(\theta)) \le \sup_{\theta\in[0,2\pi)} n(\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2) \le \sup_{\theta\in[0,2\pi)} 2n\max(\widehat{\sigma}_1^2, \widehat{\sigma}_2^2) = O_p(1)$$

이므로

$$\sup_{\theta\in[0,2\pi)} \left(\int_0^1 \frac{B(y,\theta)}{\phi(\Phi^{-1}(y))}\,\Phi^{-1}(y)\,dy \right)^2 < \infty$$

이 성립한다.

□

실제로 H_0에서 $P_n - a_n^0$도 식(2.2)의 우변과 같은 극한분포를 갖으리라고 예상되고 이를 보이기 위해서는 충분히 큰 n에 대해서 $P_n - a_n^0$의 꼬리부분(tail parts)이 $n\to 0$ 일 때 0으로 수렴함을 보여야한다.

참고문헌

[1] Cox, D. R. and Small, N. J. H. (1978). "Testing multivariate normality," *Biometrika*, 65, 263-272.

[2] D'Agostino, R. B. and Stephens, M. A. (1986). *Goodness-of-fit Techniques,* Marcel Dekker, New York.

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

[4] Horsewell, R. L. and Looney, S. W. (1992). "A comparison of tests for multivariate normality that are based on measure of multivariate skewness and kurtosis," *Journal of Statistical Computation and Simulation*, 42, 21-38.

[5] Huber, P. J. (1985). "Projection pursuit," *The Annals of Statistics*, 13, 435-475

[6] Kim, N. (1997). "이변량 정규분포의 적합도 검정을 위한 통계량의 극한분포에 대한 연구," 한국통계학회논문집, 4, 863-879.

[7] Machado, S. G. (1983). "Two statistics for testing multivariate normality," *Biometrika*, 70, 713-718.

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

[9] Mardia, K. V. (1970). "Measures of multivariate skewness and kurtosis with applications," *Biometrika*, 57, 519-530.

[10] Mardia, K. V. (1974). "Applications of some measures of multivariate skewness and kurtosis for testing normality and robustness studies," *Sankhya* A, 36, 115-128.

[11] Mardia, K. V. (1975). "Assessment of multivariate normality and the robustness of Hotelling's T^2 test," *Applied Statistics*, 24, 163-171.

[12] Mardia, K. V. (1980). "Tests of univariate and multivariate normality," In *Handbook in Statistics*, Ed. P. R. Krishnaiah, 279-320. Amsterdam, North-Holland.

[13] Mardia, K. V. and Foster, K. (1983). "Omnibus test of multinormality based on skewness and kurtosis," *Communications in Statistics - Theory and Methods*, 12, 207-221.