The Korean Communications
in Statistics Vol. 10 No. 3, 2003
pp. 765-777

Multivariate Normality Tests Based on Principal Components[1)]

Namhyun Kim[2)]

Abstract

In this paper, we investigate some measures as tests of multivariate normality based on principal components. The idea was proposed by Srivastava and Hui(1987). They generalized Shapiro-Wilk statistic for multivariate cases. We show the null distributions of the statistics do not depend on the unknown parameters and mention the asymptotic null distributions. Also power performance of the tests are assessed in a Monte Carlo study.

keywords : Shapiro-Wilk statistic, skewness, multivariate normality, principal components.

1. 서론

많은 다변량 해석기법은 모집단이 다변량 정규분포를 따른다는 가정 하에서 여러 가지 추론의 방법을 제안하고 있다. 따라서 다변량 정규성 검정을 위한 여러 가지 통계적 방법들이 제안되어 온 것은 매우 당연한 일이다. 이에 대한 전반적인 설명은 Gnanadesikan(1977), Mardia(1980), Cox와 Small(1978), D'Agostino와 Stephens(1986, 9.7절), Thode(2002, Chapter 9), Henze(2002) 등을 참고로 한다. 이러한 방법들은 대부분 일변량 정규성 검정통계량을 다변량으로 확장하거나 그래프를 이용하는 것을 기본으로 하고 있다.

우선 일변량 첨도와 왜도를 다변량으로 확장하는 방법이 Mardia(1970, 1974), Mardia와 Foster(1983), Malkovich와 Afifi(1973)에 의해서 제안되었고 Baringhaus와 Henze(1991, 1992), Machado(1983)는 이들의 극한분포에 대하여 연구하였다.

첨도나 왜도 등의 적률을 이용하는 검정법과 더불어 Shapiro와 Wilk(1965)가 제안한 일변량 정규분포의 검정통계량도 여러 가지 대립가설에서 우수한 검정력을 갖는다는 것이 알려져 있다(Pearson, D'Agostino와 Bowman(1977)). 따라서 Shapiro와 Wilk의 통계량을 다변량으로 일반화하는 방법이 Malkovich와 Aififi(1973), Fattorini(1986), Srivastava와 Hui(1987), Mudhokar, Srivastava와 Lin(1995), Royston(1983) 등에 의해서 제안되었다. Kim과 Bickel(2003)에서는 Shapiro와 Wilk(1965)의 검정통계량과 밀접한 관련이 있고, 같은 극한분포를 갖는 de Wet과 Venter(1972)의 일변량 정규분포의 검정통계량을 사영추적(projecion pursuit)의 개념을 이용하여

---

1) This work was supported by grant No. R04-2002-000-20014-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

2) Associate Professor, Department of Science, Hongik University, Seoul, 121-791, Korea.
E-mail : nhkim@wow.hongik.ac.kr

이변량으로 확장하고 이의 극한분포에 대해서 연구하였다. 또한 Kim(2004)에서는 이를 다변량으로 확장하는 실제적인 방법을 제안하였다.

그리고 Henze와 Zirkler(1990), Henze와 Wagner(1997)은 Epps와 Pulley(1983)가 제안한 empirical characteristic function을 이용한 일변량 정규분포의 검정법을 다변량으로 확장하고 이의 일치성(consistency)과 극한분포에 대하여 연구하였다. 그리고 Zhu, Wong과 Fang(1995), Liang, Li, Fang과 Fang(2000)은 projection pursuit을 이용한 검정법을 제안하였다.

Srivastava와 Hui(1987)는 Shapiro와 Wilk의 검정통계량을 주성분(principal component)을 이용하여 다변량으로 확장하였다. 본 논문에서는 이러한 방법을 이용하여 다변량으로 일반화된 검정통계량에 대해서 살펴보고 모의실험을 통하여 다른 통계량과의 검정력을 비교해 보고자 한다.

2. 주성분을 이용한 통계량

$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$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-차원 정규분포라고 하자. 본 논문에서는 자료가 다변량 정규분포에 따른다는 가정

$$H_d : \boldsymbol{X}\text{의 분포가 } N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ \boldsymbol{\mu}\text{와 } \boldsymbol{\Sigma}\text{는 미지,를 따른다.} \tag{2.1}$$

을 검정하고자 한다.

$\boldsymbol{X}$를 평균이 $\boldsymbol{\mu}$이고 공분산행렬이 $\boldsymbol{\Sigma}$인 d-변량 확률변수라고 하고 $\boldsymbol{\Gamma}=(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_d)$를 $\boldsymbol{\Gamma}'\boldsymbol{\Sigma}\boldsymbol{\Gamma}= \boldsymbol{D}_\lambda= \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$를 만족하는 직교행렬이라고 하자. 여기서 $'$는 전치(transpose)를 의미하고 $\mathrm{diag}(\lambda_1, \ldots, \lambda_d)$는 대각선에 $\lambda_1, \ldots, \lambda_d$를 갖는 대각행렬을 의미한다. 만일 $\boldsymbol{X}$가 정규분포라면 $\boldsymbol{\gamma}_1'\boldsymbol{X}, \ldots, \boldsymbol{\gamma}_d'\boldsymbol{X}$은 d개의 주성분(principal components)이고 각각 평균이 $\boldsymbol{\gamma}_1'\boldsymbol{\mu}, \ldots, \boldsymbol{\gamma}_d'\boldsymbol{\mu}$이고 분산이 $\lambda_1, \ldots, \lambda_d$인 독립인 분포를 갖는다. 공분산행렬 $\boldsymbol{\Sigma}$가 미지일 경우는 데이터로부터 추정하고 근사적으로 독립인 주성분을 갖는다.

$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$이 d-변량 확률변수 $\boldsymbol{X}$의 분포에서의 크기 n인 확률표본일 때 $\overline{\boldsymbol{X}}$와 $\boldsymbol{S}$를 각각 표본평균, 표본공분산행렬, 즉

$$\overline{\boldsymbol{X}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i,\quad \boldsymbol{S}=\frac{1}{n-1}\sum_{j=1}^{n}(\boldsymbol{X}_j-\overline{\boldsymbol{X}})(\boldsymbol{X}_j-\overline{\boldsymbol{X}})' \tag{2.2}$$

라고 하자. $\boldsymbol{X}$의 분포는 $(d-1)$차원의 초평면(hyperplane)에 집중되어 있지 않다고 가정하고 $n \geq d+1$이라고 하면, 표본공분산행렬 $\boldsymbol{S}$는 거의 확실하게(almost surely) 정칙행렬(nonsingular matrix)이 된다(Eaton and Perlman(1973)).

$\boldsymbol{H}=(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_d)$를 $\boldsymbol{H}'\boldsymbol{S}\boldsymbol{H}= \boldsymbol{D}_w= \mathrm{diag}(w_1, \ldots, w_d)$인 직교행렬이라고 하고

$$y_{ij}= \boldsymbol{h}_i'\boldsymbol{x}_j,\ i=1, \ldots, d,\ j=1, \ldots, n,$$

이라고 하자. 그러면 y_{ij}는 j번째 표본의 i번째 주성분을 의미한다. 따라서 귀무가설 H_d에서

$(y_{i1}, \ldots, y_{in})$, $i=1, \ldots, d$는 근사적으로 독립인 표본이다.

i번째 표본에 대해서 일변량 Shapiro-Wilk 통계량은

$$W(i) = \frac{1}{(n-1)w_i}\left(\sum_{j=1}^{n} a_j y_{i(j)}\right)^2, \quad i=1, \ldots, d,$$

로 정의된다. 여기서 a_j는 Shapiro와 Wilk(1965)에 주어진 상수이고 $y_{i(j)}$는 y_{ij}의 순서통계량으로 $y_{i(1)} \le y_{i(2)} \le \cdots \le y_{i(n)}$을 만족한다. 변수 i에 대하여 $W(i) <$ 이면 정규분포의 가정을 기각하게 되므로 Srivastava와 Hui(1987)은

$$W_m = \min_{1 \le i \le d} W(i) < c_m \tag{2.3}$$

이면 H_d를 기각하는 검정법을 제안하였다.

이러한 방법은 물론 기타의 일변량 정규성 검정통계량에도 적용이 가능하다. 예를 들어 일반적으로 정규성 검정에 자주 사용되는 일변량 왜도

$$b_{1i} = \frac{n\left[\sum_{j=1}^{n}(y_{ij}-\overline{y_i})^3\right]^2}{\left[\sum_{j=1}^{n}(y_{ij}-\overline{y_i})^2\right]^3}, \quad i=1, \ldots, d,$$

에 대해서도 주성분을 이용하여 다변량으로 확장하는 것이 가능할 것이다. b_{1i}가 크면 정규분포를 기각하게 되므로 각 성분의 최대값을 취하여 통계량

$$B_1^* = \max_{1 \le i \le d} b_{1i} \tag{2.4}$$

를 고려할 수 있고 $B_1^* > c^*$이면 H_d를 기각한다.

Srivastava와 Hui(1987)은 $W(i)$를 정규분포화하는 변환을 이용하여 (Shapiro와 Wilk(1968) 참조) W_m의 백분위수(percentiles)를 구하는 방법을 제안하고, 모의실험을 통하여 제안한 방법으로 구한 기각값이 어느 정도 정확한 유의확률(p-value)을 갖는다는 것을 보여 주었다. 또한 모의실험에서의 유의확률을 통하여 제안된 통계량의 분포가 귀무가설에서의 미지의 모수에 의존하지 않음을 보였다. 본 논문에서는 이러한 사실을 좀 더 이론적인 측면에서 살펴보고, W_m과 B_1^*의 극한분포를 살펴보려한다. 또한 주성분을 이용하여 다변량으로 일반화한 두 통계량의 검정력을 모의실험을 통하여 기존의 통계량과 비교하고자 한다.

(2.1)의 복합귀무가설 H_d를 위한 검정통계량 $T_n = T_n(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$이 모든 $\boldsymbol{b} \in \boldsymbol{R}^d$와 정칙행렬 $\boldsymbol{A} \in \boldsymbol{R}^{d \times d}$에 대해서

$$T_n(\boldsymbol{AX}_1 + \boldsymbol{b}, \ldots, \boldsymbol{AX}_n + \boldsymbol{b}) = T_n(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \tag{2.5}$$

을 만족할 때 T_n이 affine invariance의 성질을 갖는다고 한다. 이 경우 T_n의 분포는 H_d에서 $\boldsymbol{\mu}$와 $\boldsymbol{\Sigma}$에 의존하지 않는다. 그러나 알려진 바와 같이 주성분은 affine invariance의 성질을 만족하지는 않는다. 따라서 W_m과 B_1^*도 역시 마찬가지이다. 하지만 W_m과 B_1^*의 분포는 $\boldsymbol{\mu}$와 $\boldsymbol{\Sigma}$에

무관하며 아래의 정리를 얻는다.

정리 1. W_m과 B_1^*의 귀무가설 H_d에서의 분포는 $\boldsymbol{\mu}$와 $\boldsymbol{\Sigma}$에 의존하지 않는다.

증명. 식(2.2)의 S가 $X+b$에 대해서 불변(invariant)이고 $\sum a_j=0$으로부터 $W_m(X)=W_m(X+b)$는 자명하다. 따라서 W_m의 분포는 $\boldsymbol{\mu}$에 의존하지 않는다.

다음으로 W_m의 분포가 $\boldsymbol{\Sigma}$에 의존하지 않음을 보이자. W_m은 위치불변(location invariant)이므로 $\boldsymbol{\mu}=\mathbf{0}$이라고 가정하자. W_m 또는 $W(i)$는 $\boldsymbol{h}_i'\boldsymbol{x}_j/\sqrt{w_i}$의 함수이다. $D_{1/\sqrt{w}}=diag(1/\sqrt{w_1},\ldots,1/\sqrt{w_d})$라고 하고 Z를 $N_d(\mathbf{0},I)$를 따르는 확률벡터라고 하면

		0.01	0.02	0.05	0.10	0.5	0.9	0.95	0.98	0.99
$n=20$		0.8532	0.8698	0.8899	0.9064	0.9448	0.9669	0.9713	0.9756	0.9779
		0.8480	0.8664	0.8897	0.9061	0.9446	0.9670	0.9720	0.9766	0.9793
		0.8456	0.8668	0.8886	0.9066	0.9445	0.9662	0.9712	0.9761	0.9782
		0.8548	0.8700	0.8891	0.9064	0.9449	0.9680	0.9720	0.9765	0.9792
$n=50$		0.9285	0.9367	0.9469	0.9543	0.9726	0.9829	0.9849	0.9873	0.9884
		0.9290	0.9365	0.9471	0.9544	0.9724	0.9830	0.9851	0.9872	0.9885
		0.9264	0.9365	0.9474	0.9550	0.9730	0.9830	0.9850	0.9871	0.9885
		0.9280	0.9367	0.9462	0.9542	0.9723	0.9832	0.9853	0.9876	0.9885

<표 1> 여러 ρ에서 W_m-통계량의 근사백분위수 ($d=2$)

	α	0.01	0.02	0.05	0.10	0.5	0.9	0.95	0.98	0.99
$n=20$	$\rho=0$	0.8220	0.8397	0.8666	0.8834	0.9251	0.9501	0.9554	0.9607	0.9636
	$\rho=0.3$	0.8340	0.8484	0.8683	0.8852	0.9252	0.9499	0.9549	0.9598	0.9631
	$\rho=0.6$	0.8229	0.8435	0.8655	0.8839	0.9255	0.9504	0.9558	0.9603	0.9638
	$\rho=0.9$	0.8247	0.8432	0.8666	0.8824	0.9259	0.9500	0.9549	0.9600	0.9635
$n=50$	$\rho=0$	0.9173	0.9257	0.9356	0.9445	0.9638	0.9750	0.9772	0.9796	0.9811
	$\rho=0.3$	0.9172	0.9264	0.9366	0.9446	0.9639	0.9752	0.9772	0.9798	0.9812
	$\rho=0.6$	0.9184	0.9268	0.9381	0.9456	0.9642	0.9753	0.9774	0.9797	0.9812
	$\rho=0.9$	0.9179	0.9262	0.9364	0.9443	0.9641	0.9751	0.9772	0.9795	0.9809

<표 2> 여러 ρ에서 W_m-통계량의 근사백분위수 ($d=5$)

의 $W(i)$분포는 $\boldsymbol{D}_{1/\sqrt{w}}\boldsymbol{H}'\boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}$ 또는 $\boldsymbol{D}_{1/\sqrt{w}}\boldsymbol{H}'\boldsymbol{S}^{1/2}\boldsymbol{S}^{-1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}$의 함수와 같은 분포를 갖는다. 여기서 $\boldsymbol{A}^{1/2}$은 행렬 $\boldsymbol{A}$의 대칭 양정치 제곱근(symmetric positive definite square root)을 말한다. 잘 알려진 바와 같이 $(n-1)\boldsymbol{S}$는 모수 $\boldsymbol{\Sigma}$와 $(n-1)$을 갖는 Wishart 분포 $W_d(\boldsymbol{\Sigma}, n-1)$을 따른다. 따라서 $(n-1)\boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2}$은 $W_d(I, n-1)$을 따르고 $\boldsymbol{S}^{-1/2}\boldsymbol{\Sigma}^{1/2}$의 분포도 $\boldsymbol{\Sigma}$에 의존하지 않는다. 또한 정의에 의해서 $\boldsymbol{D}_{1/\sqrt{w}}\boldsymbol{H}'\boldsymbol{S}\boldsymbol{H}\boldsymbol{D}_{1/\sqrt{w}} = \boldsymbol{I}$이므로 $\boldsymbol{D}_{1/\sqrt{w}}\boldsymbol{H}'\boldsymbol{S}^{1/2}$의 분포도 $\boldsymbol{\Sigma}$에 의존하지 않는다. B_1^*에 대해서도 같은 방법을 적용할 수 있고 정리는 성립한다. ■

실제로 $n=20, 50$, $\rho=0, 0.3, 0.6, 0.9$, $d=2, 5$의 조합에서 $N=5000$개의 d-변량 정규분포에서의 표본을 추출하여 W_m과 B_1^*의 값을 계산하고 그 결과로부터 각 분포에서의 백분위수를 구하였다(<표 1-4>). 그 결과 W_m과 B_1^*의 분포는 ρ에 의존하지 않음을 확인할 수 있다.

		0.01	0.02	0.05	0.10	0.5	0.9	0.95	0.98	0.99
$n=20$		0.002919	0.006201	0.01602	0.03275	0.2214	0.8437	1.1821	1.7266	2.1912
		0.003145	0.006020	0.01447	0.03175	0.2288	0.8842	1.2347	1.6773	2.1418
		0.003915	0.006976	0.01720	0.03331	0.2335	0.8896	1.2305	1.7635	2.2236
		0.003499	0.006035	0.01637	0.03338	0.2276	0.8693	1.2139	1.7140	2.0636
$n=50$		0.001209	0.002723	0.007670	0.01622	0.1048	0.3999	0.5470	0.7650	0.9566
		0.001620	0.003198	0.007832	0.01686	0.1124	0.4180	0.5620	0.7870	0.9823
		0.001461	0.002978	0.008026	0.01648	0.1084	0.4069	0.5669	0.7911	0.9724
		0.001210	0.003073	0.007371	0.01536	0.1076	0.4010	0.5424	0.7560	0.9495

<표 3> B_1^*-통계량의 근사백분위수 $(d=2)$

$d=5$		0.01	0.02	0.05	0.10	0.5	0.9	0.95	0.98	0.99
$n=20$		0.05330	0.07045	0.1184	0.1630	0.5028	1.3434	1.7237	2.3319	2.9462
		0.04997	0.06909	0.1111	0.1599	0.4959	1.2829	1.6474	2.2029	2.6149
		0.05278	0.07359	0.1108	0.1583	0.4868	1.3172	1.7168	2.3161	2.8597
		0.05660	0.07524	0.1194	0.1647	0.5050	1.3111	1.6528	2.2774	2.8191
$n=50$		0.02862	0.03722	0.05695	0.08252	0.2402	0.6151	0.8021	1.0186	1.2233
		0.02632	0.03724	0.05486	0.08176	0.2366	0.6250	0.7923	1.0333	1.2927
		0.02678	0.03731	0.05640	0.07756	0.2357	0.6133	0.7967	1.0516	1.2977
		0.02832	0.03776	0.05649	0.07914	0.2415	0.6103	0.7986	1.0388	1.2381

<표 4> B_1^*-통계량의 근사백분위수 $(d=5)$

n	$\alpha=0.1$	$\alpha=0.05$	$\alpha=0.02$	$\alpha=0.01$
100	1.3782	1.7694	2.3621	2.9135
200	1.4375	1.8314	2.4082	2.7474
400	1.5295	1.9990	2.5307	2.9639
600	1.4865	1.9328	2.4644	2.8613
800	1.5321	1.9603	2.4386	2.9554
1000	1.5540	1.9954	2.5279	2.9737
∞	1.6	2.1	2.7	3.1

<표 5> W_m-통계량의 기각값 k_α : $\Pr(n(1-W_m)-a_n \geq k_\alpha)=\alpha$ ($d=2$)

n	$\alpha=0.1$	$\alpha=0.05$	$\alpha=0.02$	$\alpha=0.01$
100	1.8723	2.2968	2.7964	3.1153
200	1.9869	2.4931	3.0530	3.5725
400	2.0618	2.4899	3.1244	3.4965
600	2.0661	2.5058	3.0642	3.6586
800	2.1103	2.5066	3.0329	3.4851
1000	2.0885	2.5613	3.1180	3.4897
∞	2.2	2.7	3.2	3.5

<표 6> W_m-통계량의 기각값 k_α : $\Pr(n(1-W_m)-a_n \geq k_\alpha)=\alpha$ ($d=5$)

다음으로 W_m의 극한분포에 대해서 고려해 보자. 잘 알려진 바와 같이 (de Wet과 Venter(1972), Leslie, Stephens과 Fotopolous(1986))

$$n(1-W(i))-a_n \xrightarrow{d} \sum_{m=3}^{\infty}(Z_{im}-1)/m \equiv T_i \tag{2.6}$$

이 성립한다. 여기서 $Z_{i1},\ldots,Z_{\in},\ldots,$ $i=1,\ldots,d$은 독립이며 같은 분포를 따르는 (i.i.d.) $N(0,1)$인 확률변수들이고 a_n은 de Wet과 Venter(1972)에 주어져있는 상수이다. 식(2.6)의 우변을 확률변수 T_i라고 하면, 귀무가설 H_d에서 $(y_{i1},\ldots,y_{in})$, $i=1,\ldots,d$는 근사적으로 독립인 표본이므로

$$n(1-W_m)-a_n \xrightarrow{d} \max_i T_i \tag{2.7}$$

가 성립한다. 식(2.7)을 모의실험을 통하여 확인하기 위해서 W_m의 근사상위백분위수와 $\max_i T_i$의 분포의 상위백분위수를 $d=2$, $d=5$에서 비교하였다(<표 7>, <표 8>). $\max_i T_i$의 상위백분

위수는 de Wet과 Venter(1972)에 주어져 있는 T_i의 분포로부터 구하였다. 그 결과 식(2.7)은 비록 수렴속도는 느리지만 성립함을 볼 수 있다.

B_1^*에 대해서는 $\sqrt{\frac{n}{6}}\sqrt{b_{1i}}$가 근사적으로 정규분포 $N(0,1)$을 따른다는 것이 알려져 있으므로

$$\frac{n}{6} b_{1i} \xrightarrow{d} \chi^2(1)$$

$$\frac{n}{6} B_1^* \xrightarrow{d} \max_{1 \le i \le d} V_i, \quad V_1, \ldots, V_d \overset{i.i.d.}{\sim} \chi^2(1)$$

이 성립한다. 즉

$$P\left(\frac{n}{6} B_1^* \le x\right) \to (P(V_1 \le x))^d$$

이 성립한다. 이를 이용하여 B_1^*에 대한 근사적인 기각값을 구할 수 있다.

3. 검정력 비교

식(2.3)의 W_m-통계량과 식(2.4)의 B_1^*의 검정력을 표본크기 $n=20$, 50, 유의수준 $\alpha=0.05$에서 모의실험(simulation)을 통하여 살펴보았다. 이 때 각각의 대립가설의 분포에서 $N=1000$개의 표본을 S-plus 6.1을 이용하여 추출하였다. Henze와 Zirkler(1990)는 몇 가지 위치-척도 불변인 다변량 정규분포의 검정을 위한 통계량의 검정력을 비교하였고 Malkovich와 Afifi(MA)의 일반화된 Shapiro-Wilk 통계량(Malkovich and Afifi(1973))과 이것의 수정된 형태인 Fattorini(FA) 통계량(Fattorini(1986)), 그리고 Mardia(1970)의 다변량 왜도(MS)도 비교 대상에 포함하였다. 그들은 (i) 주변분포가 서로 독립인 분포 (ii) 혼합정규분포(mixtures of normal distributions) 등을 고려하였다. <표 7>, <표 8>에서 $N(0,1)$, $C(0,1)$, $Logis(0,1)$, $\exp(1)$은 각각 표준정규분포, 코쉬분포, 로지스틱분포, 지수분포를 나타낸다. χ_k^2과 t_k는 자유도가 k인 카이제곱분포와 t분포를 나타낸다. $\Gamma(a,b)$는 확률밀도함수가

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

인 감마분포이고 $B(a,b)$는 확률밀도함수

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

인 베타분포이고 $LN(a,b)$는 확률밀도함수

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

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. $NMIX_2(x,\delta,\rho_1,\rho_2)$는

$$xN_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho_1\\ \rho_1 & 1\end{pmatrix}\right) + (1-x)N_2\left(\begin{pmatrix}\delta\\ \delta\end{pmatrix}, \begin{pmatrix}1 & \rho_2\\ \rho_2 & 1\end{pmatrix}\right)$$

인 이변량 혼합정규분포를 말한다.

<표 7>, <표 8>의 검정력은 $N=1000$개의 표본 중 유의한 표본의 백분위를 소수 첫째자리에서 반올림한 것이다. 여기에서는 W_m-통계량의 검정력을 Henze와 Zirkler(1990)에 주어진 MA, FA-통계량의 검정력과 비교하고 B_1^*의 검정력을 MS와 비교하였다.

<표 7>의 결과를 보면 우선 FA-통계량은 고려된 거의 대부분의 분포에서 W_m이나 MA보다 우수하거나 비슷한 검정력을 갖는다는 것을 볼 수 있다. 또한 W_m-통계량은 고려된 대부분의 대립가설에서 MA보다는 우수하거나 비슷한 검정력을 가짐을 볼 수 있다. 일반적으로 W_m-통계량이 MA보다 우수한 경우에는 검정력의 차이가 매우 큰 경우가 많으나 그 반대의 경우에는 검정력의 차이가 미미하다. 특히 주변분포중 하나가 $N(0,1)$인 경우나 혼합정규분포인 경우에 MA보다 매우 우수한 경우가 많이 나타나는 것으로 파악된다. 이 때는 FA보다도 W_m이 우수한 검정력을 갖는 경우를 찾아 볼 수 있다. 이러한 경향은 차원이 높아질 경우에도 (<표 8>, $d=5$) 어느 정도 유지됨을 볼 수 있으나 $d=5$, $n=50$인 차원이 높고 표본크기가 클 경우에는 W_m의 검정력이 다른 통계량에 비해서 현저히 떨어지는 것을 ($\Gamma(5,1)^5$, $(t_5)^5$, $Logis(0,1)^5$, $N(0,1)^4 * \exp(1)$ 등) 볼 수 있다. B_1^*에 대해서는 MS가 일반적으로 좀 더 좋은 검정력을 가짐을 볼 수 있고 이는 차원이 커지면서 그 정도가 더 확실하다. 전반적으로 FA와 MS가 우수한 검정력을 보여주고 있고 특히 MS는 차원이 클 때 타 통계량과 달리 검정력의 감소가 그리 심하지 않음을 보여주고 있다.

4. 결론 및 토의

$\boldsymbol{X}$의 분포가 다변량 정규분포를 따른다는 (2.1)의 복합귀무가설을 위한 검정통계량은 기본적으로 식(2.5)의 affine invariance의 성질을 갖는 것이 바람직하다. 왜냐하면 $\boldsymbol{X}$가 정규분포일 때 $\boldsymbol{AX}+\boldsymbol{b}$ ($\boldsymbol{b}\in R^d$, $\boldsymbol{A}\in \boldsymbol{R}^{d\times d}$, $\boldsymbol{A}$는 정칙행렬)도 역시 정규분포를 따르기 때문이다. 그리고 그 결과 검정통계량의 분포가 미지의 $\boldsymbol{\mu}$와 $\boldsymbol{\Sigma}$에 의존하지 않는다.

또한 다변량 정규분포의 검정은 차원이 높을 때에도 실제적으로 이용하기 편리하여야 한다. 많은 통계량들이 $\boldsymbol{X}$가 d-차원 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 사실을 이용하여 일변량 정규분포의 검정통계량을 다변량으로 일반화하고 있다. $U_n(Z_1,\ldots,Z_n)$을 일변량 정규분포의 검정통계량이라고 하자. 그리고 U_n의 값이 작을 때 귀무가설을 기각하고 U_n은 affine transformation ($aZ+b$)에 대해서 불변(invariant)이라고 하자. 그러면

$$T_n(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)=\min_{\boldsymbol{c}\in \boldsymbol{R}^d,\, \boldsymbol{c}\neq 0} U_n(\boldsymbol{c}'\boldsymbol{X}_1,\ldots,\boldsymbol{c}'\boldsymbol{X}_n) \tag{4.1}$$

은 affine invariance의 성질을 갖는 합리적인 통계량이고 이론적인 연구의 가치가 충분하다. 예를 들어 Malkovich and Afifi(1973), Fattorini(1986), Kim and Bickel(2002) 등에서 제안한 통계량이 대부분 이런 형태이다. 그러나 이와 같은 통계량의 가장 큰 단점은 (4.1)의 해석적인 최소값 또는

최소값을 갖는 벡터 c를 찾아내는 것이 매우 어려워 근사적인 해를 찾아야 한다는 것이다.

Srivastava와 Hui(1987)가 제안한 방법은 결국 모든 방향의 c를 고려하는 대신에 주성분의 방향만을 고려하자는 것이다. 주성분만을 고려하면 통계량의 affine invariance는 성립하지 않게 되나 역시 귀무가설에서의 분포는 μ와 Σ에 의존하지 않는다(**정리 1**). 식(4.1)의 방법으로 일반화된 통계량의 경우는 극한분포의 유도가 상당히 복잡하거나 몇몇 통계량의 경우는 극한분포에 대해서 알려진 바가 없다. 그러나 주성분을 고려하는 경우에는 각 성분이 점근적으로 독립이 되므로 일변량에서의 극한분포의 결과로부터 근사분포에 대한 정보를 얻어낼 수가 있다. 그리고 차원이 높을 때에도 실제적으로 사용하는데 크게 무리가 없다.

그러나 주성분만을 고려함으로써 다변량 자료의 주성분이외의 방향, 즉 의미 있는 일차결합에 대한 정보를 손실하게 되고 따라서 특히 자료의 수가 많고 차원이 높을 경우에는 검정력에 있어서 그리 좋은 결과를 얻지 못한다. 따라서 주성분을 이용한 검정의 방법은 공식적인 검정에 앞서 그래프 등을 이용한 자료의 탐색이나 자료의 진단(diagnosis)을 통한 정규분포로의 변환에 정보를 제공하는 도구로서 의미가 있다고 생각된다.

참고문헌

[1] Baringhaus, L., and Henze, N. (1991). "Limit distributions for measuresof multivariate skewness and kurtosis based on projections," *Journal of multivariate analysis*, 38, 51-69.

[2] Baringhaus, L., and Henze, N. (1992). "Limit distribution for Mardia's measure of multivariate skewness," *The Annals of Statistics*, 20, 1889-1902.

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

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

[5] de Wet, T. and Venter, J. H. (1972). "Asymptotic distributions of certain test criteria of normality," *South African Statistical Journal*, 6, 135-149.

[6] Eaton, M. R. and Perlman, M. D. (1973). "The non-singularity of generalized sample covariance matrices," *The Annals of Statistics*, 1, 710-717.

[7] Epps, T. W. and Pulley, I. B. (1983). "A test for normality based on the empirical characteristic function," *Biometrika*, 70, 723-726.

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

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

[10] Henze, N. and Wagner, T. (1997). "A new approach to the BHEP tests for multivariate normality," *Journal of multivariate analysis*, 62, 1-23.

[11] Henze, N. (2002) "Invariant tests for multivariate normality: A critical review," *Statistical*

Papers, 43, 467-506.

[12] Henze, N. and Zirkler, H. (1990). "A class of invariant and consistent tests for multivariate normality," *Communications in Statistics-Theory and Methods*, 19, 3539-3617.

[13] Kim, N. (2004). "An approximate Shapiro-Wilk statistic for testing multivariate normality," *The Korean Journal of Applied Statistics*, 17, *-*.

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

[15] Leslie, J. R., Stephens, M. A. and Fotopolous, S. (1986). "Asymptotic distribution of the Shapiro-Wilk W for testing for normality," *The Annals of Statistics*, 14, 1497-1506.

[16] Liang, J., Li, R., Fang, H. and Fang, K.-T. (2000). "Testing multinormality based on low-dimensional projection," *Journal Statistical planning and Inference*, 86, 129-141.

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

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

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

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

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

[22] 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.

[23] Mudholkar, G. S., Srivastava, D. K. and Lin, C. T. (1995). "Some p-variate adaptations of the Shapiro-Wilk test of normality," *Communications in Statistics-Theory and Methods*, 24, 953-985.

[24] Pearson, E. S., D'Agostino, R. B. and Bowman, K. O. (1977). "Tests for departure from normality. Comparison of powers," *Biometrika*, 64, 231-246.

[25] Royston, J. P. (1983). "Some techniques for accessing multivariate normality based on the Shapiro-Wilk W," *Applied Statistics*, 32. 121-133.

[26] Shapiro, S. S. and Francia, R. S. (1972). "An approximate analysis of variance test for normality," *Journal of the American statistical Association*, 67, 215-216.

[27] Shapiro, S. S. and Wilk, M. B. (1965). "An analysis of variance test for normality (complete samples)," *Biometrika*, 52, 591-611.

[28] Shapiro, S. S. and Wilk, M. B. (1968). "Approximations for the null distribution of the W statistic," *Technometrics*, 10, 861-866.

[29] Srivastava, M. S. and Hui, T. K. (1987). "On assessing multivariate normality based on Shapiro-Wilk W statistic," *Statistics & Probability Letters*, 5, 15-18.

[30] Thode, Jr. H. C. (2002). *Testing for Normality*, Marcel Dekker, New York.

[31] Zhu, L., Wong, H. L. and Fang, K. (1995). "A test for multivariate normality based on sample entropy and projection pursuit," *Journal of Statistical Planning and Inference*, 45, 373-385.

[2003년 7월 접수, 2003년 11월 채택]

	$n=20$					$n=50$				
대립가설	W_m	MA	FA	B_1^*	MS	W_m	MA	FA	B_1^*	MS
$N(0,1)^2$	5	5	5	5	6	5	5	5	6	5
$\exp(1)^2$	76	76	86	67	80	98	100	100	96	100
$LN(0,0.5)^2$	53	53	59	54	60	90	92	97	87	97
$C(0,1)^2$	96	96	96	91	93	100	-	-	98	-
$\Gamma(5,1)^2$	25	22	25	26	25	53	58	67	59	68
$(\chi_5^2)^2$	42	43	44	42	43	84	84	93	77	92
$(\chi_{15}^2)^2$	16	18	17	19	19	40	42	46	43	49
$(t_2)^2$	66	69	68	65	67	92	94	95	87	91
$(t_5)^2$	21	24	22	24	25	42	46	40	39	46
$B(1,1)^2$	8	2	6	1	0	27	4	77	0	0
$B(1,2)^2$	17	9	19	6	7	45	35	86	18	23
$B(2,2)^2$	4	2	3	1	1	7	2	15	1	0
$Logis(0,1)^2$	13	16	15	12	17	20	21	16	23	24
$N(0,1)*\exp(1)$	47	52	63	42	47	83	87	99	84	93
$N(0,1)*\chi_5^2$	36	26	28	34	23	85	61	73	78	61
$N(0,1)*t_5$	16	16	16	18	17	29	24	19	28	23
$N(0,1)*B(1,1)$	9	4	6	3	2	46	4	56	2	2
$NMIX_2(0.5,2,0,0)$	7	4	4	3	3	19	4	17	2	2
$NMIX_2(0.5,4,0,0)$	80	4	51	3	3	100	5	100	3	2
$NMIX_2(0.5,2,0.9,0)$	19	27	29	15	32	52	54	66	20	80
$NMIX_2(0.5,0.5,0.9,0)$	17	21	20	19	22	45	33	29	25	30
$NMIX_2(0.5,0.5,0.9,-0.9)$	43	47	51	35	42	74	76	83	53	65

<표 7> 각 분포에서 W_m, MA, FA, B_1^*, MS 통계량의 검정력 비교
(유의수준 $\alpha=0.05$, $n=20,50$, $d=2$)

	$n=20$					$n=50$				
대립가설	W_m	MA	FA	B_1^*	MS	W_m	MA	FA	B_1^*	MS
$N(0,1)^5$	6	5	5	4	5	4	5	6	5	5
$\exp(1)^5$	67	61	65	61	82	96	97	100	93	100
$LN(0,0.5)^5$	50	47	49	51	63	86	90	96	85	100
$C(0,1)^5$	100	99	99	98	99	100	-	-	100	-
$\Gamma(0.5,1)^2$	88	86	90	86	98	100	-	-	99	-
$\Gamma(5,1)^5$	17	15	15	17	20	39	45	54	43	70
$(\chi_5^2)^5$	33	30	31	32	40	69	73	84	68	96
$(\chi_{15}^2)^5$	13	14	15	12	15	27	31	34	31	49
$(t_2)^5$	79	80	81	73	86	98	99	100	95	99
$(t_5)^5$	23	28	29	26	34	46	55	56	46	64
$B(1,1)^5$	3	1	1	3	0	5	0	1	1	0
$B(1,2)^5$	6	4	5	5	3	14	5	14	7	10
$B(2,2)^5$	4	2	2	3	1	3	1	1	1	0
$Logis(0,1)^5$	12	13	13	14	15	19	27	27	18	35
$N(0,1)^4 * \exp(1)$	17	21	21	17	19	38	56	72	41	62
$N(0,1)^{4*}\chi_5^2$	23	10	10	24	10	70	29	34	63	29
$N(0,1)^4 * t_5$	11	10	10	12	9	20	19	19	18	18

<표 8> 각 분포에서 W_m, MA, FA, B_1^*, MS 통계량의 검정력 비교
(유의수준 $\alpha=0.05$, $n=20,50$, $d=5$)