응용통계연구 제 18권 1호, 2005년, pp. 99-113

구형 대칭성 검정에 대한 연구 *

박철용[1)]

요 약

이 논문에서는 카이제곱 구형 대칭성 검정을 제안한다. 이 검정은 검정통계량과 점근적 유의확률을 쉽게 계산할 수 있는 장점이 있다. 이 통계량의 구형 대칭성 가정하의 극한 분포를 도출하고 유한표본에서 잘 부합되는지 모의실험을 통해 살펴본다. 또한 다양한 대립분포에서 기존의 구형 대칭성 검정과 검정력을 비교하는 모의실험을 수행하며 마지막으로 실제 자료 분석 예제를 제공한다.

주요용어: 극좌표, 다변량 대칭, 카이제곱 검정

1. 서론

단변량 변수의 대칭성 검정은 비교적 단순하며 많이 연구되어 왔으나 다변량 변수의 대칭성 검정은 복잡하고 잘 알려져 있지 않다. 다시 말해 단변량 대칭성에 대한 연구는 Csörgö & Heathecote(1987), Schuster & Barker(1987), Arcones & Giné(1991) 등에서 알 수 있듯이 상대적으로 활발하게 진행되어 왔으나, 다변량 대칭성은 정의될 수 있는 형태가 다양함에도 불구하고 여러 형태의 대칭성을 탐색하는 검정방법에 대한 연구가 상당히 미진했다. 예를 들어 Beran(1979)의 타원형 대칭성(elliptical symmetry)에 대한 검정, Romano(1989)가 고려한 여러 가지 형태의 붓스트랩(bootstrap) 검정 중 하나인 구형 대칭성(spherical symmetry) 검정, Baringhaus(1991)의 구형 대칭성에 대한 von Mises 형태의 검정, Heathcote, Rachev & Cheng(1995)의 대각대칭성(diagonal symmetry)에 대한 경험적 특성함수(empirical characteristic function)에 근거한 붓스트랩 검정, Koltchinskii & Li(1998)의 구형 대칭성에 대한 Kolmogorov-Smirnov 형태의 붓스트랩 검정, 그리고 최근에 발표된 Manzotti, Perez & Quiroz(2002)와 Schott(2002)의 타원형 대칭성 검정 등이 다변량 대칭성에 관련된 대표적인 연구결과라고 할만큼 그 연구가 그리 활발하지 못했다.

여러 다변량 대칭분포 중 타원형 대칭분포는 다변량 정규분포를 준모수(semi-parametric) 다변량 분포로 확장한 것으로 실제 자료분석에서 유용하게 사용될 수 있다. 왜냐하면 다변량 정규분포를 가정한 통계절차보다 로버스트 분석을 행할 수 있는 기초가 되는 분포이기 때문이다. 구형 대칭분포는 타원형 대칭분포의 특수한 경우로서 실제 자료분석에서는 타원형 대칭분포보다 응용성이 떨어진다. 그러나 구형 대칭성 검정이 바로 타원형 대칭성 검정으로 변환될 수 있는 이론적 밀접성 때문에 타원형 대칭성 검정의 선행 연구과제로서 많

* 본 연구는 한국과학재단 목적기초연구(R05-2003-000-10926-0)지원으로 수행되었음.

1) (704-701) 대구시 달서구 신당동 1000번지, 계명대학교 자연과학대학 통계학과, 부교수
E-mail: cypark1@kmu.ac.kr

이 연구된다. 다시 말해 원자료 대신에 평균벡터가 0이고 공분산행렬이 단위행렬인 구형화된 자료(spherized data)에 구형 대칭성 검정을 적용하는 것이 일종의 타원형 대칭성 검정이 될 수 있는 것이다. 따라서 이 논문에서는 이론적 도출이 용이한 구형 대칭성 검정에 대한 연구로 한정하지만 이론적으로 정밀화하는 과정을 거치면 타원형 대칭성 검정으로 연장될 수 있을 것이다.

기존의 연구에서 구형 대칭성 검정으로 사용될 수 있는 것은 Beran(1979), Romano(1989), Baringhaus(1991), Koltchinskii & Li(1998) 등이다. 이 연구들은 실제 적용에 있어 많은 어려움을 가지고 있다. 왜냐하면 각 통계량들이 계산하기 복잡하며 특히 Beran(1979)의 연구를 제외하면 모두 검정통계량의 점근분포를 쉽게 계산할 수 없어 점근적 유의확률을 얻기 힘들기 때문이다. 이런 어려움을 극복하기 위해서 이 논문에서는 검정통계량을 쉽게 계산할 수 있을 뿐만 아니라 또한 이 검정통계량의 점근적 유의확률을 쉽게 계산할 수 있는 카이제곱 검정을 제시하고자 한다. 또한 기존의 연구에서는 자신들이 제안한 통계량을 벗어난 모의실험이 거의 행해지지 않았다. 따라서 이 논문에서는 우리가 제안한 검정통계량 뿐만 아니라 Beran(1979)과 Baringhaus(1991)도 포함시켜 다양한 모의실험을 수행하고자 한다. 다시 말해 구형 대칭인 분포에서 이들 통계량들의 유한표본 분포가 점근분포에 잘 적합되는지 살펴보는 모의실험과 함께 구형 대칭이 아닌 여러 다양한 분포에서 이들의 검정력을 비교하는 모의실험을 수행하게 된다.

이 논문은 다음과 같이 구성되어 있다. 2절에서는 이 논문에서 제안하는 검정통계량을 자세히 소개하고 이 검정통계량의 극한 귀무분포(limiting null distribution)를 유도한다. 3절에서는 우리 검정통계량의 극한 귀무분포가 유한표본에서 얼마나 정확성(accuracy)을 가지는지 간단한 모의실험을 통해 살펴보고 Beran(1979)과 Baringhaus(1979)의 점근 귀무분포가 유한표본에서 얼마나 정확한지도 함께 살펴본다. 또한 다양한 대립 분포에서 기존의 두 가지 검정통계량과 검정력을 비교하는 모의실험을 수행한다. 4절에서는 실제 자료분석 예제를 제공하고 5절에서는 이 연구의 결과를 요약하고 결론을 내린다.

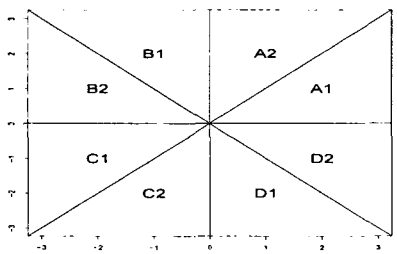
2. 검정통계량과 극한 귀무분포

이 절에서는 먼저 구형 대칭성에서 어떻게 이 논문에서 제안하는 검정통계량이 유도될 수 있는지 직관적인 면에서 소개하고, 다음에 구체적인 검정통계량과 극한 귀무분포를 이론적으로 유도한다. 이 논문에서 사용하는 구형 대칭성의 정의는 다음과 같다: 모든 직교행렬(orthogonal matrix) Γ에 대해 ΓX와 X의 분포가 같으면 X를 구형대칭이라 한다. 평균이 0 공분산이 $\sigma^2 I$인 다변량 정규분포, 다변량 t분포 등이 잘 알려진 구형 대칭분포이다. 다음 절의 모의실험에서는 구형대칭인 분포족(family of distributions)으로서 Ernst(1998)의 다변량 일반화 라플라스 분포(multivariate generalized Laplace distribution)를 고려하는데 이것은 다변량 정규 분포와 다차원 공(ball)에서의 균일분포 등을 포함하게 된다.

이차원에서의 직교행렬

$$\Gamma(\theta) = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$$

을 사용하면 확률벡터 X의 축을 θ만큼 반시계 방향으로 회전시킬 수 있다. 다시 말해 $\Gamma(\theta)X$는 원래 확률벡터 X의 좌표축을 θ만큼 반시계 방향으로 회전시킨 새로운 좌표축을 가진다. 따라서 구형대칭인 X는 원래 좌표축에서나 회전된 좌표축에서나 분포가 동일하게 된다. 따라서 다음과 같은 절반사분면(half-quadrant) A1, A2, ..., D1, D2의 확률을 쉽게 계산할 수 있다.



원래 좌표축의 1, 2 사분면은 각각 반시계 방향으로 π만큼 회전한 좌표축의 3, 4 사분면에 해당되기 때문에 구형 대칭성 하에서 1, 2 사분면의 확률은 각각 3, 4 사분면의 확률과 동일하다. 마찬가지 방법으로 $\pi/2$만큼 회전한 경우를 생각하면 1, 3 사분면의 확률은 각각 2, 4 사분면의 확률과 동일하게 된다. 따라서 각 사분면의 확률이 동일하게 되어 그 값이 1/4이 된다. 마찬가지로 $\pi/4$만큼 회전한 경우를 생각하면 A1, B1, C1, D1의 확률은 A2, B2, C2, D2의 확률과 동일하게 되어 각각의 확률이 1/8이 되는 것이다. 이와 같이 사분면을 동일한 각도를 가지도록 반분시켜 나가면 각 영역이 동일한 확률을 가지도록 만들 수 있다. 이렇게 동일 확률을 가지는 영역에 속하는 관찰도수와 기대도수를 비교하는 카이제곱 통계량으로 구형 대칭성을 검정할 수 있는 것이다.

이 아이디어는 쉽게 p차원으로 연장될 수 있다. p차원 확률벡터 $X = (X_1, \ldots, X_p)^t$에 대해 극좌표(polar coordinates) 값을 다음과 같이 구할 수 있다.

$$\begin{aligned} R &= \sqrt{X^t X} \\ X_1 &= R\cos(\theta_1) \\ &\vdots \qquad\qquad \vdots \\ X_{p-1} &= R\sin(\theta_1)\cdots\sin(\theta_{p-2})\cos(\theta_{p-1}) \\ X_p &= R\sin(\theta_1)\cdots\sin(\theta_{p-2})\sin(\theta_{p-1}) \end{aligned}$$

구형 대칭성이 만족되면 극좌표 값 $\theta_1, \ldots, \theta_{p-1}, R$은 서로 독립이며 θ_i의 확률밀도함수가 $\sin^{p-1-i}\theta_i$에 비례하게 된다(Muirhead (1982)의 정리 1.5.5 참조). 따라서 각 각도가 취하는 영역($0 \le \theta_{p-1} \le 2\pi$, $0 \le \theta_i \le \pi$, $i = 1, \ldots, p-2$)을 동일한 확률을 가지는 구간으로 분할하여 사각 칸(rectangular cells)을 구성하면 구형 대칭성 하에서 동일확률을 가지는 분할이 된다. 다만 카이제곱 검정을 보다 더 정교하게 만들기 위해 각도뿐만 아니라 (원점에서의) 반지름 R도 포함시킨다. 그런데 일반적으로 R의 분포는 구형 대칭인 분포의 형태에 따라

달라지기 때문에 고정구간에 의해 동일확률을 가지도록 나누는 것은 불가능하여 순서통계량에 근거한 확률구간(random interval)을 사용하게 된다.

구체적인 검정통계량의 모양을 알아보기 전에 필요한 표기법을 정의한다. 우선 p차원 다변량분포 $f(x)$에서의 확률표본을 $X_1, \ldots, X_n$으로 나타낸다. 단, 여기서 $X_i = (X_{i1}, \ldots, X_{ip})^t$, $i = 1, \ldots, n$이다. 이 때 우리가 검정하고자 하는 가설은 H_0 : "$f(x)$는 *구형대칭이다*" 라는 가설이 된다. 또한 $I_{i1}^{(d)}, \ldots, I_{id}^{(d)}$를 극좌표 상의 각도 θ_i가 취하는 영역($0 \le \theta_{p-1} \le 2\pi$, $0 \le \theta_j \le \pi$, $j = 1, \ldots, p-2$)을 d개의 동일확률로 만드는 분할이라고 놓는다. 다시 말해 $I_{ij}^{(d)} \cap I_{ik}^{(d)} = \emptyset$, $\forall j \ne k$,

$$\bigcup_{j=1}^{d} I_{ij}^{(d)} = \begin{cases} [0, \pi], & \text{만약 } i = 1, \ldots, p-2 \\ [0, 2\pi], & \text{만약 } i = p-1 \end{cases}$$

이고 $P(\theta_i \in I_{ij}^{(d)}) = 1/d$, $\forall j = 1, \ldots, d$를 만족한다. 마찬가지로 $R_1^{(d)}, \ldots, R_d^{(d)}$는 관찰벡터의 반지름 $R_1, \ldots, R_n$이 n/d개씩 들어가는 확률분할(random partition)로 다음과 같이 정의된다.

$$R_i^{(d)} = \begin{cases} (0, R_{(n/d)}], & \text{만약 } i = 1 \\ (R_{(n(i-1)/d)}, R_{(n \times i/d)}], & \text{만약 } i = 2, \ldots, d-1 \\ (R_{(n(d-1)/d)}, \infty), & \text{만약 } i = d \end{cases}$$

여기서 $R_{(1)} \le \cdots \le R_{(n)}$은 $R_1, \ldots, R_n$의 순서통계량으로 n이 d의 배수라고 가정한다. 만약 n이 d의 배수가 아니면 $n \times k/d$ 대신에 이것을 초과하지 않는 최대 정수인 $[n \times k/d]$로 대체하면 된다($k = 1, \ldots, d-1$).

구형 대칭성을 탐색하기 위해 이 논문에서 제시하는 검정의 절차는 다음과 같이 요약될 수 있다. 우선 각 관찰벡터의 극좌표를 구해야 하는데 i번째 관찰벡터 X_i에서 계산된 극좌표 벡터를 $(R_i, \theta_{i1}, \ldots, \theta_{i(p-1)})^t$라고 나타낸다. 물론 여기서 극좌표 상의 각도가 취하는 영역은 $0 \le \theta_{i(p-1)} \le 2\pi$, $0 \le \theta_{ik} \le \pi$, $k = 1, \ldots, p-2$이다. 다음 절차는 극좌표 값을 범주형 벡터 $T_i = (T_{i1}, \ldots, T_{ip})^t$로 치환하는 것이다. 여기서 T_{ij}는 1에서 d_j까지의 범주를 취하며 특정 범주에 속하는 확률이 (최소한 근사적으로) 동일하도록 다음과 같이 만든다.

$\theta_{ij} \in I_{jk}^{(d_j)}$ $(j = 1, \ldots, p-1)$이면 $T_{ij} = k$로 놓고, $R_i \in R_k^{(d_p)}$이면 $T_{ip} = k$로 놓는다.

이 범주형 벡터로 부터 p차원 $d_1 \times \cdots \times d_p$ 분할표를 구성하면 각 칸에 속할 확률이 (최소한 근사적으로) 동일하게 되며 분할표의 특정 칸 $\pi = (\pi_1, \ldots, \pi_p)$에 속하는 관찰도수 U_π는 다음과 같이 나타낼 수 있다.

$$\begin{aligned} U_\pi &= \sum_{i=1}^{n} I(T_i = \pi) \\ &= \sum_{i=1}^{n} I\left(\theta_{i1} \in I_{1\pi_1}^{(d_1)}, \ldots, \theta_{i(p-1)} \in I_{(p-1)\pi_{p-1}}^{(d_{p-1})}, R_i \in R_{\pi_p}^{(d_p)}\right) \end{aligned}$$

이 관찰도수에서 피어슨-피셔(Pearson-Fisher) 카이제곱 통계량

$$X^2 = \sum_{\pi} \left(U_\pi - n \Big/ \prod_{i=1}^{p} d_i \right)^2 \Big/ \left(n \Big/ \prod_{i=1}^{p} d_i \right)$$

을 계산하여 검정통계량으로 사용하게 된다. 이 검정통계량은 각 관찰도수의 (근사적) 기대도수가 $n/\prod d_i$로 동일하기 때문에 계산하기 아주 용이하다.

앞에서 범주수 $d_1, d_2, \ldots, d_p$를 결정하는 것이 자의적이다. 최적의 범주수를 결정하는 것은 쉬운 문제가 아니기 때문에 여기서는 통상적으로 적용될 수 있는 규칙을 제안하고자 한다. 그것은 관찰도수의 기대도수 $n/\prod d_i$가 가능하면 5를 넘도록 범주수를 선택하고 이것이 불가능하면 기대도수가 최대한 커지도록 범주수를 결정하는 것이다. 그런데 차원이 크면 카이제곱 검정에는 일반적으로 차원의 저주(curse of dimensionality) 문제가 발생하기 때문에, 이 경우 표본크기가 아주 크지 않으면 $d_1 = \cdots = d_p = 2$로 선택할 수밖에 없는 경우가 종종 발생하게 될 것이다.

앞에서 제안한 카이제곱 검정통계량의 구형 대칭성 하의 극한 분포는 다음과 같다.

정리 2.1 $X_1, \ldots, X_n$이 구형 대칭분포에서의 확률표본일 때, $n \to \infty$이면 다음이 성립한다.

$$X^2 \xrightarrow{d} \chi^2 \left(\prod_{i=1}^{p} d_i - d_p \right)$$

증명: 표기법의 편의를 위해서 관찰도수 U_π로 구성된 열벡터 U가 표준 순서에 의해 나열되어 있다고 가정하자. 다시 말해 π_p가 제일 먼저 1에서 d_p까지 변하며, 다음으로 π_{p-1}이 두번째로 1에서 d_{p-1}까지 변하며, 마지막으로 π_1이 1에서 d_1까지 변하는 순서에 따라 U_π가 나열되어 관찰도수 전체의 열벡터 U가 구성된 것이다. 따라서 $\pi_p = k$에 해당되는 관찰도수들을 앞에서와 마찬가지 방법으로 나열한 열벡터를 U_k라고 표기하면 $U^t = (U_1^t, \ldots, U_{d_p}^t)$가 된다.

앞에서 $n_k \equiv \sum_\pi I(\pi_p = k) = \sum_{i=1}^{n} I(R_i \in R_k^{(d_p)})$는 (근사적으로) n/d_p가 되게 사전에 고정된 값을 가지므로 $U_1, \ldots, U_{d_p}$는 서로 독립이며 시행횟수가 각각 $n_1, \ldots, n_{d_p}$이며 성공의 확률이 각각 $q = e_1/\prod_{i=1}^{p-1} d_i$인 다항분포를 따른다. 여기서 e_1는 1이 $\prod_{i=1}^{p-1} d_i$개 반복되는 열벡터이다. 여기서 q의 제곱근 벡터 $\sqrt{q} \equiv e_1/\sqrt{\prod_{i=1}^{p-1} d_i}$를 정의하자. 그러면 $n \to \infty$일 때 다음이 성립한다.

$$\left(U_k - n_k\, e_1 \Big/ \prod_{i=1}^{p-1} d_i \right) \Big/ \sqrt{n_k \Big/ \prod_{i=1}^{p-1} d_i} \xrightarrow{d} N\left(0, I - \sqrt{q}\sqrt{q}^t\right)$$

그런데 $n_k = n/d_p + O(1)$이기 때문에 $n \to \infty$일 때

$$\left(U_k - n\, e_1 \Big/ \prod_{i=1}^{p} d_i \right) \Big/ \sqrt{n \Big/ \prod_{i=1}^{p} d_i} \xrightarrow{d} N\left(0, I - \sqrt{q}\sqrt{q}^t\right)$$

와

$$V \equiv \left(U - n\,e_2 \Big/ \prod_{i=1}^{p} d_i\right) \Big/ \sqrt{n \Big/ \prod_{i=1}^{p} d_i} \xrightarrow{d} N\left(0, \mathrm{diag}\left(I - \sqrt{q}\sqrt{q}^{\,t}, \ldots, I - \sqrt{q}\sqrt{q}^{\,t}\right)\right)$$

가 성립한다. 여기서 e_2는 1이 $\prod_{i=1}^{p} d_i$개 있는 열벡터이며 $\mathrm{diag}(I-\sqrt{q}\sqrt{q}^{\,t}, \ldots, I-\sqrt{q}\sqrt{q}^{\,t})$는 대각에 $I-\sqrt{q}\sqrt{q}^{\,t}$ 행렬이 d_p개 반복되는 행렬이다. 그런데 $\mathrm{diag}(I-\sqrt{q}\sqrt{q}^{\,t}, \ldots, I-\sqrt{q}\sqrt{q}^{\,t})$는 계수가 $d_p(\prod_{i=1}^{p-1} d_i - 1)$인 대칭 멱등행렬(idempotent matrix)이기 때문에 $X^2 = V^tV$의 극한분포는 $\chi^2(\prod_{i=1}^{p} d_i - d_p)$가 된다(예를 들어, Rao(1973)의 3b.4의 (vii) 참조). □

3. 모의실험

이 절에서는 다양한 모의실험을 수행한다. 우선 구형 대칭성 가정 하에서 검정통계량의 점근분포가 유한표본에서 얼마나 정확한지 알아보는 모의실험을 수행한다. 이 모의실험에서는 단순히 우리가 제안한 카이제곱 통계량뿐만 아니라 Beran(1979)과 Baringhaus(1991)도 포함시켜 비교하게 된다. 이 때 사용하는 구형대칭 분포족은 Ernst(1998)에서 제시한 다변량 일반화 라플라스 분포(multivariate generalized Laplace distribution)이다. 그 다음에는 구형 대칭에서 벗어나는 여러 형태의 대립분포에서 검정력을 비교하는 모의실험을 수행하게 된다. 여기에 포함되는 분포는 두 변수끼리의 상관계수가 거의 0에 가까우면서 주변 분포는 거의 대칭에 가까운 분포들이다.

우선 구형대칭인 분포족으로 사용할 다변량 일반화 라플라스 분포를 간략히 설명하도록 하자. 이 분포는 보다 일반적인 타원형대칭인 분포로서 Ernst(1998)가 난수생성 알고리즘을 제시하였는데, 여기서는 구형 대칭분포만 사용한다. 구체적으로 구형대칭인 다변량 일반화 라플라스 분포는 다음과 같은 확률밀도함수를 가지며 이 분포를 $MGL(\lambda)$라고 나타내도록 한다.

$$f(x;\lambda) = \frac{\Gamma(p/2)}{2\pi^{p/2}\Gamma(p/\lambda)} \exp\left\{-(x^tx)^{\lambda/2}\right\}$$

다변량 일반화 라플라스 분포는 λ가 작으면 꼬리가 긴 분포가 되며 λ가 커지게 되면 꼬리가 짧은 분포가 되는데 구체적으로 $\lambda = 2$이면 다변량 정규분포, $\lambda = 1$이면 라플라스 분포 형태이며 $\lambda \to \infty$이면 다차원 공(ball)에서의 균일분포에 수렴하게 되어 다양한 형태의 구형 대칭분포를 포함하게 된다. 다차원 공에서의 균일분포를 $MGL(\infty)$라고 나타내도록 한다.

다음으로 이 모의실험에서 사용될 Beran(1979)과 Baringhaus(1991)의 검정통계량을 간단하게 소개한다. Beran(1979)의 검정통계량은 다음과 같다.

$$S_n = \sum_{k=1}^{K_n} \sum_{m=1}^{M_n} \left[n^{-1/2} \sum_{i=1}^{n} a_k(A_i) b_k(B_i) \right]^2$$

여기서 $A_i = \mathrm{rank}(R_i)/(n+1)$(즉 $R_1, \ldots, R_n$ 중 R_i의 순위를 $n+1$로 나눈 것), $B_i = (\theta_{i1}, \ldots, \theta_{i(p-1)})$, a_k는 $[0,1]$에 정의된 르베그 측도(Lebesgue measure)와 상수 함수에 직

교인 직교정규함수이며, b_k는 $[0,\pi]^{p-2}\times[0,2\pi]$에 정의된 르베그 측도와 상수함수에 직교인 직교정규함수이다. 이 검정통계량은 구형 대칭성과 간단한 다른 정칙조건(regularity conditions) 하에서 $n\to\infty$일 때 $\lim K_n=\lim M_n=\infty$이면 평균이 K_nM_n이고 분산이 $2K_nM_n$인 정규분포로 수렴하게 된다. 그런데 이 검정은 직교정규함수 a_k, b_k를 선정하는 방법에 따라 검정통계량이 달라지고 또 이 직교정규함수들이 무한히 많아져야만 정규분포로 수렴한다는 점 때문에 실제 적용 상에 다소 문제가 있을 수 있다. 실제 모의실험에서 우리는 a_k에는 5차까지의 직교다항함수(orthogonal polynomial function)를 사용하며, b_k에는 이것 대신에 $Z_i=(Z_{i1},\ldots,Z_{ip})^t=X_i/R_i$에 정의되는 2^p-1개의 함수 b'_k를 사용한다. 이 함수의 구체적인 모양은 $k=1,\ldots,2^p-1$의 이진법 전개(binary expansion)를 $(l_1,\ldots,l_p)$라고 나타낼 때

$$b'_k(Z_i)=\exp\left(\sum_{j=1}^{p} l_j/2^{p+1}\log 3\right)\prod_{j=1}^{p} Z_{ij}^{l_j}$$

이 된다. b_k 대신에 b'_k을 사용하는 이유는 Z_i에 정의된 직교함수를 만들기 훨씬 간편하기 때문이다. 다음으로 Baringhaus(1991)의 검정통계량은 다음과 같다.

$$T_n=\frac{1}{n}\sum_{i,j=1}^{n} h(Z_iZ_j)\min\left(1-\frac{\mathrm{rank}(R_i)-1}{n},1-\frac{\mathrm{rank}(R_j)-1}{n}\right)$$

여기서 $h(t)$ 함수에는 여러 가지가 사용될 수 있지만 이 모의실험에서는 Baringhaus(1991)의 식 (3.3)에 소개된

$$h(t)=1+4(\pi p)^{-1}\left[\Gamma((p+1)/2)/\Gamma(p/2)\right]^2-2\pi^{-1}\left[\arccos t+(1-t^2)^{1/2}\right],\ -1\le t\le 1$$

함수를 사용한다.

이 검정통계량의 점근분포는 복잡하여 직접 사용하기 어렵기 때문에 모의실험을 통해 근사적 분포를 생성시켜 사용하도록 하겠다. 근사적 분포는 다변량 표준정규분포에서 표본을 100만 개 뽑아 각각 T_n을 계산하고 이것의 (100만 개) 순서 통계량에서 근사적으로 $i/10000, i=1,\ldots,9999$ 위치에 해당되는 값들을 뽑아 사용한다. 이 근사적 분위수 값들은 모의실험과 실제 자료분석 예제에서 Baringhaus(1991)의 근사적 유의확률을 계산할 때 사용되게 된다.

구형대칭인 다변량 일반화 라플라스 분포 $MGL(\lambda)$에서의 모의실험은 다음과 같이 실행되었다. 고려된 λ는 1, 2, 5와 ∞, 고려된 표본크기 n은 100과 200이며 고려된 차원 p는 2, 3과 4이다. 해당 $MGL(\lambda)$에서 차원이 p인 표본크기 n인 표본을 1000개 뽑아 X^2, S_n, T_n 통계량들과 이 검정통계량들의 점근 유의확률을 계산하였다. 이 때 X^2은 $p=2$일 때 $(d_1=6,d_2=4)$, $p=3$일 때 $(d_1=2,d_2=4,d_3=3)$, $p=4$일 때 $(d_1=d_2=d_4=2,d_3=4)$를 사용하여 각 칸의 기대도수가 $n=100$일 때 4.16, 4.16, 3.13으로 너무 작지 않게 만들었다. (여기서 $p=3,4$인 경우 $d_{p-1}=2d_1$를 한 이유는 θ_{p-1}가 취하는 범위 $[0,2\pi]$가 θ_1이 취하는 범위 $[0,\pi]$보다 두 배 넓기 때문이다.) 점근 유의확률을 계산할 때 X^2과 S_n는 이론적으로 유도된 점근분포를 이용하였고, T_n은 앞에서 설명하였던 모의실험을 통한 근사적 분포를 이용

하였다. 유의수준이 $\alpha = .01, .05, .1$일 때 1000개의 표본 중 귀무가설 기각율을 정리한 것이 표 3.1, 표 3.2와 표 3.3이다. 이 표에서 알 수 있는 것은 X^2와 T_n의 귀무가설 상대기각율

표 3.1: $p = 2$일 때 1000개의 표본 중 귀무가설 기각율

		$n = 100$			$n = 200$		
λ	stat	$\alpha = .01$	$\alpha = .05$	$\alpha = .1$	$\alpha = .01$	$\alpha = .05$	$\alpha = .1$
	X^2	.009	.050	.099	.014	.054	.099
1	S_n	.012	.034	.063	.017	.046	.083
	T_n	.006	.033	.076	.007	.038	.081
	X^2	.007	.041	.093	.007	.046	.086
2	S_n	.016	.052	.085	.012	.052	.093
	T_n	.011	.048	.101	.007	.048	.085
	X^2	.006	.044	.089	.015	.053	.106
5	S_n	.012	.038	.066	.021	.052	.094
	T_n	.011	.060	.107	.011	.044	.088
	X^2	.014	.058	.090	.011	.047	.097
∞	S_n	.016	.051	.088	.021	.053	.095
	T_n	.013	.052	.113	.009	.044	.093

표 3.2: $p = 3$일 때 1000개의 표본 중 귀무가설 기각율

		$n = 100$			$n = 200$		
λ	stat	$\alpha = .01$	$\alpha = .05$	$\alpha = .1$	$\alpha = .01$	$\alpha = .05$	$\alpha = .1$
	X^2	.010	.050	.010	.008	.054	.108
1	S_n	.006	.023	.052	.007	.036	.069
	T_n	.011	.057	.110	.009	.036	.077
	X^2	.008	.047	.096	.006	.049	.107
2	S_n	.006	.035	.058	.007	.041	.078
	T_n	.009	.030	.081	.010	.044	.090
	X^2	.009	.047	.090	.006	.050	.104
5	S_n	.010	.038	.071	.009	.040	.082
	T_n	.012	.045	.100	.012	.049	.098
	X^2	.009	.056	.105	.011	.049	.113
∞	S_n	.007	.027	.056	.020	.043	.078
	T_n	.011	.047	.098	.017	.042	.091

은 대체로 명목 기각율 α에 부합되지만, S_n은 $p = 3, 4$에서 $n = 100$일 때 다소 보수적인 검

표 3.3: $p = 4$일 때 1000개의 표본 중 귀무가설 기각율

		$n = 100$			$n = 200$		
λ	stat	$\alpha = .01$	$\alpha = .05$	$\alpha = .1$	$\alpha = .01$	$\alpha = .05$	$\alpha = .1$
1	X^2	.010	.039	.077	.013	.059	.110
	S_n	.003	.024	.048	.012	.043	.083
	T_n	.012	.049	.091	.007	.050	.105
2	X^2	.014	.060	.107	.010	.051	.095
	S_n	.009	.035	.064	.007	.037	.077
	T_n	.014	.051	.102	.007	.057	.109
5	X^2	.011	.046	.093	.014	.053	.098
	S_n	.003	.026	.054	.008	.046	.087
	T_n	.005	.059	.119	.008	.050	.104
∞	X^2	.010	.047	.093	.005	.039	.094
	S_n	.010	.024	.043	.008	.043	.081
	T_n	.011	.050	.091	.014	.061	.101

정이 되는 경향이 있다는 것이다. 그림의 수가 너무 많아 이 논문에서 포함시키지는 않았지만 분위수대조도를 그려보면 X^2, T_n의 경우 점근분포가 고려된 유한표본에서 대체로 잘 부합되며 S_n의 경우 $n = 100$일 때 다소 보수적인 경향을 나타냈다.

다음으로 여러 가지 비구형 대칭인 대립가설 분포에서 검정력을 비교하는 모의실험을 수행하였다. 처음으로 고려된 분포는 나선형(spiral) 분포이다. 모수가 b인 나선형 분포의 확률밀도함수는

$$f(x_1, x_2; b) \propto \{1 + \cos[2(\theta - b \cdot r)]\} \exp(-r^2/2)$$

인데 여기서 r과 θ는 (x_1, x_2)의 극좌표 값이다. 이 분포의 산포도는 대칭인 두 개의 나선형을 이루는데 b가 커짐에 따라 꼬리가 길어진다. 이 모의실험에서는 $b = 2$를 사용하였는데 이 때 4차 적률까지 이변량 표준정규분포와 아주 비슷하게 되기 때문이다. X^2에서 $(d_1 = 6, d_2 = 4)$를 사용하고 500개의 표본에서 계산된 세 가지 통계량의 검정력을 요약한 것이 표 3.4이다. 이 분포에 대한 검정력은 S_n, X^2인 경우 아주 뛰어나고 T_n은 검정력이 많

표 3.4: 나선형 분포인 경우 500개 표본 중 검정력

	$n = 100$			$n = 200$		
stat	$\alpha = .01$	$\alpha = .05$	$\alpha = .1$	$\alpha = .01$	$\alpha = .05$	$\alpha = .1$
X^2	.498	.770	.864	.978	.996	.998
S_n	.800	.914	.952	1.00	1.00	1.00
T_n	.018	.106	.200	.046	.212	.368

이 떨어지는 것을 알 수 있다.

두 번째로 고려되는 분포의 확률밀도함수는

$$f(x_1,x_2)=\begin{cases}1/\pi, & \text{만약 } r\in[0,1)\text{이고 } \theta\in[0,\pi/2)\cup[\pi,3\pi/2)\\ & \quad \text{혹은 } r\in[1,\sqrt{2})\text{이고 } \theta\in[\pi/2,\pi)\cup[3\pi/2,2\pi)\\ 0, & \text{그외}\end{cases}$$

로서 여기서 r과 θ는 앞에서와 마찬가지로 (x_1,x_2)의 극좌표 값이다. 이 분포는 θ의 분포가 $[0,2\pi)$ 상에서 거의 균일분포가 되지만 r에 따라 다른 모양을 가지도록 만든 분포이다. X^2에서 $(d_1=6,d_2=4)$를 사용하고 500개의 표본에서 계산된 세 가지 통계량의 검정력을 요약한 것이 표 3.5이다. 이 분포에 대한 검정력은 S_n,X^2,T_n 모두 아주 뛰어나지만 T_n에서

표 3.5: θ가 거의 균일분포인 경우 500개 표본 중 검정력

	$n=100$			$n=200$		
stat	$\alpha=.01$	$\alpha=.05$	$\alpha=.1$	$\alpha=.01$	$\alpha=.05$	$\alpha=.1$
X^2	1.00	1.00	1.00	1.00	1.00	1.00
S_n	1.00	1.00	1.00	1.00	1.00	1.00
T_n	.021	.620	.820	.858	.996	1.00

약간 떨어진다. 이 모의실험에서는 우리의 검정통계량의 경우 정확한 좌표축을 알고 있는 장점이 있기 때문에 이 장점을 없애기 위해 랜덤하게 축을 회전시킨 자료에 대해 검정통계량을 계산하였다.

세 번째 분포는 4차원에서의 불완비 블록 설계(incomplete block design)에서 생성되는 분포이다. 다시 말해 $\tau=(\tau_1,\tau_2,\tau_3,\tau_4)$가

$$\begin{pmatrix}1&1&1&1\\1&1&-1&-1\\1&-1&1&-1\\1&-1&-1&1\\-1&1&1&-1\\-1&1&-1&1\\-1&-1&1&1\\-1&-1&-1&-1\end{pmatrix}$$

에서 랜덤하게 뽑은 행이고 $Z=(Z_1,Z_2,Z_3,Z_4)$가 다변량 표준정규 확률벡터라고 했을 때 확률벡터 $(\tau_1|Z_1|,\ldots,\tau_4|Z_4|)$가 따르는 분포를 의미한다. 이 분포는 4개 중 3개의 원소만 고려하면 서로 독립인 다변량 정규분포를 따르게 되지만 4개 변수를 모두 고려하면 발생할 수 없는 영역이 생기게 된다. X^2에 $(d_1=d_2=d_4=2,d_3=4)$를 적용하고 500개 표본에서 계산된 세 통계량의 검정력을 요약한 것이 표 3.6이다 이 분포에 대해 S_n은 전혀 검정력

표 3.6: 불완비 블록 설계인 경우 500개 표본 중 검정력

	$n=100$			$n=200$		
stat	$\alpha=.01$	$\alpha=.05$	$\alpha=.1$	$\alpha=.01$	$\alpha=.05$	$\alpha=.1$
X^2	.040	.126	.194	.078	.176	.248
S_n	.004	.020	.048	.008	.026	.056
T_n	.012	.082	.136	.024	.110	.188

이 없고 X^2, T_n은 다소간의 검정력을 가지지만 X^2의 검정력이 다소 높다. 이 모의실험에서도 랜덤하게 회전된 자료에 대해 X^2이 계산되었는데 회전되지 않은 자료에 X^2을 적용하면 검정력이 1이 되는 것은 자명하다.

네 번째로 고려된 분포는 Johnson & Kotz(1972) 299쪽에 소개된 (대칭)분포이다. 이 분포의 확률밀도함수는

$$f(x)=\left\{1+C\cdot x_1x_2\cdots x_p\exp(-\sum_{i=1}^{p}x_i^2/2)\right\}(2\pi)^{-p/2}\exp\left(-\sum_{i=1}^{p}x_i^2/2\right)$$

로서 p가 짝수일 때 대칭인 분포이다. 여기서 C는 상수값으로 Johnson & Kotz에서는 1로 잡았지만 이 모의실험에서는 최대값인 $\exp(p/2)$을 잡아 비구형 대칭성을 확대시켰다. $p=4$일 때 X^2에 $(d_1=d_2=d_4=2, d_3=4)$를 적용하고 500개 표본에서 계산된 세 가지 통계량의 검정력을 요약한 것이 표 3.7이다 이 분포의 경우 S_n, T_n의 검정력은 유의수준에 비해 별로

표 3.7: Johnson & Kotz의 분포인 경우 500개 표본 중 검정력

	$n=100$			$n=200$		
stat	$\alpha=.01$	$\alpha=.05$	$\alpha=.1$	$\alpha=.01$	$\alpha=.05$	$\alpha=.1$
X^2	.038	.126	.188	.064	.204	.324
S_n	.010	.032	.050	.014	.042	.100
T_n	.010	.040	.100	.016	.062	.108

나을 것이 없지만 X^2만 다소간의 검정력을 가진다.

4. 실제 자료분석 예제

이 절에서는 앞에서 제안한 검정통계량을 실제 자료분석에 적용하는 예제를 제시하도록 하겠다. 이 예제에서 사용되는 자료는 단변량 시계열 분석에서 나오는 잔차이며 이 잔차의 적절성을 판단하는 기준으로 구형 대칭성 검정을 사용하고자 한다. 구형 대칭성 검정을 단변량 잔차의 적절성을 판단하는 하나의 기준으로 사용할 수 있는 근거는 다음과 같다. 만

약 적절한 시계열 모형이 설정되었다면 이 모형에서 나오는 잔차는 근사적으로 무상관이 되며 평균이 0이 된다. 따라서 이 잔차를 연속적으로 겹치지 않게 잘라서 다변량 관찰벡터를 만들었을 때 이것은 근사적으로 평균벡터가 0, 공분산 행렬이 단위행렬에 비례하게 된다. 구형 대칭성 검정을 따르는 확률벡터는 평균벡터가 0, 공분산 행렬이 단위행렬에 비례하기 때문에 구형 대칭성 검정으로 잔차의 적절성을 판단할 수 있는 근거가 마련되는 것이다.

이 예제에서 사용되는 자료는 간헐천 자료(geyser data)이다. 이 자료는 미국 와이오밍 주의 옐로스톤 국립공원의 Old Faithful Geyer의 분출에 관련된 자료로서 이 예제에서는 연속적인 분출사이의 대기시간을 관측한 299개의 시계열 자료를 사용한다(구체적인 설명은 Azzalani and Bowman(1990) 참조). 이 시계열 자료에 S-Plus의 AR 절차를 적용하였는데 이 절차는 아카이케 정보기준(Akaike information criterion)에 의해 최적의 자기상관 시계열 모형을 찾아준다. 자기상관 계수의 추정치를 구하는 방법으로 율-워커 방정식(Yule-Walker equation)을 적용하였을 때 AR(2) 모형이 최적으로 선택되었다. 이 잔차에 대해 자기상관함수와 편자기상관함수를 그려보면 시차 25까지 모두 신뢰구간 안에 들어 있어, 일견 AR(2) 모형이 적절한 모형처럼 보인다.

이 잔차에 대해 우리가 제안한 카이제곱 검정을 적용하도록 하겠다. 앞에서 설명된대로 연속되는 잔차를 겹치지 않게 3개씩 잘라서 다변량 관찰치를 구성하면 총 99개의 3차원 다변량 관찰벡터를 얻을 수 있다. 이 자료에 대해 $(d_1 = 2, d_2 = 4, d_3 = 3)$을 사용하여 카이제곱 통계량을 구하면 32.15가 되고 점근적 유의확률이 .0565가 된다. 따라서 이 잔차의 적절성이 약간 의심되는 결과가 제시된 것이다. 실제로 현시점 잔차 e_t와 전시점 잔차 e_{t-1} 사이의 산포도를 그려보면 e_{t-1}이 커지면 e_t의 분산이 커지는 현상이 나타나 잔차간의 무상관성이 깨지는 것을 알 수 있다. 이 자료에 대해 Beran(1979)과 Baringhaus(1991)의 절차를 적용하면 점근적 유의확률이 각각 .0372, .329가 된다. 따라서 Beran의 절차는 잔차의 적절성이 다소 의심되는 결과를 제시하지만, Baringhaus의 절차는 그렇지 못하다는 것을 알 수 있다.

5. 결론

이 논문에서는 구형 대칭성을 검색하는 카이제곱 검정 통계량을 제안하고 있다. 이 검정은 특징은 통계량을 쉽게 계산할 수 있으며 카이제곱 극한분포를 가지기 때문에 점근 유의확률을 쉽게 계산할 수 있어 현장에서 쉽게 사용할 수 있다는 점이다. 이 카이제곱 검정 통계량의 극한분포를 이론적으로 유도하였으며 이 극한분포가 유한표본 하에서 얼마나 정확한지 알아보는 모의실험을 수행하였다. 또한 여러 가지 비구형 대칭인 분포에서 Beran(1979)과 Baringhaus(1991)의 통계량과 검정력을 비교하는 모의실험을 실행하였으며, 시계열 잔차 자료에 이 카이제곱 검정을 적용하는 실제 자료분석 예제도 제시하였다.

그 모의실험과 실제 자료분석 예제에서 다음과 같은 사실을 관측할 수 있었다.

첫 번째로 우리의 검정통계량이 계산하기 쉽고 유한표본에서도 잘 부합되는 극한분포를 가지고 있다는 사실이다. 이에 비해 Beran(1979)의 통계량은 선택되는 직교정규함수에

의존하며 계산이 다소 복잡하며, 점근분포도 $n = 100$일 때 다소 보수적인 검정이 되는 경향이 있었다. Baringhaus(1991)의 통계량은 계산은 그렇게 어렵지 않지만 점근분포가 복잡하여 쉽게 이용할 수 없는 단점이 있다. 그래서 이 모의실험에서는 다변량 표준정규분포에서 생성된 근사적 분포를 이용하였는데 유한표본에서 상당히 정확하였지만 표본을 뽑을 때마다 달라질 수 있는 자의성 때문에 실제 적용에는 다소간의 무리가 있을 수 있다.

두 번째로 우리의 검정통계량이 고려된 여러 가지 비구형 대칭 분포에서 Beran(1979)과 Baringhaus(1991)에 비해 결코 떨어지지 않는 검정력을 가진다는 것을 알 수 있었다. 특히 불완비 불록 설계와 Johnson & Kotz(1972) 292쪽에 소개된 대칭분포에서는 우리의 검정통계량이 가장 좋은 검정력을 보였으며, 특히 후자의 분포에서는 우리의 통계량만 검정력을 보였다. 또한 실제 자료분석 예제에서는 우리의 검정통계량과 Beran의 절차만 다소간의 검정력을 보였으며 Baringhaus의 절차는 거의 검정력을 보여주지 못했다.

이 연구에서는 Beran(1979)와 Baringhaus(1991)만 모의실험에 포함되고 Romano(1989)와 Koltchinskii & Li(1998)는 포함되지 않았다. 이들이 제외된 가장 큰 현실적인 이유는 이들이 붓스트랩에 근거한 검정통계량을 사용하여 이들을 포함한 모의실험에 많은 계산시간이 요구되었기 때문이었다. 따라서 이 두 가지 검정을 포함하여 다양한 비구형 대칭분포에서 검정력을 비교하는 모의실험이 추후연구로서 요망된다고 할 수 있다.

참고문헌

Arcones, M. and Giné, E. (1991). Some bootstrap tests of symmetry for univariate continuous distributions, *Annals of Statistics*, **19**, 1496-1511.

Azzalini, A. and Bowman, A.W. (1990). A look at some data on the old faithful geyser, *Applied Statistics*, **39**, 357-365.

Baringhaus, L. (1991). Testing for spherical symmetry of a multivariate distribution, *Annals of Statistics*, **19**, 899-917.

Beran, R. (1979). Testing for ellipsoidal symmetry of a multivariate density, *Annals of Statistics*, **7**, 150-162.

Csörgö , S. and Heathecote, C.R. (1987). Testing for symmetry, *Biometrika*, **74**, 177-186.

Ernst, M.D. (1998). A multivariate generalized Laplace distribution, *Computational Statistics*, **13**, 227-232.

Heathecote, C.R., Rachev, S.T., and Cheng, B. (1995). Testing multivariate symmetry, *Journal of Multivariate Analysis*, **54**, 91-112.

Johnson, N.L. and Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*, John Wiley & Sons, Inc, New York.

Koltchinskii, V.I., and Li, L. (1998). Testing for spherical symmetry of a multivariate distribution, *Journal of Multivariate Analysis*, **65**, 228-244.

Manzotti, A., Perez, F.J., and Quiroz, A.J. (2002). A statistic for testing the null hypothesis of elliptical symmetry, *Journal of Multivariate Analysis*, **81**, 274-285.

Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory*, John Wiley & Sons, New York.

Rao, C.R. (1973). *Linear Statistical Inference and its Applications*, second edition, John Wiley & Sons, New York.

Romano, J.P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses, *Annals of Statistics*, **17**, 141-159.

Schott, J.R. (2002). Testing for elliptical symmetry in covariance-matrix-based analyses, *Statistics and Probability Letters*, **60**, 395-404.

Schuster, E.F. and Barker, R.C. (1987). Using the bootstrap in testing symmetry versus asymmetry, *Communications in Statistics - Simulation and Computation*, **16**, 69-84.

[2004년 5월 접수, 2004년 10월 채택]

A Test for Spherical Symmetry *

Cheolyong Park [1)]

ABSTRACT

In this article, we propose a chi-squared test of spherical symmetry. The advantage of this test is that the test statistic and its asymptotic p-value are easy to compute. The limiting distribution of the test statistic is derived under spherical symmetry and its accuracy, in finite samples, is studied via simulation. Also, a simulation study is conducted in which the power of our test is compared with those of other tests for spherical symmetry in various alternative distributions. Finally, an illustrative example of application to a real data is provided.

Keywords: Chi-squared test; Multivariate symmetry; Polar coordinates

* This work was supported by grant No. R05-2003-000-10926-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

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