Proceedings of Joint Conference of Korean Data And Information Science Society and The Korean Data Analysis Society, April 28 - 29, 2006, pp. 203-212

A Test for Multivariate Normality Focused on Elliptical Symmetry Using Mahalanobis Distances*

Cheolyong Park[1]

Abstract

A chi-squared test of multivariate normality is suggested which is mainly focused on detecting deviations from elliptical symmetry. This test uses Mahalanobis distances of observations to have some power for deviations from multivariate normality. We derive the limiting distribution of the test statistic by a conditional limit theorem. A simulation study is conducted to study the accuracy of the limiting distribution in finite samples. Finally, we compare the power of our method with those of other popular tests of multivariate normality under two non-normal distributions.

Keywords : Chi-squared test, Conditional limit theorem.

1. 서론

다변량 정규성(multivariate normality)은 다변량 분산분석, 판별분석, 정준분석 등과 같이 많은 다변량 통계절차에서 기본적으로 가정되는 사항이다. 그리고 Leoney(1995)에서도 밝혔듯이 이러한 다변량 절차들의 성능은 다변량 정규성에서의 특정 이탈에 대해 정도의 차이는 있을지 모르지만 영향을 받을 수밖에 없는 것이다. 실제로 Mardia, Kent & Bibby(1979)에서는 평균벡터에 관련된 가설검정은 왜도(skewness)에 영향을 받기 쉬우며, 공분산행렬에 관계된 가설검정은 첨도(kurtosis)에 영향을 받기 쉽다고 밝히기도 하였다.

통계학 실무자들이 주어진 다변량 자료의 정규성을 검정하기 위해서 사용할 수 있는 절차의 절대수가 결코 부족한 것은 아니다. 사실은 Leoney(1995)과 Gnanadesikan(1977) 등의 참고문헌을 살펴보더라도 그 절차가 50 종류는 충분히 되는 것을 알 수 있다. 왜냐하면 다변량 정규성에서 이탈되는 비정규성의 행태는 무수히 많기 때문이다. 따라서 어떤 절차에서는 다변량 정규성이 전혀 문제가 없는 듯이 보이지만 다른 절차에서는 다변량 정규성이 상당히 의심되는 결과가 제시되는 현상이 벌어지는 것은 어떻게 보면 너무나 당연한 일인지도 모른다. 그래서 Andrews, Gnanadesikan &

*This research was conducted by Bisa Research Grant of Keimyung University in 2005.

[1]Associate Professor, Department of Statistics, Keimyung University, Taegu 704-701, Korea.
E-mail : cypark1@kmu.ac.kr

Warner(1973)에서는 하나가 아닌 복수의 다변량 정규성 검정을 동시에 사용하는 것이 필요하다는 결론을 내리고 있다.

이와 같은 현실에서 Romeu, Ozturk(1993)나 Manzotti, Quiroz(2001) 등과 같은 모의실험 연구는 다양한 다변량 정규성 검정의 검정력을 비교하여 나름대로의 지침을 제시하고 있다. 비록 이 연구들이 모든 이탈을 다룰 수 없는 한정된 모의실험 연구라고 하지만 전반적으로 보았을 때 Mardia(1970), Ozturk, Romeu(1992), Baringhaus, Henze(1988) 등에서 제시된 절차들이 우수한 성능을 보이는 것으로 나타났다. 이 방법들은 검정통계량을 계산하기도 용이할 뿐만 아니라 나름대로 점근분포(asymptotic distribution)도 쉽게 계산할 수 있는 장점이 있어 통계학 실무자도 쉽게 접근할 수 있는 검정이라고 판단된다.

이 연구에서 제안하고자 하는 것은, 기존에 제시된 여러 우수한 다변량 정규성 검정들과 함께 사용되었을 때 특히 유용한, 특수 비정규성 다변량 구조에 초점을 둔 검정법이다. 구체적으로 이 연구에서 제시하는 검정은 다변량 정규성에서 관측되는 타원형 대칭성(elliptical symmetry)에 초점을 두고 정규성에서의 이탈을 정밀화하기 위해 마할라노비스 거리를 이용하는 카이제곱 검정이다. 구체적인 절차는 다변량 자료의 척도화 잔차(scaled residuals)와 마할라노비스 거리(Mahalanobis distances)를 계산하고, 척도화 잔차를 기초로 각 사분면(quadrant)으로 나누고 마할라노비스 거리제곱을 기초로 카이제곱 분위수(quantile)로 나눈 칸(cell)을 구성하여 관측도수를 계산한 후 기대도수와 비교하는 카이제곱 검정을 수행하는 것이다.

이 논문은 다음과 같이 구성되어 있다. 2장에서는 구체적인 검정절차를 자세히 소개하고 관찰도수 벡터와 카이제곱 검정통계량의 점근분포를 유도한다. 3장에서는 2장에서 유도된 카이제곱 검정통계량의 점근분포가 유한표본(finite samples)에서 얼마나 정확한지 알아보는 모의실험을 제공한다. 또한 두 개의 비정규 분포에서 검정력을 비교하는 모의실험을 실시한다. 이 모의실험에는 이 연구에서 제안된 검정 뿐만 아니라 기존 연구에서 검정력이 우수한 것으로 드러난 Mardia(1970), Ozturk, Romeu(1992), Baringhaus, Henze(1988)가 포함된다.

2. 구체적 절차와 주요 결과들

이 절에서 반복되는 표기법이 여러 가지 나오기 때문에 그것을 먼저 소개하자. I, e 및 0은 각각 단위행렬(identity matrix), 모든 원소가 1인 벡터 및 모든 원소가 0인 벡터나 행렬을 나타낸다. 차수를 나타낼 때는 첨자로서 나타내고 차수가 문맥상 분명하면 생략하기로 한다. 모든 벡터는 열벡터이지만 편의상 논문에서 원소를 나열할 때는 행벡터로 나타내기도 한다. 또한 평균이 μ이고 공분산행렬이 Σ인 p-차원 다변량 정규분포를 $N_p(\mu, \Sigma)$, 자유도가 k인 카이제곱분포의 누적분포함수를 $F_k(\cdot)$로 나타내기로 한다.

이 연구에서 제안하는 절차를 구체적으로 살펴보면 다음과 같다. 우선 원자료 벡터 $Y_1, \ldots, Y_n$는 다변량 정규분포에서의 확률표본이라고 가정한다. 이 원자료 벡터에서 척도화 잔차(scaled

residuals) $Z_1, \ldots, Z_n$을 계산한다. $Y_1, \ldots, Y_n$의 표본평균벡터를 $\overline{Y}=\sum_{i=1}^{n} Y_i/n$, 표본공분산행렬을 $S=\sum_{i=1}^{n}(Y_i-\overline{Y})(Y_i-\overline{Y})^t/n$라고 놓았을 때, 척도화 잔차는

$$Z_i=R(Y_i-\overline{Y}), \quad i=1, \ldots, n$$

과 같이 계산된다. (이 논문에서는 검정통계량의 점근분포 유도 과정에서 표기의 편의성을 위해 분모가 n인 표본공분산행렬을 사용한다.) 여기서 $R=R(S)$는 $RSR^t=I$를 만족하는 S의 함수로서 이렇게 계산된 척도화 잔차들은 표본평균이 0이고 표본공분산행렬이 단위행렬 I가 되는 것이다. 이렇게 척도화된 잔차를 구할 때 사용되는 함수 $R=R(S)$에는 여러 가지가 사용될 수 있다. 예를 들어 주성분 방법에서는 $R=D\Gamma$ 형태로 나타내는 것인데 여기서 Γ는 직교행렬(orthogonal matrix), D는 대각행렬(diagonal matrix)이다. 그리고 Gram-Schmidt 방법은 R을 대각선의 값이 양수인 하삼각행렬(lower triangular matrix)로 취하는 방법이고, 기타 일상적으로 사용되는 방법으로 $R=S^{-1/2}$가 있다. 여기서 우리가 유용하게 이용할 수 있는 통계적 사실은 Huffer, Park (2002)의 Lemma 3.1에 의해 $RSR^t=I$를 만족하는 S의 함수이면 어떤 함수를 사용하더라도 척도화 잔차의 분포에는 영향이 없다는 것이다. 따라서 지금부터 특별히 언급하지 않을 경우 Gram-Schmidt 방법을 사용하며 이 경우 $R(I)=I$라는 성질을 만족하게 되는 것을 알 수 있다.

다음으로 앞에서 계산된 척도화 잔차를 이용하여 기대도수가 근사적으로 동일한 칸들을 다음과 같이 생성하게 된다. 우선 처음으로 고려되는 칸 형태는 척도화 잔차의 각 변수가 양수인지 여부로 결정되는 칸이다. 다시 말해 여기서는 칸의 기본형태가

$$G=\{(z_1, \ldots, z_p)^t : z_i>0 \text{ 혹은 } z_i\leq 0 \ \forall \ i\}$$

로 주어지게 되어 총 $g\equiv 2^p$개의 칸 $G_1, \ldots, G_g$가 생성된다. 이를 이차원에서 생각하면 (0인 점을 빼고 생각한다면) 사분면(quadrant)이 되는 것을 쉽게 알 수 있다. 다음으로 고려되는 칸 형태는 척도화 잔차의 거리제곱, 즉 원자료의 마할라노비스 거리제곱(squared Mahalanobis distances)을 분할하는 칸으로서 카이제곱 분위수에 의해 각 칸에 속할 확률이 근사적으로 동일하게 만든다. 다시 말해 이 칸의 형태는

$$S_j=\{z\in R^p : q_{j-1}\leq z^tz < q_j\}, \quad j=1, \ldots, c \tag{2.1}$$

로서 각 칸에 척도화 잔차의 거리제곱이 속할 확률이 근사적으로 동일하게 되는 것이다. 여기서 $q_j=F_p^{-1}(j/c)$는 자유도가 p인 카이제곱 분포의 (j/c)-분위수이며 $q_0=0, q_c=\infty$로 놓는다. 이 두 가지 형태의 칸을 결합하여 최종적인 칸

$$\Lambda_{ij}=G_i\cap S_j,\quad i=1,\ldots,2^p;j=1,\ldots,c$$

이 완성되게 된다. 앞으로 Λ_{ij}의 개수를 $K\equiv gc=2^pc$로 표기하여 사용하도록 한다.

앞에서 생성된 칸들에 의해 관찰도수

$$U_{ij}^{(n)}=\sum_{k=1}^{n}I(z_k\in\Lambda_{ij}),\quad i=1,\ldots,g;\ j=1,\ldots,c$$

가 계산되며 Pearson-Fisher 카이제곱 통계량

$$X^2=\sum_{i=1}^{g}\sum_{j=1}^{c}(U_{ij}-np_0)^2/(np_0)$$

이 계산되어 검정통계량으로 사용될 수 있다. 여기서 $p_0=1/K=1/(gc)$는 척도화 잔차가 특정 칸에 속하게 될 근사적 확률로서, 모두 동일한 값을 가지도록 칸을 생성했기 때문에 제안된 카이제곱 통계량을 쉽게 계산할 수 있음을 알 수 있다.

이제 주요 정리를 나타내는데 필요한 표기법을 소개하도록 하겠다. $U_n=(U_{ij}^{(n)})$은 K-벡터인 관찰도수 벡터를 나타낸다. 이 벡터는 척도화 잔차의 함수이며 이것을 명확히 할 필요가 있을 때는 $U_n(Z_1,\ldots,Z_n)$라고 표기하도록 하겠다. 결과를 쉽게 나타내기 위해서 U_n의 원소들은 표준순서에 의해 나열되었다고 가정한다. 다시 말해 먼저 i가 1에서 g까지 먼저 변하고 그 다음에 j가 1에서 c까지 변하는 순서이다. 여기서 $G_1,\ldots,G_g$가 지정되는 순서인 i도 표준순서, 즉 $z_1>0$ 에서 $z_1\le 0$이 가장 먼저 바뀌고, 다음으로 $z_2>0$ 에서 $z_2\le 0$로 두 번째 빨리 변하고, $z_p>0$ 에서 $z_p\le 0$로 가장 나중에 바뀐다고 가정한다. 따라서 이 순서에 의하면

$$G_1=\{(z_1,\ldots,z_p)^t:\ z_i>0\ \forall i\},\ G_g=\{(z_1,\ldots,z_p)^t:\ z_i\le 0\ \forall i\}$$

가 되는 것을 쉽게 알 수 있다.

또한 D_1와 D_2을 $g=2^p$ 요인설계(factorial design)의 주효과(main effects) 및 1차 상호작용효과(interaction effects) 나타내는 설계행렬(design matrix)이라고 정의하자. 다시 말해 D_1은 $g\times p$ 행렬로서 i번째 열이 $(-e_{2^{i-1}},e_{2^{i-1}})$벡터를 2^{p-i} 되풀이하여 만들어지며, D_2 는 $g\times p(p-1)/2$ 행렬로서 D_1으로부터 모든 가능한 다른 두 열을 곱하여 만들어진다.

마지막으로

$$A=\begin{pmatrix}a_1D_1\\a_2D_1\\\vdots\\a_cD_1\end{pmatrix},\quad B=\begin{pmatrix}b_1e_ge_p^t\\b_2e_ge_p^t\\\vdots\\b_ce_ge_p^t\end{pmatrix},\quad C=\begin{pmatrix}c_1D_2\\c_2D_2\\\vdots\\c_cD_2\end{pmatrix}\tag{2.2}$$

를 정의한다. 여기서

$$a_i = E[z_1 I(Z \in \Lambda_{1i})],\ b_i = E[z_1^2 I(Z \in \Lambda_{1i})] - p_0,\ c_i = E[z_1 z_2 I(Z \in \Lambda_{1i})]$$

로서 $Z=(z_1, \ldots, z_p)^t$는 분포가 $N_p(0, I)$인 확률변수이며 $\Lambda_{1i} = G_1 \cap S_i$로서 G_1 및 S_i는 각각 (2.2) 직전과 (2.1)에 정의되어 있다.

이제 관찰도수 벡터 U_n의 점근분포와 카이제곱 검정통계량 X^2의 극한분포를 다음과 같이 유도할 수 있다.

정리 1. $Y_1, Y_2, \ldots, Y_n$이 평균이 μ이고 정칙인(nonsingular) 공분산행렬인 Σ를 가지는 다변량 정규분포 $N_p(\mu, \Sigma)$에서의 확률표본이면, $n \to \infty$일 때 다음이 성립한다.

$$n^{-1/2}(U_n - np_0 e) \to N_K(0, \Psi)$$

여기서

$$\Psi = p_0 I - p_0^2 ee^t - AA^t - BB^t/2 - CC^t$$

로서 A, B 및 C는 (2.2)에 정의되어 있으며 $p_0 = 1/K = 1/(gc)$, $g = 2^p$이다.

증명: Huffer, Park(2002)의 Lemma 3.1에 의해 관찰도수 벡터 $U_n(Z_1, \ldots, Z_n)$은 보조통계량(ancillary statistic)이며 따라서 $\theta = (\mu, \Sigma)$의 완비충분통계량(complete and sufficient statistic)인 $(\overline{Y}, S)$와 독립이 된다. (앞의 절차 설명에서도 설명했듯이 척도화 잔차를 계산할 때 $R(S)$를 구하는 방법이 다양하지만 척도화 잔차의 분포는 동일하기 때문에 Gram-Schmidt 방법을 사용하는 것으로 가정하고 증명한다.) 따라서 모든 θ에 대해

$$\mathcal{L}_\theta(U_n(Z_1, \ldots, Z_n)) = \mathcal{L}_{\theta_0}(U_n(Z_1, \ldots, Z_n))$$
$$= \mathcal{L}_{\theta_0}(U_n(Z_1, \ldots, Z_n) | \overline{Y} = 0, S = I) = \mathcal{L}_{\theta_0}(U_n(Y_1, \ldots, Y_n) | \overline{Y} = 0, S = I)$$

가 성립한다. 여기서 $\theta_0 = (0, I)$는 고정된 값이며, 마지막 등식은 $\overline{Y} = 0, S = I$ 조건이 만족되면 척도화 잔차 $Z_i = R(S)(Y_i - \overline{Y})$는 $R(I) = I$에 의해 Y_i와 일치하기 때문이다.

Park(1995)의 조건부 극한이론(conditional limit theorem)을 적용하기 위해서는 앞의 식의 조건인 $\overline{Y} = 0, S = I$를 정준충분통계량(canonical sufficient statistic)으로 나타낼 필요가 있다. 따라서 임의의 p-벡터 $y = (y_1, \ldots, y_p)$에 대해 $s(y) = (y, d(y), r(y))$라고 정의하자. 여기서 $d(y) = (y_1^2, \ldots, y_p^2)$는 제곱으로 구성되는 벡터이며 $r(y) = (y_1 y_2, \ldots, y_{p-1} y_p)$는 서로 다른 두 원소의 곱으로 구성되는 벡터이다. 그러면 정준충분통계량은 $\sum_{i=1}^{n} s(y_i)$가 되며 앞의 식에서 주어진

조건은 $\sum_{i=1}^{n} s(y_i)/n=(0_p, e_p, 0_{p(p-1)/2})$가 된다. 따라서 앞의 식은

$$\mathcal{L}_{\theta_0}\Big(U_n(Y_1,\ldots,Y_n)\,|\,\sum_{i=1}^{n} s(y_i)/n=(0_p, e_p, 0_{p(p-1)/2})\Big)$$

가 된다. 다변량 정규분포에서는 Park(1995)의 Corollary 1의 조건을 만족하는 것을 쉽게 보일 수 있다. 따라서 $n\to\infty$일 때

$$n^{-1/2}(U_n-n p_0 e) \to N_K(0, E_1-E_2E_3^{-1}E_2^t)$$

가 성립한다. 여기서

$$E_1=\mathrm{Cov}_{\theta_0}(U_1(Y_1)),\ E_2=\mathrm{Cov}_{\theta_0}(U_1(Y_1), s(Y_1)),\ E_3=\mathrm{Cov}_{\theta_0}(s(Y_1))$$

이다. 그런데

$$E_1=p_0 I-p_0^2 ee^t,\ E_2=(A,B,C),\ E_3=\mathrm{diag}(I_p, 2I_p, I_{p(p-1)/2})$$

를 만족하는 것을 쉽게 알 수 있다. 따라서

$$\Psi=E_1-E_2E_3^{-1}E_2=p_0 I-p_0^2 ee^t-AA^t-BB^t/2-CC^t$$

가 성립하여 증명이 완료된다. □

정리 2. 정리 1의 조건하에서 $n\to\infty$일 때 다음이 성립한다.

$$X^2 \to W_1+(1-v_1)W_2+(1-v_2)W_3+(1-v_3)W_4$$

여기서 W_1, W_2, W_3, 및 W_4는 서로 독립인 카이제곱 확률변수로서 각각의 자유도가 $K-2-p-p(p-1)/2$, p, 1 및 $p(p-1)/2$이며

$$v_1=g\sum_{i=1}^{c} a_i^2/p_0,\ v_2=pg\sum_{i=1}^{c} b_i^2/(2p_0),\ v_3=g\sum_{i=1}^{c} c_i^2/p_0$$

로서 a_i, b_i 및 c_i는 (2.2)에서 정의되어 있으며 $g=2^p$이다.

증명: $X^2=(U_n-np_0e)^t(U_n-np_0e)/(np_0)$가 만족하기 때문에 X^2의 점근분포는 $\sum_{i=1}^{d^*}\lambda_i W_i$가 된다. 여기서 λ_i는

$$H=\Psi/p_0=I-p_0 ee^t-AA^t/p_0-BB^t/(2p_0)-CC^t/p_0 \tag{2.3}$$

의 고유값이며 W_i는 서로 독립인 자유도가 1인 카이제곱 확률변수이다. 그런데 D_1와 D_2가 $g=2^p$ 요인설계의 주효과 및 1차 상호작용효과를 나타내는 설계행렬이기 때문에 $D_1 e=0, D_2 e=0, D_1^t D_2=0$이 되며 간단한 연산을 통해 $\sum_{i=1}^{c} b_i=0$가 되는 것을 알 수 있다. 따라서 AA^t, BB^t, CC^t 및 ee^t는 서로 직교한다. 그런데

$$A^tA=u_1 I_p,\ (BB^t)^2=u_2 BB^t,\ C^tC=u_3 I_{p(p-1)/2}$$

를 만족하기 때문에 AA^t/u_1, BB^t/u_2, CC^t/u_3는 각각 계수가 $p, 1, p(p-1)/2$인 멱등행렬(idempotent matrix)이 된다. 여기서

$$u_1=g\sum_{i=1}^{c} a_i^2,\ u_2=pg\sum_{i=1}^{c} b_i^2,\ u_3=g\sum_{i=1}^{c} c_i^2$$

이다. 따라서 (2.3)의 행렬 H의 고유값은 $1-u_1/p_0$이 p번, $1-u_2/(2p_0)$가 1번, $1-u_3/p_0$가 $p(p-1)/2$번, 0이 1번 되풀이 된다. 따라서 증명이 완료된다. □

그런데 정리 2의 상수인 v_1, v_2 및 v_3을 계산하는 공식은 다음과 같이 간편화 할 수 있다. 먼저 $i=0, 1, \ldots, c$에 대하여

$$a_i^*=F_{p+1}(q_i)-F_{p+1}(q_{i-1}),\ b_i^*=F_{p+2}(q_i)-F_{p+2}(q_{i-1}) \tag{2.4}$$

를 정의하자. 여기서 $F_{p+1}(\cdot)$와 $F_{p+2}(\cdot)$는 각각 자유도가 $p+1$ 및 $p+2$인 카이제곱 분포의 누적분포함수이며, q_i는 자유도가 p인 카이제곱 분포의 (i/c)-분위수로서 (2.1)에 정의되어 있다. 그러면 여러 번의 연산을 통해

$$v_1=\frac{2c}{\pi}\sum_{i=1}^{c} a_i^{*2},\ v_2=\frac{pc}{2}\sum_{i=1}^{c}(b_i^*-1/c)^2,\ v_3=\frac{4c}{\pi^2}\sum_{i=1}^{c} b_i^{*2}$$

가 성립한다는 것을 보일 수 있다.

4. 모의실험

이 절에서는 두 가지 유형의 모의실험을 수행한다. 먼저 정리 2에 유도된 X^2의 극한분포가 유한표본에서 얼마나 정확한지 알아보는 모의실험을 수행한다. 그리고 두 개의 비정규 분포에서 이 연구에서 제안한 카이제곱 검정과 Mardia(1970), Ozturk, Romeu(1992), Baringhaus, Henze(1988)의 검정력을 비교하는 모의실험을 실행한다.

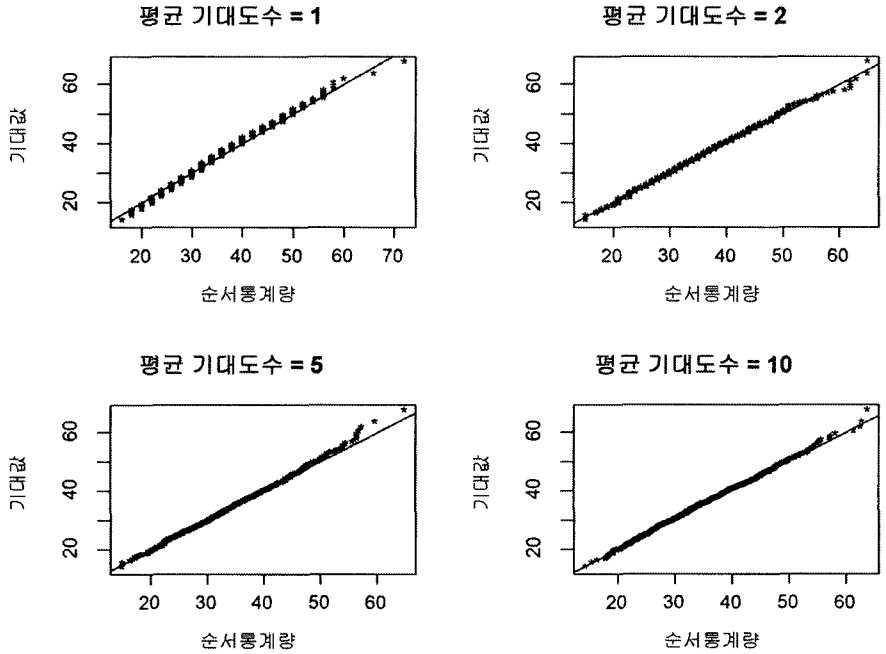
먼저 X^2의 극한분포의 정확성을 알아보는 모의실험을 수행한다. 셀의 수인 K가 너무 작지 않지만 유사한 결과를 제공하기 때문에 여기서는 하나의 모의실험 결과만 제시하기로 한다. 이 모의실험에서 사용하는 값은 $p=3, c=5$로서 이 경우 카이제곱 검정통계량의 극한분포는

$$\chi^2(32)+0.261\chi^2(3)+0.179\chi^2(1)+0.373\chi^2(3) \tag{3.1}$$

가 된다. 네 가지 표본크기를 고려했는데 평균 기대도수 np_0가 1, 2, 5 및 10인 경우에 해당되는 표본크기 40, 80, 200 및 400이다.

구체적인 모의실험 절차는 다음과 같다. 각 표본크기 n에 대해 $N_3(0, I)$에서 표본크기 n인 1000개의 표본을 생성하여 각각의 카이제곱 검정통계량을 계산한다. 이 1000개 카이제곱 통계량으로 순서통계량을 계산하여 순서통계량의 기대값과 함께 산포도를 그리는 것이다. 카이제곱 확률변수의 선형결합인 분포에서 순서통계량의 기대값을 직접 계산하는 것은 번거롭기 때문에 여기서는 모의실험에 의존하여 기대값을 계산한다. 구체적으로 카이제곱 통계량의 극한분포 (3.1)에서 표본크기 일백만인 확률표본을 생성하여 그 표본의 500번째 순서통계량에서 출발하여 매 1000번째 순서통계량을 취하는 방법이다. 모의실험의 결과인 카이제곱 확률도(chi-square probability plot)를 정리한 것이 <그림 1>이다.

각 카이제곱 확률도에는 원점을 지나면 기울기가 1인 직선을 그려놓았는데 이것은 순서통계량



<그림 1> 네 가지 표본크기일 때의 카이제곱 확률도

과 기대값이 정확히 일치하는 이상적인 경우에 해당되는 참고 직선이다. 이 그림들을 살펴보면 극한분포가 아주 좋은 근사분포라는 것을 알 수 있다. 심지어 평균 기대도수가 1인 경우에도 약간의 이산성을 제외하면 극한분포의 정확성이 우수한 것을 알 수 있다.

다음으로 이 연구에서 제안한 카이제곱 검정과 Mardia(1979), Ozturk, Romeu(1992), Baringhaus, Henze(1988)의 검정력을 비교하는 모의실험을 실행한다. 결과를 나타낼 때 편의를 위해 카이제곱 검정은 X^2, Mardia의 왜도와 첨도 검정은 각각 Skew와 Kurt, Ozturk, Romeu (1992), Baringhaus, Henze(1988)의0 검정은 각각 OR과 BH로 나타내기로 한다.

첫 번째 고려된 분포는 다음과 같은 확률밀도함수를 가지고 있다.

$$f(y_1, y_2) = \frac{1}{2\pi}\{1+\cos[2(\theta-2r)]\}\exp(-r^2/2) \tag{3.2}$$

여기서 r과 θ는 (y_1, y_2)의 극좌표 값인 반지름과 각도이다. 이 확률밀도함수의 산포도는 두 개의 대칭인 나선을 가지며 $N_2(0, I)$와 4차 적률(moments)까지 비슷한 값을 가진다. X^2을 계산할 때 $c=5$를 사용하고 1000 번의 반복실험을 하였을 때 다섯 가지 검정의 검정력을 요약한 것이 <표 1>이다.

<표 1> 다섯 가지 다변량 정규성 검정의 (3.2) 분포에 대한 검정력

	$n=100$			$n=200$		
α	.01	.05	.1	.01	.05	.1
X^2	.420	.671	.788	.870	.948	.971
Skew	.009	.048	.095	.014	.055	.113
Kurt	.004	.017	.051	.005	.025	.072
OR	.028	.110	.180	.046	.109	.191
BH	.029	.112	.176	.048	.163	.285

이 표에서 카이제곱 검정은 탁월한 검정력을 가지고 있으며 나머지 네 가지 검정 중 OR과 BH만 다소간의 검정력을 가지고 있음을 알 수 있다.

다음으로 고려된 분포는 다음과 같은 확률밀도함수를 가진다.

$$f(y_1, y_2, y_3, y_4) = \frac{1}{2\pi^2}\exp(-r^2/2)I\big(y_1y_2y_3y_4(r^2-m)<0\big) \tag{3.3}$$

여기서 $r^2=\sum_{i=1}^{4} y_i^2$이며 $m=F_4^{-1}(1/2)$으로 자유도가 4인 카이제곱 분포의 중앙값이다. 이 분포는 $N_4(0, I)$의 확률밀도함수의 각 사분면에서 반지름 $\sqrt{m}$ 기준으로 안쪽이나 바깥쪽 중 한 쪽만 나타날 수 있도록 만든 분포이다. X^2을 계산할 때 $c=2$를 사용하고 1000 번의 반복실험을 하였을 때 다섯 가지 검정의 검정력을 요약한 것이 <표 2>이다.

<표 2> 다섯 가지 다변량 정규성 검정의 (3.3) 분포에 대한 검정력

	$n=100$			$n=200$		
α	.01	.05	.1	.01	.05	.1
X^2	.014	.080	.148	.058	.178	.269
Skew	.010	.049	.084	.012	.040	.090
Kurt	.004	.024	.070	.009	.035	.087
OR	.014	.068	.118	.008	.051	.108
BH	.013	.067	.124	.028	.120	.219

이 표에서는 다섯 가지 검정 중 카이제곱 검정과 BH 만이 다소간의 검정력을 가지며 카이제곱 검정이 BH보다 약간 우위를 가지고 있는 것을 알 수 있다.

참고문헌

[1] Andrews, D. Gnanadesikan, R., and Warner, J. (1973). Methods for assessing multivariate normality, In: P.R. Krishnaiah (Ed.) *Proceedings of the International Symposium on Multivariate Analysis,* Vol. 3, Academic Press, New York.

[2] Baringhaus, L. and Henze, N. (1988). A consistent test for multivariate normality based on the empirical characteristic function, *Metrika*, 35, 339-348.

[3] Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations,* Wiley, New York.

[4] Huffer, F.W., and Park, C. (2002). The limiting distribution of a test for multivariate structure, *Journal of Statistical Planning and Inference,* 105, 417-431.

[5] Leoney. S. (1995). How to use tests for univariate normality to assess multivariate normality, *The American Statistician,* 39, 75-79.

[6] Manzotti, A., Quiroz, A.J. (2001). Spherical harmonics in quadratic forms for testing multivariate normality, *Test,* 10, 87-104.

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

[8] Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Analysis,* Academic Press, New York.

[9] Ozturk, A. and Romeu, J.L. (1992). A new method for assessing multivariate normality with graphical applications, *Communications in Statistics - Simulations and Computation,* 21, 15-34.

[10] Park, C. (1995). Some remarks on the chi-squared test with both margins fixed, *Communications in Statistics - Theory and Methods,* 24, 653-61.

[11] Romeu, J.L. and Ozturk, A. (1993). A comparative study of goodness-of-fit tests for multivariate normality, *Journal of Multivariate Analysis,* 46, 309-34.