Methodological Contributions to Functional Data Analysis

Deniz Ozisik

1 Introduction and Literature Review

Functional data analysis (FDA) deals with the analysis and theory of data that are in the form of functions, images and shapes, or more general objects. First generation functional data typically consist of a random sample of independent real-valued functions, $X_1(t), \ldots, X_n(t)$, on a compact interval $I = [0, T]$ on the real line. In its most general form, under an FDA framework, each sample element of functional data is considered to be a random function.

- TWO-SAMPLE FUNCTIONAL DATA TESTING PROBLEM AND ISSUES OF CONCERN

- UTILITY OF TWO-SAMPLE TESTS

- FUNCTIONAL DATA MODELING - SOME APPLICATIONS AND

1.1 Aims/Objectives

Including statement of the problem.

2 Two-Sample Tests for Functional Data

2.1 Introduction

We restrict ourselves to hypothesis testing in the two sample setting. Suppose we have i.i.d. samples $y_{1i}(t), i = 1, \ldots, n_1$ from Population 1, and $y_{2i}(t), i = 1, \ldots, n_2$ from Population 2. Let $\mu_1(t) = E[y_{1i}(t)]$ and $\mu_2(t) = E[y_{2i}(t)]$. We assume finite second moments. We want to test the null hypothesis

$$H_0 : \mu_1(t) = \mu_2(t), \text{ for all for all } t \in T$$

or

$$E[y_{1i}(t)] = E[y_{2i}(t)], \text{ for all for all } t \in T$$

Two sample hypothesis testing for functional data has been approached in many contexts; ranging from testing for specific types of differences, such as 25 differences in the mean or covariance functions, to testing for overall differences in the cumulative density functions. To detect differences in the mean functions of two independent samples of curves, Ramsay and Silverman (2005) introduced a pointwise t-test, Zhang et al. (2010) presented an L^2norm based test, Horvath et al. (2013) proposed a test based on the sample means of the curves, and Staicu et al. (2014) developed a pseudo likelihood ratio test. Extension to k independent samples of curves was discussed in Cuevas et al. (2004), Est evez-Perez and Vilar (2008), and Laukaitis and Rackauskas (2005), who proposed ANOVA-like testing procedures for testing the equality of mean functions. Recent research also focused on detecting differences in the covariance functions of independent samples of curves: see the

factor-based test proposed by Ferraty et al. (2007), the regularized M-test introduced by Kraus and Panaretos (2012), and the chi-squared test proposed by Fremdt et al. (2012).

Throughout the paper, we use test statistics such that do not have known distributions. That is why we use the permutation method to obtain the null distribution. The process is as follows:

- Generating a large number of random samples under the null hypothesis.
- Calculating the test statistic for each sample.
- Taking the maximum of the test statistics for each sample.
- Repeating the above steps many times to obtain a large number of maximum statistics.
- Plotting the distribution of the maximum statistics to obtain the null distribution.

2.2 Existing test procedures

2.2.1 $\max -T$-Statistics

Perhaps the easiest test statistic we can try is performing a pointwise two sample t-test, and obtain the maximum of the absolute value of t-statistics over all t, the $\max -T$-statistic:

$$\max -T = \sup_{t \in T} |T(t)|,$$

where $T(t)$ denotes the value of the two sample t-statistic based on $y_{11}(t), y_{12}(t), \ldots, y_{1n_1}(t)$ and $y_{21}(t), y_{22}(t), \ldots, y_{2n_2}(t)$. One can derive an approximate distribution of the max-T statistic, but I will utilize the permutation method to get a null distribution.

2.2.2 Hotelling's T-Squared Test

Hotelling's T-Squared is the multivariate counterpart of the T-test. Suppose that we have vectors of functions evaluated on a grid $\boldsymbol{y}_{1i}$, $i = 1, \ldots, n_1$ from Population 1 and $\boldsymbol{y}_{2i}$, $i = 1, \ldots, n_2$ from Population 2, all with dimension p. Then, our null hypothesis becomes $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, where $\boldsymbol{\mu}_i$ now is an p-dimensional vector. If $p < n - 2$, we can use the two-sample Hotelling's T-squared test.

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)' \hat{\Sigma}^{-1} (\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2),$$

where,

$$\hat{\Sigma} = \frac{(n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2}{n_1 + n_2 - 2}$$

This test has many attractive properties, as it is UMP invariant, admissible, and robust. But there are some assumptions that must be satisfied in order for this test to be valid. First, the data must be Gaussian with a common covariance structure. That is, $\boldsymbol{y}_1 \sim N(\boldsymbol{\mu}_1, n_1^{-1}\Sigma_1)$ and $\boldsymbol{y}_2 \sim N(\boldsymbol{\mu}_2, n_2^{-1}\Sigma_2)$, with $\Sigma_1 = \Sigma_2 = \Sigma$. Also, in order for the estimated covariance $\hat{\Sigma}$ to be non-singular, we must have that $n-2 > p$, where $n = n_1 + n_2$. Under the null hypothesis,

$$F = \frac{n - k}{k(n - 1)} T^2 \sim F(k, n - k),$$

where $n = n_1 + n_2 - 1$.

2.2.3 Truncated Hotelling's T-Squared Test

Truncated Hoteeling's T-Squared Test is a modification of Hotelling's T-Squared Test. One of the conditions of the Hotellin T-Squared Test as mentioned earlier is that $p < n - 2$. This condition is not always satisfied as we often have $n \ll p$ in the functional data setting, and the equal covariance assumption may not always hold, we try to modify the ordinary two-sample T^2-statistic. In order to overcome this problem, we can truncate the data.

In the Hotelling's T-squared test, we have the following covariance matrix:

$$\hat{\Sigma} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2},$$

where S_1 and S_2 are the sample covariance matrices of the two samples. We notice that this sample covariance matrix is symmetric and non-negative definite. So we can decompose it as:

$$\hat{\Sigma} = \hat{\boldsymbol{V}}\hat{\boldsymbol{D}}\hat{\boldsymbol{V}}' = \sum_{j=1}^{p} \hat{\lambda}_j \hat{\boldsymbol{v}}_j \hat{\boldsymbol{v}}_j{}',$$

where the $\hat{\lambda}_j$'s are the eigenvalues and the $\hat{v}_j$'s are the corresponding eigenvectors. They proposed to keep only the first k components, for some $k \leq p$, and generally $k \ll p$. The adaptive Neyman methodology provides a rationale for selecting a value of k. Then, we substitute in $\hat{\Sigma}$ above with k replacing p into the T^2-statistic to obtain the test statistic:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)' \left(\sum_{j=1}^{k} \hat{\lambda}_j \hat{\boldsymbol{v}}_j \hat{\boldsymbol{v}}_j{}' \right) (\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)$$

2.2.4 Truncated Hotelling's T-Squared Test with F-Transform

Our goal is that the $c_k \tilde{T}_k^2$ all be comparable in magnitude under H_0. One way of deriving the normalizing constant ck is as follows. In the case $p \leq n - 2$, recall that

$$\frac{n - p - 1}{(n - 2)p} \sim F_{p,n-p-1}$$

In our problem we want to transform the truncated statistic $\tilde{T}_k^2$. It is obvious that transformation above will not work in the FDA setting, because the covariance term in $\tilde{T}_k^2$ is of rank $k \leq p$. However, since we have a rank k covariance matrix in $\tilde{T}_k^2$, we may try multiplying $\tilde{T}_k^2$ by

$$c_k = \frac{n - k - 1}{(n - 2)k}$$

So our test statistic will be,

$$S^* = \max_k c_k \, \tilde{T}_k^2 = \max_k \frac{n-k-1}{(n-2)k} \tilde{T}_k^2$$

2.2.5 Truncated Hotelling's T-Squared Test with χ^2-Transform

Another way of obtaining the normalization makes use of the following fact: Let $\hat{\Sigma} = \sum_{j=1}^{p} \hat{\lambda}_j \hat{\boldsymbol{v}}_j \hat{\boldsymbol{v}}_j{}'$ be the sample pooled covariance and let Σ be the true covariance matrix. In general, the sample eigenvalues $\hat{\lambda}_j$ and the sample eigenvectors $\hat{\boldsymbol{v}}_j$ of sample covariance matrix $\hat{\Sigma}$ are consistent estimators of true eigenvalues λ_j and true eigenvectors v_j, respectively, for $j = 1, \ldots, p$. Then under $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, for large sample sizes n_1 and n_2, the statistic $\tilde{T}_k^2$ is approximately distributed as a χ^2 distribution with k degrees of freedom.

It follows that $\tilde{T}_k^2$ has mean k and variance $2k$. Hence, we can normalize $\tilde{T}_k^2$ by $(\tilde{T}_k^2)/\sqrt{2k}$, so that $c_k = (1 - k/\tilde{T}_k^2)/\sqrt{2k}$. Then our test statistic becomes,

$$T^* = \max_k \frac{1}{\sqrt{2k}} (\tilde{T}_k^2 - k)$$

2.2.6 Adaptive Neyman Test

Adaptive Neyman Test is developed by Fan (1996). If there is a vague prior indicating that large absolute values of θ are located mainly on the first m component, then one would test only the first m-dimensional subproblem, leading to the test statistic: $\sum_{j=1}^{m} X_j^2$. The parameter m must be determined. Based on the power consideration, Fan (1996) proposed using:

$$\hat{m} = \underset{m:1 \le m \le n}{argmax} = \frac{1}{\sqrt{m} \sum_{j=1}^{m} (X_j^2 - 1)}$$

This leads to the adaptive Neyman test statistic:

$$T_{AN}^* = \frac{1}{\sqrt{2m}} \sum_{j=1}^{\hat{m}} (X_j^2 - 1)$$

For independent heteroscedastic error case, we assume that the error terms are independent and identically distributed with mean zero and variance σ^2. $\epsilon_j(t) \sim N(0, \sigma_1^2(t))$ and $\epsilon'_j(t) \sim N(0, \sigma_2^2(t))$ for all j and t. Consider the summarized curves:

$$\bar{X}(t) = \frac{1}{n_1}\sum_{j=1}^{n_1} X_j(t), \quad \bar{Y}(t) = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j(t)$$

and

$$\hat{\sigma}_1^2(t) = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}\{X_j(t) - \bar{X}(t)\}^2, \hat{\sigma}_2^2(t) = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}\{Y_j(t) - \bar{Y}(t)\}^2$$

Denote the standardized difference by

$$Z(t) = \frac{\bar{X}(t) - \bar{Y}(t)}{\sqrt{n_1^{-1}\hat{\sigma}_1^2(t) + n_2^{-1}\hat{\sigma}_2^2(t)}}$$

and let $\mathbf{Z} = (\mathbf{Z}(\mathbf{t}), \ldots, \mathbf{Z}(\mathbf{T}))'$. Now the Fourier transform can be applied to the standardized difference vector $\mathbf{Z}$ to compress useful signals into low frequencies. Let $\mathbf{Z}^*$ be the Fourier transform of $\mathbf{Z}$. One then can apply the adaptive Neyman test statistic to the $\mathbf{Z}^*$ vector.

With some standardization, they were able to get a asymptotic distribution for the $T^{\star}_{AN}$ and perform the test. We will not need this as we can apply the permutation methodology to obtain a null distribution.

2.2.7 Shen and Faraway (2004)

They considered a modified F test in the general functional regression setting, but the method can be specialized for this problem. Their proposed test statistic is,

$$\mathcal{F} = \frac{rss_0 - rss_1}{rss_1/(n-2)},$$

where

$$rss_1 = \sum_{i=1}^{n_1}\int (y_{1i}(t) - \bar{y}_1(t))^2 dt + \sum_{i=1}^{n_2}\int (y_{2i}(t) - \bar{y}_2(t))^2 dt,$$

$$rss_0 = \sum_{i=1}^{n_1} \int (y_{1i}(t) - \bar{y}(t))^2 dt + \sum_{i=1}^{n_2} \int (y_{2i}(t) - \bar{y}(t))^2 dt,$$

$$\bar{y}_s(t) = \frac{1}{n_s} \sum_{i=1}^{n_s} y_{si}(t),$$

$$\bar{y}(t) = \frac{n_1 \bar{y}_1(t) + n_2 \ bary_s(t)}{n_1 + n_2}$$

Assume that the t_k are equally spaced. Then we can approximate the integrals in (2.5) and (2.6) by a rectangular quadrature rule. That is, we have

$$rss_1 = \frac{1}{p} \sum_{j=1}^{n_1} \sum_{k=1}^{p} (y_{1j}(t_k) - \bar{y}_1(t_k))^2 + \frac{1}{p} \sum_{j=1}^{n_2} \sum_{k=1}^{p} (y_{2j}(t_k) - \bar{y}_2(t_k))^2,$$

$$rss_0 = \frac{1}{p} \sum_{i=1}^{2} \sum_{j=1}^{n_i} \sum_{k=1}^{p} (y_{ij}(t_k) - \bar{y}(t_k))^2 dt,$$

as approximations to the integrals above. The authors present a distributional properties of the $\mathcal{F}$ statistic and provide some results.

2.3 Proposed tests: Tests based on correlated p-values

2.4 Simulation Study

2.4.1 Simulation design

The simulation was ran based on the Cuevas et al (2004) and also .

We have considered an artifical example with $[a, b] = [0, 1]$ and 2 levels in four cases.

(M1) $m_i = t(1-t),\ i = 1, 2, 3$

(M2) $m_i = t^i(1-t)^{6-i},\ i = 1, 2, 3$

(M3) $m_i = t^{i/5}(1-t)^{6-i/5},\ i = 1, 2, 3$

(M4) $m_i = 1 + \frac{i}{50},\ i = 1, 2, 3$

Case M1 corresponds to a situation where H_0 is true; M2 and M3 provide examples, with H_0 false, of monotone functions with different increase patterns. Whereas in M2 the m_i are quite separated, in M3 the differences are less apparent (so the testing problem should be harder).Finally, M4 is an example where the functional approach is unnecessarily complicated as the functions are in fact constant.

For each choice $M1, \ldots, M4$ of the mean functions are ran under the model,

$$X_{ij} = m_i(t) + e_{ij}(t),\ j = 1, \ldots, 10$$

There are 2 different types of error values, $e_{ij}(t)$. In the first subgroup the $e_{ij}(t_r)$ are iid random variables $N(0, \sigma)$.

In the second subgroup, the $e_{ij}(t)$ is a standard Brownian process with dispersion parameter σ. These values are, for the brownian case, $\sigma_1 = 0.2$, $\sigma_2 = 1$, $\sigma_3 = 1.8$, $\sigma_4 = 2.6$, $\sigma_5 = 3.4$, $\sigma_6 = 4.2$ and $\sigma_7 = 5$. In the case of independent errors the values of the dispersion parameter are $\sigma_{k\star} = \sigma_k \times 0.45$, for $k = 1, \ldots, 6$.

The 4 models each are ran with 12 different error values. Each one ran 1,000 times. Then we count the proportion of times we get a significant value. In summary, the performance of our procedure has been evaluated in the 56 different situations obtained by combining the underlying models M1, M2, M3, M4 the error structure (independent or brownian) and the value of the parameter σ.

2.4.2 Simulation results and discussion

Table 1: The number of rejections for 1000 runs for different statistical tests using 4 different model types, using the independent error structure.

Test		$\sigma_1^\star$	$\sigma_2^\star$	$\sigma_3^\star$	$\sigma_4^\star$	$\sigma_5^\star$	$\sigma_6^\star$
$\max -T$	M1	0.052	0.058	0.052	0.05	0.049	0.05
	M2	1	0.808	0.421	0.233	0.089	0.066
	M3	1	1	1	1	0.991	0.53
	M4	0	0	0	0	0	0
Truncated with F	M1	0.049	0.048	0.04	0.05	0.052	0.051
	M2	0.4	0.246	0.086	0.063	0.026	0.019
	M3	0.4	0.4	0.385	0.293	0.059	0.028
	M4	1	1	1	1	1	1
Truncated with Chi	M1	0.045	0.044	0.048	0.053	0.044	0.053
	M2	0.948	0.383	0.187	0.106	0.08	0.063
	M3	1	0.984	0.791	0.503	0.144	0.082
	M4	1	1	1	1	1	1
Cuevas et al, homo	M1	0.002	0.005	0.004	0	0.005	0.004
	M2	1	0.789	0.2	0.065	0.011	0.001
	M3	1	1	1	0.892	0.094	0.007
	M4	1	1	1	1	1	1
Cuevas et al, hete	M1	0.006	0.006	0.012	0.009	0.012	0.01
	M2	1	0.87	0.337	0.139	0.021	0.012
	M3	1	1	1	0.957	0.17	0.029
	M4	1	1	1	1	1	1
Shen and Faraway	M1	0.007	0.0012	0.004	0.01	0.007	0.004
	M2	1	0.845	0.292	0.105	0.021	0.007
	M3	1	1	1	0.942	0.153	0.024
	M4	1	1	1	1	1	1
Gorecki and Smaga	M1	0.048	0.059	0.05	0.052	0.053	0.054
	M2	1	0.979	0.697	0.417	0.11	0.074
	M3	1	1	1	0.995	0.42	0.137
	M4	1	1	1	1	1	1

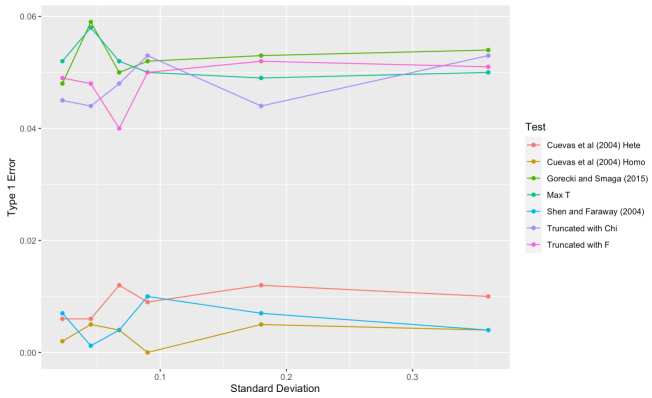

Figure 1: The number of rejections for 1000 runs for different statistical tests using 4 different model types, using the independent error structure.

Table 2: The number of rejections for 1000 runs for different statistical tests using 4 different model types, using the Brownian pendent error structure.

Test		$\sigma_1^\star$	$\sigma_2^\star$	$\sigma_3^\star$	$\sigma_4^\star$	$\sigma_5^\star$	$\sigma_6^\star$
$\max -T$	M1	0.05	0.046	0.057	0.059	0.055	0.044
	M2	1	0.919	0.603	0.398	0.145	0.084
	M3	1	1	1	0.978	0.356	0.115
	M4	0	0	0	0	0	0
Truncated with F	M1	0.025	0.011	0.021	0.028	0.026	0.024
	M2	0.089	0.005	0.001	0	0	0
	M3	0.4	0.169	0.021	0.008	0	0
	M4	1	1	1	1	1	1
Truncated with Chi	M1	0.056	0.054	0.053	0.062	0.054	0.047
	M2	0.918	0.321	0.163	0.127	0.059	0.062
	M3	1	1	1	0.994	0.613	0.288
	M4	1	1	1	1	1	1
Cuevas et al, homo	M1	0.053	0.038	0.038	0.024	0.035	0.0051
	M2	0.371	0.075	0.042	0.049	0.033	0.0464
	M3	1	0.569	0.159	0.081	0.05	0.049
	M4	1	1	1	1	1	0.996
Cuevas et al, hete	M1	0.055	0.058	0.065	0.068	0.064	0.061
	M2	0.476	0.094	0.076	0.07	0.07	0.071
	M3	1	0.64	0.197	0.128	0.092	0.078
	M4	1	1	1	1	1	0.998
Shen and Faraway	M1	0.05	0.045	0.044	0.055	0.047	0.046
	M2	0.359	0.083	0.061	0.064	0.045 0.042	
	M3	1	0.551	0.162	0.083	0.051	0.053
	M4	1	1	1	1	1	0.997
Gorecki and Smaga	M1	0.046	0.048	0.053	0.059	0.061	0.053
	M2	0.45	0.089	0.071	0.058	0.056	0.05
	M3	1	0.654	0.16	0.086	0.057	0.044
	M4	1	1	1	1	1	0.996

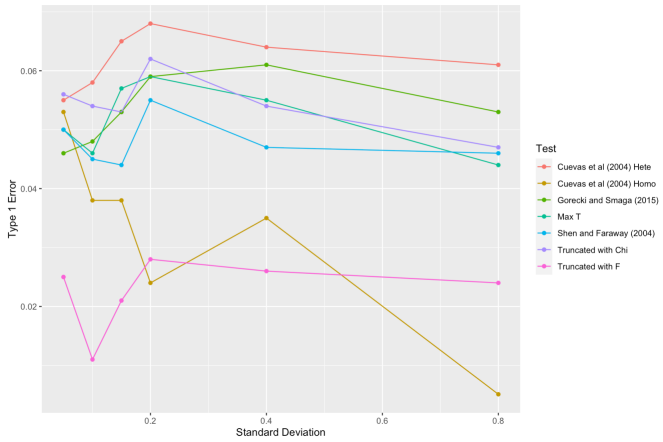

Figure 2: The number of rejections for 1000 runs for different statistical tests using 4 different model types, using the brownian pendent error structure.

2.5 Application

3 Modeling Physical Activity Patterns in African Americans Using Functional Principal Component Analysis

Reference

Xu, S.Y., Nelson, S., Kerr, J. et al. Modeling Temporal Variation in Physical Activity Using Functional Principal Components Analysis. Stat Biosci 11, 403–421 (2019).

3.1 Introduction

Xu and et al (2019)

They implemented a functional principal components mixed model which uses dimension reduction via principal components to model the minute-level activity counts. Furthermore, unlike the approach in [28] which averaged an individual's daily activity records, the use of mixed models allows inclusion of repeated daily records in the model.

4 Modeling Physical Activity Patterns in African Americans Using Functional Mixed Models

4.1 Introduction

We apply the Functional Data Analysis (FDA) to fit the physical activity data measured by the accelerometer. FDA is capable of handling data that discretely collected from either fixed or random time grid, and the time grid can be dense, sparse or neither, which allow us here to largely explore the abundant information from the accelerometer data. In our analysis, we summarize the data into one-hour epoch by aggregating measurements originally recorded in 15-seconds epochs, providing 24 observations per day individually.

The main characteristic of FDA is that the vector of high resolution data is modelled as a unique functional object typically defined in terms of a spline basis. This functional object provides a mathematical framework which enables exploratory data analysis and inference to be performed using statistical techniques analogous to standard multivariable methods (e.g. ANOVA, regression models, principal component and cluster analysis).

Ullah and Finch (2013)

Smoothing is the first step in any FDA, and its purpose is to convert raw discrete data points into a smoothly varying function. This emphasizes patterns in

the data by minimizing short-term deviations due to observa- tional errors, such as measurement errors or inherent system noise. Overall, B-spline smoothing was the most popular smoothing technique used (25 papers), presumably be- cause of its simplicity and flexibility for tackling a wide range of nonparametric and semiparametric modeling sit- uations. A common approach towards B-spline smoothing is to construct a large number of knots (as the smoothing parameter) to reduce the effective degrees of freedom and increase smoothness in the overall function estimate. Other smoothing techniques adopted in the published studies included use of Fourier smoothing (8 papers), regression splines (6), kernel smoothing (7) and so on.

Ramsay and Silverman [9] emphasize that the choice of smoothing technique is dependent upon the underlying behavior of the data being analyzed. Ideally, the smoother should reflect or have features that match those of the data. For example, Fourier smoothers are traditionally used when the data are cyclical or periodic. Environmental diurnal ozone and NOx cycles [71,116], trends in ecologic- ally meaningful water quality variates in ecology [81], cash flows in finance [92] and fetal heart rate monitoring in medicine [18,19] are examples of the application of Fourier smoothers. Splines (regression splines, polynomial splines, B-spline) are typically chosen to represent noncyclical nonperiodic data [25,51,84], and wavelet bases are chosen to represent data displaying discontinuities and/or rapid changes in behavior

Keser(2015)

As Turkey is surrounded by three seas and the elevation increases from west to east, both the ge- ographical situation and the landform affect humidity. According to this, it is expected that humidity curves of coastal areas and hinterlands differ. For this aim, humidity curves of 35 cities, whose data for 2000-2010 years are complete, were separated into two groups as coastal areas and hinterlands. In order to see if two curve groups have the same functional curve, in other words, in order to test if

humidity mean functions of coastal areas and hinterlands are statistically different, functional t-tests, which were created on the basis of permutation tests and Westfall and Young approach, were used simultaneously to improve the validity of results.

In this study, it is found out that using either Fourier or B-Spline basis for modeling the differences does not have any considerable effect on p-values. However, smoothing the data makes interpretation easier for both approaches. Therefore, using basis functions for smoothing instead of interpolation can be suggested.

For the analysis of the data, they used functional t-tests.

Fan and et al (2015)

Now we outline the notation and basic ideas of FDA as follows. Consider for each individual $i = 1, \ldots, N$, physical activity data are recorded at timepoints $t_j, j = 1, \ldots, l$. We denote the physical activity measurements as $\boldsymbol{Y} = \boldsymbol{y}(t)$, for the individual i, the physical activity profile can be represented as $\boldsymbol{Y}_i = (y_i(t_1), y_i(t_2), \ldots, y_i(t_l))'$, which is a function of time t. The activity profile $\boldsymbol{y}_i(t)$ of individual i is a function of time t, which can be estimated by Y_i.

To estimate the activity function $y_i(t)$ from the activity counts Y_i, we use an ordinary linear square smoother. Specifically, let $\phi_k(t), k = 1, \ldots, K$ be a series of K basis functions, such as B-spline basis functions and Fourier basis functions, $\boldsymbol{\phi}$ is a N by K matrix containing basis function values $\phi_k(t_j)$. Then we can estimate the physical activity profile from the basis system with least-square smoother that is,

$$\hat{\boldsymbol{y}}_i(t) = \boldsymbol{\phi}(\boldsymbol{\phi}'\boldsymbol{\phi})^{-1}\boldsymbol{y}_t(t),$$

where $\boldsymbol{\phi}(t) = (\phi_1(t), \ldots, \phi_K(t))'$. The estimate $\hat{y}_i(t)$ smoothes activity patterns overtime.

By building up the functional object, we have the smoothed activity profile for each individual over time. The Fourier basis functions is considered for the periodic nature of physical activity data, B-spline basis functions is also applicable as imple-

mented by other study. By building up the functional object, we have the smoothed activity profile for each individual over time. The Fourier basis functions is considered for the periodic nature of physical activity data, B-spline basis functions is also applicable as implemented by other study.

In the paper by Fan and et al (2015) they considered the Fourier basis functions: $\phi_0(t) = 1$, $\phi_{2r-1}(t) = \sin(2\pi rt/N)$ and $\phi_{2r}(t) = \cos(2\pi rt/N)$ for $r = 1, \ldots, (K-1)/2$, where K is taken as a positive odd integer. They claimed that one may use B-spline basis functions, but the activity data are likely to be periodic and Fourier basis functions make more sense.

To test if there are activity differences among the three grade adolescent girls, we use a functional version of the univariate F-statistic. The functional version of the univariate F-statistic is defined as follows:

$$F(t) = \frac{Var\{\hat{y}(t)\}}{\sum\{y_i(t) - \hat{y}(t)\}^2/n},$$

where $\hat{y}(t)$ are the predicted values from the functional linear model.

Sera and et al (2017)

In this study we applied FDA to model daily profiles of PA in a large sample of seven year old children. Subsequently, we used functional analysis of variance (FANOVA) to examine time and place of measurement, demographic and behavioural characteristics that may explain the variability of daily PA profiles. Our overall objective was to understand temporal patterns of PA according to characteristics of the child, their family and wider environment in order to inform public health interventions designed to increase activity levels in primary school aged children. In this analysis we considered data only from singleton children who wore the accelerometer for at least ten hours a day from 7:00 to 22:00.

The first step is to model observed accelerometer data by means of latent smooth

functions $y_i(t)$ assumed to be smooth over time t. The linear predictor

$$y_i(t) = \boldsymbol{\phi}'(t)\boldsymbol{c}_i,$$

where, c_i is a $k \times 1$ vector of coefficients, and $\boldsymbol{\phi}(t)$ is a k–dimensional basis function system as our starting point.

Within this framework, there are several basis functions $\phi_k(t)$ among which one can choose, e.g., Fourier, exponential,truncated power functions, orthogonal polynomials and splines. Here we considered spline functions given their computational efficiency. Among these we used a fourth order cubic B–spline basis function system. Numerically, B–splines are attractive because they require an amount of computation that increases linearly with the number of observations. [17] Another desirable property of cubic splines is that they are the smoothest possible interpolant through any set of data. [28] This property implies that estimated cubic splines yield the interpolant function that minimize the curvature (i.e. the integral for the second derivative) of the objective function.

Then the paper uses the functional ANOVA in order to analyze the relationship between covariates.

Lin and et al (2022)

To account for the hierarchical structure of the data (visits within subjects) and its longitudinal nature in both predictors (PA) and health outcomes, we applied a longitudinal FPCA model to decompose densely sampled PA data, and a (functional) mixed effects regression model to explore the association between predictors and outcomes.

4.2 Activity Level Data

From the papers above, it can be seen that the in order to analyze the physical activity data at hand, we first need to transform the data. In literature, the most

commonly used transformations are:

- Fourier Transform
- B-spline
- Wavelet Transform

Keser (2014) did the analysis using both Fourier and Wavelet. They reported that the initial transformation did not change the results of the study. Therefore, it is a good idea to try all of the transformations on the data.

Initially, I will take the step count as the response variable measured hourly, between 07/06/2016 to 07/13/2016. Therefore there is a total of 169 observations.

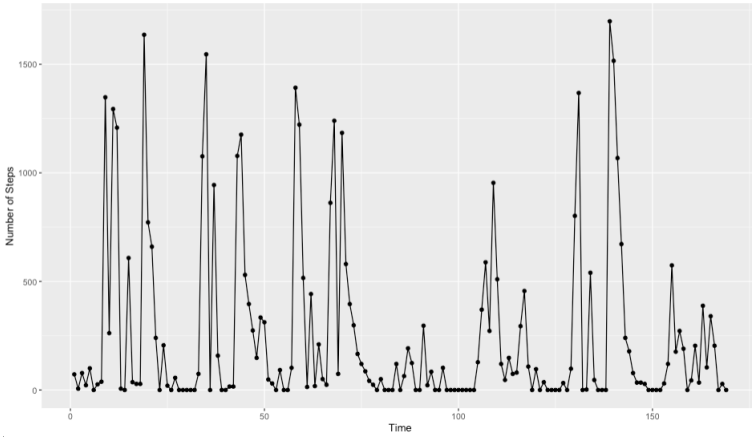

Figure 3: Number of steps taken each hour between 07/06/2016 to 07/13/2016.

Note: In most of the literature, ones included above, restrict their analysis to measurements taken during the day. Most of the analysis is done between 6-7am and 11pm. The entire data is not included.

4.3 Fourier Transform

In order to transform the data above using Fourtier transform, I initiialy used to built in function fft, fast discrete fourier transform. Then I used the mod function to ger the magnitude of each Fourier transform.

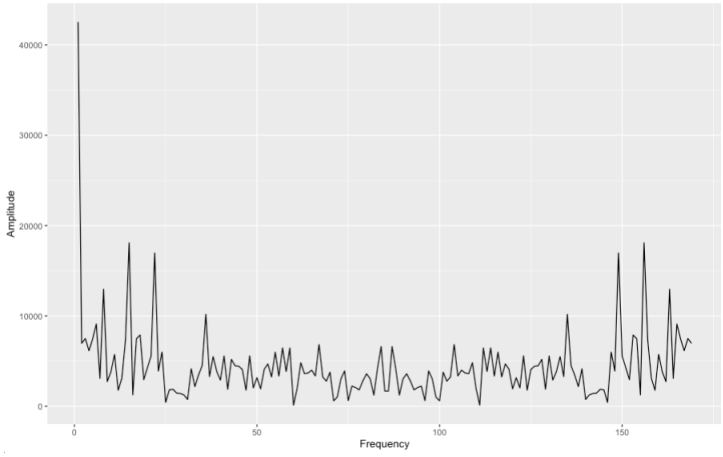

Figure 4: Fourier transform of the number of steps taken each hour between 07/06/2016 to 07/13/2016.

4.4 B-Spline

4.5 Wavelet Transform

References