The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.4.567
30(4), 567–578

Non-parametric approach for the grouped dissimilarities using the multidimensional scaling and analysis of distance

Seungchan Nam[a] · Yong-Seok Choi[a,1]

[a]Department of Statistics, Pusan National University

(Received May 8, 2017; Revised June 29, 2017; Accepted July 2, 2017)

Abstract

Grouped multivariate data can be tested for differences between two or more groups using multivariate analysis of variance (MANOVA). However, this method cannot be used if several assumptions of MANOVA are violated. In this case, multidimensional scaling (MDS) and analysis of distance (AOD) can be applied to grouped dissimilarities based on the various distances. A permutation test is a non-parametric method that can also be used to test differences between groups. MDS is used to calculate the coordinates of observations from dissimilarities and AOD is useful for finding group structure using the coordinates. In particular, AOD is mathematically associated with MANOVA if using the Euclidean distance when computing dissimilarities. In this paper, we study the between and within group structure by applying MDS and AOD to the grouped dissimilarities. In addition, we propose a new test statistic using the group structure for the permutation test. Finally, we investigate the relationship between AOD and MANOVA from dissimilarities based on the Euclidean distance.

Keywords: dissimilarity, multidimensional scaling, analysis of distance, permutation test, analysis of variance

1. 서론

최근 빅 데이터시대에 자연과학뿐만 아니라 경영·경제학, 의학, 사회과학, 체육학 등 여러 분야로부터 나오는 자료들은 그 형태가 매우 다양하다. 그 중에서도 여러 개의 변수로 이루어진 자료를 다변량자료라고 한다. 만약 이 자료가 그룹화되어 있다면 일반적으로 다변량 분산분석(multivariate analysis of variance)을 통해 그룹 간 차이를 검정할 수 있다. 그러나 자료가 정규성(normality)의 가정을 만족하지 못할 경우라든지 공분산행렬(covariance matrix)의 동질성(homogeneity)이 위반되는 경우 그리고 개체보다 변수의 수가 더 많은 경우 등에는 현실적으로 다변량 분산분석을 적용하기 어렵다.

이러한 경우에 그룹 간 차이를 검정하기 위하여 Clarke (1993)은 각 개체 간 유사성(similarity)들의 순위를 활용한 순열검정(permutation test)을 제안하였다. 이후로 Gower와 Krzanowski (1999)는 비유사성(dissimilarity)으로부터 그룹 구조를 파악할 수 있는 거리분석(analysis of distance)을 제안하였고 순열검정으로 다변량 분산분석을 대신할 수 있다고 하였다.

[1]Corresponding author: Department of Statistics, Pusan National University, 2, Busandaehak-ro, 63 beon-gil, Geumjeong-Gu, Busan 46241, Korea. E-mail: yschoi@pusan.ac.kr

거리분석은 그룹화된 비유사성에 다차원척도법(multidimensional scaling)을 적용시킨 후 그룹 간과 그룹 내의 구조를 파악하는 기법으로써 다변량 분산분석이 적용되기 어려운 상황에 유용하다. 이때 비유사성의 측도로 다양한 거리들을 활용할 수 있으며 만약 유클리드 거리를 사용할 경우 거리분석과 다변량 분산분석은 수리적으로 매우 밀접한 연관성을 가진다. 순열검정은 거리분석을 통해 계산된 그룹 구조의 정보를 이용하여 그룹 간의 유의미한 차이를 검정할 수 있는 비모수적 접근법이다.

따라서 본 연구에서는 다양한 거리를 활용한 비유사성에 다차원척도법과 거리분석을 적용하여 그룹 구조를 파악하고, 그로부터 그룹 간의 차이에 대하여 순열검정을 하기 위한 새로운 검정통계량을 제안하려 한다. 덧붙여 유클리드 거리를 활용한 비유사성을 통해 거리분석과 다변량 분산분석과의 수리적 연관성을 고찰하고자 한다.

이에 2장에서는 그룹화된 비유사성과 다차원척도법의 기초이론을 소개하려 한다. 3장에서는 거리분석과 순열검정의 이론을 소개하면서 다변량 분산분석과의 관계에 대해 설명하고자 한다. 4장에서는 모의실험을 통해 새로운 검정통계량과 기존의 것과의 검정력을 비교하고 5장에서는 다변량 분산분석을 적용할 수 없는 자료에 대한 활용 사례를 제시하려한다. 6장의 결론에서는 본 연구를 정리·요약하려 한다.

2. 그룹화된 비유사성과 다차원척도법

2.1. 그룹화된 다변량자료와 비유사성

이 절에서는 그룹화된 다변량자료와 비유사성을 설명하고자 한다. 그룹화된 다변량자료는 그룹의 수와 개체의 수, 그리고 변수의 수에 따라 행렬로 표현이 가능하다. 그룹의 수를 g라 하고 그룹별 개체 수를 n_r, $r = 1, \ldots, g$라 하자. 이때 전체 개체 수는 $n = \sum_{r=1}^{g} n_r$이고 변수의 수를 p라 할 때 크기가 $n \times p$인 그룹화된 자료행렬은

$$\mathbf{X} = \left[\mathbf{X}_1^t \ \mathbf{X}_2^t \ \cdots \ \mathbf{X}_g^t\right]^t \tag{2.1}$$

와 같이 표현할 수 있다. 위의 식에서 $\mathbf{X}$는 크기가 $n_r \times p$인 부분행렬 $\mathbf{X}_r$로 이루어져 있다. 그리고 식 (2.1)에서 r번째 부분행렬 $\mathbf{X}_r$의 i번째 행을 $\mathbf{x}_{ri}$, $i = 1, \ldots, n_r$라 한다면 $\mathbf{X}_r = (\mathbf{x}_{r1}, \mathbf{x}_{r2}, \ldots, \mathbf{x}_{rn_r})^t$와 같다. 여기서 $\mathbf{x}_{ri} = (x_{ri1}, \ldots, x_{rip})^t$는 r번째 그룹의 i번째 개체를 나타내는 행벡터이다. 그리고 s번째 그룹의 j번째 개체를 나타내는 행벡터를 $\mathbf{x}_{sj} = (x_{sj1}, \ldots, x_{sjp})^t$라 한다면 다양한 거리들을 사용하여 비유사성 $d_{rs}(i, j)$를 계산할 수 있다. 대표적으로 다음과 같은 유클리드 거리

$$d_{rs}(i,j) = \left[(\mathbf{x}_{ri} - \mathbf{x}_{sj})^t (\mathbf{x}_{ri} - \mathbf{x}_{sj})\right]^{\frac{1}{2}} = \left[\sum_{h=1}^{p} (x_{rih} - x_{sjh})^2\right]^{\frac{1}{2}} \tag{2.2}$$

와 시티-블록(city-block) 거리

$$d_{rs}(i,j) = \sum_{h=1}^{p} |x_{rih} - x_{sjh}| \tag{2.3}$$

등이 있다. 식 (2.2)와 (2.3)에서 x_{rih}는 r번째 그룹에서 i번째 개체의 h번째 변수를 나타내고 x_{sjh}는 s번째 그룹에서 j번째 개체의 h번째 변수를 나타낸다. 또한 식 (2.1)의 $\mathbf{X}$를 다음과 같은 비유사성행렬

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} & \cdots & \mathbf{D}_{1g} \\ \mathbf{D}_{21} & \mathbf{D}_{22} & \cdots & \mathbf{D}_{2g} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{D}_{g1} & \mathbf{D}_{g2} & \cdots & \mathbf{D}_{gg} \end{bmatrix} \tag{2.4}$$

로 변환할 수 있다. 위의 식에서 $\mathbf{D}$는 g^2개의 부분행렬 $\mathbf{D}_{rs} = (d_{rs}(i,j))$, $r,s = 1,\ldots,g$; $i = 1,\ldots,n_r$; $j = 1,\ldots,n_s$로 이루어진 비유사성 행렬이며 $\mathbf{D}_{rs}$는 $\mathbf{X}_r$와 $\mathbf{X}_s$사이의 비유사성 행렬을 의미한다.

2.2. 다차원척도법

이 절에서는 Choi (2014), Cox와 Cox (2001) 그리고 Torgerson (1958)을 참고로 하여 다차원척도법의 기초이론을 소개하고자 한다. 다차원척도법이란 고차원의 다변량자료로부터 계산되는 비유사성을 저차원 공간에 기하적으로 나타내어 개체들 간의 관계를 탐색적으로 살펴보는 다변량 그래프적 기법이다 (Choi, 2014). 일반적으로 n개의 개체 간 거리는 그들 간의 비유사성을 측정한 것이므로 다차원척도법은 n개의 개체 간 비유사성을 나타내는 행렬 $\mathbf{D} = (d_{kl})$, $k,l = 1,\ldots,n$을 구한 후 이 $\mathbf{D}$를 저차원 공간에 나타내는 기법이라고 볼 수 있다. 이때 고차원 유클리드 공간에서의 개체 간 비유사성 d_{kl}과 차원 축소된 저차원 공간에서의 비유사성 δ_{kl}사이의 관계가 서로 일치되도록 하는 것이 다차원척도법의 주목적이라고 볼 수 있다.

Kruskal과 Wish (1978)에 따르면 다차원척도법은 d_{kl}을 측정하는 척도에 따라 계량형과 비계량형 다차원척도법으로 나눌 수 있다. 흔히 계량형 다차원척도법에서 비유사성 d_{kl}의 측정척도는 개체 간의 실제 측정거리 또는 고전적인 유클리드거리를 나타내며 비계량형 다차원척도법에서는 거리들의 크기 순서를 나타낸다. 본 연구에서는 계량형 다차원척도법의 기법 중 하나인 고전적 척도법만을 다루도록 하겠다. Choi (2014) 그리고 Cox와 Cox (2001)를 참고로 하여 고전적 척도법의 절차에 대해 정리하면 다음과 같다.

[1단계] n개의 개체 간 비유사성 d_{kl}로 구성된 비유사성행렬 $\mathbf{D} = (d_{kl})$, $k,l = 1,\ldots,n$을 구한다.

[2단계] 비유사성행렬 $\mathbf{D} = (d_{kl})$로부터 행렬 $\mathbf{A} = (a_{kl})$을 구한다. 여기서 $a_{kl} = -d_{kl}^2/2$을 만족한다.

[3단계] $\mathbf{I}_n$은 단위행렬이고 $\mathbf{J}_n$은 구성 원소가 모두 1인 행렬일 때 중심화행렬을 $\mathbf{H} = \mathbf{I}_n - \mathbf{J}_n/n$라 한다면 행렬 $\mathbf{A}$로부터 다음과 같은 이중-중심화된 행렬 $\mathbf{Z}$를 구한다.

$$\mathbf{Z} = (z_{kl}) = \mathbf{HAH}, \tag{2.5}$$

여기서 $z_{kl} = a_{kl} - \bar{a}_{\cdot l} - \bar{a}_{k\cdot} + \bar{a}_{\cdot\cdot}$을 만족하고 $\bar{a}_{\cdot l}, \bar{a}_{k\cdot}, \bar{a}_{\cdot\cdot}$는 각각 행렬 $\mathbf{A}$의 l번째 열 평균, k번째 행 평균, 모든 구성 원소들의 평균이다.

[4단계] 다음과 같이 행렬 $\mathbf{Z}$의 스펙트럼분해

$$\mathbf{Z} = \mathbf{V\Lambda V}^t \tag{2.6}$$

를 계산한다. 식 (2.6)에서 $\mathbf{\Lambda}$는 $\mathbf{Z}$의 고유값 λ_k, $k = 1,\ldots,n$을 대각원소로 하는 대각행렬이다. 이때 고유값은 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$의 관계를 가진다. $\mathbf{V}$는 고유값에 대응하는 고유벡터들인 $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$으로 열을 구성하고 있는 행렬이며 $\mathbf{V}^t\mathbf{V} = \mathbf{VV}^t = \mathbf{I}$를 만족하는 직교행렬이다.

[5단계] 식 (2.6)으로부터 다음과 같은 행렬을 계산한다. 식 (2.7)은 $\mathbf{Z} = \mathbf{CC}^t$의 성질을 만족한다.

$$\mathbf{C} = \mathbf{V}\lambda^{\frac{1}{2}} = \left[\mathbf{v}_1\lambda_1^{\frac{1}{2}}, \ldots, \mathbf{v}_n\lambda_n^{\frac{1}{2}}\right]. \tag{2.7}$$

식 (2.7)의 행렬 $\mathbf{C}$는 n차원 다차원척도법 그림의 좌표점을 나타낸다. 일반적으로 고전적 척도법 [4단계]에서 구해진 고유값들을 식 $(\sum_{u=1}^{q}\lambda_u / \sum_{u=1}^{n}\lambda_u)\times 100\%$에 대입하여 $q\,(\leq n)$차원 다차원척

도법 그림의 적합도(goodness-of-fit)를 계산할 수 있다. 그리고 적절한 q차원을 선택하여 고유값 $\lambda_1, \ldots, \lambda_q (q \leq n)$과 이에 대응하는 고유벡터를 가지고 크기가 $n \times q$인 행렬 $\mathbf{C}_{(q)} = \mathbf{V}_{(q)} \mathbf{\Lambda}_{(q)}^{1/2} = (\mathbf{v}_1 \lambda_1^{1/2}, \ldots, \mathbf{v}_q \lambda_q^{1/2})$을 계산할 수 있다. 이 식에서 $\mathbf{C}_{(q)}$는 n차원으로부터 차원 축소된 q차원 다차원척도법 그림의 좌표점을 의미하는데 본 연구에서는 시각적으로 해석하기에 용이한 2차원 다차원척도법 그림을 활용하겠다. 다음 장에서는 고전적 척도법을 통해 계산된 $\mathbf{C}$를 활용하여 어떻게 그룹 구조를 파악할 수 있는지 살펴보고자 한다.

3. 그룹화된 비유사성에 대한 비모수적 접근법

3.1. 거리분석

Cox와 Cox (2001)는 Gower와 Krzanowski (1999)가 제안한 거리분석을 그룹화된 비유사성행렬에 적용시켜서 그룹 구조를 파악할 수 있다고 하였다. 이 절에서는 Cox와 Cox (2001) 그리고 Gower와 Krzanowski (1999)를 참고하여 거리분석의 이론을 정리하고자 한다. 거리분석은 비유사성행렬을 이용하여 그룹 내와 그룹 간의 구조를 파악하는 기법이라고 볼 수 있다. 우선 g개의 그룹으로 나누어져 있는 그룹화된 다변량자료행렬 $\mathbf{X}$가 있다고 생각해보자. 계산의 편의상 새로운 비유사성 행렬을 $\widetilde{\mathbf{D}} = (d_{rs}^2(i,j)/2)$, $i = 1, \ldots, n_r$; $j = 1, \ldots, n_s$라 정의하면 식 (2.5)와 $\mathbf{Z} = \mathbf{C}\mathbf{C}^t$의 관계로부터

$$-\mathbf{H}\widetilde{\mathbf{D}}\mathbf{H} = \mathbf{C}\mathbf{C}^t \tag{3.1}$$

를 만족한다. 그리고 $\mathbf{N} = \mathrm{diag}(n_1, \ldots, n_g)$라 하고 크기가 $n \times g$인 행렬 $\mathbf{G} = \left(\mathbf{G}_1^t \ \mathbf{G}_2^t \cdots \mathbf{G}_g^t\right)^t$라 하자. 여기서 크기가 $n_r \times g$인 부분행렬 $\mathbf{G}_r = (g_{ir})$, $i = 1, \ldots, n_r$; $r = 1, \ldots, g$을 다음과 같이 정의하자.

$$g_{ir} = \begin{cases} 1, & i\text{번째 개체가 } r\text{번째 그룹에 속할 경우,} \\ 0, & \text{그 외의 경우.} \end{cases}$$

이때 그룹별 평균좌표는 $\bar{\mathbf{C}} = \mathbf{N}^{-1}\mathbf{G}^t\mathbf{C}$와 같이 계산 가능하므로

$$\bar{\mathbf{C}}\bar{\mathbf{C}}^t = \mathbf{N}^{-1}\mathbf{G}^t\mathbf{C}\mathbf{C}^t\mathbf{G}\mathbf{N}^{-1} \tag{3.2}$$

가 만족한다. 식 (3.2)에서 우변의 $\mathbf{C}\mathbf{C}^t$를 식 (3.1)로 대체하고 정리하면

$$\bar{\mathbf{C}}\bar{\mathbf{C}}^t = -\mathbf{N}^{-1}\mathbf{G}^t\widetilde{\mathbf{D}}\mathbf{G}\mathbf{N}^{-1} + \frac{1}{n}\mathbf{1}_g\mathbf{1}_n^t\widetilde{\mathbf{D}}\mathbf{G}\mathbf{N}^{-1} + \frac{1}{n}\mathbf{N}^{-1}\mathbf{G}^t\widetilde{\mathbf{D}}\mathbf{1}_n\mathbf{1}_g^t - \frac{1}{n^2}\mathbf{1}_g\mathbf{1}_n^t\widetilde{\mathbf{D}}\mathbf{1}_n\mathbf{1}_g^t \tag{3.3}$$

와 같다. 비유사성행렬 $\widetilde{\mathbf{D}}$는 식 (2.4)와 같이 크기가 $n_r \times n_s$인 부분행렬 $\widetilde{\mathbf{D}}_{rs}$, $r, s = 1, \ldots, g$로 이루어지는데 식 (3.3)의 우변에서 $\mathbf{F} = (f_{rs}) = \mathbf{N}^{-1}\mathbf{G}^t\widetilde{\mathbf{D}}\mathbf{G}\mathbf{N}^{-1}$라고 하면 $f_{rs} = \widetilde{\mathbf{D}}_{rs}/n_r n_s$가 성립한다. 그리고 그룹별 평균좌표 $\bar{\mathbf{C}}$에서 r번째 그룹의 평균좌표와 s번째 그룹의 평균좌표 간의 거리를 ψ_{rs}, $r, s = 1, \ldots, g$라 한다면

$$\psi_{rs}^2 = 2f_{rs} - f_{rr} - f_{ss}$$

의 관계가 성립한다. 참고로 식 (3.3) 우변의 나머지 세 항들은 ψ_{rs}의 계산 시에는 생략이 가능한 부분이다. 그리고 그룹 평균 사이의 비유사성으로 구성된 행렬은 $\bar{\mathbf{\Psi}} = (\psi_{rs}^2/2)$, $r, s = 1, \ldots, g$와 같다. 그리고 $\mathbf{n} = (n_1, n_2, \ldots, n_g)^t$라 할 때 $\widetilde{\mathbf{D}}$로부터 거리분석을 통해 그룹의 구조를 다음과 같이 분해할 수 있다.

$$\frac{\mathbf{1}^t\widetilde{\mathbf{D}}\mathbf{1}}{n} = \sum_{r=1}^{g} \frac{\mathbf{1}_r^t\widetilde{\mathbf{D}}_{rr}\mathbf{1}_r}{n_r} + \frac{\mathbf{n}^t\bar{\mathbf{\Psi}}\mathbf{n}}{n}. \tag{3.4}$$

식 (3.4)에서 $\mathbf{1}^t\widetilde{\mathbf{D}}\mathbf{1}/n$은 전체의 비유사성, $\sum_{r=1}^{g}\mathbf{1}_r^t\widetilde{\mathbf{D}}_{rr}\mathbf{1}_r/n_r$은 그룹 내의 비유사성 그리고 $\mathbf{n}^t\bar{\mathbf{\Psi}}\mathbf{n}/n$은 그룹 간의 비유사성을 나타내며 이들의 관계는 3.3절에서 설명하게 될 분산분석의 구조와 매우 닮아 있다. 다음 절에서는 식 (3.4)의 그룹 구조를 활용하여 순열검정을 위한 새로운 검정통계량을 제안하려 한다.

3.2. 순열검정

Gower와 Krzanowski (1999)는 다변량 분산분석이 의존해야 하는 가정들에 제약이 있는 경우는 다변량 분산분석은 적합하지 않기 때문에 자료의 분포를 가정하지 않을 때 쓰는 비모수적 검정 방법인 순열검정을 제안하였다. 이 순열검정은 거리분석에 대한 객관적인 평가를 할 수 있는데 순열검정을 하기 위해서는 먼저 적절한 검정통계량을 선택해야 한다. Clarke (1993)은 그룹화된 유사성행렬로부터 순위를 활용한 검정통계량을 제안하였는데 이는 비계량형 다차원척도법에 가깝다. 본 연구에서는 식 (3.4)를 참고로 하여 새로운 검정통계량 Γ를 다음과 같이 제안하고자 한다.

$$\Gamma = n\left(\mathbf{1}^t\widetilde{\mathbf{D}}\mathbf{1}\right)^{-1}\sum_{r=1}^{g}\frac{\mathbf{1}_r^t\widetilde{\mathbf{D}}_{rr}\mathbf{1}_r}{n_r}. \tag{3.5}$$

식 (3.5)의 검정통계량 Γ는 식 (3.4)의 그룹 구조를 나타내는 식에서 전체의 비유사성에 대하여 그룹 내의 비유사성이 차지하는 비율이라고 볼 수 있다. 이 식은 3.3절에서 설명하게 될 Wilks's Λ와 매우 밀접한 연관이 있는데 이에 대해서는 3.3절에서 좀 더 구체적으로 설명하려 한다. 우선 Manly (2007)를 참고로 하여 식 (3.5)의 Γ를 활용한 순열검정의 절차를 정리해 보면 다음과 같다.

[1단계] 식 (2.1)과 같이 크기가 $n_r\times p$인 부분행렬 $\mathbf{X}_r$, $r=1,\ldots,g$로 이루어진 원자료를 $\mathbf{X}^{(0)}$라 하자. $\mathbf{X}^{(0)}$에 대하여 초기 검정통계량 $\Gamma_{(0)}$를 구한다.

[2단계] 개체들에 대해서 그룹을 무작위로 할당한 후, 그룹별로 재정렬한 자료를 $\mathbf{X}^{(1)}$이라 하고 이에 대하여 검정통계량 $\Gamma_{(1)}$을 구한다.

[3단계] [2단계]를 m번 반복하면서 $\mathbf{X}^{(t)}$에 대한 검정통계량 $\Gamma_{(t)}$, $t=1,\ldots,m$를 구한 후 $\Gamma_{(t)}$에 대한 경험적 분포(empirical distribution)를 히스토그램으로 그린다.

[4단계] 히스토그램에서 $\Gamma_{(0)}$의 위치를 확인하여 이보다 더 극단적인 값의 비율로 p-값을 구한 다음 유의수준 α와 비교하여 그룹 간의 평균에는 차이가 없다는 귀무가설의 기각 여부를 판단한다.

위의 순열검정 절차에 따라 검정을 할 때 적절한 m을 설정하는 것도 중요하다. 예를 들어서 g개의 그룹이 있고 각 그룹별로 개체가 n개씩 있다고 가정해보자. 그러면 무작위로 순열화할 수 있는 개수는 $(gn)!/[(n!)^g g!]$이다. 만약 g와 n이 점점 증가한다면 순열화할 수 있는 개수도 기하급수적으로 많아지게 된다. 따라서 모든 가능한 순열화를 해보는 것은 비현실적이다. Manly (2007)에 따르면 m이 증가할수록 p-값의 정확성은 올라가는데 일반적으로 유의수준 0.05일 때는 적어도 1,000번의 순열화가 필요하고, 유의수준이 0.01일 때는 적어도 5,000번의 순열화가 필요하다고 하였다.

3.3. 거리분석과 다변량 분산분석

2.1절에서 설명한 비유사성의 측도로 식 (2.2)와 같은 유클리드 거리를 사용한다면 거리분석과 다변량 분산분석은 수리적으로 매우 밀접한 연관 관계를 가진다. 이 절에서는 Johnson과 Wichern (2007)의 내용을 참고하여 다변량 분산분석을 설명하고 거리분석과 다변량 분산분석과의 수리적인 연관성을 살펴보고자 한다. 먼저 다변량 분산분석의 기본적인 이론에 대해서 정리해보자. 다변량 분산분석은 크게 일

원배치(one-way) 다변량 분산분석과 이원배치(two-way) 다변량 분산분석으로 나눌 수 있는데 본 연구에서는 일원배치 다변량 분산분석만을 다루도록 하겠다. 우선 $g(>2)$개의 모집단으로부터 g개의 서로 독립인 확률표본이 $\mathbf{x}_{ri} \sim N_p(\boldsymbol{\mu}_r, \boldsymbol{\Sigma})$, $r = 1, \ldots, g$; $i = 1, \ldots, n_r$라고 가정해보자. 이때 다변량 정규분포의 공분산행렬은 모집단에 관계없이 모두 같다고 가정한다. 여기서 $\mathbf{x}_{ri}$는 r번째 그룹에서 i번째 개체의 관측벡터이다. 이것을 다음과 같은 모형으로 표현할 수 있다.

$$\mathbf{x}_{ri} = \boldsymbol{\mu} + \boldsymbol{\tau}_r + \boldsymbol{\epsilon}_{ri}, \quad r = 1, \ldots, g;\ i = 1, \ldots, n_r, \tag{3.6}$$

여기서 $\boldsymbol{\mu}$는 전체평균벡터를 나타내고 $\boldsymbol{\tau}_r$은 $\boldsymbol{\tau}_r = \boldsymbol{\mu}_r - \boldsymbol{\mu}$, $\sum_{r=1}^{g} n_r \boldsymbol{\tau}_r = \mathbf{0}$을 만족하는 r번째 그룹의 그룹효과벡터를 나타낸다. 그리고 $\boldsymbol{\epsilon}_{ri}$는 서로 독립이며 $N_p(\mathbf{0}, \boldsymbol{\Sigma})$를 따르는 확률벡터이다. 그리고 관측벡터 $\mathbf{x}_{ri}$의 변동을 위의 모형에 대응하도록 분해하면 다음과 같다.

$$\mathbf{x}_{ri} = \bar{\mathbf{x}} + (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}) + (\mathbf{x}_{ri} - \bar{\mathbf{x}}_r), \tag{3.7}$$

여기에서 $\bar{\mathbf{x}} = \sum_{r=1}^{g} \sum_{i=1}^{n_r} \mathbf{x}_{ri}/n$, $\bar{\mathbf{x}}_r = \sum_{i=1}^{n_r} \mathbf{x}_{ri}/n_r$을 만족한다. 즉 $\bar{\mathbf{x}}$는 전체표본의 평균벡터, $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}$는 그룹 r의 그룹효과벡터 그리고 $\mathbf{x}_{ri} - \bar{\mathbf{x}}_r$은 그룹 내의 잔차벡터를 나타낸다. 여기서 $\bar{\mathbf{x}}$를 좌변으로 넘기고 양변에 대하여 교차곱을 구한 후 $i = 1, \ldots, n_r$와 $r = 1, \ldots, g$에 대하여 더해주면 제곱합과 교차곱행렬의 분해는 다음과 같다. Table 3.1은 아래의 식을 변동요인별로 요약한 표이다.

$$\sum_{r=1}^{g} \sum_{i=1}^{n_r} (\mathbf{x}_{ri} - \bar{\mathbf{x}})(\mathbf{x}_{ri} - \bar{\mathbf{x}})^t = \sum_{r=1}^{g} n_r (\bar{\mathbf{x}}_r - \bar{\mathbf{x}})(\bar{\mathbf{x}}_r - \bar{\mathbf{x}})^t + \sum_{r=1}^{g} \sum_{i=1}^{n_r} (\mathbf{x}_{ri} - \bar{\mathbf{x}}_r)(\mathbf{x}_{ri} - \bar{\mathbf{x}}_r)^t.$$

이제 식 (3.6)에서 그룹효과벡터들 사이에는 차이가 없다는 귀무가설 $H_0 : \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \cdots = \boldsymbol{\tau}_g = \mathbf{0}$을 검정하기 위하여 식 (3.8)과 같이 그룹 내의 변동량을 측정한 일반화된 분산 $|\mathbf{W}|$와 전체의 변동량을 측정한 일반화된 분산 $|\mathbf{B} + \mathbf{W}|$의 비를 고려한 검정통계량을 사용한다.

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|}. \tag{3.8}$$

식 (3.8)의 검정통계량은 Wilks's Λ라고 부르는데 일반적으로 매우 작은 Λ값에 대해서 귀무가설 H_0을 기각하게 된다.

2.1절에서 개체들 간의 비유사성 측도로 다양한 거리들이 사용될 수 있다고 하였다. 만약 식 (2.2)와 같은 유클리드 거리가 비유사성의 측도로 사용된다면 거리분석을 통해 계산된 식 (3.4)의 그룹구조정보와 Table 3.1의 제곱합과 교차곱행렬들은 매우 밀접한 연관 관계를 갖고 있다. 우선 유클리드 거리와 분산의 기초적인 관계를 살펴보기 위해 임의의 관측값을 x_i, $i = 1, \ldots, n$라 정의하면

$$\frac{1}{n} \sum_{i<j}^{n} (x_i - x_j)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3.9}$$

가 성립한다. 여기서 $\bar{x} = \sum_{i=1}^{n} x_i/n$이다. 식 (3.9)를 보면 개체 간의 비유사성을 유클리드 거리로 계산할 경우 편차의 제곱합으로 귀결될 수 있음을 알 수 있다. 이 관계를 기초로 하여 Gower와 Krzanowski (1999)는 자료의 모든 개체 짝들 간의 비유사성들로 이루어진 행렬을 만들어서 다변량 분산분석에 접근할 수 있다고 하였다. 비유사성과 다변량 분산분석과의 관계는 Table 3.2를 통해서 보다 명확하게 알 수 있다. Table 3.2는 3.1절과 Table 3.1을 토대로 관점에 따른 그룹 구조에 대하여 정리한 것이다.

Table 3.2를 보면 개체들 간의 비유사성 측도로 유클리드 거리를 사용할 경우 거리분석을 통해 계산된 그룹의 구조는 다변량 분산분석을 통해 계산된 제곱합과 교차곱행렬에서 대각원소의 합과 일치한다.

Table 3.1. One-way MANOVA table

Source of variation	Matrix of sum of squares and cross products	Degrees of freedom
Between	$\mathbf{B} = \sum_{r=1}^{g} n_r(\bar{\mathbf{x}}_r - \bar{\mathbf{x}})(\bar{\mathbf{x}}_r - \bar{\mathbf{x}})^t$	$g-1$
Within	$\mathbf{W} = \sum_{r=1}^{g}\sum_{i=1}^{n_r}(\mathbf{x}_{ri} - \bar{\mathbf{x}}_r)(\mathbf{x}_{ri} - \bar{\mathbf{x}}_r)^t$	$\sum_{r=1}^{g} n_r - g$
Total	$\mathbf{B}+\mathbf{W} = \sum_{r=1}^{g}\sum_{i=1}^{n_r}(\mathbf{x}_{ri} - \bar{\mathbf{x}})(\mathbf{x}_{ri} - \bar{\mathbf{x}})^t$	$\sum_{r=1}^{g} n_r - 1$

Table 3.2. Comparison of group structure by viewpoint

Source of variation	Viewpoint of dissimilarity	Viewpoint of total variance
Between	$\mathbf{n}^t\bar{\mathbf{\Psi}}\mathbf{n}/n$	$\mathrm{tr}(\mathbf{B})$
Within	$\sum_{r=1}^{g}\mathbf{1}_r^t\widetilde{\mathbf{D}}_{rr}\mathbf{1}_r/n_r$	$\mathrm{tr}(\mathbf{W})$
Total	$\mathbf{1}^t\widetilde{\mathbf{D}}\mathbf{1}/n$	$\mathrm{tr}(\mathbf{B}+\mathbf{W})$

즉, 변동요인별로 살펴보면 $\mathbf{n}^t\bar{\mathbf{\Psi}}\mathbf{n}/n = \mathrm{tr}(\mathbf{B})$, $\sum_{r=1}^{g}\mathbf{1}_r^t\widetilde{\mathbf{D}}_{rr}\mathbf{1}_r/n_r = \mathrm{tr}(\mathbf{W})$, $\mathbf{1}^t\widetilde{\mathbf{D}}\mathbf{1}/n = \mathrm{tr}(\mathbf{B}+\mathbf{W})$가 성립한다. 이 관계로부터 앞서 제안한 식 (3.5)의 Γ는 $\mathrm{tr}(\mathbf{W})/\mathrm{tr}(\mathbf{B}+\mathbf{W})$와 같아짐을 알 수 있다. 이 식은 총분산의 비를 의미하며 일반화된 분산의 비인 식 (3.8)과 형태가 매우 비슷하다. 따라서 그룹 간의 차이 검정 시 제곱합과 교차곱행렬로부터 두 가지의 방법으로 검정이 가능하다. 하나는 일반적인 다변량 분산분석을 통해 Wilks's Λ를 계산하여 검정하는 방법이고, 또 하나는 Γ를 계산하여 순열검정을 하는 방법이다. 특히 Γ를 활용한 순열검정은 다변량 분산분석을 수행할 수 없는 다변량자료에 대해서도 그룹 간 차이 검정이 가능하다는 장점을 가지고 있다. 이 두 가지 검정통계량의 검정력에 대해서는 4장에서 좀 더 구체적으로 다룰 것이다.

4. 모의실험

이 장에서는 모의실험을 통해 식 (3.5)의 Γ와 식 (3.8)의 Wilks's Λ 간 검정력을 비교하려 한다. Table 4.1은 여섯 가지 경우에 대한 모의실험 설정 및 제1종 오류율을 나타낸다. 여섯 가지 경우 각각에 대하여 다변량 정규분포를 따르는 난수를 1,000번 반복 생성하였다. 난수 생성 시 그룹별 평균벡터는 모두 $(0,0,0,0)^t$으로 같게 설정하였고 공분산행렬의 경우 다변량 분산분석의 가정 중 하나인 공분산행렬의 동질성을 만족시키기 위하여 식 (4.1)과 같은 공분산행렬

$$\mathbf{\Sigma} = \begin{bmatrix} 1.000 & 0.500 & 0.250 & 0.125 \\ 0.500 & 1.000 & 0.500 & 0.250 \\ 0.250 & 0.500 & 1.000 & 0.500 \\ 0.125 & 0.250 & 0.500 & 1.000 \end{bmatrix} \tag{4.1}$$

을 모의실험의 모든 경우에 고정하여 사용하였다. 현실적으로 다변량자료의 변수들은 서로 단위가 다른 경우가 많기 때문에 자료에 대하여 표준화를 시켜야 한다. 따라서 공분산행렬의 대각원소들은 1로 통일을 시켰다. 또한 변수들 간에는 완전히 독립되어 있다기보다는 어느 정도 상관이 되어 있는데 가까이 있는 변수들 간에는 상대적으로 강한 상관이 있도록 공분산행렬에 상관 패턴을 적용하였다. 그리고 한 번 생성한 난수에 대하여 Γ와 Wilks's Λ의 p-값을 구할 때 유의수준 $\alpha = 0.05$로 하였다. Γ의 경우는 계산 시 비유사성의 측도로 유클리드 거리를 사용하였으며 순열검정을 해야 한다. 이때 3.2절의 순열검정 절

Table 4.1. Settings for simulation study and type I error rates

Settings	(a)	(b)	(c)	(d)	(e)	(f)
Number of groups	2	2	2	3	3	3
Number of observations	30	50	70	30	50	70
Number of variables	4	4	4	4	4	4
Γ	0.058	0.047	0.059	0.051	0.042	0.043
Wilks's Λ	0.055	0.045	0.049	0.046	0.046	0.040

Table 4.2. The statistical powers of Γ and Wilks's Λ

	c	1	2	3	4	5	6	7	8	9	10
(a)	Γ	0.092	0.183	0.336	0.581	0.753	0.879	0.949	0.987	0.998	0.999
	Wilks's Λ	0.071	0.125	0.221	0.370	0.559	0.738	0.870	0.941	0.981	0.995
(b)	Γ	0.091	0.232	0.517	0.741	0.921	0.979	0.995	1.000	1.000	1.000
	Wilks's Λ	0.066	0.146	0.327	0.571	0.761	0.925	0.982	0.995	1.000	1.000
(c)	Γ	0.120	0.311	0.645	0.892	0.980	0.996	0.999	1.000	1.000	1.000
	Wilks's Λ	0.089	0.210	0.449	0.749	0.928	0.990	0.998	1.000	1.000	1.000
(d)	Γ	0.062	0.125	0.247	0.436	0.629	0.797	0.897	0.970	0.994	0.998
	Wilks's Λ	0.058	0.078	0.145	0.279	0.430	0.602	0.751	0.882	0.950	0.985
(e)	Γ	0.079	0.181	0.420	0.680	0.849	0.962	0.993	1.000	1.000	1.000
	Wilks's Λ	0.063	0.112	0.243	0.447	0.695	0.874	0.958	0.992	1.000	1.000
(f)	Γ	0.086	0.259	0.559	0.832	0.953	0.993	1.000	1.000	1.000	1.000
	Wilks's Λ	0.062	0.149	0.368	0.648	0.868	0.963	0.995	0.999	1.000	1.000

차에 따라 $m = 1000$으로 설정하여 경험적 분포를 구하고 이를 통해 초기 Γ의 p-값을 구한다. 이런 과정을 1,000번 생성되는 난수 각각에 대하여 적용한 후 두 검정통계량에 대한 제1종 오류율을 계산하였다. Table 4.1을 보면 전체적으로 제1종 오류율이 0.050내외인 것을 확인할 수 있다. 이는 두 검정통계량으로 그룹 간의 차이를 검정하는 것이 어느 정도 타당한 검정이라고 볼 수 있다.

이제 두 검정통계량의 검정력을 비교하기 위하여 Table 4.1의 설정 (a)–(c)과 같이 그룹 수가 두 개인 경우에 대해서는 첫 번째 그룹은 평균벡터를 $\boldsymbol{\mu}_1 = (0,0,0,0)^t$로 고정하고, 두 번째 그룹의 평균벡터는 $\boldsymbol{\mu}_2^{(c)} = 0.1\times c\times \mathbf{1}_4$, $c = 1,\ldots,10$과 같이 평균벡터의 각 원소가 0.1씩 증가하도록 변화를 주면서 $c = 1,\ldots,10$까지 총 10번에 걸쳐 두 그룹 간의 차이 검정을 하였다. Table 4.1의 설정 (d)∼(f)과 같이 그룹 수가 세 개인 경우에 대해서는 두 그룹인 경우와 마찬가지로 첫 번째 그룹은 평균벡터를 $\boldsymbol{\mu}_1 = (0,0,0,0)^t$로 고정하였고, 나머지 두 그룹의 평균벡터는 $\boldsymbol{\mu}_2^{(c)} = -0.05\times c\times \mathbf{1}_4$, $\boldsymbol{\mu}_3^{(c)} = 0.05\times c\times \mathbf{1}_4$, $c = 1,\ldots,10$과 같이 평균벡터의 각 원소가 ± 0.05씩 증감하도록 변화를 주면서 $c = 1,\ldots,10$까지 총 10번에 걸쳐 세 그룹 간의 차이를 검정하였다. 여섯 가지 설정에 대하여 평균벡터의 변화에 따른 모의실험결과는 Table 4.2와 Figure 4.1과 같다. 결과를 보면 전반적으로 평균의 차이가 증가할수록, 그룹별 개체 수가 증가할수록 검정력이 강해짐을 알 수 있다. 그리고 모든 경우에 대하여 Γ의 검정력이 Wilks's Λ의 검정력보다 근소한 차이로 더 강하다는 것을 알 수 있다. 또한 변수의 개수가 개체수보다 더 큰 자료와 같이 다변량 분산분석을 수행할 수 없는 상황에 대하여 모의실험을 통해 Γ의 제1종 오류율과 검정력을 확인해본 결과 앞선 모의실험의 결과와 유사함을 알 수 있었다. 따라서 다변량 분산분석의 기본 가정을 만족하지 못하는 자료일 경우에는 Γ를 활용한 순열검정을 통해서 그룹 간의 차이에 대한 검정을 생각해볼 수 있다.

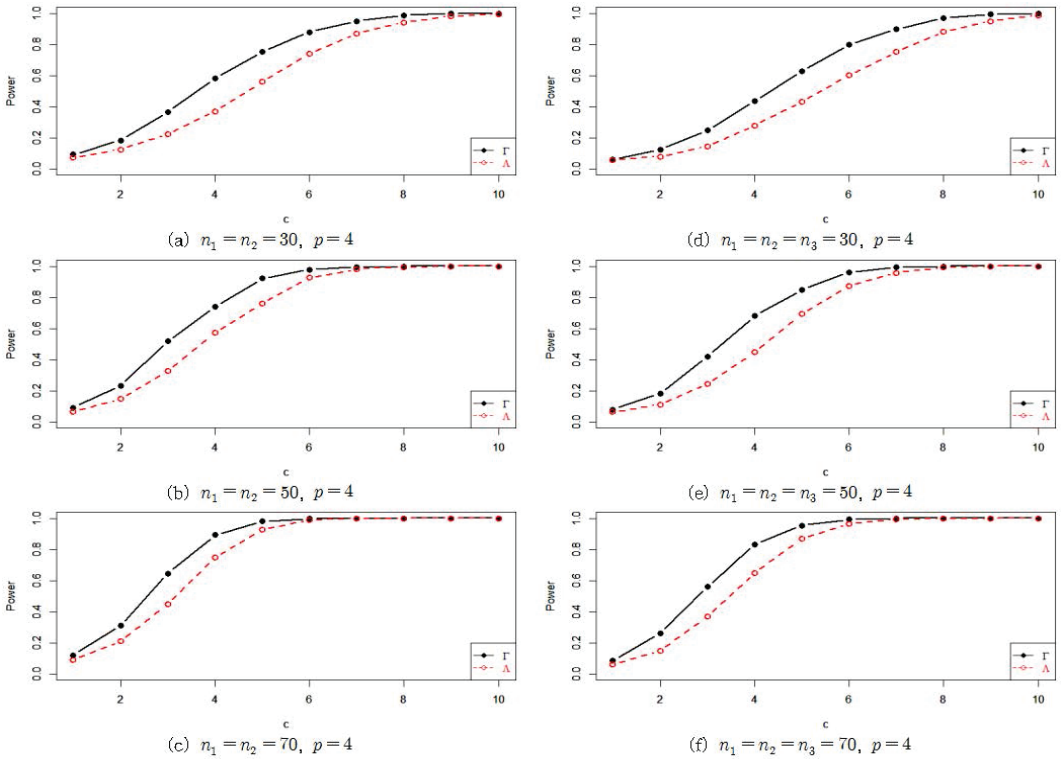

Figure 4.1. Comparison of the statistical powers between Γ and Wilks's Λ. (a)–(c) are comparisons between two groups, (d)–(f) are comparisons between three groups.

5. 활용사례

브라질월드컵 결승 진출국 선수 자료는 축구통계전문 사이트인 whoscored.com에서 제공하는 자료로 한국 기준으로 2014년 6월 13일부터 7월 14일까지 브라질에서 진행된 제 20회 FIFA월드컵축구대회 결승전 진출국인 독일과 아르헨티나의 선수 자료이다. 출전경기 수와 출전시간을 기준으로 골키퍼를 제외한 수비수, 미드필더, 공격수 각 포지션별로 8명씩 선별하였고, 경기력에 영향을 줄 수 있는 27개의 변수들을 포함하여 그 크기가 24×27인 그룹화된 다변량자료로 가공하였다. 먼저 변수들의 단위가 서로 다르므로 표준화를 시킨 후 식 (2.3)의 시티-블록 거리를 비유사성 측도로 활용하여 다차원척도법을 적용시키면 Figure 5.1을 얻을 수 있다. 이 2차원 다차원척도법 그림의 적합도는 57.38%로 높은 적합도라고 보기는 어렵지만 해석에는 큰 문제가 없다. Figure 5.1에서 제1축은 적합도가 54.72%로 이 축을 기준으로 우측에 군집되어 있는 선수들은 공격적 성향이 강한 선수들이며 특히 가장 우측에 모여 있는 Messi, Mueller, Kroos, Di Maria, Oezil과 같은 선수들은 공격력이 탁월한 선수들이라는 사실을 알 수 있다. 제1축을 기준으로 좌측에 군집되어 있는 선수들은 수비적 성향이 강한 선수들임을 알 수 있다. Kroos나 Oezil과 같이 포지션이 미드필더인 선수들 중 절반가량이 우측에 포진되어 있는데 이는 공격적 성향이 강한 미드필더들임을 말해준다. 반대로 나머지 절반가량은 좌측에 포진되어 있는데 수비 쪽 역할을 수행한 수비형 미드필더들이 주를 이루고 있다. Schweinsteiger나 Mascherano는 미드필더이지만 거의 수비적으로 핵심 역할을 수행한 선수임을 짐작할 수 있다. 제2축의 경우는 적합도가 2.66%로 매

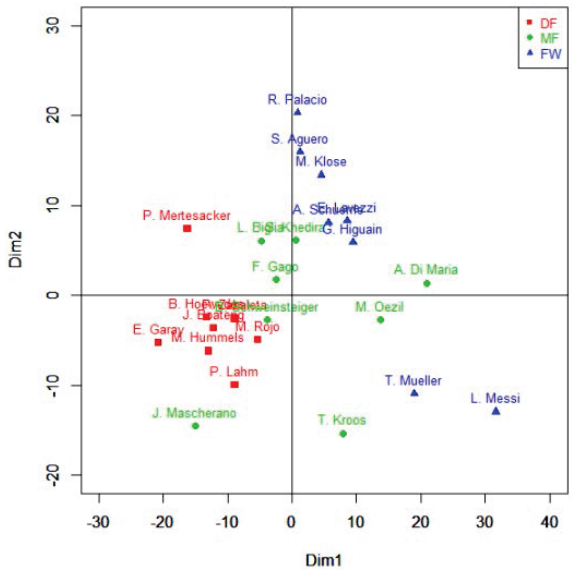

Figure 5.1. Two-dimensional MDS map for World Cup final data.

Table 5.1. Group structure for World Cup final data

Total	Within	Between
9968.870	6572.168	3396.702

Table 5.2. Results of permutation test for World Cup final data

Γ	Minimum value	Maximum value	p-value
0.659	0.762	0.981	0.000

우 낮아서 해석상 큰 의미가 없다고 볼 수 있다.

이 자료는 각 그룹별 부분행렬들의 크기가 8×27행렬로 개체 수보다 변수 수가 더 많으므로 이들의 공분산행렬은 비정칙 행렬이 되고 다변량 정규성 검정이나 공분산행렬의 동질성 검정을 할 수가 없다. 따라서 다변량 분산분석은 적합하지 않은 자료이다. 이 경우 그룹화된 비유사성에 거리분석과 순열검정을 적용하여 그룹 간에 유의미한 차이를 검정할 수 있다. 거리분석을 통해 계산된 그룹 구조는 Table 5.1과 같고 이로부터 검정통계량을 계산하면 $\Gamma = 0.659$이며 순열검정을 통해 산출된 경험적 분포의 요약결과는 Table 5.2와 같다. Table 5.2를 보면 경험적 분포의 최소값은 0.762로 0.659보다 더 작은 값은 존재하지 않는다. 따라서 p-값은 0.000이고 그룹별 평균벡터의 차이가 없다는 귀무가설을 기각할 수 있다.

이 절에서 살펴본 검정방법은 다변량 분산분석을 적용하기 어려운 자료임에도 불구하고 다양한 거리를 활용한 비유사성으로부터 다차원척도법, 거리분석 그리고 순열검정을 차례로 적용하여 그룹 간의 차이를 검정할 수 있다는 점에서 매우 유용한 방법이라고 볼 수 있겠다.

6. 결론

그룹화된 다변량자료는 일반적으로 다변량 분산분석을 통해 그룹 간 차이를 검정할 수 있다. 하지만 현실적으로는 다변량 분산분석의 기본 가정을 만족하지 않는 자료들이 많다. 이에 그룹화된 다변량자료로

부터 다양한 거리를 활용하여 비유사성을 계산한 후 비모수적으로 접근하면 그룹 간의 차이를 검정할 수 있다. 본 연구에서는 비유사성에 다차원척도법과 거리분석을 차례로 적용하여 그룹 구조를 분해하였다. 그리고 그룹 구조정보로부터 새로운 검정통계량을 제안하였고 이를 활용한 순열검정을 통해 그룹 간의 차이를 검정하였다. 또한 비유사성의 측도로 유클리드 거리를 활용하여 거리분석과 다변량 분산분석의 수리적 연관성을 고찰하였고 새롭게 제안한 검정통계량과 기존의 검정통계량과의 검정력을 비교하였다.

모의실험결과 새롭게 제안한 검정통계량은 기존의 검정통계량에 비견할 정도의 검정력이 있음을 확인하였다. 또한 활용사례를 통해 유클리드 거리뿐만 아니라 다양한 거리를 활용하여도 그룹 구조를 파악할 수 있음을 알 수 있었고 특히 다변량 분산분석을 수행하기 어려운 자료라 할지라도 비모수적으로 접근하여 그룹 간의 차이를 검정할 수 있다는 장점도 확인하였다.

본 연구에서는 계량형 다차원척도법에 국한된 알고리즘으로부터 출발하여 비모수적 검정으로 귀결되었고 기존의 모수적 검정법과도 비교를 하였다. 차후에는 비유사성이 비계량형 다차원척도법의 알고리즘을 따를 경우 기존의 어떤 기법과 연계될 수 있을지에 관한 연구가 더 필요하다고 생각된다.

References

Choi, Y. S. (2014). *Walk in Multidimensional Scaling*, Free Academy, Kyungki.

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure, *Australian Journal of Ecology*, **18**, 117–143.

Cox, T. F. and Cox, M. A. A. (2001). *Multidimensional Scaling*, Chapman & Hall/CRC, London.

Gower, J. C. and Krzanowski, W. J. (1999). Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance, *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **48**, 505–519.

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*, Prentice Hall, New Jersey.

Kruskal, J. B. and Wish, M. (1978). *Multidimensional Scaling*, Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-011, Sage Publications, Beverly Hills and London.

Manly, B. F. J. (2007). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, Chapman & Hall/CRC, London.

Torgerson, W. S. (1958). *Theory and Methods of Scaling*, Wiley, New York.

다차원척도법과 거리분석을 활용한 그룹화된 비유사성에 대한 비모수적 접근법

남승찬[a] · 최용석[a,1]

[a]부산대학교 통계학과

(2017년 5월 8일 접수, 2017년 6월 29일 수정, 2017년 7월 2일 채택)

요 약

일반적으로 그룹화된 다변량자료는 다변량 분산분석(multivariate analysis of variance; MANOVA)을 사용하여 그룹 간 차이를 검정할 수 있다. 그러나 만약 다변량 분산분석의 기본적인 가정이 위배되면 이 방법은 적절하지 못하다. 이 경우 다양한 거리로부터 그룹화된 비유사성을 계산한 후 다차원척도법(multidimensional scaling; MDS), 거리분석(analysis of distance; AOD) 그리고 비모수적 기법인 순열검정(permutation test)을 적용하여 문제를 해결할 수 있다. 다차원척도법은 비유사성으로부터 개체들의 좌표를 계산해주며 거리분석은 이 좌표를 활용하여 그룹 구조를 파악하는데 유용하다. 특히 비유사성의 측도로 유클리드 거리를 사용하면 거리분석은 다변량 분산분석과 수리적으로 매우 밀접한 연관관계를 맺는다. 따라서 본 연구에서는 그룹화된 비유사성에 다차원척도법과 거리분석을 적용하여 그룹 내와 그룹 간의 구조를 파악하고 순열검정을 위한 새로운 검정통계량을 제안하려 한다. 덧붙여 유클리드 거리를 활용한 비유사성을 통해 거리분석과 다변량 분산분석과의 수리적 연관성을 고찰하고자 한다.

주요용어: 비유사성, 다차원척도법, 거리분석, 순열검정, 다변량 분산분석

[1]교신저자: (46241) 부산광역시 금정구 부산대학로 63번길 2, 부산대학교 통계학과.
E-mail: yschoi@pusan.ac.kr