응용통계연구 (2009)
22(6), 1355–1364

다변량 자료에서 위치모수에 대한 로버스트 검정

소선하[1] · 이동희[2] · 정병철[3]

[1]우리은행 리스크모델 적합성검증팀, [2]경기대학교 경영학과, [3]서울시립대학교 통계학과

(2009년 5월 접수, 2009년 10월 채택)

요 약

본 논문에서는 다변량 자료의 위치모수에 대한 로버스트 검정 방법으로 유사등변성과 고붕괴성을 만족하는 MVE와 MCD 추정량에 근거한 로버스트 검정방법을 제안하였다. 일반적으로 이들 추정방법은 낮은 효율성으로 인하여 통계적 추론보다는 잠재적 이상치의 발견과 같은 탐색적분석에서 사용된다. 우리는 검정력을 높이기 위하여 MVE와 MCD 추정량에 근거한 일단계 재가중절차를 사용했는데, 가중치 선정과 관련된 임계값을 조절함으로써 현실적으로 사용가능한 높은 효율성과 정확성을 갖춘 검정방법을 제시하였다. 모의실험 결과 본 연구에서 제안한 검정법은 모분포에 관계없이 모두 명목유의수준을 제대로 유지하고 검정력도 높게 나타났으며, 이상치를 포함하고 있는 사례를 이용하여 실제로 모평균에 대한 가설검정을 수행한 결과 기존 방법과는 달리 영향을 받지 않았다.

주요용어: 고붕괴점추정, 이상치, 재가중방법, 최소공분산행렬값, 최소부피타원체, 공간중위수.

1. 서론

다변량 자료에서 위치모수(location parameter)에 대한 통계적 추론과정은 대부분 다변량 정규모집단으로부터 추출된 랜덤표본임을 가정한다. 이때 위치모수(즉 모평균벡터)에 대한 추정량으로서 표본평균벡터는 최소분산불편성을 가지며, 동시에 다변량 정규모집단 가정하에서 최우추정량과 동일하다는 바람직한 통계적 성질을 가진다. 그러나 이들 랜덤표본이 다변량 정규분포에 비하여 꼬리가 긴(heavy-tailed) 분포처럼 정규모집단으로부터 추출된 경우가 아니거나 다변량 정규모집단으로부터 추출된 랜덤표본이더라도 이상치(outlier)가 포함되었다면 표본평균벡터는 효율성에 심각한 문제가 발생하게 되어 이를 이용한 위치모수의 추정에 있어서도 심대한 영향을 끼치게 된다. 더불어 표본공분산 역시 이상치에 매우 민감하기 때문에 위치모수에 대한 통계적 추론 과정 역시 영향을 받게 된다 (Huber, 1981). 그러므로 이상치가 존재하는 다변량 자료에 있어서 위치모수는 물론 공분산과 같은 산포모수(dispersion parameter)에 대한 로버스트 추정량의 사용이 필수적이다.

본 연구에서는 다변량 자료에서 로버스트 추정량을 이용한 위치모수의 로버스트 검정절차를 제안하고자 한다. 이 경우 여러 가지 로버스트 추정량이 제안되어 있다. 하지만 우리가 주목하는 로버스트 추정량은 유사등변성(affine equivariance property)과 고붕괴성(high-breakdown property)을 만족하는 것으로 한정하고자 한다. 유사등변성이란 $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$를 각각 랜덤표본 $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)'$을 이용한 위치모수 벡터와 산포모수행렬에 대한 추정량이라 할 때 다음과 같은 조건을 만족함을 일컫는다.

$$\hat{\boldsymbol{\mu}}(\mathbf{A}X + \mathbf{b}) = \mathbf{A}\hat{\boldsymbol{\mu}}(X) + \mathbf{b},\ \hat{\boldsymbol{\Sigma}}(\mathbf{A}X + \mathbf{a}) = \mathbf{A}\hat{\boldsymbol{\Sigma}}(X)\mathbf{A}',$$

이 논문은 2007년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음(KRF-2007-314-C00039).

[3]교신저자: (130-743) 서울시 동대문구 전농동, 서울시립대학교 통계학과, 조교수. E-mail: bcjung@uos.ac.kr

즉 어떠한 선형변환에도 관계없이 추정결과가 유지될 수 있다는 것으로 추정량이 갖추어야 할 중요한 성질 중 하나이다. 또한 고붕괴성이란 자료 가운데 거의 절반 가까이 오염되어 있어도 추정량이 붕괴되지 않는다는 것을 의미한다. 이러한 추정량의 점근적 붕괴점(asymptotic breakdown point)은 0.5이다. 반면 표본평균의 경우 n개의 표본 가운데 단 하나의 관찰치가 매우 커지거나 매우 작아지게 되면 추정량 역시 이에 따라 변화하여 결국 단 하나의 값이 무한대로 접근하면 붕괴하게 된다. 따라서 표본평균의 붕괴점은 $1/n$이며, 이때 점근적 붕괴점, 즉 $\lim_{n\to\infty} 1/n$은 0이다.

이와 같은 유사등변성과 고붕괴성은 로버스트 추정량이 가져야할 중요한 성질이다. 그러나 대부분 이들 성질을 동시에 만족하는 추정량은 매우 낮은 효율성(efficiency)를 갖기 때문에 추정이나 검정 등 통계적 추론에서 정확한 결과를 기대하기 어렵다 (Davies, 1992). 따라서 이제까지 이들 로버스트 추정량은 통계적 추론을 위한 확증적 분석(confirmatory analysis)보다는 자료 가운데 의심되는 이상치를 사전에 살펴보기 위한 탐색적 분석(exploratory analysis)을 위한 도구로써 사용되었다 (Rousseeuw와 van Zomeren, 1990; Fung, 1993). 본 연구에서는 이렇듯 낮은 효율성으로 인하여 통계적 검정절차에 사용되지 않았던 유사등변성을 갖는 고붕괴점추정량을 위치모수에 대한 가설검정에 적용하기 위한 방법을 제안하고자 한다.

2절에서는 유사등변성을 갖는 고붕괴점추정량에 대한 소개와 함께 간단한 모의실험을 통해 기존의 일단계 재가중방법(one-step reweighting)에 기반한 로버스트 검정절차의 문제점을 살펴보고 이를 개선하기 위한 방법을 제안하고자 한다. 3절에서는 모의실험을 통하여 본 논문에서 제안한 로버스트 검정방법과 다변량 자료에서 대표적인 모평균벡터에 대한 검정방법인 Hotelling의 T^2검정방법의 효율성을 파악하고자 한다. 4절에서는 실제 사례를 가지고 우리가 제안한 검정방법의 결과를 살펴보고, 마지막 5절에서는 연구의 결론 및 추후 연구과제에 대해 언급하고자 한다.

2. 다변량 자료의 위치모수에 대한 로버스트 검정방법

먼저 $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$이 평균이 $\boldsymbol{\mu}_p$이고 공분산 행렬이 $\boldsymbol{\Sigma}_p$를 갖는 모집단에서 얻어진 크기가 p인 n개의 표본벡터라 했을 때, 단일 모집단에서 다음과 같은 모집단 평균벡터에 대한 검정을 고려해보자.

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \qquad vs. \qquad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0. \tag{2.1}$$

만일 $\mathbf{x}_i$가 p-변량 정규분포로부터 추출된 랜덤벡터임을 가정한다면, 단일 모집단에서 평균벡터에 대한 가설 (2.1)에 대한 검정은 표본평균벡터 $\bar{\mathbf{x}}$와 귀무가설하에서의 모평균벡터 사이의 "Mahalanobis 거리"에 근거를 둔 다음과 같은 Hotelling의 T^2 통계량을 이용하게 된다 (김기영과 전명식, 2002).

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0), \tag{2.2}$$

여기서 $\mathbf{S}$는 표본공분산행렬을 나타내며, 귀무가설이 맞다는 가정하에서 $(n-1)p/(n-p)F(p, n-p)$분포를 따르게 된다. 그러나 앞에서 언급했듯이 표본평균벡터와 표본공분산행렬은 이상치에 매우 민감하므로 기저분포가 다변량정규분포에서 많이 벗어나거나 이상치를 포함하게 되면 정확한 검정결과를 제공하지 못하게 된다 (Somorčík, 2006). 그러므로 다변량 자료에 대한 로버스트 추론과 관련된 연구는 이와 같은 상황에서 영향을 덜 받는 모평균벡터와 모분산행렬에 대한 로버스트 추정량을 얻기 위한 방법에 집중해 왔다. 이제까지 제안된 로버스트 추정량 가운데 대표적인 방법인 Rousseeuw (1985)가 제안한 최소부피타원체(Minimum Volume Ellipsoid; MVE)와 Butler 등 (1993)의 최소공분산행렬값(Minimum Covariance Determinant; MCD)을 중심으로 논의를 진행하고자 한다.

관찰벡터 $\mathbf{x}$에 대해 $\mathbf{d}(\mathbf{x}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$를 $d(\mathbf{x}_i, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})\,(i = 1, \ldots, N)$를 원소로 갖는 벡터라 하고, $\hat{\sigma}$를 로버스트 척도 추정량(robust scale estimate)이라 하자. 여기서 $d(\mathbf{x}_i, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$는 $(\mathbf{x}_i - \hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})$인 다변량

거리를 나타낸다. 이때 다변량 위치모수와 산포모수에 대한 추정량 $\hat{\boldsymbol{\mu}}$과 $\hat{\boldsymbol{\Sigma}}$을 다음을 최소화하도록 정의하자.

$$\min \hat{\sigma}(\mathbf{d}(\mathbf{x}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})) \tag{2.3}$$

이와 같은 기준에서 $\hat{\sigma}$을 표본중위수를 사용하는 경우가 가장 단순한 경우인데, 이때 얻어진 추정량이 MVE 추정량이다 (Rousseeuw, 1985). 이는 최소한 관찰자료의 절반을 포함하는 여러 타원체(ellipsoid)들 가운데 최소부피를 갖는 추정량으로, 절반 가까이 자료가 오염되더라도 추정량이 붕괴되지 않는 고붕괴점을 갖는다. 반면 동일한 수준의 고붕괴점을 갖는 또 다른 로버스트 추정량이 MCD 추정량이다. 이는 (2.3) 기준 하에서 $d_i = d(\mathbf{x}_i, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$를 크기순으로 나타낸 $d_{(1)} \leq \cdots \leq d_{(n)}$에 대하여 MVE가 사용하는 표본중위수 대신 절사제곱합

$$\hat{\sigma} = \sum_{i=1}^{h} d_{(i)} \tag{2.4}$$

를 사용하는데 이 기준에 의한 추정량 $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$을 MCD추정량이라 하며 (Butler 등, 1993), 특히 $h = (n-p)/2$인 경우에 MVE와 같은 수준의 고붕괴점을 갖게된다.

이들 로버스트 추정량 MVE와 MCD는 관찰자료의 50% 가까이가 오염돼 있더라도 영향을 받지 않는 특징을 가지고 있지만, 추정량의 낮은 수렴속도때문에 매우 낮은 효율성을 갖는다고 알려져 있다 (Davies, 1992). 따라서 일반적으로 사용하는 정규근사에 의한 통계적 추론을 사용하기 어려우며, 그의 점근적 성질(asymptotic property) 역시 아직 알려져 있지 않다 (Hawkins와 Olive, 2002). 그러므로 이와 같이 고붕괴점을 갖는 추정량을 기반으로 하여 확률적 분포에 기반한 통계적 추론을 하는 것은 현실적으로 쉽지 않은 문제이다. 사실 Rousseuw와 van Zomeren (1990)과 Fung (1993)의 언급처럼 로버스트 추정결과는 자료에 포함되어 있는 이상치를 판별하기 위한 탐색적 자료분석 도구로는 훌륭한 결과를 제공하지만 통계적 추론을 위한 확증적 도구로써는 미흡한 것이 사실이다. 그러나 이들 로버스트 추정량에 기반한 통계적 검정절차가 불가능한 것은 아니다. 최근 Maronna 등 (2006)은 고붕괴점 추정결과를 바탕으로 하는 일단계 재가중과정(one-step reweighting)에 의해 MVE와 MCD 추정량의 편이(bias)는 줄이고 효율성은 높일 수 있다는 결과를 제시한 바 있다. 일단계 재가중방법이란 고붕괴점 추정결과를 이용하여 일정한 기준을 벗어나는 관찰치를 이상치로 고려하여 절사한 후 남은 관찰치만을 대상으로 표본평균벡터와 표본공분산행렬을 재추정하여 이를 최종 추정량으로 제공하는 것이다. 우리는 이와 같은 MVE와 MCD를 이용한 일단계 재가중방법을 다변량 위치모수벡터에 대한 가설검정방법으로 확장해 보았는데, 그 절차는 다음과 같다.

단계 1 MVE 또는 MCD를 이용한 추정량 $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$를 계산한다.

단계 2 각 i번째 관찰벡터에 대해 다음과 같이 정의된 로버스트 거리 RD_i를 산출한다.

$$\mathrm{RD}_i = \sqrt{(\mathbf{x}_i - \hat{\boldsymbol{\mu}})' \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})}.$$

단계 3 앞에서 계산된 RD_i에 기반하여 i번째 관찰벡터에 대해 다음과 같이 가중치 w_i를 계산한다.

$$w_i = \begin{cases} 1, & \mathrm{RD}_i < \sqrt{\chi^2_{p,\beta}}, \\ 0, & \text{otherwise}, \end{cases} \tag{2.5}$$

여기서 $\chi^2_{p,\beta}$는 자유도가 p인 카이제곱분포의 β-분위수이다. 이와 같은 w_i를 이용한 가중벡터 $w_1\mathbf{x}_1, \ldots, w_n\mathbf{x}_n$을 이용한 가중표본평균벡터 $\bar{\mathbf{x}}_w$와 가중표본공분산행렬 $\mathbf{S}_w$를 계산한다.

표 **2.1.** 3-변량 모평균벡터에 대한 명목유의수준 0.05에서의 검정방법의 유의수준 추정결과

	정규분포				코시분포			
n	50	100	200	500	50	100	200	500
T^2	.039	.052	.054	.052	.015	.022	.019	.014
MVE기반의 T_w^2	.099	.081	.102	.117	.104	.118	.105	.140
MCD기반의 T_w^2	.254	.140	.126	.122	.134	.127	.106	.144

단계 4 이제 앞에서 계산된 가중표본평균벡터와 가중표본공분산행렬을 이용한 Hotelling의 T^2 통계량은 다음과 같이 정의된다.

$$T_w^2 = n_w(\bar{\mathbf{x}}_w - \boldsymbol{\mu}_0)'\mathbf{S}_w^{-1}(\bar{\mathbf{x}}_w - \boldsymbol{\mu}_0), \tag{2.6}$$

여기서 $n_w = \sum_{i=1}^n w_i$이다. 결국 가중표본평균벡터와 가중표본공분산행렬은 이들 추정량에 근거하여 이상치가 제거된 "좋은"관찰치만을 이용하여 검정하는 방법이다. 귀무가설하에서 T_w^2는 $(n_w-1)p/(n_w-p)F(p, n_w-p)$분포를 따르는 것으로 고려하여 가설 (2.1)에 대한 검정을 실시한다.

이러한 검정절차의 단계3과 같은 재가중방법은 Maronna 등 (2006)이 제안한 것으로 β의 값으로 이들의 방법에 따라 0.95와 0.975를 사용하였다. 이는 RD_i^2가 다변량정규분포하에서 근사적으로 자유도가 p인 카이제곱분포를 따른다는 사실을 이용한 것으로 모든 자료벡터가 "좋은"값으로만 구성되었을때도 근사적으로 $1-\beta$비율 정도의 자료는 제거하게 된다는 점에 착안한 것이다. Maronna 등 (2006)에 의하면 MCD나 MVE 모두 이와 같은 일단계 재가중방법을 통해 평균제곱오차(mean squared error) 측면에서 추정량의 효율성을 높일 수 있다고 주장하였다. 하지만 검정문제에서는 여전히 고붕괴점 추정량이 갖는 낮은 효율성으로 인하여 문제가 발생할 수 있는데, 이를 구체적으로 살펴보기 위하여 다음과 같이 간단한 모의실험을 실시하였다. 먼저 표본크기를 50, 100, 200 및 500인 경우를 고려하여 귀무가설이 맞다는 가정하에서 평균이 0인 3-변량 정규분포와 코시분포에서 표본벡터들을 생성하였다. 총 1000번의 반복을 통하여 명목유의수준(nominal significance level) 0.05에서 각 검정의 추정된 유의수준을 계산하여 그 결과를 나타낸 것이 표 2.1이다.

명목유의수준이 0.05인 경우 이항분포에서 정규분포로의 근사를 이용하면 1000번의 반복을 통해 추정된 유의수준이 0.036보다 작거나 0.064보다 크게 나타날 가능성은 5% 미만이다. 표 2.1을 살펴보면 정규모집단에서는 예측했던대로 Hotelling의 T^2는 명목유의수준을 제대로 유지한 반면 코시분포에서는 명목유의수준을 과소추정하는 경향을 보여주었다. 또한 로버스트 추정량인 MVE와 MCD를 이용하여 일단계 재가중과정 후 Hotelling의 T^2를 적용한 결과는 두 분포 모두 표본크기에 관계없이 명목유의수준을 과대추정하는 결과를 보여주고 있다. 특히 과대추정하는 수준이 분포에 상관없이 일정하게 나타남을 확인할 수 있다. 이때 단계3의 β값을 0.95와 0.975로 나누어 실험한 결과 그 차이가 거의 나타나지 않아 모든 과정에서 β값을 0.975로 설정하였다.

이와 같은 모의실험결과 로버스트 추정량에 기반한 검정절차가 정규모집단을 벗어난 상황에서 전통적인 방법에 비하여 영향을 덜 받는 것은 사실이지만, 실제 검정에 활용하기 위해서는 효율성이 많이 떨어진다는 사실을 알 수 있다. 이를 해결하는 가장 좋은 방법은 이들 로버스트 추정량의 점근적 분포에 근거하여 검정절차를 수행하는 것이지만, 이와 관련된 규명은 아직까지 이루어지지 않은 상황이다 (Hawkins와 Olive, 2002).

본 논문에서는 다양한 모의실험을 통해 MCD와 MVE에 근거한 로버스트 검정법의 문제점과 그 해결책을 살펴보았다. 우선 고붕괴점을 갖는 로버스트 추정량들이 갖는 낮은 효율성이다. He와 Portnoy

(1992)가 밝혔듯이 회귀모형에서 수렴률(convergence rate)이 낮은 고붕괴점추정량에 근거한 잔차를 이용한 재가중최소제곱추정량(reweighted least squares estimates)은 초기추정량의 수렴률을 따른다. 이와 같은 모습이 다변량 자료의 위치모수에 대한 검정절차에서도 동일하게 나타나는 것을 모의실험을 통해 확인할 수 있다. 실제로 MCD의 경우 식 (2.4)에서 h를 전체 표본크기 n에 근접시켜감에 따라 추정량의 수렴률을 높일 수 있으나, 이에 따라 붕괴점은 감소하게 된다. 따라서 자료에 포함된 이상치의 비율 등을 정확히 모르는 상태에서 이들 로버스트 추정량이 갖는 고붕괴성을 포기하면 로버스트 추정량의 사용목적 자체가 무의미하게 된다.

이와 같은 경우 고붕괴성을 유지하면서 고려해볼 수 있는 방법은 재가중과정, 즉 식 (2.5)와 같이 주어진 기준을 변경하는 것이다. 여러 가지 모의실험 결과 MCD와 MVE를 이용한 재가중 과정에서 식 (2.5) 대신 다음과 같이 정의한 가중치 w_i^*를 사용할 경우 검정의 효율성을 높일 수 있음을 확인할 수 있었다.

$$w_i^* = \begin{cases} 1, & \mathrm{RD}_i < \chi^2_{p,\beta}, \\ 0, & \text{otherwise.} \end{cases} \tag{2.7}$$

식 (2.7)와 같이 가중치를 조절하는 것은 결국 식 (2.5)에서 1을 갖는 가중치를 보다 많이 확보하도록 기준을 설정한 재가중방법이다. 모의실험결과 식 (2.5)에서 정의된 가중치 w_i를 사용할 경우 표본의 크기에 따라 약간의 차이는 있지만 심지어 정규모집단에서도 약 20% 내외의 표본이 0의 가중치가 부여되어 많은 수의 관측치가 잠재적 이상치로 판별되는 현상을 볼 수 있었다. 이와 같은 현상은 벡터자료의 특성으로 나타나는 것으로 보인다. 즉 표본벡터 가운데 하나의 관찰치가 이상치로 포함되는 경우에도 벡터를 구성하는 모든 관찰치에 대한 가중치가 모두 0으로 부여되기 때문에 원래 설정된 기준인 $1-\beta$ 비율 이상의 자료가 검정과정에서 제거되는 것이다. 가중치를 조절하기 위하여 고려할 수 있는 또다른 방법은 식 (2.5)에서 1보다 큰 상수 c를 이용한 $c\sqrt{\chi^2_{p,\beta}}$를 사용하는 것이다. 모의실험 결과 c가 2보다 크게 되면 다변량 정규분포를 따르는 경우에는 명목유의수준을 잘 유지하는 것으로 나타났으나, 코시분포에서는 표본의 크기와 차원의 수에 따라 유의수준이 다르게 나타나는 것을 확인할 수 있었다. 즉 차원이 작은 경우에는 명목유의수준보다 크게 나타나고 반면 차원의 수가 큰 경우에는 명목유의수준보다 작게 나타나는 것을 보게 되는데, 어느 특정 차원이나 표본크기에 대한 명목유의수준을 유지하게 되면 다른 차원과 표본크기에서는 이를 유지할 수 없게 되는 현상이 나타나게 된다. 본 논문에서는 식 (2.7)에서 정의한 w_i^*을 이용한 재가중방법의 효율성을 모의실험을 통하여 살펴보고자 한다.

3. 모의실험

모의실험에서는 전통적인 다변량 자료에서의 위치모수 검정방법인 Hotelling의 T^2-검정(T^2), (2.7)에서 정의한 가중치 w_i^*를 이용한 MVE 기반의 검정(RMVE) 및 MCD에 근거한 가중치 w_i^*를 이용한 검정(RMCD)을 비교하였다.

모의실험은 정규분포와 코시분포 등 2가지 분포를 고려하여 실시하였다. 정규분포의 경우 유의수준의 비교를 위하여 $N_p(\mathbf{0}, \mathbf{I})$로부터 $\mathbf{x}_i$를 생성시켰으며, 검정력을 비교하기 위하여 다른 변수의 평균은 모두 0으로 고정한 채 첫 번째 변수 $\mathbf{x_1}$만 평균을 0에서부터 1.0까지 0.2씩 증가시켜가며 실험하였다. 코시분포의 경우에도 동일한 방법으로 표본을 생성하였다. 본 연구에서는 모의실험에서 상관을 고려하지 않았는데, 그 이유는 본 연구에서 제안한 검정방법의 기초를 이루는 추정량들이 모두 유사등변성을 만족하기 때문이다.

모든 실험에서 50, 100, 200, 500으로 증가시켜가며 실험하였으며 1,000번의 반복을 통하여 각 검정법의 추정된 유의수준과 검정력을 계산하였다. 다음의 그림 3.1과 3.2는 각각 3-변량과 5-변량 자료에 대

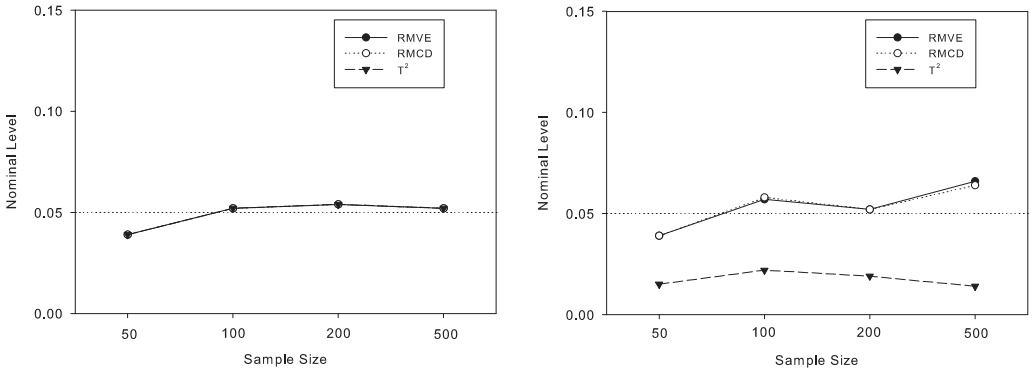

그림 3.1. 3-변량 다변량 자료의 모평균벡터에 대한 명목유의수준 0.05에서의 표본크기에 따른 유의수준 추정결과: (A) 다변량정규분포, (B) 코시분포

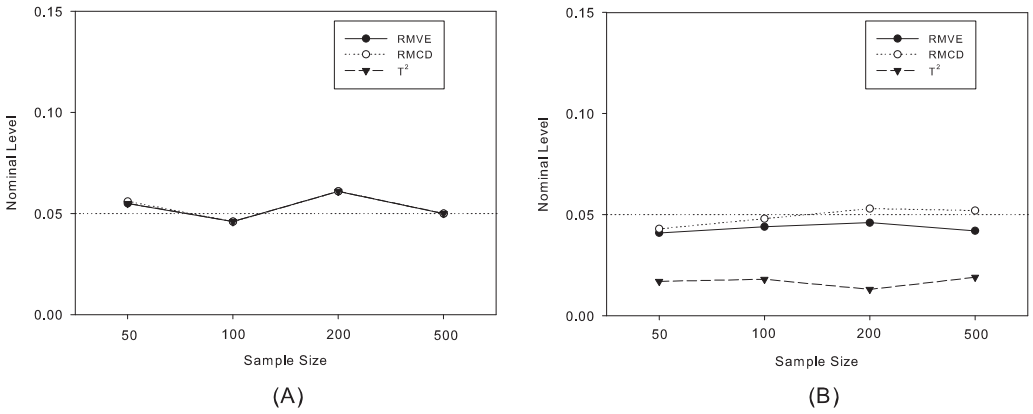

그림 3.2. 5-변량 다변량 자료의 모평균벡터에 대한 명목유의수준 0.05에서의 표본크기에 따른 유의수준 추정결과: (A) 다변량정규분포, (B) 코시분포

해 귀무가설하에서 모평균벡터에 대한 가설검정을 실시한 결과를 나타낸 것이다.

그림 3.1과 3.2를 살펴보면 새로운 재가중방법의 기준을 사용한 RMVE와 RMCD의 경우 다변량 정규모집단뿐만 아니라 꼬리가 두터운 코시분포에 대해서도 표본크기에 상관없이 명목유의수준을 제대로 유지하는 것으로 나타났다. 특히 다변량 정규모집단의 경우에는 세 방법 모두 거의 동일한 유의수준을 가지고 있어 그림에서는 겹쳐 보이고 있다. 반면 Hotelling의 T^2는 정규모집단에서는 명목유의수준을 제대로 유지하지만 코시분포에서는 명목유의수준을 과소추정하고 있음을 알 수 있다. 단지 p가 증가할 경우 RMCD가 RMVE에 비해 명목유의수준이 높게 나타남을 볼 수 있는데 그 차이는 매우 미미하였다.

그림 3.3은 코시분포를 따르는 5-변량 다변량 자료의 모평균벡터에 대한 검정의 검정력을 나타낸다. 정규모집단에서는 Hotelling의 T^2, RMVE 및 RMCD 등 3가지 검정의 검정결과가 거의 유사하게 나타나 그 결과를 제시하지 않았으며, $p = 3$인 경우는 $p = 5$인 경우와 유사한 결과가 나타나기 때문에 제시

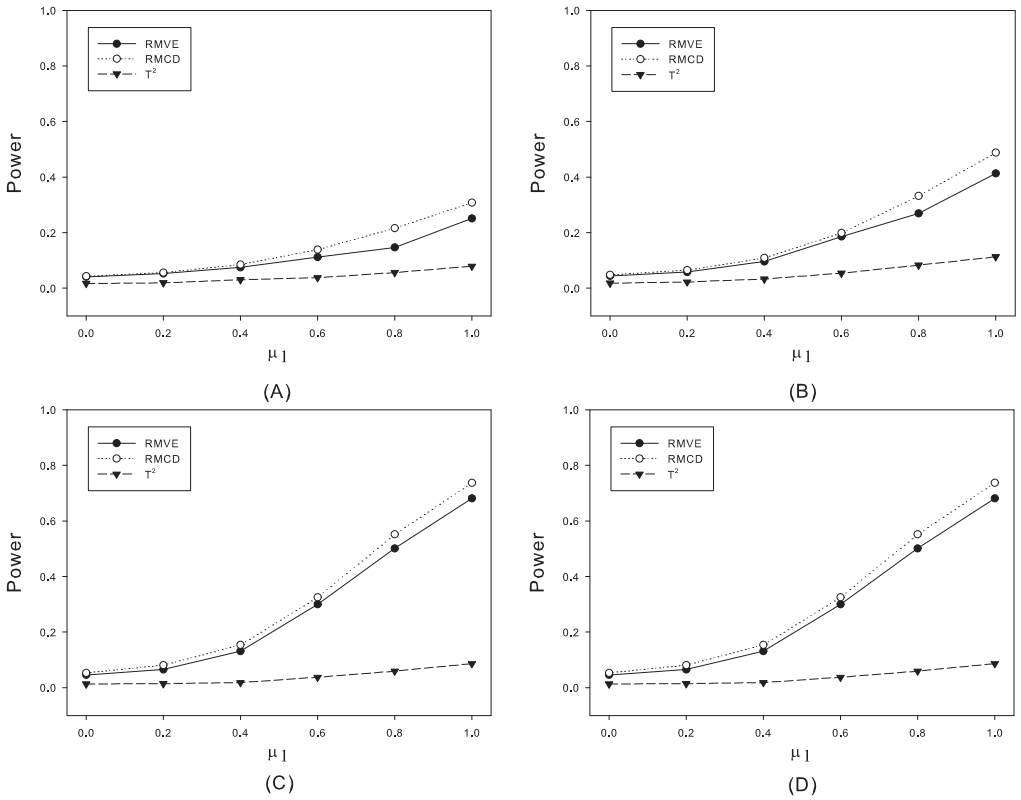

그림 3.3. 코시분포를 따르는 5-변량 다변량 자료의 모평균벡터에 대한 검정력: (A) $n = 50$, (B) $n = 100$, (C) $n = 200$, (D) $n = 500$

하지 않았다. 단 $p = 3$인 경우는 $p = 5$인 경우와 달리 RMVE와 RMCD의 결과 간에 거의 차이를 보이지 않았다. 그림 3.3을 보면 Hotelling의 T^2 통계량은 귀무가설이 틀린 경우 검정력이 낮게 나타난 반면 RMVE와 RMCD는 μ_1의 값이 0에서 멀어지고 표본크기 n이 커질수록 높은 검정력을 보이고 있다.

4. 사례연구

본 절에서는 사례를 통해 본 논문에서 제안한 다변량 자료의 위치모수에 대한 검정결과를 비교해 보고자 한다. 이 자료는 Hawkins 등 (1984)에서 사용된 것 가운데 일부를 발췌한 것이다. 원자료는 회귀분석에 사용되었으나 본 절에서는 독립변수로 사용된 부분만을 이용하여 모평균벡터에 대한 가설검정을 실시하였다.

분석에 사용된 자료는 3-변량 75개의 관찰벡터로 구성되었으며, 이 가운데 최초 14개의 관찰벡터는 이상치로 알려져 있다. 이 절에서는 이들 14개를 제외한 나머지 61개 자료값에 대해 변수별 평균을 구한 후 이를 빼주어 중심화 함으로써 모평균벡터의 값이 모두 0이라는 가설에 대한 검정을 실시하였다.

표 4.1은 원자료에 포함된 각 변수별로 전체, 이상치로 알려진 14개의 관측치(1–14) 및 이들 이상치를

표 **4.1.** Hawkins-Bradu-Kass 자료에서 사용된 변수의 평균과 표준편차

	x_1		x_2		x_3	
	평균	표준편차	평균	표준편차	평균	표준편차
전체	3.21	3.653	5.60	8.239	7.23	11.740
1–14	10.48	0.838	22.23	3.854	31.39	2.665
1–14 제외	1.54	1.064	1.78	1.073	1.69	1.034

표 **4.2.** Hawkins-Bradu-Kass 자료의 모평균벡터에 대한 검정결과

	검정통계량	p-값
T^2	16.732	<.0001
RMVE	2.137	.1054
RMCD	2.137	.1054

제외한 나머지 자료(1–14 제외)에 대한 평균과 표준편차를 나타낸 것이다. 14개의 이상치에 의해 전체 자료의 평균과 표준편차가 크게 나타나 이들에 의해 자료가 크게 왜곡돼 있음을 확인할 수 있다.

이와 같이 수정된 자료에 대해 모평균벡터가 모두 0이라는 가설에 대해 검정한 결과 Hotelling의 T^2, RMVE 및 RMCD 등 검정통계량 값과 유의확률은 표 4.2와 같이 계산되었다. 표 4.2를 살펴보면 유의수준을 0.05라 했을 때 Hotelling의 T^2는 귀무가설을 기각하지만, RMVE와 RMCD는 귀무가설을 기각하지 못하는 것으로 나타나 본 연구에서 제안한 로버스트 검정방법이 이상치에 의해 영향을 받지 않고 정확한 검정결과를 제시하고 있음을 확인할 수 있다.

5. 결론

본 연구에서는 다변량 자료의 위치모수에 대한 로버스트 검정방법으로 유사등변성과 고붕괴성을 만족하는 MVE와 MCD 추정량을 이용하는 검정방법을 제안하였다. 모의실험 결과 본 연구에서 제안한 검정법은 모분포에 관계없이 모두 명목유의수준을 제대로 유지하고 검정력도 높게 나타났으며, 사례 자료에 대한 검정결과 전통적인 검정방법이 이상치에 의해 왜곡된 검정결과를 제시하는 반면에 제안된 방법은 정확한 검정결과를 제시하고 있음을 확인할 수 있었다. 그러나 MVE와 MCD 추정량은 유사등변성과 고붕괴성을 만족하지만 앞서 언급했듯이 낮은 수렴률때문에 본 논문에서 사용한 재가중절차에 기반한 검정통계량의 점근적 분포를 제시하지는 못했는데, 이는 이후 연구과제로 남겨 두고자 한다.

사실 본 연구에서 다룬 위치모수에 대한 로버스트 검정방법들은 여러 연구자들에 의하여 연구되었다. Utts와 Hettmansperger (1980)은 순위통계량을 이용한 검정방법을 제안하였고, Brown (1983)과 Chakraborty 등 (1998)은 공간중위수를 이용한 각도(angle) 검정을 시도하였으며 Somorčík (2006)은 공간중위수를 이용한 여러 집단간 평균벡터의 동일성 검정 방법을 제안하였다. 본 연구에서는 유사등변성과 고붕괴성을 만족하는 추정량에 한정하여 로버스트 검정방법을 제안하였기 때문에 기존 방법들과의 비교문제는 다루지 못하였다. 이는 추후 연구과제로 다루고자 한다.

참고문헌

김기영, 전명식 (2002). <다변량 통계자료분석>, 자유아카데미.

Brown, B. M. (1983). Statistical uses of the spatial median, *Journal of the Royal Statistical Society B*, **45**, 25–30.

Butler, R. W., Davies, P. L. and Jhun, M. (1993). Asymptotic for minimum covariance determinant estimator, *The Annals of Statistics*, **21**, 1385–1400.

Chakraborty, B., Chaudhuri, P. and Oja, H. (1998). Operating transformation retransformation on spatial median and angle test, *Statistica Sinica*, **8**, 767–784.

Davies, P. P. (1992). The asymptotics of Rousseeuw's minimum volume ellipsoid estimator, *The Annals of Statistics*, **20**, 1828–1843.

Fung, W. K. (1993). Unmasking outliers and leverage points: A confirmation, *Journal of the American Statistical Association*, **88**, 515–519.

Hawkins, D. M., Bradu, D. and Kass, G. V. (1984). Location of several outliers in multiple regression data using elemental sets, *Technometrics*, **26**, 197–208.

Hawkins, D. M. and Olive, D. J. (2002). Inconsistency of resampling algorithm for high-breakdown estimators and a new algorithm, *Journal of the American Statistical Association*, **97**, 136–148.

He, X. and Portnoy, S. (1992). Reweighted LS estimators converge at the same rate as the initial estimator, *The Annals of Statistics*, **20**, 2161–2167.

Huber, P. J. (1981). *Robust Statistics*, John Wiley & Sons, New York.

Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*, John Wiley & Sons, New York.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point, *Mathematical Statistics and its applications (vol. B)*, W. Grossmann, G. Pflug, I. Vinczc and W. Wertz (eds.), 283–297, Reidel, Dordrecht.

Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association*, **85**, 633–639.

Somorčík, J. (2006). Tests using spatial median, *Austrian Journal of Statistics*, **35**, 331–338.

Utts, J. M. and Hettmansperger, T. P. (1980). A robust class of tests and estimates for multivariate location, *Journal of the American Statistical Association*, **75**, 939–946.

A Robust Test for Location Parameters in Multivariate Data

Sunha So[1] · Dong-Hee Lee[2] · Byoung Cheol Jung[3]

[1]Risk Model Validation Team, WooriBank
[2]Department of Business Administration, Kyonggi University
[3]Department of Statistics, University of Seoul

Abstract

This work propose a robust test for location parameters in multivariate data based on MVE and MCD with the affine equivariance and the high-breakdown properties. We consider the hypothesis testing satisfying high efficiency and high test power simultaneously to bring in the one-step reweighting procedure upon high-breakdown estimators, which generally suffer from the low efficiency and, as a result, usually used only in the exploratory analysis. Monte Carlo study shows that the suggested method retains nominal significance levels and higher testing power without regard to various population distributions than a Hotelling's T^2 test. In an example, a data set containing known outliers does not make an influence toward our proposal, while it renders a Hotelling's T^2 useless.

Keywords: High-breakdown estimation, minimum covariance determinant, minimum volume ellipsoid, outliers, reweighting, spatial median.

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund)(KRF-2007-314-C00039).

[3]Corresponding author: Assistant professor, Department of Statistics, University of Seoul, Jeonnong-Dong 90, Dongdaemun-Gu, Seoul 136-743, Korea. E-mail: bcjung@uos.ac.kr