The Korean Journal of Applied Statistics
2022, Vol. 35, No. 5, 579–591

DOI: https://doi.org/10.5351/KJAS.2022.35.5.579
Print ISSN1225-066X / Online ISSN 2383-5818

Application of functional ANOVA and functional MANOVA

Mijeong Kim[1,a]

[a]Department of Statistics, Ewha Womans University

Abstract

Functional data is collected in various fields. It is often necessary to test whether there are differences among groups of functional data. In this case, it is not appropriate to explain using the point-wise ANOVA method, and we should present not the point-wise result but the integrated result. Various studies on functional data analysis of variance have been proposed, and recently implemented those methods in the package `fdANOVA` of R. In this paper, I first explain ANOVA and multivariate ANOVA, then I will introduce various methods of analysis of variance for univariate and multivariate functional data recently proposed. I also describe how to use the R package `fdANOVA`. This package is used to test equality of weekly temperatures in Seoul and Busan through univariate functional data ANOVA, and to test equality of multivariate functional data corresponding to handwritten images using multivariate function data ANOVA.

Keywords: ANOVA, MANOVA, functional ANOVA, functional data, functional MANOVA

1. 서론

곡선이나 함수로 표현되는 관측치, 즉 함수 데이터는 다양한 분야에서 수집된다. 예를 들면, 어느 지역에서 1년 동안 매 달 측정된 기온은 12개의 원소를 가진 벡터로 표현될 수 있다. 벡터의 각 원소에 해당되는 값은 시간 순서에 따라 측정된 값이므로 연관성이 높을 것이고, 인접 지역의 1년 동안의 기온에 해당하는 데이터와 높은 연관성을 가질 가능성이 높다. 이러한 시간적 공간적 연관성을 무시할 수 없는 특성을 가진 함수 데이터가 집단 별로 차이가 있는지에 대한 분석 방법에 대한 연구가 이루어져왔다. 예를 들어, 특정 질병을 가진 사람들이 정상인 사람들과 신체의 몇몇 부위에서 측정한 값들이 차이가 있는지 연구를 진행해야 한다면, 단변량 분산분석(Analysis of Variance; ANOVA)을 적용하는 것은 적합하지 않다. 또한, 때로는 좀 더 높은 차원의 함수 데이터의 집합을 비교해야할 경우도 있다. 예를 들면, A 지역의 1년 동안 매 주 측정된 기온, 습도, 강수량이 B 지역에서 측정된 값과 차이가 있는지 확인하는 경우, 3개의 함수 데이터가 합쳐진, 다변량 함수 데이터가 되기 때문에 문제는 좀 더 어려워진다. 이 경우에도 다변량 분산분석(Multivariate analysis of variance; MANOVA)를 수행하기에는 적합하지 않다.

단순히 생각하면, 단변량 함수의 집합 간 비교는 각 점에서 ANOVA를 통해 구하면 될 것이라고 생각할 수 있다. 이러한 관점에서 Ramsay와 Silverman (2005)은 점 별(point-wise) $F(t)$ 통계량을 제안하였으나, 이 방법은 점별로 검정 통계량을 제시하기 때문에 통합된 결론을 내릴 수 없는 단점이 있다. 또한 ANOVA는 정규성과 등분산성의 가정을 하기 때문에, 함수 데이터의 각각의 값에서 이러한 가정이 성립하기 어려울 때도

This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korean Government (NRF-2020R1F1A1A01074157).

[1] Department of Statistics, Ewha Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. E-mail: m.kim@ewha.ac.kr

Published 31 October 2022 / journal homepage: http://kjas.or.kr

많을 것이다. 이러한 단점을 보완하고자, Faraway (1997), Cuevas 등 (2004), Zhang과 Chen (2007)은 각 점에서 구한 값들을 적분하는 방식으로 하나의 검정 통계량을 제시하였고, 또한 bootstrap 방법을 이용하여 정규성 가정에 위배되더라도 이용 가능한 검정 통계량을 제시하였다. Shen과 Faraway (2004), Zhang (2011)은 정규성 가정 하에서 이용 가능하지만, 좀 더 bias가 보정된 분산분석 검정 통계량을 제시하였다. 때로는 함수 데이터의 차원이 클 경우, spline 등의 방법을 이용하여 차원 축소를 해야하는 경우가 종종 발생하는데, 이런 경우 적용 가능한 검정 통계량을 Górecki와 Smaga (2015)이 제시하였다. Górecki와 Smaga (2017)은 다변량 함수 데이터에 적용 가능한 분산분석 검정 통계량을 제시하였다. Górecki와 Smaga (2019)는 단변량 및 다변량 함수 데이터에 적용가능한 검정 통계량을 소프트웨어 R에서 구할 수 있도록 `fdANOVA` 패키지에 함수로 구현하였다. 단변량 및 다변량 함수 데이터의 분산분석을 설명하기위 위해 2장에서는 함수 데이터가 아닌 하나의 값을 가진 종속 변수에 대한 단변량 분산분석 및 다변량 분산분석 방법을 설명하였다. 3장에서는 단변량 및 다변량 함수 데이터의 분산분석 방법을 설명하였다. 4장에서는 R 패키지 `fdANOVA`를 이용하여 서울의 25개 구의 주별 기온과 부산 15개 구와 1개의 군의 주별 기온이 집단 별로 차이가 있는지 단변량 함수 분산분석을 수행하고, 손글씨 데이터 중 숫자 0과 1에 해당하는 이미지를 다변량 함수로 변환하여 다변량 함수 분산분석을 수행하였다. 5장에서는 이 논문의 전반적인 내용을 요약 및 정리하였다.

2. 분산분석

이 장에서는 단변량 분산분석과 다변량 분산분석의 기본 가정과 검정 통계량에 대해서 알아보도록 한다.

2.1. 단변량 분산분석 (ANOVA)

$l(>2)$개의 집단의 평균을 비교하고자 할 때, 다음과 같은 가정 하에서 분산 분석을 수행할 수 있다.

1. 정규성 가정: 각 모집단에서 종속 변수는 정규 분포를 따른다.
2. 분산의 동질성 가정: 종속 변수의 분산은 모든 집단에서 동일하다.
3. 독립성 가정: 각각의 모집단에서 표본들은 독립적으로 수집된다.

귀무 가설은 다음과 같이 각 집단의 평균이 같다고 설정한다. μ_i는 i ($i = 1, \ldots, l$)번째 집단의 평균이다.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_l. \tag{2.1}$$

대립가설은 '적어도 하나의 집단이 다른 집단의 평균과 다르다'이다. 종속 변수가 단변량이고, 요인이 한 개인 경우 일원분산분석(one-way ANOVA)을 이용하고, 요인이 두 개인 경우 이원 분산 분석(two-way ANOVA)을 이용한다. 예를 들어, 근무연수에 따라 직업 만족도의 차이가 있는지 검정하는 경우는 일원분산분석, 근무연수와 작업환경에 따른 직업만족도 차이를 검정 하는 경우에는 이원분산분석을 이용한다. 요인이 두 개 이상인 경우, 각각의 요인으로 인한 종속 변수의 차이 뿐만 아니라 요인들 간의 교호작용으로 인한 종속 변수의 차이도 검정하기도 한다. ANOVA에서 이용하는 검정 통계량은 집단 간의 변동 대비 집단 내의 변동의 비율로 계산하는데, 이러한 통계량을 ANOVA-type 검정 통계량(ATS)이라고 한다. 데이터가 정규분포를 따를 때, ATS는 F-분포를 따른다. ATS가 비교적 큰 값을 가질 때 집단 간의 차이가 크다고 판단하여 귀무가설을 기각한다. 다변량이나 함수형으로 확장한 경우에도 ATS를 이용하거나 ATS를 약간 변형하여 사용하기도 한다. R에서 함수 aov를 이용하여 단변량 분산분석을 수행할 수 있다.

2.2. 다변량 분산분석

다변량 분산 분석(MANOVA)은 종속 변수가 두 개 이상인 경우 집단간 평균 벡터를 비교하는 방법이다. 종속변수들 간의 상관성이 높은 경우 주로 이용한다. 정규성 가정과 등분산 가정이 적합할 때와 그렇지 않은

Table 1: MANOVA

Source	Matrix	Degree of freedom
Between sum of squares	$\boldsymbol{B}_T$	$l-1$
Within sum of squares	$\boldsymbol{W}$	$n-l$
Total	$\boldsymbol{T}$	$n-1$

경우로 나누어서 설명하고자 한다.

2.2.1. 정규성 가정과 등분산 가정이 적합할 때

다음과 같은 가정을 한다.

1. 정규성 가정: 모든 종속 변수들은 다변량 정규 분포를 따른다.
2. 분산의 동질성 가정: 각 집단의 공분산 행렬은 동일하다.
3. 독립성 가정: 각각의 모집단에서 표본들은 독립적으로 수집된다.

모든 종속 변수가 다변량 정규 분포를 따르는 데이터는 일반적으로 얻기 어려우므로, 변수 변환을 통해 정규 분포에 가까운 형태로 만들어주는 작업을 하기도 하나, 고차원 데이터의 경우에는 변수 변환이 어려울 수 있다.

귀무가설은 '모든 집단의 평균 벡터가 같다'이고, 대립가설은 '적어도 하나의 집단은 다른 평균 벡터 값을 갖는다'이다. 예를 들어, 세 가지 학습 방법에 따라 학생들의 국어, 수학, 영어, 과학의 학업 성취도의 차이가 있는지 확인하려면 다변량 분산분석을 이용할 수 있다.

다변량 분석을 시행하지 않고, 각각의 과목에 대해서 일원분산분석을 할 경우, 다중 검정으로 인한 오류율이 증가하는 문제가 생긴다. 또한 종속 변수의 수가 관측치의 수에 비해 상대적으로 클 경우에 다중공성선 문제가 발생할 수 있기 때문에, 종속 변수가 많다면 충분한 관측치가 확보되어야 한다.

분산분석표는 Table 1과 같다. $\boldsymbol{B}_T$는 집단 간의 변동, $\boldsymbol{W}$는 집단 내의 변동이고 $\boldsymbol{T}$는 전체 변동에 해당하는 행렬이다. l은 집단의 수, n은 관측치의 수이다. 행렬 계산 방법은 Don (2018)을 참고하기 바란다. 다음과 같은 검정통계량을 이용하여 가설 검정을 한다.

1. Wilk's Lambda: $\Lambda = \frac{|\boldsymbol{W}|}{|\boldsymbol{B}_T+\boldsymbol{W}|}$.
2. Hotelling-Lawley Trace: $U = \mathrm{tr}(\boldsymbol{B}_T\boldsymbol{W}^{-1}) = \mathrm{tr}(\boldsymbol{W}^{-1}\boldsymbol{B}_T)$.
3. Pillai's Trace: $\boldsymbol{V} = \mathrm{tr}(\boldsymbol{B}_T\boldsymbol{T}^{-1}) = \mathrm{tr}(\boldsymbol{T}^{-1}\boldsymbol{B}_T)$.
4. Roy's Maximum Root: Largest eigenvalue of $\boldsymbol{B}_T\boldsymbol{W}^{-1}$.

종속 변수가 다변량 정규분포를 따른다는 가정하에 귀무가설 하에서 각각의 통계량을 이용한 값이 근사적으로 χ^2, F와 같은 특정 분포를 따름을 이용하여 만든 검정 방법이다. 어느 검정 통계량이 절대적으로 우월하다고는 할 수는 없으므로, 분석시 여러 가지 방법을 수행하고 비교하는 것이 좋을 것이다. 다변량 분산분석 검정통계량의 근사적 분포에 대한 자세한 내용은 Olive (2017)의 Chapter 10을 참고하기 바란다. R에서 함수 manova를 이용하여 다변량 분산분석을 수행할 수 있다.

2.2.2. 정규성 가정과 등분산성 가정이 적합하지 않을 때

다변량 분산분석에서는 정규성 가정과 등분산성 가정을 적용하기 어려운 경우가 많다. Srivastava와 Kubokawa (2013)은 정규성 가정을 만족하지 않는 경우에 적용 가능한 다변량 분산분석 방법을 제안하였다. Friedrich와

Pauly (2018)는 Srivastava와 Kubokawa (2013)이 제시한 방법이 데이터의 singularity 문제가 있는 경우 이용하기 어려운 점을 지적하고, singularity에도 강건한(robust) 방법을 제안하였다. 이 절에서는 Friedrich와 Pauly (2018)가 제안한 방법에 대해서 살펴보고자 한다. i번째 집단 ($i = 1, \ldots, l$), i번째 집단 내 k번째 ($k = 1, \ldots, n_i$) 개체, s번째 처리 ($s = 1, \ldots, d$)에 속하는 관측치 X_{iks}에 대해서 다음과 같은 모형을 고려한다.

$$\boldsymbol{X}_{ik} = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_{ik}. \tag{2.2}$$

주어진 집단 i, 개체 k에 대해서, $\boldsymbol{X}_{ik} = (X_{ik1}, \ldots, X_{ikd})$이고, d개의 처리에 대한 i번째 집단의 평균은 $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{id})$이고, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^{\mathrm{T}}, \ldots, \boldsymbol{\mu}_l^{\mathrm{T}})^{\mathrm{T}}$이다. 오차항 $\boldsymbol{\epsilon}_{ik}$는 독립이고 같은 분포를 따르는 d-차원 확률 변수이며, 다음과 같은 가정을 한다.

1. $E(\boldsymbol{\epsilon}_{i1}) = 0, \quad i \in \{1, \ldots, l\}$.
2. $0 < \sigma_{is}^2 = \mathrm{var}(X_{iks}) < \infty, \quad i \in \{1, \ldots, l\}, \quad s \in \{1, \ldots, d\}$.
3. $\mathrm{cov}(\boldsymbol{\epsilon}_{i1}) = \boldsymbol{V}_i \geq 0, \quad i \in \{1, \ldots, l\}$.

위의 가정으로부터 데이터에 대한 정규성 가정과 등분산성 가정을 하지 않았음을 알 수 있다. 계산 과정에서 공분산 행렬이 singularity를 가질 때에는 ATS의 계산이 불가능하므로 이러한 점을 개선하기 위해 Modified ANOVA-type 통계량(MATS)을 제시하였다. 적절한 대비 행렬(contrast matrix) $\boldsymbol{H}$과 사영 행렬(projection matrix) $\boldsymbol{P} = \boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}})^{+}\boldsymbol{H}$를 이용하여, 가설 검정 $H_0 : \boldsymbol{P}\boldsymbol{\mu} = \boldsymbol{0}$을 수행한다. 이 때, $(\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}})^{+}$는 $\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}$의 Moore–Penrose inverse이다. MATS는 다음과 같다.

$$Q_N = N\bar{\boldsymbol{X}}_{\cdot}^{\mathrm{T}}\boldsymbol{P}\left(\boldsymbol{P}\widehat{\boldsymbol{D}_N}\right)^{+}\boldsymbol{P}\bar{\boldsymbol{X}}. \tag{2.3}$$

이 때, 총 관측치 $N = n_1 + \cdots + n_l$이고, i 번째 집단의 공변량의 추정치는 $\widehat{\boldsymbol{D}_N} = \mathrm{diag}(N/n_i \cdot \widehat{\sigma_{is}^2})$로 계산된다. 주어진 집단 i에 대해서 합동 집단 평균(pooled group means) 벡터는 $\bar{\boldsymbol{X}}_{i\cdot} = (\boldsymbol{X}_{i1} + \cdots + \boldsymbol{X}_{in_i})/n_i$이고, $\bar{\boldsymbol{X}}_{\cdot} = (\bar{X}_{1\cdot}^{\mathrm{T}}, \ldots, \bar{X}_{l\cdot}^{\mathrm{T}})^{\mathrm{T}}$이다. 데이터에 대한 특정 분포 가정을 하지 않았기 때문에 검정 통계량에 대한 유의 수준 α에 해당하는 적절한 값을 찾기 위해 bootstrap 방법을 이용하여 Q_N의 $(1-\alpha)$ 분위수(quantile)을 구하는 방법을 제시하였다. 데이터가 정규성 가정을 만족하지 않는 경우, R에서 `MANOVA.RM` 패키지에 있는 함수 MANOVA, MANOVA.wide을 이용할 수 있고, 정규성 가정과 등분산성 가정을 충족시키지 못한 경우 `MANOVA.RM` 패키지의 MANOVARM 함수를 이용해서 다변량 분산분석을 수행할 수 있다. 이 때 bootstrap resampling 방법과 횟수를 지정할 수 있다.

3. 함수 분산분석

이 장에서는 단변량 함수 분산분석과 다변량 함수 분산분석의 기본 가정과 여러 가지 검정 통계량에 대해서 알아보도록 한다.

3.1. 단변량 함수 분산분석

단변량 함수 데이터가 집단 간 차이가 있는지 검정하는 방법을 설명하고자 한다. 함수 데이터는 시간 또는 공간 t에 따라 관측될 수 있다. i 번째 집단 ($i = 1, \ldots, l$), 집단 내의 관측치 j ($j = 1, \ldots, n_i$)에 해당하는 함수 데이터를 독립 랜덤 함수 $X_{ij}(t)$로 표기한다. 이 때, t는 닫혀있고 유계(closed and bounded)인 $I = [a, b]$에 속한다. $X_{ij}(t)$는 평균 함수(mean function) $\mu_i(t)$와 분산 함수(covariance function) $\gamma(s, t)$인 확률 과정(stochastic processes)을 따르는 것으로 가정한다.

집단 간의 차이가 있는지 검정하기 위한 귀무 가설은 다음과 같다.

$$H_0 : \mu_1(t) = \cdots = \mu_l(t), \quad t \in I. \tag{3.1}$$

즉, 귀무 가설은 'l개의 집단의 평균 함수가 모두 같다'이고, 대립가설은 그 반대인 'l개의 집단의 평균 함수 중 적어도 하나는 같지 않다'이다. 이러한 가설 검정은 함수 데이터에 대한 일원 배치 분산 분석(the one-way analysis of variance for functional data; FANOVA)이라고 불린다.
FANOVA에 대해서는 다음과 같은 방법이 제시되었다.

1. 점별(point-wise) $F(t)$ 통계량 (Ramsay와 Silverman, 2005)
   모든 점 $t \in I$에서 다음과 같은 검정 통계량 $F_n(t)$를 계산한다.

$$F_n(t) = \frac{\mathrm{SSR}_n(t)/(l-1)}{\mathrm{SSE}_n(t)/(n-l)}. \tag{3.2}$$

   이 때, $\mathrm{SSR}_n(t)$과 $\mathrm{SSE}_n(t)$은 다음과 같다.

$$\mathrm{SSR}_n(t) = \sum_{i=1}^{l} n_i \left(\bar{X}_i(t) - \bar{X}(t)\right)^2, \quad \mathrm{SSE}_n(t) = \sum_{i=1}^{l}\sum_{j=1}^{n_i} \left(X_{ij}(t) - \bar{X}_i(t)\right)^2. \tag{3.3}$$

   $\bar{X}(t) = (1/n)\sum_{i=1}^{l}\sum_{j=1}^{n_i} X_{ij}(t)$, $\bar{X}_i(t) = (1/n_i)\sum_{j=1}^{n_i} X_{ij}(t)$, $i = 1, \ldots, l$이다. 식 (3.2)는 주어진 t에 대해서 구한 일원분산분석 검정 통계량이다. 각각의 점 t에서 검정 통계량을 구할 수 있고, 함수 데이터의 검정 통계량은 하나의 수치가 아닌 t에 따른 함수값으로 표현된다.

2. L^2-norm에 근거한 검정 방법

- Faraway (1997), Zhang과 Chen (2007)
  점별 집단 간 변동(pointwise between-subject variation)을 적분한 값 $\int_I \mathrm{SSR}_n(t)dt$을 통계량으로 이용한다. 데이터가 다변량 정규분포를 따를 경우, χ^2 분포를 이용하여 가설검정을 할 수 있다. 다변량 정규분포를 따르지 않을 경우에는 bootstrap L^2-norm에 근거한 검정 방법을 이용할 수 있다.
- Cuevas 등 (2004)
  다음 통계량을 검정 통계량으로 이용하였다.

$$\sum_{1\le i\le j\le l} n_i \int_I \left(\bar{X}_i(t) - \bar{X}_j(t)\right)^2 dt.$$

  귀무가설의에서 검정통계량의 limit random expression과 관련된 Gaussian process를 재샘플링하는 부트스트랩 방법을 이용하였다. Cuevas 등 (2004)은 등분산과 이분산일 경우 모두 이용 가능한 검정 방법을 제시하였다.

3. F-type 검정 방법 (Shen과 Faraway, 2004; Zhang, 2011)
   점별 집단 간 변동을 적분한 값과 점별 집단 내 변동을 적분한 값을 이용한 방법이다.

$$\frac{\int_I \mathrm{SSR}_n(t)dt/(l-1)}{\int_I \mathrm{SSE}_n(t)dt/(n-l)}. \tag{3.4}$$

   이 통계량은 귀무가설 하에서 근사적으로 F-분포를 따르는데, F-분포의 자유도를 구하는 방법에 따라 두 가지 추정 방법이 제시되었다. 편향(bias) 보정 없이 naive하게 구하는 방법과, 편향을 보정한 추정 방법이 사용 가능하다. 편향을 보정한 검정 방법에는 bootstrap F-type 검정 방법이 이용되었다.

4. Górecki와 Smaga (2015).

   $X_{ij}(t)$를 다음과 같이 smoothing한 후, 식 (3.4)에 해당하는 검정 통계량을 계산한다.

$$X_{ij}(t) \approx \sum_{m=1}^{K} c_{ijm}\phi_m(t), \ \ t \in I. \tag{3.5}$$

   이 때, $X_{ij}(t)$는 I 위에서 함수식의 곱이 적분 가능한 모든 값을 포함하는 Hilbert space에서 정의되었다고 가정한다. basis 함수 ϕ_m은 $L_2(I)$에 속하고, c_{ijm}, $m = 1, \ldots, K$ 분산이 유한한 랜덤 변수이다. K는 충분히 큰 값을 선택하여 $X_{ij}(t)$이 근사가 잘 되도록 해야하며, BIC, AIC와 같은 information criterion을 이용하여 구하게 된다. Górecki와 Smaga (2015)는 식 (3.4)의 근사값을 계산하였다.

5. 무작위로 선택된 벡터를 이용한 검정 방법 (Cuesta-Albertos와 Febrero-Bande, 2010)

   Cuesta-Albertos와 Febrero-Bande (2010)는 다변량 정규분포로부터 반복적으로 무작위로 벡터를 선택하고, 선택된 벡터들로부터 False Discovery Rate (FDR)을 조절하여 가설 검정에 필요한 새로운 p-value를 정의하는 방법을 이용하였다. 관측된 함수 데이터가 Hilbert space $\mathcal{H}$에 속한다고 가정하고, $\mathcal{H}$에 속하는 다변량 정규 분포를 따르는 벡터 $\boldsymbol{h}$를 무작위로 선택한다. 다음과 같은 귀무가설을 설정하였다.

$$H_0^h : < \boldsymbol{\mu}_1, \ \boldsymbol{h} > = \cdots = < \boldsymbol{\mu}_l, \ \boldsymbol{h} > . \tag{3.6}$$

   위의 식에서 $< \cdot, \cdot >$은 scalar product로써 $< \boldsymbol{a}, \boldsymbol{b} > = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{b}$로 계산된다. Cuesta-Albertos와 Febrero-Bande (2010)는 귀무가설 (3.1) H_0가 기각될 경우, 위의 귀무가설 (3.6) H_0^h가 기각됨을 보였다. FDR을 계산하기 위해서는 반복하여 벡터를 선택하여야 하는데, 이 때 반복 횟수 k를 지정하여야 하는 문제가 생긴다. Cuesta-Albertos와 Febrero-Bande (2010)는 k를 30에 가까운 값을 제안하였으나, Górecki와 Smaga (2019)에서는 k가 30보다 훨씬 클 때, 좋은 결과를 얻을 때가 있음을 언급하였다. 자세한 검정 방법은 Cuesta-Albertos와 Febrero-Bande (2010)와 Górecki와 Smaga (2019)을 참고하길 바란다.

R에서 `fdANOVA` 패키지에 있는 함수 fanova.tests를 이용하여 단변량 함수 분산분석을 시행할 수 있다. fanova.tests 함수의 syntax는 다음과 같다.

```
fanova.tests(x = NULL, group.label, test = "ALL",
        params = NULL,
      parallel = FALSE, nslaves = NULL).
```

검정 방법(test)을 지정하지 않으면, 모든 방법을 이용할 수 있다. 또는 다음과 같은 검정 방법 중에서 선택하여 이용할 수 있다.

- "FP": 위에서 나열한 방법 중 4번 Górecki와 Smaga (2015)가 제시한 검정 방법이다.
- "CH": 위에서 나열한 방법 중 2번 L^2-norm에 근거한 검정 방법 중 두 번째에 소개된 방법으로써 등분산 가정에 적합한 방법이고, bootstrap resampling을 이용하여 가설검정이 수행된다.
- "CS": 위에서 나열한 방법 중 2번 L^2-norm에 근거한 검정 방법 중 두 번째에 소개된 방법으로써 이분산 가정에 적합한 방법이고. bootstrap resampling을 이용하여 가설검정이 수행된다.
- "L2N": 위에서 나열한 방법 중 2번 L^2-norm에 근거한 검정 방법 중 첫 번째에 소개된 방법으로써 정규분포 가정 하에 수행된 naive한 방법으로 불리는 방법이다 (Faraway, 1997; Zhang과 Chen, 2007; Zhang, 2014).
- "L2B": 위에서 나열한 방법 중 2번 L^2-norm에 근거한 검정 방법 중 첫 번째에 소개된 방법으로써 정규분포 가정 하에 수행된 방법으로 bias가 보정된 방법이다 (Faraway, 1997; Zhang과 Chen, 2007; Zhang, 2014).

- "FN": 위에서 나열한 방법 중 3번 F-type 검정 방법으로 편향(bias)이 보정되지 않은 naive한 방법으로 불리는 방법이다 (Shen과 Faraway, 2004).
- "FB": 위에서 나열한 방법 중 3번 F-type 검정 방법으로 편향(bias)이 보정된 방법이다 (Zhang, 2011).
- "Fb": 위에서 나열한 방법 중 3번 F-type 검정 방법으로 bootstrap 검정을 이용한 방법이다 (Zhang, 2014).
- "GPF": 점별 F-test 검정 통계량을 적분하여 하나의 값으로 나타낸 globalizing 방법이다 (Zhang과 Liang, 2014).
- "Fmaxb": Fmax bootstrap 방법 (Zhang 등, 2019).
- "TRP": 위에서 나열한 방법 중 5번 무작위로 선택된 벡터에 근거한 검정 방법이다 (Cuesta-Albertos와 Febrero-Bande, 2010).

3.2. 다변량 함수 분산분석

Górecki와 Smaga (2017)이 다변량 함수 데이터의 분산분석 방법(multivariate analysis of variance problem for functional data; FMANOVA)에 대해 제안하였다. 데이터는 다변량 함수 $\boldsymbol{X}_{ij}(t) = (X_{ij1}(t), \ldots, X_{ijp}(t))^{\mathrm{T}} \in \mathrm{SP}_p(\boldsymbol{\mu}_i, \boldsymbol{\Gamma})$, $i = 1, \ldots, l$, $j = 1, \ldots, n_i$, $t \in I$의 형태를 갖고 있다. 여기서 $\mathrm{SP}_p(\boldsymbol{\mu}, \boldsymbol{\Gamma})$는 평균벡터가 $\boldsymbol{\mu}(t)$, $t \in I$이고 공분산 함수 $\boldsymbol{\Gamma}(s,t)$, $s, t \in I$인 p-차원 확률 과정의 집합이다.

l개의 집단 간의 평균 벡터 $\boldsymbol{\mu}_i (i = 1, \ldots, l)$가 차이가 있는지 검정하기 위한 귀무 가설을 다음과 같이 설정하였다.

$$H_0 : \boldsymbol{\mu}_1(t) = \cdots = \boldsymbol{\mu}_l(t), \ \ t \in I. \tag{3.7}$$

$\boldsymbol{X}_{ij}$는 구간 I에서 두 번 적분 가능한 $L_2^p(I)$ p-차원 Hilbert space에 속한다고 가정한다. 단변량 함수 데이터의 경우 식 (3.5)와 유사하게 다변량 함수 데이터도 다음과 같이 표현할 수 있다.

$$\boldsymbol{X}_{ij}(t) \approx \begin{pmatrix} \boldsymbol{c}_{ij1} \\ \vdots \\ \boldsymbol{c}_{ijp} \end{pmatrix} \boldsymbol{\phi}(t) = \boldsymbol{c}_{ij} \boldsymbol{\phi}(t). \tag{3.8}$$

위의 식에서 $\boldsymbol{c}_{ijm} = (c_{ijm1}, \ldots, c_{ijmK_m}, 0, \ldots, 0) \in \mathcal{R}^{\mathrm{KM}}$, $\boldsymbol{\phi}(t) = (\phi_1(t), \ldots, \phi_{\mathrm{KM}}(t))^{\mathrm{T}}$, $t \in I$, $i = 1, \ldots, l$, $j = 1, \ldots, n_i$, $m = 1, \ldots, p$, $\mathrm{KM} = \max\{K_1, \ldots, K_p\}$이다. K_m 값은 Górecki와 Smaga (2015)이 제안한 함수 분산분석 방법과 같이 AIC와 같은 Information criterion 방법을 이용해서 구할 수 있다. 집단 내 변동 $\boldsymbol{E}$와 집단 간 변동에 해당하는 $\boldsymbol{H}$를 다음과 같이 계산할 수 있다.

$$E = \sum_{i=1}^{l} \sum_{j=1}^{n_i} \int_I \Big(\boldsymbol{X}_{ij}(t) - \bar{\boldsymbol{X}}_i(t)\Big)\Big(\boldsymbol{X}_{ij}(t) - \bar{\boldsymbol{X}}_i(t)\Big)^{\mathrm{T}} dt, \tag{3.9}$$

$$H = \sum_{i=1}^{l} n_i \int_I \Big(\bar{X}_i(t) - \bar{X}(t)\Big)\Big(\bar{X}_i(t) - \bar{X}(t)\Big)^{\mathrm{T}} dt. \tag{3.10}$$

이 때, $\bar{\boldsymbol{X}}_i(t) = (1/n_i) \sum_{j=1}^{n_i} \boldsymbol{X}_{ij}(t)$, $i = 1, \ldots, l$이고, $\bar{X}(t) = 1/n \sum_{i=1}^{l} \sum_{j=1}^{n_i} \boldsymbol{X}_{ij}(t)$, $t \in I$이다. 식 (3.8)에서 데이터의 근사값을 이용하기로 하였으므로, basis 함수 $\boldsymbol{\phi}$를 이용하여 $\boldsymbol{E}$와 $\boldsymbol{H}$의 근사값을 구할 수 있다. 다변량 분산분석에서 이용된 방법과 비슷한 방식으로 다음과 같은 검정 통계량을 계산한다.

1. Wilk's Lambda: $W = \frac{|\boldsymbol{E}|}{|\boldsymbol{E}+\boldsymbol{H}|}$.

Table 2: Result of univariate functional ANOVA of temperature data of Seoul and Busan

Test method	Test statistic	p-value
CH	4245.748	0
CS	4245.748	0
L2N	1656.877	0
L2B	1656.877	0
L2b	1656.877	0
FN	302.720	0
FB	302.720	0
Fb	302.720	0
GPF	275.152	0
Fmaxb	821.482	0
TRP	NA	p-value ANOVA = NA (without permutation)
	NA	p-value ATS = 0 (without permutation)
	NA	p-value WTPS = 0 (using B = 10000 permutations)

2. Hotelling-Lawley Trace: LH = $\text{tr}(\boldsymbol{H}\boldsymbol{E}^{-1}) = \text{tr}(\boldsymbol{E}^{-1}\boldsymbol{H})$.

3. Pillai's Trace: $P = \text{tr}(\boldsymbol{H}(\boldsymbol{H}+\boldsymbol{E})^{-1})$.

4. Roy's Maximum Root: R = Largest eigenvalue of $\boldsymbol{H}\boldsymbol{E}^{-1}$.

등분산 다변량 분산분석에서는 정규 분포 가정을 하였으나, 이 경우에는 특별한 분포 가정을 하지 않았으므로, W, LH, P, R에 해당하는 분포는 permutation 검정 방법을 이용하여 추정할 수 있다.

Górecki와 Smaga (2017)는 Cuesta-Albertos와 Febrero-Bande (2010) 방법을 다변량으로 확장하였다. 다음과 같은 가설 검정을 설정하였다.

$$H_0^S : \left(< \mu_{11}, s_1 >, \ldots, < \mu_{1p}, s_p >\right)^{\mathrm{T}} = \cdots = \left(< \mu_{l1}, s_1 >, \ldots, < \mu_{lp}, s_p >\right)^{\mathrm{T}}. \tag{3.11}$$

Górecki와 Smaga (2017)는 모든 $\boldsymbol{S} = (s_1, \ldots, s_p)^{\mathrm{T}} \in \mathcal{H} \times \cdots \mathcal{H}$에 대해서, H_0가 기각되면, H_0^S가 기각됨을 보였다.

Górecki와 Smaga (2017)의 Table 2에 따르면, Monte Carlo 시뮬레이션 결과 Roy's maximum root test (R)을 이용했을 때1종 오류를 범할 가능성이 높아지므로 권장하지 않는다. Wilk's lambda (W) test the Lawley-Hotelling trace (LH) test와 the Pillai trace (P) test를 쓰는 것이 좋으나, 이 세 가지 통계 검정 방법이 항상 같은 결과를 제시하지는 않으므로, 다양한 검정 방법을 통해 결과를 모두 확인하는 것을 권장한다. R에서 `fdANOVA` 패키지의 함수 fmanova.ptbfr를 이용하여 다변량 함수 분산분석을 시행할 수 있다. 이 함수에서는 특정 검정 방법을 지정하는 옵션은 없으며, 이 함수를 실행하면 위에서 언급한 네 가지 검정 방법 (W, LH, P, R)의 결과를 모두 보여준다.

4. 데이터 분석

4.1. 함수적 단변량 분산분석

서울과 부산의 기온은 차이가 크다. 위도 상으로 부산이 서울보다 낮기 때문에 부산은 전반적으로 서울보다 기온이 높다. Figure 1에서는 2020년 1주부터 20주까지의 서울 25개 구의 매주 기온 평균과 부산의 15개 구와 1개의 군의 매주 기온 평균을 그래프로 표현하였다. 52주의 함수 데이터를 모두 이용하려고 하였으나 25개의

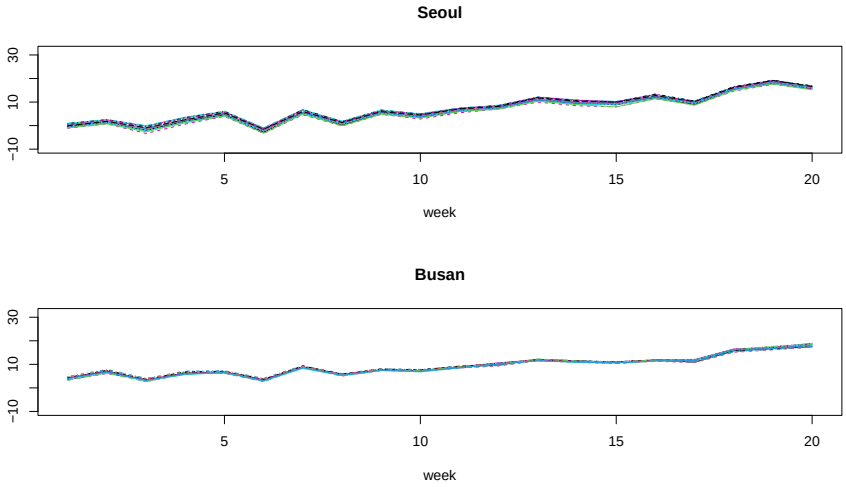

Figure 1: *Upper plot represents 25 Seoul Gu's average weekly temperatures during first 20 weeks in 2020. Bottom plot represents 15 Gu's and 1 Gun of Busan during first 20 weeks in 2020.*

서울 데이터와 16개의 서울 데이터로 분석하기에는 52주에 해당하는 벡터의 차원이 비교적 큰 편이라 20주의 함수 데이터를 이용하였다. 서울과 부산으로 나누어진 두 집단의 20주의 평균 기온차이가 같은지 함수적 단변량 분산분석을 통해 확인하고자 한다.

R에서 `fdANOVA` 패키지의 함수 fanova.tests를 이용하여 분석하였다. 3.1절에 설명한 모든 검정 방법을 이용한 결과는 Table 2에 제시하였다. 모든 검정 방법에서 결과는 귀무가설(두 집단의 매 주 평균 기온은 같다)을 기각하는 것으로 나타났다. 즉, 서울과 부산의 매주 평균 기온 함수는 다르다고 결론 내릴 수 있다.

4.2. 함수적 다변량 분산분석

MNIST 데이터셋은 손으로 쓴 숫자들로 이루어진 데이터 셋으로써, LeCun 교수의 웹사이트(http://yann.lecun.com/exdb/mnist/)에 공개되어 있다. 이 웹사이트에는 확장자가 .idx3-ubyte인 파일을 제공하고, Kaggle에서는 같은 데이터를 csv 형식으로 변환한 파일(https://www.kaggle.com/datasets/oddrationale/mnist-in-csv)을 제공하고 있다. 하나의 이미지는 28×28의 pixel로 구성되어 있고, 하나의 pixel은 0에서 255까지의 값 중 하나의 값으로 표시된다. 0은 하얀색, 255는 검정색이고, 그 사이의 값은 회색인데, 0에 가까울 수록 연하고, 255에 가까울 수록 진한색이다. Figure 2는 MNIST 데이터셋에 있는 0과 1의 이미지의 예를 보여주고 있다. 이러한 이미지를 다변량 함수 데이터로 변환하기 위해, 세로로 총 28개의 pixel을 10, 9, 9개의 구분하여 세 가지 윈도우를 만들고, 각각의 윈도우에서 가로로 1부터 28번째 pixel을 하나의 함수로 고려하였다. 가장 진한 부분에 해당하는 위치를 함수값으로 표현하고, 고정된 X 좌표에서 여러 개의 pixel이 최댓값을 동시에 갖고 있는 경우, 그 중 가장 낮은 위치를 선택하도록 하였다. 고정된 X 좌표에서 Y 값이 모두 0인 경우에는 결측값으로 처리하지 않기 위해 $y = 0$으로 채웠다. 이런식으로 하나의 이미지는 길이가 28인 3가지의 함수로 나타내어진다. Figure 2의 각각의 이미지는 세개의 함수로 변환되었고, 이것을 그래프로 그린 것이 Figure 3이다.

R에서 `fdANOVA` 패키지의 함수 fmanova.ptbfr 을 이용하여 (1) 0으로 분류된 데이터의 집합과 1로 분류된 데이터를 서로 다른 집합으로 구별하는지 확인하고, (2) 0으로 분류된 데이터를 두 가지 데이터 셋으로 나누어서 다변량 함수 분산분석을 하였다.

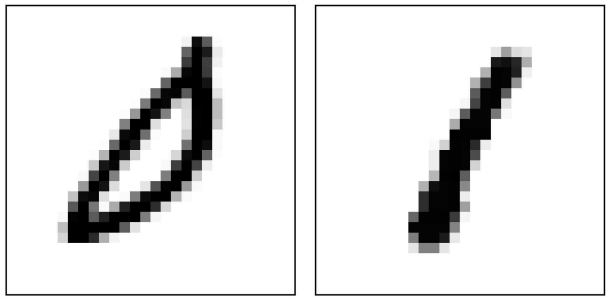

Figure 2: *Handwritten images of 0 and 1 in MNIST data set.*

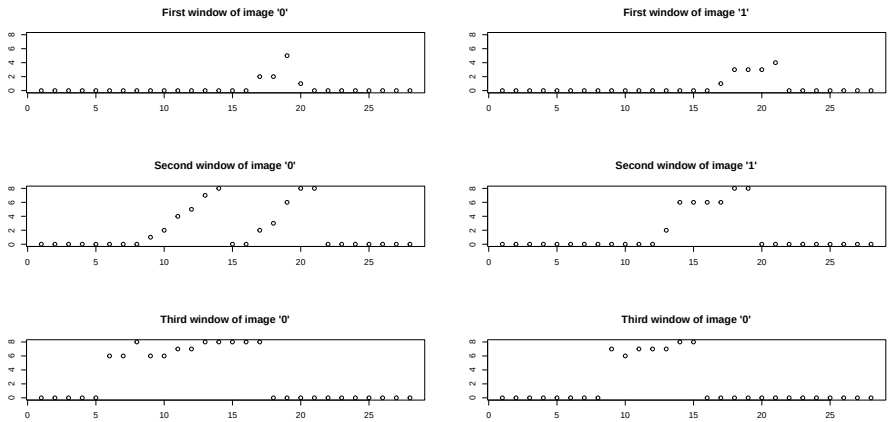

Figure 3: *Multivariate functional data corresponding to 0 and 1.*

MNIST 데이터셋에서 train 데이터 중 0과 1을 1,000개씩 선택하여 다변량 함수로 만들고, 각각의 집단의 평균값이 같은지 비교하여 보았다. zero_one은 0과 1을 함수값으로 변환한 데이터셋이고, zero_one.label은 어느 집합에 속하는지 표기한 labeling을 한 벡터이다. 분석 시간을 단축하기 위해 병렬처리를 하기 위해 parallel=TRUE로 옵션을 지정하여 분석하였다. 다음은 R 코드이다.

```
set.seed(123),
(fmanova1 <- fmanova.ptbfr(zero_one, zero_one.label,parallel=TRUE)),
summary(fmanova1).
```

분석 결과, $W = 0.3602647$ (p-value = 0), LH = 1.342745 (p-value = 0), $P = 0.8043858$ (p-value = 0), $R = 0.931352$ (p-value = 0)로 네 가지 검정 방법 모두 귀무가설(두 집단의 평균 함수는 같다)을 기각하게 된다. 즉, 두 집단이 다른 평균 함수값을 갖는다고 결론 내릴 수 있다.

이번에는 0이라고 쓰여진 이미지를 1,000개씩 두 개의 집합으로 나누어서 비교하고자 한다. 위의 R 코드와 비슷한 방식으로 입력하면 되므로, R 코드는 생략한다. 결과는 $W = 0.9980427$ (p-value = 0.191), LH = 0.001960025 (p-value = 0.191), $P = 0.001958468$ (p-value = 0.191), $R = 0.00089063$ (p-value = 0.373)으로 나왔다. 유의 수준 0.05 기준으로 했을 때, 네 가지 검정 방법의 경우 모두 귀무가설(두 집단의 평균 함수는 같다)을 기각할 수 없다. 즉, 두 집단이 동일하다고 결론 내릴 수 있다. 특이한 점은 3.2절에서 언급한 것처럼 Roy's

maximum root test (R)의 검정 결과의 p-value가 다른 검정 방법과 비교했을 때 상대적으로 크다는 점이다.

5. 결론

지금까지 단변량 함수 분산분석 방법과 다변량 함수 분산분석 방법에 대해서 살펴 보았다. 단변량 ANOVA가 쉽고 주로 쓰이는 방법이기 때문에, 함수 데이터에 대해서 종종 점별 분산분석(point-wise ANOVA)를 구하여 통합되지 않은 결과를 보여주는 경우가 있다. 이럴 경우, 모든 점에서 정규성 가정과 등분산성 가정을 만족해야 사용할 수 있다는 점, 또한 Family-wise error rate이 커질 가능성이 있는 단점이 있으므로 권장하지 않는다. 대신에 최근에 제안된 다양한 함수 분산분석을 이용하면 좀 더 합리적인 의사결정을 내릴 수 있을 것으로 생각한다. R에서 `fdANOVA` 패키지를 이용하여 단변량 및 다변량 함수 분산분석을 수월하게 할 수 있으므로, 유용하게 이용할 수 있을 것으로 기대한다. 다양한 검정 통계량이 제안되었지만, 어느 방법이 절대적으로 우월하다고 할 수 없으므로 가능한 모든 검정 방법을 이용하여 결과를 비교하는 것을 권한다.

References

Cuevas A, Febrero M, and Fraiman R (2004). An anova test for functional data, *Computational statistics & Data Analysis*, **47**, 111–122.

Cuesta-Albertos JA and Febrero-Bande M (2010). A simple multiway ANOVA for functional data, *Test:Journal of the Spanish Society of Statistics and Operations Research*, **19**, 537–557.

Don HSRA (2018). A relationship between the One-Way MANOVA test statistic and the hotelling lawley trace test statistic, *International Journal of Statistics and Probability*, **7**, 124–131.

Faraway JJ (1997). Regression analysis for a functional response, *Technometrics*, **39**, 254–261.

Friedrich S and Pauly M (2018). MATS: Inference for potentially singular and heteroscedastic MANOVA, *Journal of Multivariate Analysis*, **165**, 166–179.

Górecki T and Smaga Ł (2015). A comparison of tests for the one-way ANOVA problem for functional data, *Computational Statistics*, **30**, 987–1010.

Górecki T and Smaga Ł (2017). Multivariate analysis of variance for functional data, *Journal of Applied Statistics*, **44**, 2172–2189.

Górecki T and Smaga Ł (2019). fdANOVA: an R software package for analysis of variance for univariate and multivariate functional data, *Computational Statistics*, **34**, 571–597.

Olive DJ (2017). *Robust multivariate analysis*, Springer International Publishing.

Ramsay JO and Silverman BW (2005). *Functional Data Analysis 2nd Edition*, Springer-Verlag, New York.

Shen Q and Faraway J (2004). An F test for linear models with functional responses, *Statistica Sinica*,**14**, 1239–1257.

Srivastava MS and Kubokawa T (2013). Tests for multivariate analysis of variance in high dimension under non-normality, *Journal of Multivariate Analysis*, **115**, 204–216.

Zhang JT (2011). Statistical inferences for linear models with functional responses, *Statistica Sinica*, **21**, 1431–1451.

Zhang JT (2014). Analysis of variance for functional data, *Monographs on Statistics and Applied Probability*, **127**, 127.

Zhang JT and Chen J (2007). Statistical Inferences for functional data, *The Annals of Statistics*, **35**, 1052–1079.

Zhang JT and Liang X (2014). One-way ANOVA for functional data via globalizing the pointwise F-test, *Scan-*

dinavian Journal of Statistics, **41**, 51–71.

Zhang JT, Cheng MY, Wu HT, and Zhou B (2019). A new test for functional one-way ANOVA with applications to ischemic heart screening, *Computational Statistics & Data Analysis*, **132**, 3–17.

Received July 19, 2022; Revised August 24, 2022; Accepted August 25, 2022

단변량 및 다변량 함수 데이터에 대한 분산분석의 활용

김미정[1,a]

[a]이화여자대학교 통계학과

요 약

함수 데이터는 다양한 분야에서 수집되고 있으며, 집단 간의 함수 데이터를 비교해야하는 경우가 종종 발생한다. 이럴 경우 점별 분산분석 방법을 이용하여 설명하기에는 무리가 있으며, 통합된 결과를 제시할 필요가 있다. 이에 대한 다양한 연구가 제안되었으며, 최근에 R 패키지 `fdANOVA`로 구현되었다. 이 논문에서 우선 분산분석 및 다변량 분산분석을 설명하고, 최근에 제안된 다양한 단변량 및 다변량 함수 데이터 분산분석을 설명하고자 한다. 또한 R 패키지 `fdANOVA`의 사용 방법을 설명하고, 이 패키지를 이용하여 서울과 부산 지역의 주별 기온을 단변량 함수 데이터 분산분석을 통해 비교하고, 손글씨 이미지를 다변량 함수 데이터로 변환하여 다변량 함수 데이터 분산분석을 이용하여 비교하고자 한다.

주요용어: 다변량 분산분석, 다변량 분산분석, 다변량 함수 분산분석, 분산분석

이 논문은 연구재단 연구 과제 (NRF-2020R1F1A1A01074157)에 의하여 수행되었음.

[1](03760) 서울시 서대문구 이화여대길 52, 이화여자대학교 통계학과. E-mail: m.kim@ewha.ac.kr