Computational Methods for Differential Equations
http://cmde.tabrizu.ac.ir
Vol. 7, No. 4 (Special Issue), 2019, pp. 608-615

Statistical inference for Functional data: two sample Behrens-Fisher problem

Hassan Sharghi Ghale-Joogh*
Department of Statistics,
Shahid Beheshti University, Tehran, Iran.
E-mail: h_sharghi@sbu.ac.ir

Abstract With modern technology development, functional data analysis (FDA) has received considerable recent attention in many scientific fields. The estimation of mean in FDA is of interest, because it is not only important by itself but it is a prelude to other issues such as dimension reduction and modeling of functional data In this paper we construct a two-sample Behrens-Fisher problem when data are functions and obtain the asymptotic properties of the test statistic as data dimension increases with the sample size. The power of the proposed test is also investigated. The proposed test used for inference about the differences in the mean temperature functions of the Western and the South-western weather stations of Iran.

Keywords. Behrens-Fisher problem, Weather stations of Iran, Functional data, Increasing dimension.
2010 Mathematics Subject Classification. 62M07, 62G99, 62H15.

1. Introduction

Functional data analysis (FDA) has recently experienced intense development due to the great progress made in measuring and collecting data that are in the form of of curves or images. The main feature of these data is high dimensionality which makes it different from data arising from classical statistical studies involves scalar or vector observations. Therefore, studying the interrelations between these data is challenging. The data in these fields may be observed discretely as realizations of a continuous time stochastic process that belongs to a suitable infinite dimensional Hilbert space, typically $L^2(\mathcal{I})$, the space of all square-integrable functions on $\mathcal{I}$.

A comprehensive introduction to FDA has been provided by Ramsay and Silverman (2002, 2005) and Kokoszka and Reimherr (2017). Many recent developments are also reviewed by Horvath and Kokoszka (2012) and Hsing and Eubank (2015).

In functional data analysis, inference about the mean function is fundamental not only as a critical quantity for understanding elementary aspects of functional data but also as an indispensable ingredient for many advanced FDA procedures. Statistical inference about mean of functional data is critical, because it is useful for make decisions in for example growth curve analysis in biology, lifetime testing problems in biostatistics, and follow-up studies for monitoring disease progression in medicine.

The classical problem of testing equality of the means of two populations with unequal covariances is referred to as the Behrens-Fisher problem and has attracted a

lot of interest for several decades. However, it is challenging due to lack of an exact solution fulfilling the requirements of good tests. In the case of functional data, this problem has been considered by many researchers. In particular, the Behrens-Fisher problem in FDA has been considered by many researchers. Ramsey and Silverman (2005) adapted general F- and t-test for functional data in order to point-wisely compare mean functions. A test based on L^2 norm for this problem was proposed by Zhang et al. (2010). Pini and Vantini (2016) proposed a test procedure for compare two mean functions based on permutation test. Zhang and Ling (2014) by extending the classical point-wise F-test to functional data, studied the one-way anova problem obtained random expressions of the test statistic and the test power asymptotically. An generalization of Mahalanobis distance to make statistical inference for functional data was introduced by Ghiglcetti et al. (2017).

Testing equality of means for multivariate data based on Hotelling's T^2-statistic, is unusable in the case that the data dimension, d, is larger than the sample size, n. When using Hotelling's T^2-statistic for high-dimensional data in which d is larger than n, the inverse of the sample covariance matrix may not exist and therefore cannot be used in this setting. For relatively recent contributions of Hotelling's T^2 test for functional and high-dimensional data; see, for example, Park and Ayyala (2013), Gregory et al. (2015) and Dong et al. (2016).

Bai and Saranadasa (1996) to avoids the singularity problem, that may happen in the covariance estimation when d is larger than n, replaced $(\bar{X}_1 - \bar{X}_2)^T S^{-1}(\bar{X}_1 - \bar{X}_2)$ in Hotelling's T^2-statistic by $\|\bar{X}_1 - \bar{X}_2\|_\varepsilon$, where $\bar{X}_1$ and $\bar{X}_2$ are the two sample means, S is the sample covariance matrix and $\|.\|_\varepsilon$ denotes the Euclidean norm in R^d.

We construct a testing procedure in the context of FDA which is based on a new distance in L^2, and derive the asymptotic power of the test under local alternatives. The proposed test takes into account the covariance structure of the processes, and weights the differences between the two sample means along each of the first few principal components proportional to their importance. However, this is not the case for the test based on the L^2-norm, where those differences are weighted equally along each of the principal components. The test has also better performance compared to Hotelling's T^2 test.

The paper is organized as follows. We first introduce the notation required, explain the problem and then propose a test statistic for the Behrens-Fisher problem and derive the asymptotic properties of the test statistic as d increases with the sample size, N in section 2. In section 3, the proposed method is applied to a real data involving the Iranian temperature dataset.

2. Main results

We consider two sets of independent observations, $X_{ij}(t),\ t \in \mathcal{I},\ j = 1, 2, ..., N_i$, $i = 1, 2$, defined over a compact interval $\mathcal{I}$, such that they follow the models

$$X_{1j}(t) = \mu_1(t) + \epsilon_{1j}(t), \quad 1 \le j \le N_1,$$
$$X_{2j}(t) = \mu_2(t) + \epsilon_{2j}(t), \quad 1 \le j \le N_2,$$

where $\mu_1(.)$ and $\mu_2(.)$ are the common mean functions of the two populations, and $\epsilon_{1j}(.)$ and $\epsilon_{2j}(.)$ are random error functions satisfying $E[\epsilon_{1j}(t)] = 0$ and $E[\epsilon_{2j}(t)] = 0$. For any functions $f, g \in L^2(\mathcal{I})$, the inner product between f and g is defined as $\langle f, g\rangle = \int_{\mathcal{I}} f(t)g(t)\mathrm{d}t$. We wish to test whether the two mean functions are equal:

$$H_0 : \mu_1 = \mu_2 \ \ versus \ \ H_1 : \mu_1 \neq \mu_2.$$

To do that, we impose the following standard assumptions (see e.g., Horvath and Kokoszka, 2012; and Zhang, 2013).

(1) For each $i = 1, 2$, $\epsilon_{i1}, \epsilon_{i2}, ..., \epsilon_{in_i}$ are independent and identically distributed. Moreover, $\{\epsilon_{1j},\ 1 \leq j \leq N_1\}$ and $\{\epsilon_{2j},\ 1 \leq j \leq N_2\}$ are independent.
(2) For each $i = 1, 2$, $\int \mu_i^2(t)dt < \infty$ and $E\|\epsilon_{i1}\|^4 < \infty$, where $\|.\|$ denotes the L^2-norm.
(3) As $min(N_1, N_2) \to \infty$ we have $\frac{N_1}{N} \to \theta$ such that $\theta \in (0, 1)$, where $N = N_1 + N_2$.
(4) $d = o(N^{\eta - \frac{1}{2}})$, for some $\frac{1}{2} \leq \eta < 1$.

Under this assumptions, the covariance functions

$$\gamma_i(s, t) = E\{\epsilon_{i1}(s)\epsilon_{i1}(t)\},\ i = 1, 2,$$

are square-integrable on $\mathcal{I}^2$. Define the pooled covariance function

$$\gamma_\theta(s, t) = (1 - \theta)\gamma_1(s, t) + \theta\gamma_2(s, t).$$

The linear operator corresponding to γ_θ is defined by

$$(\Gamma_\theta\phi)(t) = \int \gamma_\theta(s, t)\phi(s)ds,$$

taking ϕ to $\Gamma_\theta\phi$. The pooled covariance function $\gamma_\theta(s, t)$ is a symmetric, square-integrable function and has a spectral decomposition of the form

$$\gamma_\theta(s, t) = \sum_{j=1}^{\infty} \nu_j\phi_j(s)\phi_j(t), \tag{2.1}$$

where $\nu_1, \nu_2, ...$and $\phi_1, \phi_2, ...$ are eigenvalues and their corresponding eigenfunctions, respectively, such that

$$\nu_j\phi_j(t) = \int \gamma_\theta(s, t)\phi_j(s)\mathrm{d}s,\ j \geq 1.$$

Let $\bar{X}_1(t)$ and $\bar{X}_2(t)$ denote the mean functions of the two samples. The statistical inference about equality of the population mean functions is based on distance between their unbiased estimators that are the two sample means. In general, distance between two functions can be defined as follows:

$$d_B^2(f, g) = \int\int (f - g)(s)B(s, t)(f - g)(t)\mathrm{d}t\mathrm{d}s, \tag{2.2}$$

where $B(\cdot, \cdot)$ is a positive-definite, symmetric, square-integrable function. It is worth noting that if $B(\cdot, \cdot)$ is taken to be $1(\cdot, \cdot)$ and $\gamma_\theta^{-1}(\cdot, \cdot)$ then $d_B(f, g)$ is equal to the usual L^2 and Mahalanobis distance, respectively. Although the distance $d_{\gamma_\theta^{-1}}(f, g)$ considers the correlations and variability expressed by the covariance structure, as we mentioned in the introduction, it is not suitable for an infinite dimensional space

such as $L^2(\mathcal{I})$. To test H_0, we can take $B(\cdot,\cdot)$ in (2.2) to be the pooled covariance function, namely

$$d^2_{\Gamma_\theta}(\bar{X}_1,\bar{X}_2) = \int\int (\bar{X}_1-\bar{X}_2)(s)\gamma_\theta(s,t)(\bar{X}_1-\bar{X}_2)(t)\mathrm{d}s\mathrm{d}t$$
$$= \sum_{j=1}^{\infty}\nu_j\langle \bar{X}_1-\bar{X}_2,\phi_j\rangle^2, \tag{2.3}$$

where the last equality obtained from using (2.1). The distance was also employed by Yuan and Cai (2010) to propose a reproducing kernel Hilbert space (RKHS) approach to the functional linear regression. The use of eigenvalues, ν_j, weighted the differences between the two sample means along each of the eigenfunctions in (2.3) is not the only possible choice, and some functions of ν_j denoted by ω_j would be also valid. For example, ω_j can be taken to be ν_j^ℓ for $\ell > 0$. Therefore, in the following we replace the eigenvalues, ν_j by ω_j. We will show that its corresponding distance can also work for testing equality of the means of two populations. To appreciate why, we have

$$\frac{N_1N_2}{N}d^2_{C^\omega_\theta}(\bar{X}_1,\bar{X}_2) = \frac{N_1N_2}{N}\sum_{j=1}^{\infty}\omega_j\int\int(\bar{X}_1-\bar{X}_2)(s)\left[\phi_j(s)\phi_j(t)\right](\bar{X}_1-\bar{X}_2)(t)\mathrm{d}s\mathrm{d}t$$
$$= T_{C^\omega_\theta} + \frac{N_1N_2}{N}\sum_{j=d+1}^{\infty}\omega_j\langle\bar{X}_1-\bar{X}_2,\phi_j\rangle^2, \tag{2.4}$$

where

$$T_{C^\omega_\theta} = \frac{N_1N_2}{N}\sum_{j=1}^{d}\omega_j\langle\bar{X}_1-\bar{X}_2,\phi_j\rangle^2. \tag{2.5}$$

We expect that the distance is small under H_0, but it is large when H_1 is valid. We will test the null hypothesis using the first term on the right-hand side of (2.4) and one can show that the second term there tends to zero in probability. We impose the following assumptions on the weights ω_j.

(5) $\sum_{j=1}^{\infty}\omega_j < \infty$.

(6) $\frac{1}{\nu_\iota\omega_\iota} = o(N^{\frac{1}{15}})$, where $\nu_\iota\omega_\iota = min_{j=1,2,\ldots,d}(\nu_j\omega_j)$.

(7) $\sup_{j=1,2,\ldots,p}|\omega_j-\hat{\omega}_j| = O_p(n^{-\frac{1}{2}})$.

Now, we show that the first term on the right-hand side of (2.4) tends to a chi-squared mixture.

Theorem 2.1. *If H_0 and Assumptions (1)-(7) hold, then*

$$\sup_t\left|P_r\left(\frac{N_1N_2}{N}\sum_{j=1}^{d}\omega_j\langle\bar{X}_1-\bar{X}_2,\phi_j\rangle^2\le t\right) - P_r\left(\sum_{j=1}^{d}\omega_j\nu_j\chi^2_{j(1)}\le t\right)\right|$$
$$\le \frac{(\sum_{j=1}^{d}\omega_j)^{3/2}(const.\,m_1^{\frac{3}{2}}+const.\,m_2^{\frac{3}{2}})}{(\nu_\iota\omega_\iota)^{3/2}\sqrt{k}}\to 0, \tag{2.6}$$

where the constants only depend on $E\|\epsilon_{i1}\|^3$, $i = 1, 2$, *the* $\chi^2_{j(1)}$ *are i.i.d. chi-squared random variables with* 1 *degree of freedom and* $\nu_\imath\omega_\imath = \min_{j=1,\ldots,d}(\nu_j\omega_j)$.

Proof. See Sharghi G-J and Hosseini-Nasab (2018) □

One can show that the empirical version of first term on the right-hand side of (2.4) in which ν_j and ϕ_j are replaced by $\hat{\nu}_j$ and $\hat{\phi}_j$, respectively, also tends to a chi-squared mixture.

2.1. **Asymptotic power under local alternatives.** To obtain the asymptotic power of the first term on the right-hand side of (2.4), we leave out the case in which the alternative is fixed, since the associated power can easily tend to 1 as $n \longrightarrow \infty$. We investigate the power of $T_{C^\omega_\theta}$ under the alternatives that tend to the null hypothesis with a rate slightly slower than $n^{-\frac{1}{2}}$ so that the difference between the two population means decreases as n increases. Hence we consider the following local alternative:

$$H_1 : \ \mu_1(t) - \mu_2(t) = n^{-\frac{\beta}{2}}\eta(t), \tag{2.7}$$

where β is some constant satisfying $0 \leq \beta < 1$ and $\eta(t)$ is any fixed real function such that $\eta \in L^2$ and $0 < \|\eta\| < \infty$. We can write

Theorem 2.2. *If* H_1 *and Assumptions (1)-(7) hold, then we have*

$$T_{C^\omega_\theta} \xrightarrow{P} \infty,$$

as long as $\langle \eta, \phi_j \rangle \neq 0$ *for some* j.

Proof. Suppose that We can write

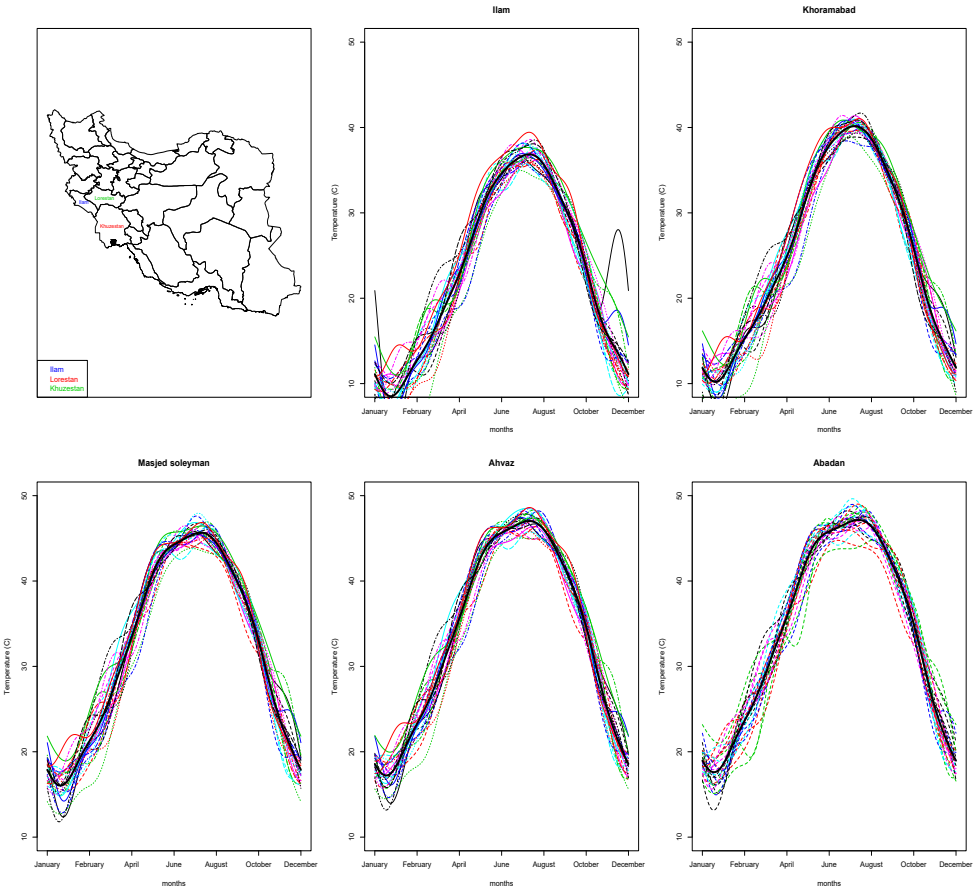
$$\begin{aligned}
T_{C^\omega_\theta} &= \frac{N_1N_2}{N}\sum_{j=1}^{d}\omega_j\langle \bar{X}_1 - \bar{X}_2, \phi_j\rangle^2 \\
&= \frac{N_1N_2}{N}\sum_{j=1}^{d}\omega_j\left(\langle \bar{U}_1 - \bar{U}_2, \phi_j\rangle + \langle \mu_1 - \mu_2, \phi_j\rangle\right)^2 \\
&= \frac{n_1n_2}{n}\sum_{j=1}^{d}\omega_j(\langle \bar{U}_1 - \bar{U}_2, \phi_j\rangle^2 + 2\langle \bar{U}_1 - \bar{U}_2, \hat{\phi}_j\rangle\langle \mu_1 - \mu_2, \phi_j\rangle + \langle \mu_1 - \mu_2, \phi_j\rangle^2) \\
&= \sum_{j=1}^{d}\omega_j^2\chi_j^2(1) + O_p(\frac{N_1N_2}{N^{1+\beta}}),
\end{aligned} \tag{2.8}$$

which completes the argument by $\frac{N_1N_2}{N^{1+\beta}} \to \infty$. □

3. Read data study

in this section the proposed methodology is illustrated via an application to a real functional data set collected in climatology. The Iran temperature data which showed in Figure 1, contains the daily temperature records of five Iranian weather stations over a year (365 days), involving Ilam, Khoramabad, Abadan, Ahvaz and Masjed-Soleyman in Western and South-western Iran. The data set we study consists of 26

FIGURE 1. The temperature curves contain, Ilam, Khoramabad, Abadan, Ahvaz and Masjed-Soleyman in Western and South-western Iran, obtained by averaging annual temperature over 26 years for each locations and the smoothed mean curve for each station showed by



years $(1990 - 2016)$ of daily temperatures for each stations. Following Ramsay and Silverman(2005), the discrete observations are converted to functional observations using a Fourier series basis with 31 basis functions.

Figure 1 shows the changes of temperature for 26 years, between 1990-2016, and mean curves for each station. The reconstructed temperature mean curves were displayed with solid line for each stations. As showed in Figure 1, the mean temperature functions of the stations in the Ilam and Khoramabad stations look like similar (with maximum temperature less than $40°C$) and far from the mean temperature function of the South-western weather stations contain Abadan, Ahvaz and Masjed-Soleyman (with maximum temperature more than $45°C$).

TABLE 1. P-values of the tests based on statistic $T_{C_\theta^\omega}$ applied to the Iranian temperature data set for Eastern and South-western weather stations.

station	Khoramabad		Ilam		Ahvaz		Abadan		Masjed-soleyman	
	$T_{C_\theta^\omega}$	P-value	$T_{C_\theta^\omega}$	P-value	$T_{C_\theta^\omega}$	P-value	$T_{C_\theta^\omega}$	P-value	$T_{C_\theta^\omega}$	P-value
Khoramabad	0	1								
Ilam	68.78	0	0	1						
Ahvaz	663.47	0	1403.06	0	0	1				
Abadan	753.01	0	1564.73	0	0.85	0.47	0	1		
Masjes-Soleyman	436.25	0	1012.07	0	26.19	0	38.97	0	0	1

The equality of the mean temperature curves of the Western, and the South-western, weather stations was tested in Table 1. The proposed test statistic, $T_{C_\theta^\omega}$, and also p-value are calculated for Behrens-Fisher problem of stations. As can be seen, test procedure leads to rejecting the equality of mean temperature functions between the stations. For measurement the test size we calculate test statistic and p-value for each station with itself. Table 1 shows that only mean temperature functions of Ahvaz and Abadan are equal and other stations mean curves have significant difference.

Based on the reconstructed temperature curves, the objective is to test if the mean temperature functions of the weather stations during the whole year are the same. However, for Abadan and Ahvaz stations, $T_{C_\theta^\omega}$ do not reject the null hypothesis of equality of mean functions.

REFERENCES

[1] Z. D. Bai and H. Saranadasa, *Effect of high dimension: by an example of a two-sample problem*, it Statist. Sinica, *6* (1996), 311–329.

[2] K. Dong, H. Pang, T. Tong, and M. G. Gentond, *Shrinkage-based diagonal Hotelling's tests for high-dimensional small sample size data*, J. Multiv. Anal., *143* (2016), 127–142.

[3] A. Ghiglietti, F. Ieva, and A. M. Paganoni, *Statistical inference for stochastic processes: two-sample hypothesis tests*, J. Stat. Plann. Infer., *180* (2017), 49–68.

[4] K. B. Gregory, R. J. Carroll, V. Baladandayuthapani, and S. N. Lahiri, *A two-sample test for equality of means in high dimension*, J. Amer. Statist. Assoc., *110* (2015), 837–849.

[5] L. Horváth and P. Kokoszka, Inference for functional data with applications, Springer, New York, 2012.

[6] T. Hsing and R. Eubank, Theoretical foundations of functional data analysis, with an introduction to linear operators, Wiley, 2015.

[7] P. Kokoszka, and M. Reimherr, Introduction to functional data analysis, Chapman & Hall/CRC Press, (2017).

[8] J. O. Ramsay, and B. Silverman, Applied Functional Data Analysis: Methods and Case Studies, Springer, New York, 2002.

[9] J. O. Ramsay and B. Silverman, Functional data analysis, Springer, New York, 2005.

[10] J. Park, and D. N. Ayyala, *A test for the mean vector in large dimension and small samples*, J. Stat. Plann. Infer., *143* (2013), 929-943.

[11] A. Pini and S. Vantini, *The interval testing procedure: A general framework for inference in functional data analysis*, Biom., *72* (2016), 835-845.

[12] H. Sharghi Ghale-Joogh and M-E. Hosseini-Nasab, *A two-sample test for mean functions with increasing number of projections*, Statistics, *52*(4) (2018), 852–873. 10.1080/02331888.2018.1472599.

[13] J.-T. Zhang, X. Liang, and S. Xiao, *On the two-sample Behrens-Fisher problem for functional data*, J. Stat. Theory Pract., *4* (2010), 571-587.

[14] J.-T. Zhang and X. Liang, *One-way ANOVA for functional data via globalizing the pointwise F-test*, Scand. J. Statist., *41* (2014), 51-71.
[15] J.-T. Zhang, Analysis of Variance for Functional Data., Chapman and Hall, London, 2013.
[16] M. Yuan and T. T. Cai, *A reproducing kernel Hilbert space approach to functional linear regression*, Ann. Statist., *38* (2010), 3412-3444.