응용통계연구 제 19권 2호, 2006년, pp. 241-256

EDF 통계량을 이용한 다변량 정규성검정*

김남현[1)]

요 약

EDF에 근거한 Cramér-von Mises 통계량을 합교원리를 이용하여 다변량으로 일반화한다. 그리고 제안된 통계량의 귀무가설에서의 극한분포를 적절한 공분산 함수를 가진 가우스 과정의 적분의 형태로 표현하고 통계량의 근사적인 계산방법을 고려한다. 또한 실제 자료에 제안된 통계량을 적용해보고 여러가지 대립가설에서의 검정력을 유사한 통계량과 비교해 본다.

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

1. 서론

$\boldsymbol{X}_1, ..., \boldsymbol{X}_n$을 다변량 확률변수 $\boldsymbol{X}$의 분포에서 관측한 확률표본이라고 하고 $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$를 평균이 $\boldsymbol{\mu}$이고 공분산행렬이 $\boldsymbol{\Sigma}$인 d-변량 정규분포라고 하자. 대부분의 다변량 통계기법은 자료가 다변량 정규분포에 따른다는 가정

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

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

위의 복합귀무가설 H_0의 검정을 위한 검정통계량은 기본적으로 affine 변환에 대해서 불변성(affine invariance)을 갖는 것이 바람직하다. 왜냐하면 $\boldsymbol{X}$가 정규분포일 때 $\boldsymbol{AX} + \boldsymbol{b}$($\boldsymbol{b} \in \boldsymbol{R}^d$, $\boldsymbol{A} \in \boldsymbol{R}^{d\times d}$, $\boldsymbol{A}$는 정칙행렬)도 역시 정규분포를 따르기 때문이다. 즉, $T_n = T_n(\boldsymbol{X}_1, ..., \boldsymbol{X}_n)$이 H_0의 검정통계량일 때

$$T_n(\boldsymbol{AX}_1 + \boldsymbol{b}, ..., \boldsymbol{AX}_n + \boldsymbol{b}) = T_n(\boldsymbol{X}_1, ..., \boldsymbol{X}_n)$$

을 만족하는 것이 바람직하다.

일변량 정규성 검정통계량을 위의 불변성을 갖는 다변량 정규성 검정 통계량으로 확장하는 한 가지 일반적인 방법은 Roy(1953)의 합교 원리(Roy's union-intersection principle)를 이용하는 것이다. 이것은 $\boldsymbol{X}$가 d-변량 정규분포를 따를 때, 모든 $\boldsymbol{c} \in \boldsymbol{R}^d$, $\boldsymbol{c} \neq \boldsymbol{0}$에 대해서 $\boldsymbol{c}'\boldsymbol{X}$는 일변량 정규분포를 따른다는 사실에 기반을 두고 있다. 여기서 $'$는 전치(transpose)를

* 이 논문은 2004년도 한국학술진흥재단의 지원에 의하여 연구되었음. (KRF-2004-041-C00073)
1) (121-791) 서울시 마포구 상수동 72-1, 홍익대학교 기초과학과, 부교수
E-mail: nhkim@hongik.ac.kr

의미한다. $U_n(Z_1, ..., Z_n)$을 일변량 정규분포의 검정통계량이라고 하자. 그리고 U_n의 값이 클 때 귀무가설을 기각하고 U_n은 변환 $aZ+b$에 대해서 불변(invariant)이라고 하자. 그러면

$$T_n(\boldsymbol{X}_1, ..., \boldsymbol{X}_n) = \max_{\boldsymbol{c}\in\boldsymbol{R}^d,\, \boldsymbol{c}\neq\boldsymbol{0}} U_n(\boldsymbol{c}'\boldsymbol{X}_1, ..., \boldsymbol{c}'\boldsymbol{X}_n) = \max_{\boldsymbol{c}\in\boldsymbol{R}^d,\, ||\boldsymbol{c}||=1} U_n(\boldsymbol{c}'\boldsymbol{X}_1, ..., \boldsymbol{c}'\boldsymbol{X}_n)$$

은 affine 변환에 대해서 불변성을 갖는 '사영추적형태(projection pursuit type)'의 합리적인 통계량이고 이론적인 연구의 가치가 충분하다. 여기서 $\boldsymbol{c}$를 $||\boldsymbol{c}|| = 1$로 제한해도 무방한 이유는 U_n이 변환 $aZ+b$에 대해서 불변이므로 $U_n(\boldsymbol{c}'\boldsymbol{X}_i) = U_n(\frac{\boldsymbol{c}'}{||\boldsymbol{c}||}\boldsymbol{X}_i)$이 성립하기 때문이다. 예를 들어 Malkovich & Afifi(1973), Fattorini(1986), Kim & Bickel(2003) 등에서 제안한 통계량이 이와 유사한 형태이다.

많은 다변량 정규성 검정통계량은 또한 $\boldsymbol{X}$가 $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$를 따르면 $(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})$는 χ_d^2 분포를 따른다는 사실을 이용한다. 이와 같은 통계량은 $\boldsymbol{\mu}$와 $\boldsymbol{\Sigma}$를 각각 표본평균벡터 $\bar{\boldsymbol{X}}_n = n^{-1}\sum_{i=1}^n \boldsymbol{X}_j$와 표본 공분산 행렬 $\boldsymbol{S}_n$,

$$\boldsymbol{S}_n = \frac{1}{n}\sum_{j=1}^n (\boldsymbol{X}_j - \bar{\boldsymbol{X}}_n)(\boldsymbol{X}_j - \bar{\boldsymbol{X}}_n)'$$

으로 추정하여 대치한 제곱 반지름(squared radii)

$$D_{n,j} = ||\boldsymbol{Y}_{n,j}||^2 = (\boldsymbol{X}_j - \bar{\boldsymbol{X}}_n)'\boldsymbol{S}_n^{-1}(\boldsymbol{X}_j - \bar{\boldsymbol{X}}_n)$$

의 함수이다. 여기서 $\boldsymbol{Y}_{n,j}$는 척도화된 잔차(scaled residuals)

$$\boldsymbol{Y}_{n,j} = \boldsymbol{S}_n^{-1/2}(\boldsymbol{X}_j - \bar{\boldsymbol{X}}_n)$$

이다. $D_{n,j}$는 같은 분포를 가지나 독립은 아니며, H_0에서는 $(n-1)^2/n\ Beta(d/2, (n-d-1)/2)$의 분포를 갖고 (Gnanadesikan & Kettenring(1972) 참조) 근사적으로 χ_d^2 분포를 따른다.

$D_{n,1}, ..., D_{n,2}$의 확률 plot에 대해서는 Gnanadesikan(1977)과 Singh(1993)가 언급하였다. $\hat{G}_n(t)$,

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

를 $D_{n,1}, ..., D_{n,2}$의 EDF(Empirical Distribution Function, 경험분포함수)라고 하고 G_d를 χ_d^2의 분포함수라고 할 때 Malkovich & Afifi(1973), Koziol(1982)는 경험과정(empirical process) $g_n(t)$,

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

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

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

도 제안하였다.

본 논문에서는 일변량 자료에서 정규성검정으로 사용할 수 있는 Cramér-von Mises 통계량을 다변량으로 확장하기 위해서 $\boldsymbol{X}$의 일차결합 $\boldsymbol{c}'\boldsymbol{X}$의 EDF와 합교원리를 이용하는 방법을 고려한다. 이에 대한 일반적인 사항들은 이미 잘 알려져 있으며 D'Agostino & Stephens(1986, 4장), Thode(2002, 5장), Zhu, Fang & Bhatti(1997) 등을 참고로 한다. 구체적인 통계량의 형태를 설명하면 다음과 같다. 우선 $\boldsymbol{F}_n(x;\boldsymbol{c})$을 일차결합 $\boldsymbol{c}'\boldsymbol{X}$의 EDF,

$$\boldsymbol{F}_n(x;\boldsymbol{c}) = \frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{c}'\boldsymbol{X}_i \le x) \tag{1.1}$$

라고 하자. 복합귀무가설 H_0에서는 모평균벡터 $\boldsymbol{\mu}$와 공분산행렬 $\boldsymbol{\Sigma}$가 미지이므로 이를 각각 표본평균벡터 $\bar{\boldsymbol{X}}_n = n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_j$와 표본 공분산 행렬 $\boldsymbol{S}_n$으로 추정한 분포함수(distribution function)를

$$\hat{F}(x;\boldsymbol{c}) = \Phi\left(\frac{x - \boldsymbol{c}'\bar{\boldsymbol{X}}_n}{(\boldsymbol{c}'\boldsymbol{S}_n\boldsymbol{c})^{1/2}}\right) \tag{1.2}$$

라고 하면, 일차결합 $\boldsymbol{c}'\boldsymbol{X}$에 대한 Cramer-von Mises 통계량은

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

이고, 합교원리를 적용하면 다변량 정규성 검정통계량으로

$$W_d^2 = \max_{||\boldsymbol{c}||=1} W_1^2(\boldsymbol{c}) \tag{1.3}$$

을 고려할 수 있다. 또한 분포함수 $\hat{\boldsymbol{G}}_n(t;\boldsymbol{c})$를

$$\hat{\boldsymbol{G}}_n(t;\boldsymbol{c}) = \frac{1}{n}\sum_{i=1}^{n} I\left(\frac{\boldsymbol{c}'(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_n)}{(\boldsymbol{c}'\boldsymbol{S}_n\boldsymbol{c})^{1/2}} \le \Phi^{-1}(t)\right) \tag{1.4}$$

라고 하면, W_d^2은 적절한 치환을 통하여

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

으로 나타낼 수 있다. 그리고 W_d^2은 $Z(\boldsymbol{c};\bar{\boldsymbol{X}}_n,\boldsymbol{S}_n) = \Phi\left(\frac{\boldsymbol{c}'(\boldsymbol{X}-\bar{\boldsymbol{X}}_n)}{(\boldsymbol{c}'\boldsymbol{S}_n\boldsymbol{c})^{1/2}}\right)$라고 할 때

$$W_d^2 = \max_{||\boldsymbol{c}||=1}\sum_{j=1}^{n}\left((Z(\boldsymbol{c};\bar{\boldsymbol{X}}_n,\boldsymbol{S}_n))_{(j)} - \frac{2j-1}{2n}\right)^2 + \frac{1}{12n}$$

로 계산할 수 있다. 여기서 $\|\cdot\|$은 L^2-norm을 의미하고, $(\ )_{(j)}$는 ( )안의 통계량의 순서통계량을 의미한다. 만일 척도화된 잔차(scaled residuals) $\boldsymbol{Z}_j$,

$$\boldsymbol{Z}_j = \boldsymbol{S}_n{}^{-1/2}(\boldsymbol{X}_j - \bar{\boldsymbol{X}}_n) \tag{1.5}$$

을 이용하면 $\boldsymbol{c}'\frac{1}{n}\sum \boldsymbol{Z}_j\boldsymbol{Z}_j'\boldsymbol{c} = 1$이므로 W_d^2은

$$W_d^2 = \max_{\|\boldsymbol{c}\|=1} \sum_{j=1}^{n} \left(\Phi((\boldsymbol{c}'\boldsymbol{Z})_{(j)}) - \frac{2j-1}{2n}\right)^2 + \frac{1}{12n} \tag{1.6}$$

이 된다.

본 논문에서는 W_d^2의 귀무가설에서의 근사분포와 구체적인 계산방법에 대해서 언급한다. Zhu, Fang & Bhatti(1997)은 좀 더 일반적인 상황에서 이의 분포를 고려하였다. 또한 실제자료에 제안된 통계량을 적용해 보고 여러가지 대립가설에서의 검정력을 기존의 통계량과 비교한다.

2. 제안된 검정통계량의 극한분포

정리 2.1 W_d^2은 H_0에서

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

을 만족한다. 여기서 $\hat{\boldsymbol{U}}(t;\boldsymbol{c})$는

$$\hat{\boldsymbol{U}}(t;\boldsymbol{c}) = \boldsymbol{U}(t;\boldsymbol{c}) + \boldsymbol{c}'\boldsymbol{Z}\,\phi(\Phi^{-1}(t)) + \frac{1}{2}\boldsymbol{c}'\boldsymbol{V}\boldsymbol{c}\,\phi(\Phi^{-1}(t))\Phi^{-1}(t) \tag{2.1}$$

이고

$$\boldsymbol{Z} \sim N_d(\boldsymbol{0},\boldsymbol{I}),\quad \boldsymbol{V} = (Z_{ij})_{d\times d},\quad Z_{ii} \overset{i.i.d.}{\sim} N_1(0,2),\quad Z_{ij} \overset{i.i.d.}{\sim} N_1(0,1),\quad i\neq j \tag{2.2}$$

이며 Z_{ij}도 독립이고 $\boldsymbol{Z}$와 $\boldsymbol{V}$도 역시 독립이다. 또한 $\boldsymbol{U}(t;\boldsymbol{c})$는 $\boldsymbol{X} \sim N_d(\boldsymbol{0},\boldsymbol{I})$일 때 공분산 함수

$$\mathrm{Cov}(\boldsymbol{U}(t_1,\boldsymbol{c}_1),\boldsymbol{U}(t_2,\boldsymbol{c}_2)) = \Pr(\Phi(\boldsymbol{c}_1\boldsymbol{X}) \le t_1 \ \text{and}\ \Phi(\boldsymbol{c}_2\boldsymbol{X}) \le t_2) - t_1t_2 \tag{2.3}$$

를 갖는 가우스 과정(Gaussian process)이고 $\hat{\boldsymbol{U}}(t;\boldsymbol{c})$는 공분산 함수

$$\begin{aligned}\mathrm{Cov}(\hat{\boldsymbol{U}}(t_1,\boldsymbol{c}_1),\hat{\boldsymbol{U}}(t_2,\boldsymbol{c}_2)) &= \mathrm{Cov}(\boldsymbol{U}(t_1,\boldsymbol{c}_1),\boldsymbol{U}(t_2,\boldsymbol{c}_2)) - \boldsymbol{c}_1'\boldsymbol{c}_2\phi(\Phi^{-1}(t_1))\phi(\Phi^{-1}(t_2)) \\ &\quad -\frac{1}{2}(\boldsymbol{c}_1'\boldsymbol{c}_2)^2\phi(\Phi^{-1}(t_1))\Phi^{-1}(t_1)\phi(\Phi^{-1}(t_2))\Phi^{-1}(t_2) \end{aligned}\tag{2.4}$$

을 갖는다.

증명: W_d^2은 affine 변환에 대한 불변성을 만족하므로 귀무가설에서의 분포를 고려할 때 $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Sigma} = \boldsymbol{I}$, $\|\boldsymbol{c}\| = 1$이라고 가정해도 무방하다. 우선 식(1.1)과 식(1.2)에 정의된 $\boldsymbol{F}_n(x;\boldsymbol{c})$과 $\hat{F}(x;\boldsymbol{c})$에 대하여, 추정된 경험과정(estimated empirical process)을

$$\hat{\beta}_n(x;\boldsymbol{c}) = \sqrt{n}\left(\boldsymbol{F}_n(x;\boldsymbol{c}) - \hat{F}(x;\boldsymbol{c})\right)$$

라고 하자. 그러면

$$\begin{aligned}\hat{\beta}_n(x;\boldsymbol{c}) &= \sqrt{n}\left(\boldsymbol{F}_n(x;\boldsymbol{c}) - \hat{F}(x;\boldsymbol{c})\right)\\ &= \sqrt{n}\left(\boldsymbol{F}_n(x;\boldsymbol{c}) - \Phi\left(\frac{x - \boldsymbol{c}'\boldsymbol{\mu}}{(\boldsymbol{c}'\boldsymbol{\Sigma}\boldsymbol{c})^{1/2}}\right)\right) - \sqrt{n}\left(\hat{F}(x;\boldsymbol{c}) - \Phi\left(\frac{x - \boldsymbol{c}'\boldsymbol{\mu}}{(\boldsymbol{c}'\boldsymbol{\Sigma}\boldsymbol{c})^{1/2}}\right)\right)\end{aligned}$$

이고 경험과정(empirical process)

$$\sqrt{n}\left(\boldsymbol{F}_n(x;\boldsymbol{c}) - \Phi\left(\frac{x - \boldsymbol{c}'\boldsymbol{\mu}}{(\boldsymbol{c}'\boldsymbol{\Sigma}\boldsymbol{c})^{1/2}}\right)\right) = \sqrt{n}\left(\boldsymbol{F}_n(x;\boldsymbol{c}) - \Phi(x)\right) \stackrel{\text{let}}{=} \beta_n(x;\boldsymbol{c})$$

는 식(2.3)의 공분산 함수를 갖는 가우스 과정 $\boldsymbol{U}$로 수렴함을 보일 수 있다. 즉,

$$\beta_n(x;\boldsymbol{c}) \Rightarrow \boldsymbol{U}(\Phi(x);\boldsymbol{c}) \tag{2.5}$$

임이 성립한다. 여기서 $\Rightarrow$는 약수렴(weak convergence)을 의미한다.이는 Dudley(1978) 또는 Massart(1989)를 이용하여 보일 수 있다. 특히 Massart(1989)는 식(2.5)의 강수렴(strong convergence)에 대해서 다루고 있다. 실제로 Kim & Bickel(2003)은 $d = 2$인 경우에, 그리고 Kim(2005)는 일반적인 $d-$차원에서 Massart(1989)를 이용하여 식(2.5)의 강수렴을 증명하였다. 즉

$$\sup_{0<t<1,\boldsymbol{c}} |\beta_n(\Phi^{-1}(t);\boldsymbol{c}) - \boldsymbol{U}(t;\boldsymbol{c})| \stackrel{\text{a.s.}}{=} O(n^{-1/(4d)}\log n)$$

이 성립함을 보였다.

또한

$$\begin{aligned}\sqrt{n}(\hat{F}(x;\boldsymbol{c}) - \Phi(x)) &= \sqrt{n}\left(\Phi\left(\frac{x - \boldsymbol{c}'\bar{\boldsymbol{X}}_n}{(\boldsymbol{c}'\boldsymbol{S}_n\boldsymbol{c})^{1/2}}\right) - \Phi(x)\right)\\ &= \sqrt{n}(\boldsymbol{c}'\bar{\boldsymbol{X}}_n)(-\phi(x)) + \sqrt{n}\left[(\boldsymbol{c}'\boldsymbol{S}_n\boldsymbol{c})^{1/2} - (\boldsymbol{c}'\boldsymbol{I}\boldsymbol{c})^{1/2}\right](-x\phi(x)) + o_p(1)\end{aligned}$$

이고 $\sqrt{n}(\boldsymbol{c}'\bar{\boldsymbol{X}}_n) \xrightarrow{d} \boldsymbol{c}'\boldsymbol{Z}, \sqrt{n}\left((\boldsymbol{c}'\boldsymbol{S}_n\boldsymbol{c})^{1/2} - (\boldsymbol{c}'\boldsymbol{I}\boldsymbol{c})^{1/2}\right) \xrightarrow{d} \frac{1}{2}\boldsymbol{c}'\boldsymbol{V}\boldsymbol{c}$ 임이 성립한다. 여기서 $\boldsymbol{V}$는 식(2.2)의 분포를 갖는다. Skorohod의 정리(Billingsley(1986, 25절) 또는 Shorack & Wellner(1986, 2.3절))를 이용하면

$$\sqrt{n}(\boldsymbol{c}'\bar{\boldsymbol{X}}_n^*) = \boldsymbol{c}'\boldsymbol{Z}^* + o_p(1), \boldsymbol{Z}^* \sim N(\mathbf{0}, \boldsymbol{I})$$
$$\sqrt{n}\left((\boldsymbol{c}'\boldsymbol{S}_n^*\boldsymbol{c})^{1/2} - (\boldsymbol{c}'\boldsymbol{I}\boldsymbol{c})^{1/2}\right) = \tfrac{1}{2}\boldsymbol{c}'\boldsymbol{V}^*\boldsymbol{c} + o_p(1)$$

인 확률변수들이 존재한다. 따라서

$$\sqrt{n}\left(\hat{F}(x;\boldsymbol{c}) - \Phi(x)\right) = \boldsymbol{c}'\boldsymbol{Z}^*(-\phi(x)) + \frac{1}{2}\boldsymbol{c}'\boldsymbol{V}^*\boldsymbol{c}(-x\phi(x)) + o_p(1)$$

이고 이를 통하여 추정된 경험과정 $\hat{\beta}_n$은

$$\hat{\beta}_n(x;\boldsymbol{c}) \Rightarrow \hat{\boldsymbol{U}}(\Phi(x);\boldsymbol{c}) = \boldsymbol{U}(\Phi(x);\boldsymbol{c}) + \boldsymbol{c}'\boldsymbol{Z}^*\phi(x) + \frac{1}{2}\boldsymbol{c}'\boldsymbol{V}^*\boldsymbol{c}(x\phi(x)) \tag{2.6}$$

이며 이는 식(2.1)의 확률과정과 동일하다. 또한 식(1.4)의 $\hat{\boldsymbol{G}}_n(t;\boldsymbol{c})$에 대한 경험과정을

$$\hat{\boldsymbol{U}}_n(t;\boldsymbol{c}) = \sqrt{n}(\hat{\boldsymbol{G}}_n(t;\boldsymbol{c}) - t)$$

라고 하면

$$\hat{\beta}_n(x;\boldsymbol{c}) = \hat{\boldsymbol{U}}_n(\hat{F}(x);\boldsymbol{c}) \tag{2.7}$$

이고 식(2.6)와 (2.7)로부터

$$\hat{\boldsymbol{U}}_n \Rightarrow \hat{\boldsymbol{U}}$$

이다. 따라서

$$W_d^2 = \max_{\boldsymbol{c},\boldsymbol{c}\neq\boldsymbol{0}} \int_{-\infty}^{\infty} \hat{\beta}_n(x;\boldsymbol{c}))^2 d\hat{F} = \max_{\boldsymbol{c},\boldsymbol{c}\neq\boldsymbol{0}} \int_0^1 \left(\hat{U}_n(t;\boldsymbol{c})\right)^2 dt \xrightarrow{d} \max_{\boldsymbol{c},\boldsymbol{c}\neq\boldsymbol{0}} \int_0^1 \left(\hat{U}(t;\boldsymbol{c})\right)^2 dt$$

가 성립한다.

또한

$$(\boldsymbol{c}_1'\boldsymbol{Z}, \boldsymbol{c}_2'\boldsymbol{Z}) \sim N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \boldsymbol{c}_1'\boldsymbol{c}_2\\ \boldsymbol{c}_1'\boldsymbol{c}_2 & 1\end{pmatrix}\right)$$

임을 이용하여 계산하면

$$\begin{aligned}
& \mathrm{Cov}(\boldsymbol{U}(t_1,\boldsymbol{c}_1), \boldsymbol{c}_2'\boldsymbol{Z}) = E(I(\Phi(\boldsymbol{c}_1'\boldsymbol{Z}) \le t_1)\ \boldsymbol{c}_2'\boldsymbol{Z}) \\
= \ & (\boldsymbol{c}_1'\boldsymbol{c}_2)\int_{-\infty}^{\Phi^{-1}(t_1)} x\phi(x)dx = -(\boldsymbol{c}_1'\boldsymbol{c}_2)\int_{-\infty}^{\Phi^{-1}(t_1)} d\phi(x) = -(\boldsymbol{c}_1'\boldsymbol{c}_2)\phi(\Phi^{-1}(t_1)), \\
& \mathrm{Cov}(\boldsymbol{U}(t_1,\boldsymbol{c}_1), \boldsymbol{c}_2'\boldsymbol{V}\boldsymbol{c}_2) = E(I(\Phi(\boldsymbol{c}_1'\boldsymbol{Z}) \le t_1)\,(\boldsymbol{c}_2'\boldsymbol{Z})^2) - t_1 \\
= \ & -(\boldsymbol{c}_1'\boldsymbol{c}_2)^2\int_{-\infty}^{\Phi^{-1}(t_1)} x\,d\phi(x) - t_1 = -(\boldsymbol{c}_1'\boldsymbol{c}_2)^2\phi(\Phi^{-1}(t_1))\Phi^{-1}(t_1)
\end{aligned}$$

이고 이로부터 식(2.4)의 공분산 함수를 얻는다. □

위의 정리는 일변량에서 성립하는 정리를 확장한 것으로 일변량인 경우의 증명은 Darling(1955), Sukhatme(1972), Durbin(1973), 또는 Shorack & Wellner(1986, 5.5절)을 참고로 한다.

3. 제안된 통계량의 계산방법

식(1.3)과 같이 사영추적형태로 정의되어 있는 통계량의 단점은 해석적인 최대값을 갖는 벡터 $\boldsymbol{c}$를 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)

$$U^d = \left\{ \boldsymbol{c} \in \boldsymbol{R}^d : ||\boldsymbol{c}|| = 1 \right\} \tag{3.1}$$

에서 균일하게 퍼져 있는 $\boldsymbol{c}_1, ..., \boldsymbol{c}_k$를 이용하여

$$W_d^2 \approx \max_{1 \le j \le k} W_d^2(\boldsymbol{c}_j)$$

로 W_d^2를 근사적으로 계산하는 방법이다. NTM은 $d = 2$일 때

$$(c_{i1}, c_{i2}) = \left(\cos 2\pi \frac{(2i-1)}{2k}, \sin 2\pi \frac{(2i-1)}{2k} \right), \quad i = 1, \ldots, k$$

을 얻고, 이는 직관적인 결과와도 일치한다. NTM에 대한 자세한 내용은 Fang & Wong(1993)을 참고로 한다.

또한 $\boldsymbol{c}$를 자료 $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$에 의존하도록 택하는 방법, 예를 들면

$$\boldsymbol{c}_l = \boldsymbol{S}_n^{-1/2}(\boldsymbol{X}_l - \bar{\boldsymbol{X}}_n)/||\boldsymbol{S}_n^{-1/2}(\boldsymbol{X}_l - \bar{\boldsymbol{X}}_n)|| = \boldsymbol{Z}_l/||\boldsymbol{Z}_l||, \quad l = 1, ..., n \tag{3.2}$$

으로 택하는 방법도 고려할 수 있다. $\boldsymbol{Z}_l$은 식(1.5)의 척도화된 잔차이다. 다시 말해서 식(3.1)의 단위 d-차원 구면 $\boldsymbol{U}^d$의 모든 벡터 $\boldsymbol{c}$에 대해서 최대를 고려하는 대신에 EDF가 근사적으로 $\boldsymbol{U}^d$에서의 균일분포를 따르는 표준화된 척도화 잔차(normalized scaled residuals)에 대해서 최대를 고려하자는 것이다. Fattorini(1986), 김남현(2004a), 김남현(2004b)는 이러한 방법 또는 이와 유사한 방법으로 합교원리를 이용하여 다변량으로 확장한 통계량을 근사적으로 구하였다. W_d^2의 근사통계량을 구체적으로 구하기 위하여 다음의 보조정리를 살펴보자.

보조정리 3.1 $Z_1, ..., Z_n$을 일변량 확률표본, $Z_{(1)}, ...; Z_{(n)}$을 해당하는 순서통계량이라고 할 때 통계량

$$M_n = \sum_{i=1}^{n} \left(\Phi\left(\frac{Z_{(i)} - \bar{Z}}{S} \right) - \frac{2j-1}{2n} \right)^2$$

은

$$\left(\frac{n-1}{nd_1}, \frac{-1}{nd_1}, ..., \frac{-1}{nd_1} \right) \tag{3.3}$$

에서 최대값을 갖는다. 여기서 $\bar{Z}$는 표본평균이고 S^2은 표본분산 $S^2 = n^{-1}\sum(Z_i - \bar{Z})^2$이고 $\boldsymbol{d} = (d_1, \ldots, d_n)' = \boldsymbol{H}/(\boldsymbol{H}'\boldsymbol{H})^{1/2}$, $\quad \boldsymbol{H} = (\Phi^{-1}(1/2n), \Phi^{-1}(3/2n) \ldots, \Phi^{-1}((2n-1)/2n))'$ 이다.

증명: 김남현(2004a)의 보조정리 3.1을 이용하면 통계량

$$L_n(Z_1, \ldots, Z_n) = \sum_{i=1}^{n} \left(\frac{Z_{(i)} - \bar{Z}}{S} - \Phi^{-1}\left(\frac{2n-1}{2n} \right) \right)^2$$

은 식(3.3)에서 최대값을 갖는다는 것을 보일 수 있다. 따라서 주어진 정리를 증명하기 위해서는 통계량 L_n과 M_n이 같은 점(벡터)에서 최대값을 갖음을 보이면 충분하다.

이를 위하여

$$\left(\frac{Z_{(1)}-\bar{Z}}{S},\ldots,\frac{Z_{(n)}-\bar{Z}}{S}\right)' \overset{\text{let}}{=} \boldsymbol{u}' = (u_1,\ldots,u_n)'$$

$$\left(\Phi^{-1}(1/2n),\Phi^{-1}(3/2n)\ldots,\Phi^{-1}((2n-1)/2n)\right)' \overset{\text{let}}{=} \boldsymbol{v} = (v_1\ldots,v_n)'$$

이라고 하고 $w \overset{\text{let}}{=} ||\boldsymbol{u}-\boldsymbol{v}||^2 = L_n$, $f(\boldsymbol{u}) \overset{\text{let}}{=} \sum_{i=1}^n (\Phi(u_i)-\Phi(v_i))^2 = M_n$라고 하자. 그러면 연쇄율(chain rule, Finney & Thomas(1994, 13.5절))에 의해서

$$\frac{df}{dw} = \sum_{i=1}^n \frac{\partial u_i}{\partial w}\frac{\partial f}{\partial u_i} = \sum_{i=1}^n \frac{\Phi(u_i)-\Phi(v_i)}{u_i - v_i}\,\phi(u_i)$$

이고 Φ가 증가함수이므로 $df/dw \geq 0$이다. 따라서 $f(\boldsymbol{u})$는 $w = ||\boldsymbol{u}-\boldsymbol{v}||^2$에 대한 증가함수이다. 즉, $||\boldsymbol{u}-\boldsymbol{v}||^2 \leq ||\boldsymbol{u}^*-\boldsymbol{v}||^2$이면 $\sum_{i=1}^n (\Phi(u_i)-\Phi(v_i))^2 \leq \sum_{i=1}^n (\Phi(u_i^*)-\Phi(v_i))^2$이 성립하고 따라서 L_n과 M_n은 같은 값에서 최대를 갖는다. □

위의 정리를 근거로 식(1.3)의 W_d^2을 최소화하기 위하여, Fattorini(1986) 또는 김남현(2004a)와 같이 $\boldsymbol{c}$가 조건

$$\boldsymbol{c}'(\boldsymbol{X}_l - \bar{\boldsymbol{X}}) = \frac{n-1}{n}, \quad \boldsymbol{c}'(\boldsymbol{X}_j - \bar{\boldsymbol{X}}) = -\frac{1}{n}, \quad j = 1,\ldots,n,\ j \neq l,$$

을 만족할 때 최소가 된다는 사실과 최소제곱법을 이용하면 $\boldsymbol{c}$의 근사해는

$$\boldsymbol{c}^{(l)} = n^{-1}\boldsymbol{S}^{-1}(\boldsymbol{X}_l - \bar{\boldsymbol{X}})$$

로 주어진다. 따라서 식(1.3)의 W_d^2에 Fattorini의 방법을 적용하면 W_d^2의 근사통계량으로

$$\begin{aligned} W_d^{2*} &= \max_{1\leq l\leq n}\sum_{i=1}^n\left(\Phi\left(\frac{((\boldsymbol{X}_l-\bar{\boldsymbol{X}})'\boldsymbol{S}^{-1}(\boldsymbol{X}-\bar{\boldsymbol{X}}))_{(i)}}{((\boldsymbol{X}_l-\bar{\boldsymbol{X}})'\boldsymbol{S}^{-1}(\boldsymbol{X}_l-\bar{\boldsymbol{X}}))^{1/2}}\right)-\frac{2j-1}{2n}\right)^2+\frac{1}{12n} \qquad (3.4)\\ &= \max_{1\leq l\leq n}\sum_{i=1}^n\left(\Phi\left(\frac{(\boldsymbol{Z}_l'\boldsymbol{Z})_{(i)}}{||\boldsymbol{Z}_l||}\right)-\frac{2j-1}{2n}\right)^2+\frac{1}{12n}\end{aligned}$$

을 얻을 수 있고 이는 식(1.6)의 $\boldsymbol{c}$에 식(3.2)의 c_l을 대입한 형태이기도 하다.

4. 적용예와 모의실험결과

우선 3절에서 제안한 NTM과 식(3.4)의 근사통계량 W_d^{2*}의 근사정도를 보기 위하여 모의실험을 행하였다. $d=2$일 때와 $d=3$일 때 두 가지 방법의 상대오차

$$D = \frac{W_d^2 - W_d^{2*}}{W_d^2}$$

를 고려하였다. 여기서 W_d^2은 식(1.3)의 통계량을 NTM으로 계산한 것을 의미한다. 이를 위하여 이변량 정규분포 $N_2(\mathbf{0}, \boldsymbol{I})$에서 표본크기 n인 표본 $N = 1000$개를 S-plus 6.1을 이용하여 추출하여 각 표본크기에서의 평균을 표 4.1과 표 4.2에 제시하였다. $d = 3$일 때는 단위 정사각형에서의 점들(NT-nets)을 얻기 위해서, $(n; h_1, h_2) = (987; 1, 610)$(generating vectors)을 이용해서 glp(good lattice point) set을 계산하였다(Fang & Wang(1993, 1.3절, 부록A)). 이로부터 표본크기가 커짐에 따라 두 방법의 상대오차가 점점 감소함을 볼 수 있다. 그러나 $d = 3$일 때는 상대오차가 줄어드는 속도가 $d = 2$일 때보다 현저히 둔화되고 있음을 알 수 있다. 이는 d가 클 경우에는 근사통계량의 이용이 제한적일 수 있음을 시사한다.

표 4.1: 표본크기 n인 $N = 1000$개의 표본에서 계산된 상대오차의 평균 ($d = 2$)

n	10	20	30	40	50	100
상대오차평균	0.1197	0.06385	0.04171	0.02425	0.02094	0.007

표 4.2: 표본크기 n인 $N = 1000$개의 표본에서 계산된 상대오차의 평균 ($d = 3$)

n	10	20	30	40	50	100	300	500
상대오차평균	0.3036	0.2411	0.1903	0.1656	0.1455	0.08696	0.03481	0.01990

표 4.3에는 $n = 10, 20, 30, 40, 50, 100$, $d = 2, 3, 4$, $\alpha = 0.1, 0.05, 0.01$에 대해서 W_d^2의 기각값 $k(\alpha; n, d)$가 주어져 있다. 이 값은 S-plus 6.1을 이용한 모의실험에 의해서 얻어졌으며 표본수는 $N = 5000$을 이용하였다. $d = 3$일 때 glp set에서 $(n; h_1, h_2) = (987; 1, 610)$로 위와 동일하며 $d = 4$일 때 glp set에서는 $(n; h_1, h_2, h_3) = (1459; 1, 256, 373)$을 이용하였다.

제안된 방법을 잘 알려진 자료인 'iris setosa'에 적용해보자. 이 자료는 다변량 분석에 대한 예제에서 자주 볼 수 있으며 Fisher(1936)가 처음 분석하였다. 'iris setosa'는 4개의 변수

X_1 : sepal length
X_2 : sepal width
X_3 : petal length
X_4 : petal width

를 갖고 표본크기는 $n = 50$이다. 이 자료는 Kendall(1975)에서도 찾을 수 있고 S-plus 6.1의 자료목록에서도 구할 수 있다. 네 개의 변수중 X_4는 정규분포에서 심하게 벗어나 있음을 확인할 수 있다(Shapiro-Wilk의 통계량 $W = 0.7998$, p값≈ 0).

표 4.3: W_d^2의 기각값 $k(\alpha; n, d)$: $P(W_d^2 > k(\alpha; n, d)) = \alpha$

	n	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
$d = 2$				
	10	0.175	0.198	0.246
	20	0.177	0.203	0.256
	30	0.179	0.205	0.260
	40	0.178	0.203	0.261
	50	0.177	0.200	0.259
	100	0.178	0.202	0.259
$d = 3$				
	n	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
	10	0.248	0.271	0.330
	20	0.257	0.285	0.347
	30	0.267	0.298	0.370
	40	0.278	0.315	0.393
	50	0.290	0.337	0.456
	100	0.379	0.478	0.684
$d = 4$				
	n	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
	10	0.300	0.323	0.373
	20	0.324	0.356	0.434
	30	0.355	0.398	0.505
	40	0.384	0.436	0.565
	50	0.418	0.485	0.653
	100	0.656	0.799	1.095

우선 $\boldsymbol{X} = (X_1, X_2, X_3, X_4)'$의 사변량 정규성을 검정해보자. NTM을 이용해서 W_d^2을 계산하면 $W_d^2 = 0.735 > k(0.05; 50, 4) = 0.485$로 사변량 정규성을 기각한다. 이는 'iris setosa'를 분석한 Small(1980), Royston(1983), Csörgő(1989) 등의 결과와도 동일하다. 표 4.4는 'iris setosa'와 이의 가능한 이변량과 삼변량 자료에 대한 검정결과이다($\alpha = 0.05$). W_d^2에 대한 검정 결과, 이변량, 삼변량 모두 변수 X_4가 포함되는 경우에 다변량 정규성이 기각되는 것을 볼 수 있다.

표 4.4: 'iris setosa'의 다변량 정규성 검정 ($\alpha = 0.05$, A =채택, R=기각)

변수	(X_1, X_2)	(X_1, X_3)	(X_1, X_4)	(X_2, X_3)	(X_2, X_4)	(X_3, X_4)
W_d^2	0.089	0.176	0.973	0.178	0.979	0.971
검정결과	A	A	R	A	R	R

변수	(X_1, X_2, X_3)	(X_1, X_2, X_4)	(X_1, X_3, X_4)	(X_2, X_3, X_4)	(X_1, X_2, X_3, X_4)
W_d^2	0.165	0.777	0.876	0.880	0.735
검정결과	A	R	R	R	R

다음으로 제안된 통계량의 검정력을 살펴보기 위해서 표본크기 $n = 20, 50$, 유의수준 $\alpha = 0.05$를 고려하였다. 이 때 각각의 대립가설의 분포에서 $N = 1000$개의 표본을 S-plus 6.1을 이용하여 추출하였다. 대립가설로는 Henze & Zirkler(1990)에서 고려한 대립가설의 일부로 (i) 주변분포가 서로 독립인 분포 (ii) 혼합정규분포(mixtures of normal distributions) 등을 고려하였다. 표 4.5에서 $N(0,1)$, $C(0,1)$, $Logis(0,1)$, $\exp(1)$은 각각 표준정규분포, 코쉬분포, 로지스틱분포, 지수분포를 나타낸다. χ_k^2과 t_k는 자유도가 k인 카이제곱분포와 t분포를 나타낸다. $\Gamma(a,b)$는 확률밀도함수가

$$b^{-a}\Gamma(a)^{-1}x^{a-1}\exp(-x/b),\ x > 0$$

인 감마분포이고 $B(a,b)$는 확률밀도함수

$$B(a,b)^{-1}x^{a-1}(1-x)^{b-1},\ 0 < x < 1,$$

인 베타분포이고 $LN(a,b)$는 확률밀도함수

$$(\sqrt{2\pi}bx)^{-1}\exp(-(\log x - a)^2/2b^2),\ x > 0,$$

인 대수정규분포를 나타낸다. 또한 $F_1 * F_2$는 서로독립인 주변분포 F_1과 F_2를 갖는 분포이며 F_1^2은 각각의 주변분포가 서로독립인 F_1분포임을 의미한다. $NMIX_2(\kappa, \delta, \rho_1, \rho_2)$는

$$\kappa N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho_1\\ \rho_1 & 1\end{pmatrix}\right) + (1-\kappa)N_2\left(\begin{pmatrix}\delta\\ \delta\end{pmatrix}, \begin{pmatrix}1 & \rho_2\\ \rho_2 & 1\end{pmatrix}\right)$$

인 이변량 혼합정규분포를 말한다.

Henze & Zirkler(1990)는 affine 변환에 대해서 불변인 몇가지 다변량 정규분포를 위한 검정통계량의 검정력을 비교하였고 Mardia(1970, 1974)의 왜도 $b_{1,d}$와 첨도 $b_{2,d}$, Fattorini(1986)도 비교대상에 포함하였다. Mardia의 왜도와 첨도는 일변량 왜도와 첨도를 다변량으로 확장한 것으로 아마도 다변량 정규성 검정을 위한 통계량 중 가장 자주 인용되는 것이라 생각되며(Thode(2002, 9.4절)), 통계적인 이론이나 성질이 가장 많이 연구된 방법이라고 할 수 있다(Henze(2002, 10절)). Fattorini(1986)는 일변량 Shapiro & Wilk(1965) 통계량을 다변량으로 확장한 통계량으로 여러가지 대립가설에서 비교적 우수한 검정력을 보여

주며 기본적으로는 합교원리를 이용하여 일반화하였다는 점에서 본 논문의 W_d^{2*}와 공통점을 갖는다고 할 수있다.

표 4.5에서는 제안된 통계량의 검정력과 Henze & Zirkler (1990)에 제시된 Mardia의 $b_{1,d}$, $b_{2,d}$, Fattorini(FA) 통계량의 검정력을 비교하고 있다. 또한 주어진 숫자는 $N = 1000$개의 표본 중 유의한 표본의 백분위를 소수 첫째자리에서 반올림한 것이다.

표 4.5의 검정력으로부터 3절에서 제시된 두 가지 방법의 검정력은 거의 비슷함을 볼 수 있다. 전반적으로 주변분포가 기운 분포(skewed distribution)일 경우에는 $b_{1,d}$나 FA가 제안된 통계량보다 우수한 검정력을 보여주고 주변분포가 대칭일 경우는 $b_{2,d}$가 우수한 검정력을 보여주는 경향이 있다. 그러나 제안된 통계량은 혼합정규분포에서는 매우 우수한 검정력을 보이며 이는 Henze & Zirkler(1990), 김남현(2004a), 김남현(2004b)에서 고려된 통계량 중 거의 가장 우수한 결과일 것이다. 혼합정규분포는 여러분야에서 응용범위가 넓은 분포로 일변량에 대해서는 꽤 연구가 이루어진 편이나(Thode(2002) 12장) 다변량에 대해서는 좀 더 연구가 되어야 하는 실정이다. 제시된 검정력의 결과로 볼 때 제안된 통계량을 정규분포 대 혼합정규분포의 검정을 위한 통계량으로 촛점을 맞추고 좀 더 연구하는 것이 의미있다고 생각된다.

표 4.5: 각 분포에서 $W^2_{d(ntm)}, W^{2*}_d$, FA, $b_{1,d}$, $b_{2,d}$ 통계량의 검정력 비교 (유의수준 $\alpha = 0.05$, $n = 20,\ 50,\ d = 2$)

	$n=20$					$n=50$				
대립가설	$W^2_{d(ntm)}$	W^{2*}_d	FA	$b_{1,d}$	$b_{2,d}$	$W^2_{d(ntm)}$	W^{2*}_d	FA	$b_{1,d}$	$b_{2,d}$
$N(0,1)^2$	6	5	5	6	5	6	5	5	5	5
$\exp(1)^2$	75	77	86	80	46	100	100	100	100	82
$LN(0,.5)^2$	45	50	59	60	34	90	90	97	97	68
$C(0,1)^2$	96	96	96	93	96	100	100	–	–	–
$\Gamma(5,1)^2$	16	16	25	25	14	48	45	67	68	27
$(\chi^2_5)^2$	30	32	44	43	22	80	79	93	92	48
$(\chi^2_{15})^2$	14	13	17	19	10	33	28	46	49	18
$(t_2)^2$	61	61	68	67	68	94	94	95	91	97
$(t_5)^2$	14	18	22	25	22	29	29	40	46	54
$B(1,1)^2$	9	10	6	0	36	37	35	77	0	91
$B(1,2)^2$	20	21	19	7	12	59	62	86	23	25
$B(2,2)^2$	4	5	3	1	12	10	9	15	0	44
$Logis(0,1)^2$	8	11	15	17	14	11	12	16	24	27
$N(0,1) * \exp(1)$	49	53	63	47	26	96	96	99	93	51
$N(0,1) * \chi^2_5$	19	22	28	23	12	59	55	73	61	23
$N(0,1) * t_5$	10	10	16	17	13	17	17	19	23	24
$N(0,1) * B(1,1)$	7	8	6	2	10	22	22	56	2	29
$NMIX_2(.5,2,0,0)$	8	7	4	3	9	17	17	17	2	16
$NMIX_2(.5,4,0,0)$	68	72	51	3	22	100	100	100	2	48
$NMIX_2(.5,2,.9,0)$	27	26	29	32	11	70	72	66	80	18
$NMIX_2(.5,.5,.9,0)$	16	16	20	22	14	43	40	29	30	23
$NMIX_2(.5,.5,.9,-.9)$	46	50	51	42	27	96	97	83	65	46

참고문헌

김남현 (2004a). 다변량 정규성검정을 위한 근사 Shapiro-Wilk 통계량의 일반화, <응용통계연구>, **17**, 35-47.

김남현 (2004b). 정규성 검정을 위한 다변량 왜도와 첨도의 이용에 대한 고찰, <응용통계연구>, **17**, 507-518.

Billingsley, P. (1986). *Probability and Measure*, Wiley, New York.

Csörgő, S. (1989). Consistency of some tests for multivariate normality, *Metrika*, **36**, 107-116.

Darling, D. A. (1955). The Cramér-Smirnov test in the parametric case, *Annals of Mathematical statistics*, **26**, 1-20.

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

Durbin, J. (1973). Weak convergence of the sample distribution function when parameters are estimated, *The Annals of Statistics*, **1**, 279-290.

Fang, K. -T. and Wang, Y. (1993). *Number-theoretic Methods in statistics*, Chapman & Hall, London.

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

Finney, R. L. and Thomas, Jr. G. B. (1994). *Calculus*, Add-son-Wesley, New York.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, **VII**, 179-188.

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

Gnanadesikan, R. and Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, **28**, 81-124.

Henze, N. and Zirkler, H. (1990). A class of invariant and consistent tests for multivariate normality, *Communications in Statistics - Theory and Methods*, **19**, 3595-3617.

Kendall, M. G. (1975). *Multivariate Analysis*, Griffin, London.

Kim, N. (2005). The limit distribution of an invariant test statistic for multivariate normality, *The Korean Communications in Statistics*, **12**, 71-86.

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.

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

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

Massart, P. (1989). Strong approximation for multivariate 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.

Royston, J. P. (1983). Some techniques for assessing multivariate normality based on the

Shapiro-Wilk W, *Applied statistics*, **32**, 121-133.
Shorack, G. R. and Wellner, J. A. (1986). *Empirical processes with applications to statistics*, Wiley, New York.
Singh, A. (1993). Omnibus robust procedures for assessment of multivariate normality and detection of multivariate outliers. In : *Multivariate Environment Statistics* (G.P. Patil and C. R. Rao, eds.) North-Holland, Amsterdam, 445-488.
Small, N. J. H. (1980). Marginal skewness and kurtosis in testing multivariate normality, *Applied Statistics*, **29**, 85-87.
Sukhatme, S. (1972). Fredholm determinant of a positive definite kernel of a special type and its applications, *Annals of Mathematical Statistics*, **43**, 1914-1926.
Zhu, L. -X., Fang, K.-T. and Bhatti, M.I. (1997). On estimated projection pursuit-type Cramér-von Mises statistics, *Journal of Multivariate Analysis*, **63**, 1-14.

[2005년 8월 접수, 2006년 2월 채택]

Testing Multivariate Normality Based on EDF Statistics*

Namhyun Kim[1)]

ABSTRACT

We generalize the Cramér-von Mises Statistic to test multivariate normality using Roy's union-intersection principle. We show the limit distribution of the suggested statistic is representable as the integral of a suitable Gaussian process. We also consider the computational aspects of the proposed statistic. Power performance is assessed in a Monte Carlo study.

Keywords: Multivariate normality; EDF; Cramér-von Mises statistic; Gaussian process

* This work was supported by the Korea Research Foundation Grant.(KRF-2004-041-C00073)
1) Associate Professor, Dept. of Science, Hongik University, 72-1 Sangsu-dong, Mapo-gu, Seoul, 121-791, Korea.
E-mail: nhkim@hongik.ac.kr