응용통계연구
제 4 권 제 1 호, 1991년

두 독립 모집단의 공분산 행렬에 관한 붓스트랩 추론*

김기영** 전명식**

〈요 약〉

다변량분산분석이나 판별분석 등에 있어서 검정의 대상이 되는 공분산행렬의 동일성에 대한 붓스트랩방법의 활용을 살펴보았다. 두 모집단의 공분산행렬을 Σ_1, Σ_2라 하면, 가설 $H: \Sigma_1 = \Sigma_2$은 불변성의 관점에서 $\Sigma = \Sigma_1 \Sigma_2^{-1}$의 고유값들이 모두 1이라는 것과 동등하다. 본 연구에서는 (1) $\Sigma = \Sigma_1 \Sigma_2^{-1}$의 표본고유값들에 대한 편의를 붓스트랩에 의해 정정하였으며, (2) 이들의 표본분포를 붓스트랩분포로 추정하여 검정에 활용하였으며, (3) 합동붓스트랩에 의해 바플렛의 수정우도비 검정통계량의 분포를 근사하였다.

1. 서 론

다변량 분산분석(MANOVA)이나 판별분석(Discriminant Analysis) 등에 있어서 모집단 공분산 행렬들의 동일성 여부는 통계적 방법의 사용에 있어서 매우 중요한 검증의 대상이 된다. 이제, 두 독립 모집단으로부터 p-차원 확률벡터 $X_{ij}(i=1, 2;\ j=1, \cdots, n_i)$를 다음과 같이 얻었다고 하자.

$$X_{11}, X_{12}, \cdots, X_{1n1} \quad \text{iid} \quad F_1$$
$$X_{21}, X_{22}, \cdots, X_{2n2} \quad \text{iid} \quad F_2$$

여기서 분포 F_i는 평균벡터가 μ_i이고 공분산 행렬이 Σ_i임 $(i=1, 2)$.

이때 두 공분산행렬의 동일성을 검정하는 통계적가설

$$H: \Sigma_1 = \Sigma_2 \quad \text{대} \quad K: \Sigma_1 \neq \Sigma_2$$

을 고려하기로 한다. 편의상 $\Sigma = \Sigma_1 \Sigma_2^{-1}$ 그리고 Σ의 고유값을 $\delta = (\delta_1, \delta_2, \cdots, \delta_p)$라고 표기하면 위의 가설검정문제는

* 이 논문은 1989년도 문교부 지원 학술진흥재단의 자유공모과제 학술연구 조성비에 의하여 연구되었음.

** (136-701) 서울시 성북구 안암동 고려대학교 통계학과

$$H : \delta_1 = \delta_2 = \cdots = \delta_p = 1 \quad 대 \quad K : H가\ 아님$$

으로 귀결될 수 있다. 여기서 분포 F_i로부터 얻은 표본평균벡터를 $\overline{X}_i$, 표본공산행렬을 $S_i (i = 1, 2)$로 표기하기로 한다.

$$\overline{X}_i = 1/n_i \sum_{j=1}^{n_i} X_{ij} \quad S_i = 1/n_i \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_i)(X_{ij} - \overline{X}_i)' \quad (i = 1, 2)$$

실제 위의 가설검정문제는 A를 (pxp) 정칙행렬이라 하고 b, $c \in R_p$라 할 때

$$(\overline{X}_1, \overline{X}_2, S_1, S_2) \rightarrow (A\overline{X}_1 + b, A\overline{X}_2 + c, AS_1A', AS_2A')$$

와 같은 변환에 대하여 불변(invariant)하다. 이때, 모수벡터(μ_1, μ_2, Σ_1, Σ_2)에 대한 최대 불변량은 $\delta = (\delta_1, \delta_2, \cdots, \delta_p)$이며, 따라서 임의의 불변검정통계량은 $S = S_1S_2^{-1}$의 표본고유값인 $l = (l_1, l_2, \cdots, l_p)$에만 의존하게 된다. 2절과 3절에서는 각각 고유값 $\delta = (\delta_1, \delta_2, \cdots, \delta_p)$에 대한 추정과 가설 $\Sigma_1 = \Sigma_2$에 대한 검정을 고려하여 붓스트랩(bootstrap)의 사용방법을 설명하고 그로 인한 향상점들을 모의실험을 통해 알아보기로 한다.

2. 고유값 δ의 추정

$\Sigma = \Sigma_1\Sigma_2^{-1}$의 고유값 $\delta = (\delta_1, \delta_2, \cdots, \delta_p)$의 추정량으로는 주어진 표본 $X_{i1}, X_{i2}, \cdots, X_{in_i} (i = 1, 2)$에 근거한 $S = S_1S_2^{-1}$의 표본고유값 $l = (l_1, l_2, \cdots, l_p)$를 사용할 수 있으나 이는 불편추정량이 아니다. 또한 i번째 표본고유값 l_i에 포함되어 있는 모분포의 j번째 고유값 $\delta_j (i \neq j)$에 대한 정보를 무시하고 있다는 약점도 있다. 이에 Muirhead와 Verathaworn(1985) 그리고 Leung과 Muirhead(1985, 1987)은 다변량정규성의 가정하에서 다변량분산분석(MANOVA)과 정준상관분석(Canonical correalation analysis)의 경우에 의사결정론적인 방법으로 모분포의 고유값의 추정에 대해 논하였다.

한편 Efron(1979)에 의해 제안된 붓스트랩 방법은 기저분포에 대한 모수적 분포의 가정없이도 주어진 표본으로부터 재표본(resampling)을 취하는 방법에 근거하여 추정량의 편의(bias)와 표준편차(standard deviation)를 구하는데 효과적으로 사용될 수 있다. 이러한 붓스트랩방법은 Beran과 Srivastava(1985)에 의해 일표본 모공분산행렬에 대한 추정이론에도 사용될 수 있음이 밝혀졌다. 여기서는 표본고유값의 편의를 붓스트랩방법으로 추정하여 $\delta = (\delta_1, \delta_2, \cdots, \delta_p)$의 편의정정 추정값을 구해보고자 한다. 또한 표본고유값의 표준편차(standard deviation)도 같은 방법으로 추정하여 보기로 한다.

우선 붓스트랩의 실제적인 사용에 필요한 몬테칼로(Monte Carlo) 근사방법은 다음과 같이 구해진다.

(단계 1) 주어진 값 n_i개의 관찰값으로부터 두모집단의 경험적분포 $F_{ni} (i = 1, 2)$를 만든다.

(단계 2) 두 경험적분포 F_{ni}로부터 복원무작위추출표본(붓스트랩표본이라고 통칭함) X_{i1}^*, X_{i2}^*, $\cdots$, $X_{in_i}^* (i = 1, 2)$를 얻고 그에 상응하는 $S_1^*S_2^{*-1}$의 붓스트랩 표본고유값 $l^* =$

$(l_1^*, l_2^*, \cdots, l_p^*)$을 계산한다.

(단계 3) 단계 2를 독립적으로 반복시행하여 구한 l^*들의 '붓스트랩분포'로 부터 $l=(l_1, l_2, \cdots, l_p)$의 편의와 표준오차를 구한다.

즉, 단계 3에서의 반복독립시행횟수를 B라고 했을 때, j번째 시행에서 구한 붓스트랩 표본 고유값들을 $l^{j*}=(l_1^{j*}, l_2^{j*}, \cdots, l_p^{j*})$라 하면 편의 $b=(b_1, b_2, \cdots, b_p)$와 표준편차 $SD=(SD_1, SD_2, \cdots, SD_p)$는 각각 다음과 같이 근사추정된다.(여기서 편의 b_i와 표준편차 SD_i는 각각 i번째 고유값 δ_i의 추정에 대한 것을 의미한다.)

$$b_i = 1/B \sum_{j=1}^{B} l_i^{j*} - l_i$$

$$SD_i^2 = 1/B \sum_{j=1}^{B} (l_i^{j*} - \bar{l}_i^*)^2$$

$$(\text{단 } \bar{l}_i^* = 1/B \sum_{j=1}^{B} l_i^{j*})$$

(계산능력이 좋은 컴퓨터의 이용은 단계 3의 독립반복시행에 필수적이라고 할 수 있다.)

이제 위의 붓스트랩방법에 의해 추정된 편의를 사용한 $\delta=(\delta_1, \delta_2, \cdots, \delta_p)$의 편의정정추정량은

$$\tilde{l}=(\tilde{l}_1, \tilde{l}_2, \cdots, \tilde{l}_p)$$
$$= (l_1, l_2, \cdots, l_p) - (b_1, b_2, \cdots, b_p)$$

로 주어진다.

모의실험

이제, 표본고유값 $l=(l_1, l_2, \cdots, l_p)$에 대한 편의정정 추정량 $\tilde{l}=(\tilde{l}_1, \tilde{l}_2, \cdots, \tilde{l}_p)$의 성능향상을 모의실험을 통하여 살펴보기로 한다. 자료를 생성하는 기저분포는 2, 3, 4차원 다변량정규분포 경우를 고려하였다. 표본의 크기는 $n_1=n_2=30$을 택하였으며 붓스트랩분포는 $B=300$을 사용한 몬테칼로 근사방법을 이용하였다. 고려한 고유값은 각 차원에서 3가지씩이며 그에 대한 내용은 아래 〈표 1〉에 주어져 있다. 이제 $\hat{\delta}=(\hat{\delta}_1, \hat{\delta}_2, \cdots, \hat{\delta}_p)$를 $\delta=(\delta_1, \delta_2, \cdots, \delta_p)$의 추정량이라 할 때 그의 성능을 측정하는 방법으로

$$E \sum_{i=1}^{p} (\hat{\delta}_i - \delta_i)^2$$

와

$$E \sum_{i=1}^{p} (1 - \hat{\delta}_i/\delta_i)^2$$

을 택했다. 이중 후자는 고유값의 크기를 고려한 일종의 표준화된 꼴이라 하겠다. 편의상 고유값들의 순서는 각각 $\delta_1 \leq \delta_2 \leq \cdots \leq \delta_p$와 $\hat{\delta}_1 \leq \hat{\delta}_2 \leq \cdots \leq \hat{\delta}_p$으로 잡았다. 이제, 위와 같은 기준을 500회의 독립반복시행을 통한 평균으로, 표본고유값 $l=(l_1, l_2, \cdots, l_p)$과 편의정정추정량 $\tilde{l}=$

$(\tilde{l}_1, \tilde{l}_2, \cdots, \tilde{l}_p)$에 대해 살펴보았으며 아울러 붓스트랩에 의한 편의추정량의 평균도 구했다. 이와 같은 결과는 아래 〈표 1〉에 주어져 있으며, 그 내용을 요약하면 다음과 같다. 고유값의 크기를 고려한 $E\sum_{i=1}^{p}(1-\delta_i/\delta_i)^2$을 기준으로 삼았을 때, 편의정정추정량 $\tilde{l}$은 표본고유값 l보다, 2-차원의 경우 약 10~40%, 3-차원의 경우 약 10~50%, 4차원의 경우 약 40~65%의 아주 높은 '효율성'의 증가를 보이고 있다. 또한, 전체적으로 차원에 관계없이 고유값들이 모두 같을 때 편의정정추정량의 효율이 가장 뚜렷한 반면, 차원이 커짐에 따라 편의정정에 의한 추정량의 성능향상이 더욱 두드러짐을 볼 수 있다. 이러한 제반현상의 가장 근본적인 이유는 가장 큰 고유값 δ_p에 대한 표본고유값 l_p의 상대적 과대추정에서 비롯되며 이에 대한 붓스트랩편의추정이 매우 바람직하게 되어지고 있음을 확인할 수 있다.

〈표 1〉 표본고유값 $l=(l_1, l_2, \cdots, l_p)$과 편의추정량 $\tilde{l}=(\tilde{l}_1, \tilde{l}_2, \cdots, \tilde{l}_p)$의 비교

기 준	$E(\hat{\delta}_i-\delta_i)^2$		$E(1-\delta_i/\delta_i)^2$		l_i의 편의평균
δ_i의 추정량	l_i	$\tilde{l}_i$	l_i	$\tilde{l}_i$	
$\delta_1=1$	0.124	0.121	0.124	0.121	−0.067
$\delta_2=1$	0.426	0.207	0.426	0.207	0.309
합	0.550	0.328	0.550	0.328	

기 준	$E(\hat{\delta}_i-\delta_i)^2$		$E(1-\hat{\delta}_i/\delta_i)^2$		l_i의 편의평균
δ_i의 추정량	l_i	$\tilde{l}_i$	l_i	$\tilde{l}_i$	
$\delta_1=1$	0.163	0.164	0.163	0.164	0.023
$\delta_2=10$	19.504	14.207	0.195	0.142	1.375
합	19.667	14.371	0.358	0.306	

기 준	$E(\hat{\delta}_i-\delta_i)^2$		$E(1-\hat{\delta}_i/\delta_i)^2$		l_i의 편의평균
δ_i의 추정량	l_i	$\tilde{l}_i$	l_i	$\tilde{l}_i$	
$\delta_1=5$	2.769	3.650	0.111	0.146	−0.262
$\delta_2=10$	23.603	16.728	0.236	0.167	2.144
합	26.372	20.378	0.347	0.313	

기 준	$E(\hat{\delta}_i-\delta_i)^2$		$E(1-\hat{\delta}_i/\delta_i)^2$		l_i의 편의평균
δ_i의 추정량	l_i	$\tilde{l}_i$	l_i	$\tilde{l}_i$	
$\delta_1=1$	0.202	0.152	0.202	0.152	−0.104
$\delta_2=2$	0.080	0.112	0.080	0.112	0.042
$\delta_3=1$	1.210	0.428	1.210	0.428	0.617
합	1.492	0.692	1.492	0.692	

기 준	$E(\hat{\delta}_i-\delta_i)^2$		$E(1-\hat{\delta}_i/\delta_i)^2$		l_i의 편의평균
δ_i의 추정량	I_i	$\tilde{l}_i$	l_i	$\tilde{l}_i$	
$\delta_1=1$	0.156	0.199	0.156	0.199	−0.047
$\delta_2=5$	2.912	3.919	0.116	0.157	−0.028
$\delta_3=10$	36.099	21.241	0.361	0.212	3.077
합	39.167	25.359	0.633	0.568	

기 준	$E(\hat{\delta}_i-\delta_i)^2$		$E(1-\hat{\delta}_i/\delta_i)^2$		l_i의 편의평균
δ_i의 추정량	l_i	$\tilde{l}_i$	l_i	$\tilde{l}_i$	
$\delta_1=1$	0.162	0.187	0.162	0.187	−0.026
$\delta_2=10$	11.592	12.572	0.116	0.126	−0.357
$\delta_3=10$	58.375	24.421	0.584	0.244	4.232
합	70.129	37.180	0.862	0.557	

기 준	$E(\hat{\delta}_i-\delta_i)^2$		$E(1-\hat{\delta}_i/\delta_i)^2$		l_i의 편의평균
δ_i의 추정량	l_i	$\tilde{l}_i$	l_i	$\tilde{l}_i$	
$\delta_1=1$	0.282	0.193	0.282	0.193	−0.122
$\delta_2=1$	0.076	0.094	0.076	0.094	−0.056
$\delta_3=1$	0.192	0.154	0.192	0.154	0.181
$\delta_4=1$	1.936	0.411	1.936	0.411	1.012
합	2.486	0.852	2.486	0.852	

기 준	$E(\hat{\delta}_i-\delta_i)^2$		$E(1-\hat{\delta}_i/\delta_i)^2$		l_i의 편의평균
δ_i의 추정량	l_i	$\tilde{l}_i$	l_i	$\tilde{l}_i$	
$\delta_1=1$	0.160	0.153	0.160	0.153	−0.121
$\delta_2=1$	0.364	0.312	0.364	0.312	0.135
$\delta_3=10$	9.861	12.221	0.099	0.122	0.003
$\delta_4=10$	80.449	24.949	0.804	0.249	5.744
합	90.834	37.635	1.427	0.836	

기 준	$E(\hat{\delta}_i-\delta_i)^2$		$E(1-\hat{\delta}_i/\delta_i)^2$		l_i의 편의 평균
δ_i의 추정량	l_i	$\tilde{l}_i$	l_i	$\tilde{l}_i$	
$\delta_1=1$	0.141	0.171	0.141	0.171	−0.069
$\delta_2=10$	18.568	15.375	0.186	0.154	−0.822
$\delta_3=10$	9.372	11.956	0.094	0.120	0.965
$\delta_4=10$	130.570	33.962	1.306	0.340	7.931
합	158.651	61.464	1.727	0.785	

3. 붓스트랩 검정방법

앞서 고려한 두 공분산행렬의 동일성 여부에 관한 통계적가설

$$H : \Sigma_1 = \Sigma_2 \quad \text{대} \quad K : \Sigma_1 \neq \Sigma_2$$

에 대한 붓스트랩의 사용방법을 살펴보기로 한다. 우선 $\Sigma_1\Sigma_2^{-1}$의 고유값 $\delta=(\delta_1, \delta_2, \cdots, \delta_p)$ 대한 붓스트랩신뢰영역을 사용하는 방법을 제시하기로 하며, 다음으로 다변량정규성하에서 사용되는 Bartlett에 의한 수정우도비 검정통계량의 귀무가설 H하에서의 분포를 붓스트랩방법을 통해 추정하는 방법을 살펴보기로 한다.

3. 1 고유값 δ에 대한 붓스트랩 신뢰영역

우리가 고려하고 있는 가설검정 문제는

$$H : \delta_1 = \delta_2 = \cdots = \delta_p = 1 \quad \text{대} \quad K : H\text{가 아님}$$

과 동등하며, 따라서 고유값 $\delta = (\delta_1, \delta_2, \cdots, \delta_p)$ 대한 붓스트랩 신뢰영역의 활용성을 살펴보기로 한다. 두 모분포 F_1과 F_2에 의존하는 확률행렬 $(n_1n_2/n_1n_2)^{1/2}(S_1S_2^{-1} - \Sigma_1\Sigma_2^{-1})$의 표본분포를 $J_{n1,n2}(F_1, F_2)$로 표기할 때 $J_{n1,n2}(F_1, F_2)$의 붓스트랩 추정량은 $J_{n1,n2}(F_{n1}, F_{n2})$가 되며 ($F_{n1}$과 F_{n2}는 각각 F_1, F_2의 경험적 분포)이의 폐쇄형(closed form)은 매우 복잡하나 앞 절에 서술된 몬테칼로 방법에 의해 근사계산될 수 있다. 이와 같은 붓스트랩방법의 이론적인 정당성은 다음과 같은 그의 일치성(consistency)에서 찾아 볼 수 있다.

붓스트랩방법의 일치성

우선 두 표본공분산행렬 S_1과 S_2의 함수인 $g(S_1, S_2) = S_1S_2^{-1}$를 (Σ_1, Σ_2)에 관해 Taylor 전개시키면

$$\begin{aligned} & S_1S_2^{-1} \\ &= \Sigma_1\Sigma_2^{-1} + \{(S_1 - \Sigma_1)(\partial/\partial U) + (S_2 - \Sigma_2)\}(\partial/\partial V)g(U, V) \mid U = \Sigma_1,\ V = \Sigma_2 \\ &\quad + O_p\{(n_1n_2/n_1 + n_2)\} \end{aligned}$$

가 될 것이다. 한편, 각 모집단에 대해 확률행렬 $n_i^{1/2}(S_i - \Sigma_i)$의 붓스트랩분포는 Lindeberg 중심극한정리를 이용하여, 실제분포와 같은 극한분포를 가짐이 증명된다(참조 : Beran 1985). 나아가 두 표본공분산행렬 S_1과 S_2는 서로 독립이므로 표본분포 $J_{n1,n2}(F_1, F_2)$의 붓스트랩추정값 $J_{n1,n2}(F_{n1}, F_{n2})$는 두 표본의 크기 n_1, n_2가 무한히 커짐에 따라 $n_1/(n_1+n_2)$가 $0 < \lambda < 1$로 접근할 경우, 동일한 연속극한분포 $J(F_1, F_2, \lambda)$를 가진다. 이 경우 붓스트랩분포에 근거한 신뢰영역은 표본의 크기가 커짐에 따라 올바른 신뢰확률에 수렴한다.

이제 고유값 $\delta = (\delta_1, \delta_2, \cdots, \delta_p)$에 대한 기초대칭다항식 $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_p)$를 다음과 같이 고려한다.

$$\begin{aligned} \gamma_1 &= \sum_{i=1}^{p} \delta_i \\ \gamma_2 &= \sum_{i<j} \sum \delta_i\delta_j \\ &\vdots \\ \gamma_p &= \prod_{i=1} \delta_i \end{aligned}$$

그런데 $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_p)$는 특성다항식의 근(root of characteristic polynomial)이므로 $\Sigma = \Sigma_1\Sigma_2^{-1}$미분가능한 함수이며, 따라서 delta-방법에 의해 붓스트랩방법의 일치성이 보장된다. 따라서, 표본고유값 $l = (l_1, l_2, \cdots, l_p)$에 대한 기초대칭다항식을 $r = (r_1, r_2, \cdots, r_p)$이라 할 때, 통계량

$$\max_{1\le i\le p} \mid r_i - \gamma_i \mid$$

의 실제분포와 붓스트랩분포는 같은 극한분포를 갖게된다. 이제, 붓스트랩 표본고유값 $l^* = (l_1^*, l_2^*, \cdots, l_p^*)$에 대한 기초대칭다항식을 $r^* = (r_1^*, r_2^*, \cdots, r_p^*)$이라 할 때, 조건부통계량

$$\max_{1\le i\le p} \mid r_i^* - r_i \mid$$

의 붓스트랩분포는 앞에서 설명했던 것과 마찬가지로 몬테칼로방법에 의해 구하여진다. 여기서 붓스트랩분포의 α-백분위수를 $C_n(\alpha)$라고 하면, $n\to\infty$에 따라

$$P[\max_{1\le i\le p} \mid r_i - \gamma_i \mid < C_n(\alpha)] \to \alpha$$

이며, p개의 γ_i에 대한 $(1-\alpha)\times 100\%$ 붓스트랩 동시신뢰영역은, 모든 i에 대해서,

$$r_i + C_n(1-\alpha/2) \quad < \quad \gamma_i \quad < \quad r_i + C_n(\alpha/2)$$

로 구할 수 있다. 그런데, $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_p) \to \delta = (\delta_1, \delta_2, \cdots, \delta_p)$는 1 : 1변환이므로, $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_p)$에 대한 동시신뢰영역은 $\delta = (\delta_1, \delta_2, \cdots, \delta_p)$에 대한 동시 신뢰영역으로 치환될 수 있으며 이는 앞에서의 가설검정문제에 사용될 수 있다. 즉, 귀무가설 H하에서 $\delta = (1, 1, 1, 1)$이므로 δ에 대한 붓스트랩 신뢰영역이 이를 포함하면 귀무가설을 받아들이고 그렇지 않으면 기각하는 것이다.

이제, 모의실험을 통하여, $\Sigma = \Sigma_1\Sigma_2^{-1}$의 고유값 $\delta = (\delta_1, \delta_2, \cdots, \delta_p)$에 대한 붓스트랩 신뢰영역을 앞의 가설검정문제에 사용했을 때 그의 유의수준과 검정력을 살펴보기로 한다. 이에 대한 결과는 아래 〈표 2〉에 주어져 있다. 자료를 생성하는 기저분포는 4차원 다변량정규분포 경우로 $(\delta_1, \delta_2, \delta_3, \delta_4) = (1, 1, 1, 1)$이며, 표본의 크기는 $n_1 = n_2 = 30$인 경우와 $n_1 = n_2 = 50$인 경우를 고려하였다. 검정력을 알아보기 위한 대립가설은 아래 〈표 2〉에서와 같이 K_1, K_2, K_3, K_4의 4경우를 택하였다. 사용한 방법은 다음과 같다. 통계량 $\max_{1\le i\le 4} \mid r_i - \gamma_i \mid$의 표본분포를 $B = 200$회의 재표본에 근거한 $\max_{1\le i\le 4} \mid r_i^* - r_i \mid$의 붓스트랩분포로 추정하여 δ_i의 $(1-\alpha)$ 수준의 동시신뢰영역을 구한 다음, 그 안에 귀무가설에 해당하는 $(\delta_1, \delta_2, \delta_3, \delta_4) = (1, 1, 1, 1)$가 포함되어 있으면 귀무가설을 받아들이고 그렇지 않으면 귀무가설을 기각하는 것을 500회의 독립반복시행을 통해 알아보고 이에 근거하여 유의수준을 추정하였다. 또한 검정력은 해당하는 $(\delta_1, \delta_2, \delta_3, \delta_4)$값이 붓스트랩 신뢰영역에 포함되지 않는 경우 즉 귀무가설을 기각하는 상대도수를 500회의 독립반복시행을 통해 추정하였다. 유의수준은 $\alpha = 0.05, 0.10$인 두 가지 경우를 고려하였으며, 그 내용을 요약하면 다음과 같다. 우선 붓스트랩신뢰영역의 유의수준의 정확함이 눈에 띠이며 이는 표본의 크기가 커짐에 따라 더욱 향상될 것으로 여겨진다. 또한 귀무가설에서 멀어질수록 표본의 크기가 커질 수록 검정력이 높아지는 타당한 결과를 제시하고 있다. 따라서, 모분포에 대한 다변량정규성 등의 가정없이도, 고유값 $\delta = (\delta_1, \delta_2, \cdots, \delta_p)$의 함수에 대한 신뢰영역의 구축 및 그에 대한 가설검정문제에 붓스트랩방법의 활용은 올바른 결과를 기대케 한다.

〈표 2〉 붓스트랩 신뢰영역의 유의수준과 검정력의 추정

가 설 \ 표본크기	$n_1=n_2=30$		$n_1=n_2=50$	
유의수준	0.05	0.10	0.05	0.10
$H:\delta=(1.0, 0.1, 0.1, .0)$	0.041	0.089	0.051	0.110
$K_1:\delta=(0.5, 1.0, 1.0, 1.5)$	0.068	0.118	0.080	0.140
$K_2:\delta=(1.0, 1.0, 1.5, 2.0)$	0.153	0.306	0.413	0.581
$K_3:\delta=(0.5, 1.0, 2.0, 2.5)$	0.188	0.332	0.466	0.619
$K_4:\delta=(0.5, 1.5, 2.5, 3.0)$	0.622	0.758	0.918	0.960

3. 2 수정우도비검정통계량의 붓스트랩분포

다변량정규성하에서 공분산행렬의 동일성을 검정하는 통계적가설

$$H:\Sigma_1=\Sigma_2 \quad 대 \quad K:\Sigma_1\neq\Sigma_2$$

에 대한 Bartlett의 수정우도비는 다음과 같다.

$$\wedge=\frac{\prod_{i=1}^{2}[\det(n_iS_i)^{(ni-1)/2}]}{[\det(n_1S_1+n_2S_2)]^{(ni+n2-2)/2}}\quad\frac{(n_1+n_2-2)^{p(ni+n2-2)/2}}{(n_1-1)^{p(ni-1)}(n_2-1)^{p(n2-1)}}$$

이때 수정우도비 검정통계량 $-2\log\wedge$는 귀무가설하에서 자유도가 $p(p+1)/2$인 카이제곱분포로 약수렴한다. 이제 Boos와 Brownie(1989)가 일변량의 경우 분산의 동질성(homogeneity)을 검정하기 위하여 사용했던 합동붓스트랩(pooled bootstrap) 방법을 다변량 공분산행렬의 동일성을 검정하는 통계적 가설의 경우로 연장하면 다음과 같다.

편의상, 붓스트랩의 실제적인 사용에 필요한 몬테칼로(Monte Carlo) 근사방법을 설명하기로 한다.

(단계 1) 주어진 각 n_i개의 관찰값으로부터 합동재표본추출공간 $\{X_{ij}-\overline{X}_i : i=1, 2, j=1, 2, \cdots, n_i\}$를 만든다.

(단계 2) 합동된 재표본추출공간으로부터 붓스트랩표본 $\{X_{i1}^*, X_{i2}^*, \cdots, X_{imi}^*, i=1, 2\}$를 얻고 그에 상응하는 붓스트랩 수정우도비 검정통계량 $-2\log\wedge^*$을 계산한다.

(단계 3) 단계 2를 독립적으로 반복시행하여 구한 $-2\log\wedge^*$의 값들로부터 '붓스트랩분포'의 근사분포를 구한다.

이와 같이 구한 $-2\log\wedge^*$의 붓스트랩분포는 수정우도비검정통계량의 귀무가설하에서 표본분포를 추정하는데 사용된다. 즉, $-2\log\wedge$의 값이 붓스트랩분포의 $(1-\alpha)$-백분위수보다 크면 유의수준 α에서 귀무가설 H를 기각한다. 합동붓스트랩은 조건부 공분산행렬이 같도록 하여 귀무가설상태를 만듦으로써 귀무가설하에서의 표본분포를 추정하는데 사용될 수 있다는 점에 착안하고 있다. 또한, 이러한 방법은 $k(>2)$개의 공분산행렬의 동일성여부의 검정으로도 연장이 가능하며 다변량정규성이란 가정을 필요로하지 않는 장점도 있다.

3. 3 실제 데이터 처리결과

앞에서 설명한 방법들을 실제 데이터에 사용해보고 그 결과들을 비교하여 보고자한다. 사용한 데이터는 피셔의 붓꽃자료(Fisher's iris data)의 일부로써 두품종 versicolor와 virginica의 두 변수 (X_1, X_2)=(sepal width, petal width)에 대한 공분산행렬의 동일성 여부의 통계검정을 다루기로한다(참조 : 자료 1). 우선 이변량정규성하에서 Bartlett의 수정우도비 검정통계량의 값은 $-2\log\Lambda=9.6626$으로 계산되었다. 이제 검정통계량의 분포를 자유도가 3인 카이제곱분포로 근사추정하면 p-value는 0.022가 된다. 또한 $B=1000$회의 독립반복시행에 의한 합동붓스트랩분포에 의하면 p-value는 0.027로 두 방법 사이에는 큰 차이를 보이지 않는다. 또한 $\Sigma_1\Sigma_2^{-1}$의 고유값 $\delta=(\delta_1, \delta_2)$의 붓스트랩 신뢰영역을 사용하기 위한 통계량 $\max_{i=1,\ 2}|r_i-\gamma_i|$의 값은 0.6427로 구해졌다.(귀무가설하에서는 $\gamma=(2,\ 1)$이다.) 이제, $B=1000$회의 독립반복시행에 의한 $\max_{i=1,\ 2}|r_i^*-r_i|$의 붓스트랩분포에 의하면 p-value는 0.062가 되었다. 이는 수정우도비 검정통계량을 사용한 결과와는 약간의 차이를 보이나, 어느것이 옳다고 단정지을 수는 없을 것으로 여겨지며 이에 대한 진전된 연구도 바람직하다. 단, 여기서 제안한 두가지 붓스트랩방법은 다변량정규성을 필요로하지 않으며, 합동붓스트랩은 검정통계량의 극한분포가 카이제곱분포임으로 미루어 타원형대칭분포에 더 걎합할 것으로 사료된다.

〈자료 1〉 Fisher의 붓꽃자료 일부 ((1) Andrews와 Herzberg p 5−8) 단위 : mm

Iris versicolor				Iris virginica			
sepal width	petal width	sepal width	petal width	sepal width	petal width	sepal width	petal width
32	14	32	15	33	25	27	19
31	15	23	13	30	21	29	18
28	15	28	13	30	22	30	21
33	16	24	10	25	17	29	18
29	13	27	14	25	18	36	25
20	10	30	15	32	20	27	19
22	10	29	14	30	21	25	20
29	13	31	14	28	24	32	23
30	15	27	10	30	18	38	22
22	15	25	11	26	23	22	15
32	18	28	13	32	23	28	20
25	15	28	12	28	20	27	18
29	13	30	14	33	21	32	18
28	14	30	17	28	18	30	18
29	15	26	10	28	21	30	16
24	11	24	10	28	19	38	20
27	12	27	16	28	22	28	15
30	15	34	16	26	14	30	23

31	15	23	13	34	24	31	18
30	13	25	13	30	18	31	21
26	12	30	14	31	24	31	23
26	12	23	10	27	19	32	23
27	13	30	12	33	25	30	23
29	13	29	13	25	19	30	20
25	11	28	13	34	23	30	18

본 논문의 오류를 지적하여주신 익명의 심사위원께 감사합니다.

◇ 참 고 문 헌 ◇

(1) Andrews, D. & Herzberg, A. (1985) *Data* Springer-Verlag New York.

(2) Beran, R (1985) "Bootstrap methods in statistics" *Jber. d. Dt. Math.* -verin. **86** 14-30.

(3) Beran, R. & Srivastava, M. (1985) "Bootstrap tests and confidence regions for functions of a covariance matrix" *Ann. Statist.* **13.** 95-115.

(4) Boos, D. & (1979) "Bootstrap methods for testing homogeneity of variances." *Technometrics.* **31**, 1, 69-82.

(5) Efron, B (1979) "Bootstrap method : another look at the jackknife" *Ann. Statist.* **7** 1-26.

(6) Leung P L & Muirhead R. (1987) "estimation of parameter matrices and eigenvalues in MANOVA and canonical correlation analysis". *Ann. Statist.* **15**, 4, 1651-1666.

(7) Muirhead, R & Verathaworn, T. (1985) "On estimating the latent roots of $\Sigma_1\Sigma_2$" *Multivariate Analysis*(P.R. Krishnaiah ed.)

(8) Muirhead, R & Leung, P (1985) "Estimating functions of canonical correlation coefficients" *Linear algebra and its applications* **70** : 173-183

Bootstrap inference for covariance matrices of two independent populations

Keeyoung Kim,* Myoungshic Jhun*

〈Abstract〉

It is of great interest to consider the homogeniety of covariance matrices in MANOVA of discriminant analysis. If we look at the problem of testing hypothesis, $H: \Sigma_1 = \Sigma_2$ from an invariance point of view where Σ_i are the covariance matrix of two independent p-variate distribution, the testing problem is invariant under the group of nonsingular transformations and the hypothesis becomes $H: \delta_1 = \delta_2 = \cdots = \delta_p = 1$ where $\delta = (\delta_1, \delta_2, \cdots, \delta_p)$ is a vector of latent roots of Σ. Bias-corrected estimators of eigenvalues and sampling distribution of the test statistics proposed are obtained. Pooled-bootstrap method also considered for Bartlett's modified likelihood ratio statistics.

*Department of Statistics, Korea University, Seoul 136-701, Korea