VILNIUS UNIVERSITY
FACULTY OF MATHEMATICS AND INFORMATICS
MODELLING AND DATA ANALYSIS
MASTER'S STUDY PROGRAMME

TWO SAMPLE HYPOTHESIS TESTING FOR FUNCTIONAL DATA. CASE STUDY

Master's Thesis

Author: Laura Jankovskytė
VU email address: laura.jankovskyte@mif.stud.vu.lt
Supervisor: Prof. Dr. Habil. Alfredas Račkauskas

Vilnius

2024

Abstract

In this study, functional data analysis methods are employed to examine viewership data for Lithuanian men and women. The objective is to find out if there are differences in television viewing habits between the two genders. To achieve this objective, we conducted simple pointwise Z test, L^2-Norm-Based test, and the widely used MMD-Based test, for the entire time interval and specific smaller intervals. The results showed that we should be careful when choosing tests to perform: despite all tests rejecting the null hypothesis that the means of two samples are equal, it is noteworthy that the MMD-Based test does not reject the null hypothesis during early morning and late evening time intervals. Additionally, new samples of weekly differences were constructed. Furthermore, a noval approach of constructing coefficient functions for pointwise AR(1) models was tried. The variance of AR(1) models residuals were analyzed. The application of AR(1) approach indicated potential similarities in television viewing habits between men and women.

Keywords: Functional data, television, viewership, hypothesis testing.

1 Introduction

As the world presents numerous opportunities, and the goods market continues to expand relentlessly, marketing and media play a huge role in not only businesses but people's lives as well. Even though digital media is taking over the ether of advertising, offering advantages such as narrow targeting, low cost per reached person, and so on, television (TV) still remains one of the main mass media channels with the largest reach capabilities. It is not surprising, since TV has a lot of benefits for advertisers and for viewers as well, as noted by [1]: TV provides sponsorship opportunities for advertisers seeking to reach loyal viewers; TV makes an impact because of its ability to make an emotional impact while delivering a message visually and audibly; and most importantly, people can watch television at no or low cost, which makes TV a mass media channel.

Given all those benefits, the need to analyze who the viewer is and what viewing habits he has arises. For example, Katie J. Damratoski, April R. Field, Katie N. Mizell, and Michael C. Budden in their paper *An Investigation Into Alternative Television Viewership Habits of College Students* [2] analyzed the television viewing habits of college students, using Nielsen data from 2009 in the United States. Dilek Altas and Hakan Oztunc in their paper *An Empirical Analysis of Television Commercial Ratings in Alternative Competitive Environments Using Multinomial Logit Model* [3] compare the first minute of advertisement breaks of four Turkish national TV channels to determine if there are significant differences among viewership of those channels, taking into account features such as age, gender, and socio-economic status. The Multinomial Logit model was chosen as the main method of analysis. Also, many papers are written that focus on television ratings forecasting. However, in this work, we are rather focusing on viewing habits than forecasting.

The aim of this work is to analyze whether or not men and women have the same viewing habits using total television viewership ratings in Lithuania, treated as functional data. To do that, two sample hypotheses for functional data can be applied. Many tests for problems like this can be found in the literature. The basic approach of the Z-test and more advanced L^2-Norm-Based test for functional data are described in the book *Analysis of Variance for Functional Data* by Jin-Ting Zhang [4]. In the paper *A Two Sample Test for Functional Data* by Jong Soo Lee, Dennis D. Cox, and Michele Follen [5], the noval test based on the adaptive Neyman methodology applied to Hotelling's T-squared statistics is proposed. Furthermore, Miroslaw Krzysko and Lukasz Smaga in their paper *Two-sample Tests for Functional Data Using Characteristic Functions* [6] use characteristic functions and a basis function representation. The authors constructed test statistics to check the equality of distributions based on a weighted distance between characteristic functions of random vectors obtained in basis representation. However, the most commonly used method in the literature is based on the Maximum Mean Discrepancy (MMD). For example, in the paper *A Kernel Method for the Two-Sample Problem*, written by Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Scholkopf, and Alexander Smola [7], authors used the largest difference in expectations over functions in the unit ball of a reproducing kernel Hilbert space as a test statistic and established three different tests for comparing two distributions.

In this work, two approaches are used to test whether or not men and women have the same viewing habits. By one method, let us call it traditional hypothesis testing, we applied a simple pointwise Z test, a L^2-Norm-Based test, and a commonly met in the literature MMD-Based test to test whether or not men and women samples come from equal distributions. We also divided the time interval into 4 different smaller intervals and performed these tests for those intervals to see how the tests react during a specific time of the day. Another approach, which could not have been found in the literature, was to compare coefficients obtained from autoregression of order 1 for data on weekly differences. Then, the coefficients and variance of residuals of autoregression models were compared for men and women data.

At the end of this work, results from both approaches are discussed and compared. Furthermore, recommendations for further analysis are suggested.

2 Dataset description

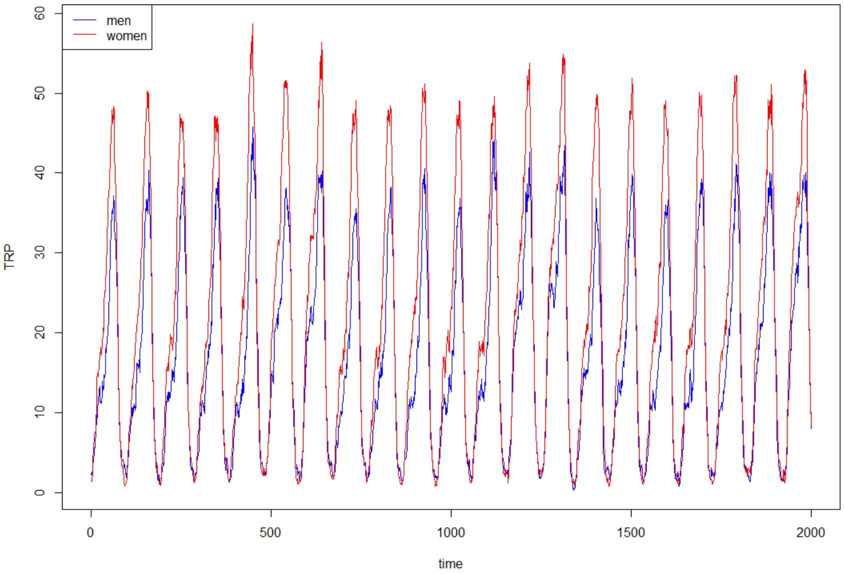
The dataset used in this paper contains Lithuanian television general viewership of men and women. The viewership is expressed as ratings and called Target Rating Points (TRP). One rating is one percent of a potential television viewer in a defined target audience. In this case, the defined target audiences are men and women from 4 years of age.

The data was collected by the market research and consulting company TNS LT. The research on television viewership includes more than 1200 respondents from 4 years old. In randomly selected households, TV meters are installed, which keep track of how long and how frequently every viewer in the household watches TV (even certain TV channels). All the households have been selected according to demographics and technical criteria so that they represent the Lithuanian residents who have TV sets. The results are processed in such a way that they represent Lithuanian population in terms of gender, age, nationality, place of living, number of household members, spoken language, TV devices at home, and TV signal receiving type [8].

A TV meter is a small electronic device that is connected to every TV in the selected household. Every member of the household must register with a remote control when watching TV. If it is not done, the TV meter always reminds you to register. The TV meter records which member is watching what channel and for how long. Every night, information from TV meters using the GSM (Global System for Mobile Communication) connection is delivered to the company, and the data is cleaned and processed. For example, very long continuous viewing of the same channel by the same person indicates that person left TV on while doing something else. Outliers like these are removed from the data. After the data is processed, it is delivered to media agencies, media channels, and other interested subjects.

In this thesis, television viewership data are treated as observations of random curves. Data of 2021, taken as 15 minutes average, is used. It is worth mentioning that, in television terms, day starts at 5 a.m. and ends at 4:59 a.m. To avoid one program being on two days (during midnight), the day ends and starts when usually no program is on - from 5 a.m. The first 2000 time points of primary data can be seen in Figure 1. We can see clear periodicity in the data: during the night, ratings are usually close to zero, and during the evening, they reach their peak. This is valid for both men and women's data.

Figure 1: Primary data: TV viewership of men and women.



3 Primary data

3.1 Smoothing

From the first sight of the data (Figure 1), we can see that women's watching is higher than men's during peak time. Also, since the data are periodic with an interval of one day, we want to reconstruct the data in such a way that one function represents one day. After reconstruction, in total, there are 371 functions measured at 96 time points. Reconstructed curves can be seen in Figure 2. When curves overlap with each other, no obvious difference between women and men in viewership can be seen.

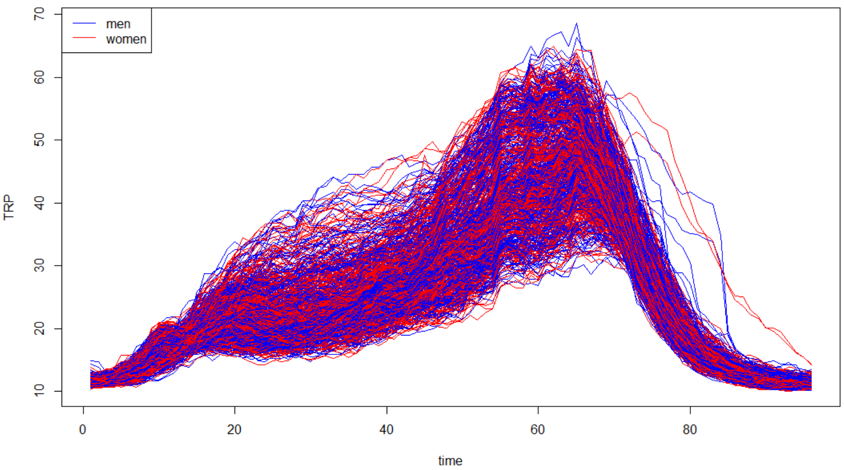
To analyze the data, the Functional Data Analysis (FDA) approach is used. Since the data is now in discrete form and probably with some measurement errors, first of all, we need to smooth the data. As explained by Jin-Ting Zhang [4], smoothing techniques allow to reconstruct discrete data in such a way that any reconstructed function could be represented at any time point and measurement errors removed as much as possible.

One of the most popular smoothing methods is basis functions expansion. The Fourier basis system is the usual choice for periodic functions, and the spline basis system, especially B-splines, works well with nonperiodic functions.

The set of basis functions $\phi_k, k = 1, ..., K$ is used, which are combined linearly. That means a function $x(t)$ can be expressed in a way:

$$x(t) \approx \sum_{k=1}^{K} c_k \phi_k(t) = \boldsymbol{c}' \boldsymbol{\phi(t)}, \quad t \in \tau.$$

Figure 2: Reconstructed data. One curve represents one day.



This expression is called basis function expansion. The parameters $c_1, c_2, ..., c_K$ are the coefficients of the expansions; vector $\boldsymbol{c}'$ is vector of K coefficients, $\boldsymbol{\phi}$ is vector of length K containing the basis functions; τ is time interval, usually $[0, 1]$ or $[0, T]$, or any other interval $[a, b]$. In our case, we leave the original time interval of a day divided into 15-minute intervals, so in total we have 96 time points. So $\tau = [1, 96]$. If we have a sample of N functions, then we have

$$x_i(t) \approx \sum_{k=1}^{K} c_{ik}\phi_k(t), \quad i = 1, ..., N, \quad t \in \tau,$$

or

$$\boldsymbol{x}(t) \approx \boldsymbol{C}\boldsymbol{\phi}(t), \quad t \in \tau,$$

where $\boldsymbol{x}(t)$ is a vector of length N containing the functions $x_i(t), i = 1, ..., N$ and $\boldsymbol{C} = (c_{ik}), \quad i = 1, .., N, \quad k = 1, ..., K$ - matrix of coefficients with N rows and K columns.

Splines are piecewise polynomials. Even though there are many different kinds of splines, for example, M-splines, I-splines, and so on, we consider and use only B-splines in this case. B-splines are also the most popular type of splines in other analysis [9]. B-splines are described by their main three characteristics: range, knots, and order. Range is simply the interval on which observations lie; in our case, the time interval τ, [1, 96]. Splines are constructed by dividing the interval of observations into smaller intervals, with boundaries at points called break points. Over any smaller interval, the spline function is a polynomial of fixed order; the most commonly used order is 4. The term degree is used to refer to the highest order of polynomials. The order of a polynomial is usually one higher than its degree, basically because there is a constant term as well. So if order is 4, then we have the highest degree of polynomials 3, which are cubic polynomials. The third characteristic, which is knots, is connected with the break points mentioned above. Every knot has the same value as a break point, but there might be multiple knots at one break point. At each break point, polynomials from the left and from the right are constrained to have a certain number of matching derivatives. The number of derivatives that must match is determined by the number of knots existing at that break point. By default, and in the vast majority of applications, there is only one knot at every break point except for the boundary values at the end of the range t. The end points have as many knots as the order of splines, meaning that the function value typically drops to zero outside the interval τ.

In our case, since the daily functions are not periodic (Figure 1), we will go with the most common case that fits the majority of the analysis: we will use B-splines and cubic polynomials (order 4). According to [9], useful rule to remember is to fix the order of the spline basis to be at least two higher than the highest-order derivative to be used. By this rule, given that we do not need to look at any derivatives, the cubic spline basis is a good choice. To smooth the data, we also need to define the number of basis functions to use. In a spline basis system, it is defined as: *number of basis functions = order + number of interior knots.* Interior knots mean knots that are placed at break points that are not at the beginning or end of the τ range. Also, we will use equally spaced knots.

We should remember that we want to keep the number of basis functions as small as possible relative to the amount of data. To choose the number of basis functions (more precisely, the number of interior knots), we rely on generalized cross-validation (GCV) and Elbow methods. GCV can be written as:

$$GCV = \frac{n * SSE}{(n - df)^2},$$

where n is the number of data points, SSE is the error sum of squares, and df is the degree of freedom measure of the smooth. In Table 1, GCV for different numbers of basis functions can be seen.

According to these methods, 9 basis functions would be most suitable for men and women data. This means, given that we take cubic splines, 5 equally spaced internal knots are chosen. As suggested by P. Hall and I. Van Keilegom in [10], exactly the same tuning parameters should be used to produce smoothed functions from raw data in both datasets, because different parameters can mask the differences between distributions and reduce the power of two sample tests that we aim to perform later. In this case, our choice of 9 basis functions for both samples should be good for smoothing.

Even though B-splines basis functions expansion is the best choice for nonperiodic functions, it performs quite poorly near the beginning or end of the range interval. The reason is that in these regions we run out of data to define them, so at the boundaries, spline function values are entirely determined by a single coefficient. In our case, we see that smoothed functions go below zero at the end. Logically, television viewership cannot be below zero. Many methods can be performed to avoid smoothed functions going below zero and performing better at the ends, but we chose to shift the measurement axis by 10, which means we added 10 to each observation to

Number of basis functions	GCV for men sample	GCV for women sample
4	25.136	42.545
5	24.748	39.194
6	7.756	10.045
7	5.923	11.999
8	5.106	7.208
9	1.999	2.546
10	1.617	2.792
11	1.539	2.677
12	1.361	1.678
13	1.132	1.288
14	1.002	1.345
15	0.962	1.327
16	0.954	1.064
17	0.853	1.004
18	0.857	1.02
19	0.78	0.922
20	0.776	0.871

Table 1: GCV for different numbers of basis functions.

avoid smoothed functions going below zero. This transformation is just a formality and is naturally repealed completely when performing further analysis.

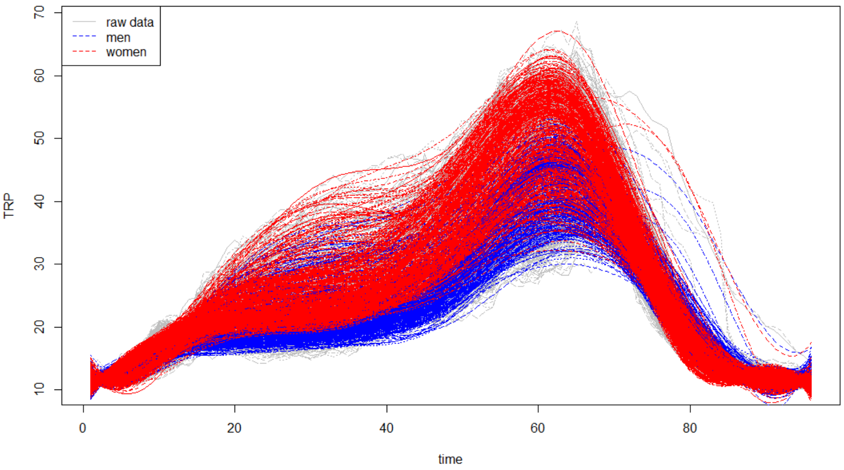
Figure 3 represents functions obtained after smoothing. Now we can clearly see that women's functions are a little bit higher than men's. However, this may not be because of different television viewing tendencies, but because there are more women who watch television than men. When considering the form of the functions, it looks like men and women functions are very similar: close to zero early in the morning, then rise little by little in the morning and during the daytime while more suddenly go up in the evening and reach peak around 65 time point (usually at this time evening news is shown on TV), and then descend quickly while reaching close to zero at night. However, since functions overlap with each other, we will see the similarities of the shapes more precisely later.

3.2 Exploratory analysis

Looking at Figure 3, we can notice that some functions are further away from others, especially at the end of the range interval. That indicates that there might be some outliers in the data. It is important to deal with outlying functions because those outliers might be outstanding days where television viewing habits were not usual, for example, national holidays, when the program changes completely and people's habits are different as well.

Outliers in functional data might be of a different nature. For example, an observation can be shifted from the other data. According to [11], outliers like these in FDA are called magnitude outliers. Another type of outliers is shape outliers - functions that are different in shape. For

Figure 3: Smoothed data. One curve represents one day.



periodic data, an outlier can be a function that has a different amplitude than other functions. Thus, identifying outliers in the data can be quite a difficult task.

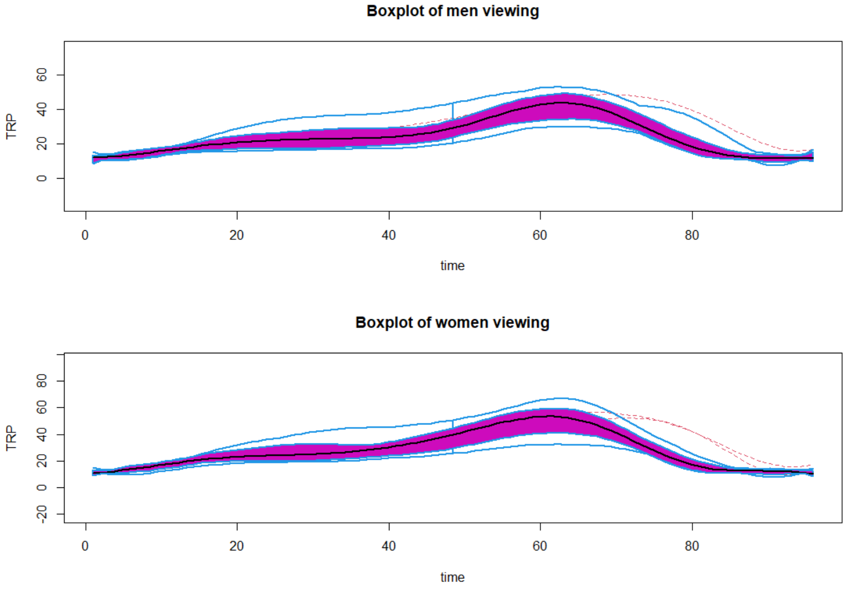
In our case, we chose to identify outliers according to the band depth. Band depth is a concept that measures how deep (or central) a function is compared to a population. In univariate data, the median would be the deepest point. The aim of depth for functional data is to measure the centrality of a given curve within a set of curves [12]. To measure outliers by depth, we use the *boxplot* function for functional boxplots in R [13]. How boxplots of men and women viewing functions look can be seen in Figure 4. The black line indicates the deepest curve, the pink plot denotes the envelope of 50% of the deepest curves, the blue line indicates the maximum non-outlying envelope, and dotted red lines are outliers [13]. Men data have one outlier, while women data have two outliers, according to the boxplots. Both datasets share a common outlier of the 362nd function, which is the 31st of December, Saturday. In the women dataset, the 139th function, which is the 22nd of May, Sunday, is also an outlier. While the 31st of December being an outlier is not surprising at all, since during New Year's Eve most of the channels change TV programs from the usual ones completely, the 22nd of May does not have an explanation as being an exclusive day. To avoid these outliers affecting later results, the data for outlier days is changed to the same day as the previous week. 31st of December data is changed to 24th of December, and 22nd of May is changed to 15th of May. After this transformation, outliers are no longer visible.

Now we can see what other descriptive statistics, like mean and variance, look like. An unbiased estimator of the mean function can be simply defined as

$$\hat{\mu}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t), \quad t \in \tau,$$

where $x_i(t)$ is our n functions sample. Also, the estimator of variance function can be defined as

Figure 4: Boxplots of men and women viewing functions.



$$\hat{\sigma}_x^2(t) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i(t) - \hat{\mu}(t))^2, \quad t \in \tau,$$

and the standard deviation function is simply $sd_x = \sqrt{\hat{\sigma}_x^2}$.

In Figure 5, sample means and +- 2 standard deviations form the mean can be seen. The blue dashed line indicates the sample mean function, while the red ones - +- 2 standard deviations functions from the mean. Now we can more clearly see the shape of the functions. As already discussed in the previous section, the shape of the mean is quite similar for both datasets, although the scale of ratings (TRP) axis differs a bit. This indicates that men and women watch television very similarly during the day. The difference can be noticed when looking at an interval of +- 2 standard deviations from the mean. Men data have a higher deviation during the morning and daytime.

In Figure 6 we can see covariance functions visualized as levelplots of both samples. It can be noticed that variance for both samples (diagonal) are highest during evening peak time. Also, both covariance functions look very similar in shape and that is a good result for later hypothesis testing.

3.3 Two sample tests for mean functions

Based on the Central Limit Theorem and the fact that the data already comes accumulated from the data of many households, we can make the assumption that both samples are approximately normal. Given the normality assumption and large sample sizes, various tests for two sample

Figure 5: Sample mean and +- 2 standard deviations from the mean of men and women functional data.

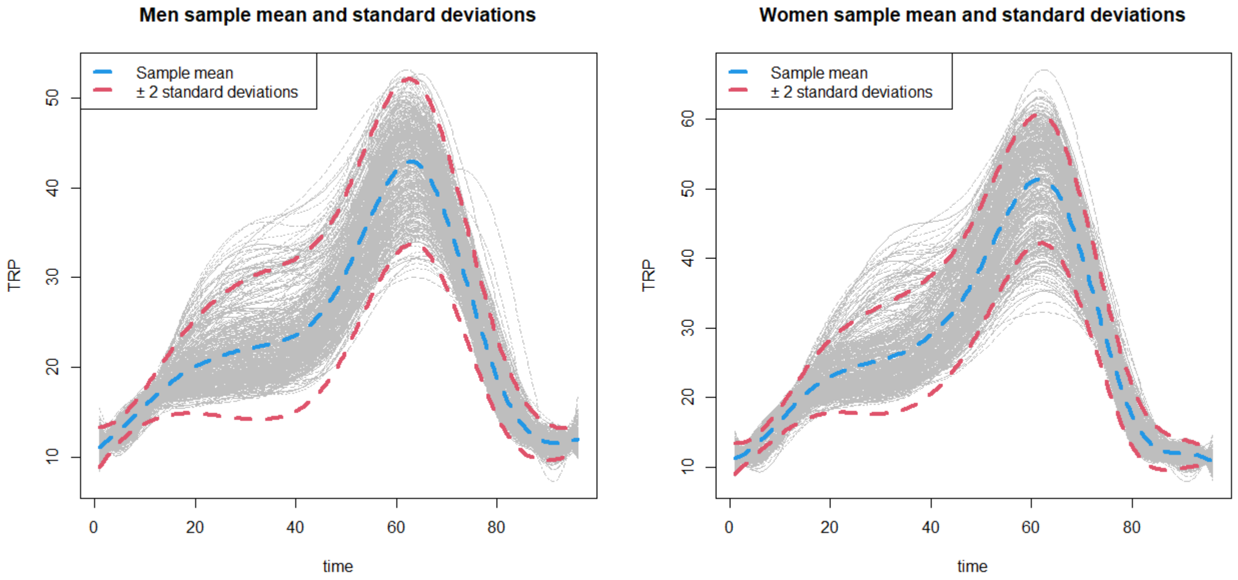
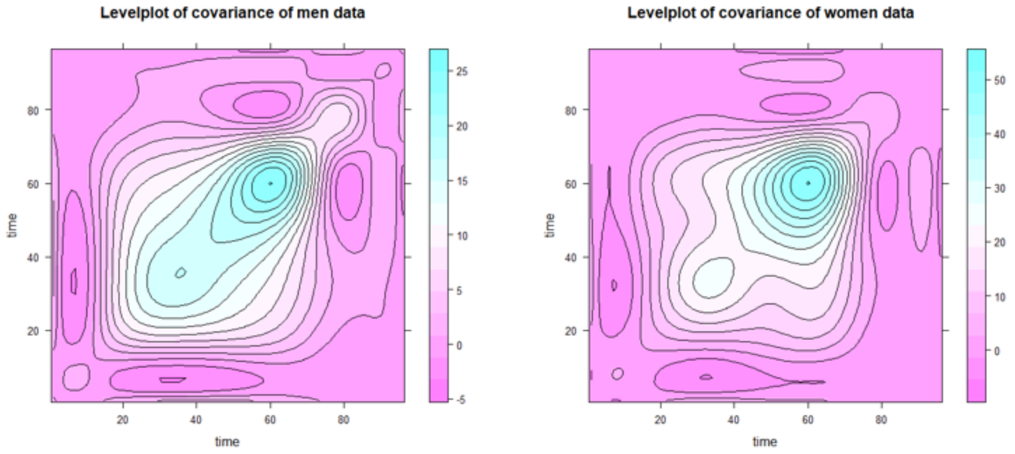


Figure 6: Levelplots of covariance functions of men and women samples.



problems can be constructed. We refer to the mean function of men's data as $\mu_m(t)$ and the mean function of women's data as $\mu_w(t)$. We wish to test:

$$\begin{aligned} H_0: &\quad \mu_m(t) = \mu_w(t), \quad t \in \tau, \quad \text{versus} \\ H_1: &\quad \mu_m(t) \neq \mu_w(t), \quad \text{for some} \quad t \in \tau, \end{aligned}$$

where $\mu_m(t)$ and $\mu_w(t)$ are mean functions of men and women samples respectively as already mention above and t is time point from the samples time period τ (in our case [1, 96]). If both samples come from the same distribution, that means that viewing habits of men and women are the same.

Pooled sample covariance function estimate is:

$$\hat{\gamma}(s,t) = (n-1)^{-1} \sum_{k=1}^{2} \sum_{j=1}^{n_k} [x_{kj}(s) - \hat{\mu}_k(s)][x_{kj}(t) - \hat{\mu}_k(t)], \quad s,t \in \tau,$$

where $n = n_m + n_w$ is total sample size of the two samples.

First, we perform the simple pointwise Z test for the two sample problem described in [4]. The idea of a pointwise test is to test the hypothesis at each time point t. The pivotal test function is:

$$z(t) = \frac{[\mu_m(t) - \mu_w(t)]}{\sqrt{(1/n_m + 1/n_w)\hat{\gamma}(t,t)}} = \frac{\Delta(t)}{\sqrt{\hat{\gamma}(t,t)}}, \quad t \in \tau,$$

where

$$\Delta(t) = \sqrt{n_m n_w / n}[\mu_m(t) - \mu_w(t)], \quad t \in \tau, \tag{1}$$

is the natural pivotal test function, which is the scaled mean function difference of the two samples. Estimators of these functions are given by the naive method (plug-in rule) if the sample sizes are large enough. If two samples are approximately Normal and sample sizes are large, then $z(t) \sim t_{n-2}, \quad t \in T$. Then, we can conclude that H_0 must be rejected when $|z(t)| > t_{n-2}(1-\alpha/2)$ or based on p-values that can be computed according to t-distribution. $t_{n-2}(1-\alpha/2)$ is $100(1-\alpha/2)$-percentile of the t-distribution with $n-2$ degrees of freedom.

The result of this test can be seen in Figure 7. Most of the regions do not fall into the critical value region, so for most of the time points, the null hypothesis is rejected. The only regions where the null hypothesis is not rejected are early mornings and late nights, where ratings are close to zero.

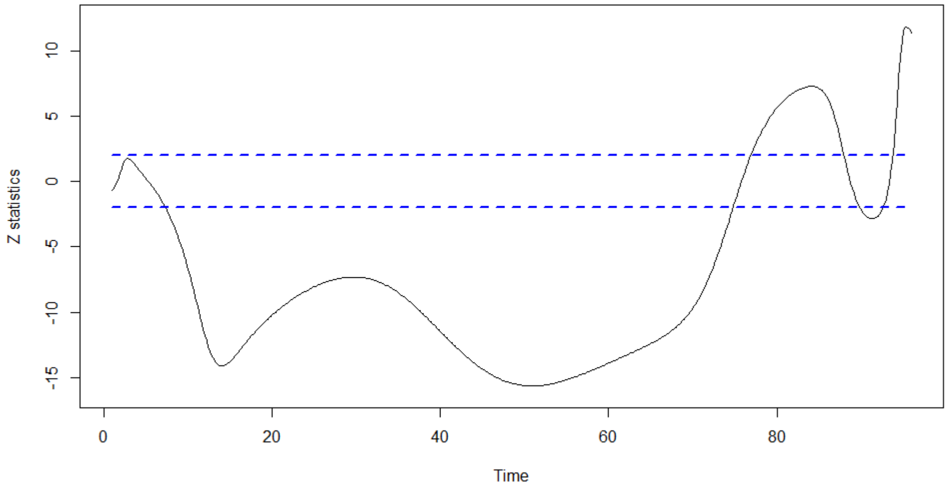
Another test we perform, which is also described in [4], is L^2-Norm-Based test. According to [4], this test works well even when the assumption of normality of samples is not satisfied but the sample sizes are large enough, even though the test does not take into account the pooled sample covariance function $\hat{\gamma}(s,t)$. L^2-Norm-Based test uses L^2-Norm of the pivotal test function $\Delta(t)$ (1) as the test statistic:

$$T_n = \int_{\tau} \Delta^2(t)dt = \frac{n_m n_w}{n} \int_{\tau} [\mu_m(t) - \mu_w(t)]^2 dt, \quad t \in \tau.$$

When the null hypothesis is valid and functional samples are approximately Normal and large enough, we get that $\Delta(t) \sim GP(0, \gamma)$. Also, we have

$$T_n \overset{d}{=} \sum_{r=1}^{k} \lambda_r A_r, \quad A_r \overset{i.i.d.}{\sim} \chi_1^2,$$

Figure 7: Pointwise Z-test for men and women samples, Z statistics. Dotted blue lines indicate critical values region.



where $\lambda_1, ..., \lambda_k$ are all the positive eigenvalues of the common covariance function $\gamma(s, t)$. Moreover, based on the Welch-Satterthwaite χ^2-approximation method, we have:

$$T_n \sim \beta\chi_d^2 \quad \text{approximately, where } \beta = \frac{tr(\gamma^{\otimes 2})}{tr(\gamma)}, \quad d = \kappa = \frac{tr^2(\gamma)}{tr(\gamma^{\otimes 2})},$$

where β is called the scale parameter and κ is called the degree of freedom. Estimators of these parameters are also given by the naive method when $tr(\gamma)$, $tr^2(\gamma)$, and $tr(\gamma^{\otimes 2})$ are replaced with their estimators. This method is suitable if both sample sizes are large enough.

After performing the L^2-Norm-Based test, p-value of 0 was obtained. Since the p-value is less than the significance level of 0.05, the null hypothesis is rejected.

However, the above-discussed tests test whether the empirical means of both samples differ significantly [14]. Thus, this type of tests are not sufficient to detect differences in distributions that may have the same mean but different variance, skewness, kurtosis, and so on.

Another test that we try to perform is the Kernel two sample test with Maximum Mean Discrepancy (MMD) as a test statistic, which is commonly met in the literature [7] [15] [16]. As described in [16], MMD is the largest difference in expectations over functions in the unit ball of a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with positive definite kernel. The MMD of two distributions with respect to the unit ball of $\mathcal{H}$ ([16], [14]) can be written as:

$$MMD[p, q, \mathcal{H}] := \sup_{f \in \mathcal{H}} |\mathbb{E}_{X\sim p}[f(X)] - \mathbb{E}_{Y\sim q}[f(Y)]|,$$

where $X = \{x_1, x_2, ..., x_m\}$ and $Y = \{y_1, y_2, ..., y_n\}$ are two samples, independently and identically distributed (i.i.d.) from p and q, which are respective Borel probability measures. $X \sim p$ means X has a distribution p. $\mathbb{E}_{X\sim p}[f(X)]$ denotes expectation with respect to p.

An empirical estimate of MMD is given by replacing population expectations with empirical expectations computed on samples X and Y ([16]):

$$MMD[X, Y, \mathcal{H}] := \sup_{f \in \mathcal{H}} \left(\frac{1}{m}\sum_{i=1}^{m} f(x_i) - \frac{1}{n}\sum_{i=1}^{n} f(y_i)\right)$$

It is expected that MMD will be small if $p = q$ and large enough if $p \neq q$. A theorem proving this can be found in [17]. Also, how the test itself is constructed can be found in [17], [18].

After performing this test, an MMD of 0.020837 and a p-value of 0.001 were obtained. Since the p-value is less than 0.05, the null hypothesis is rejected. Thus, based on the MMD test, for some time (t), men's and women's data distributions are not equal.

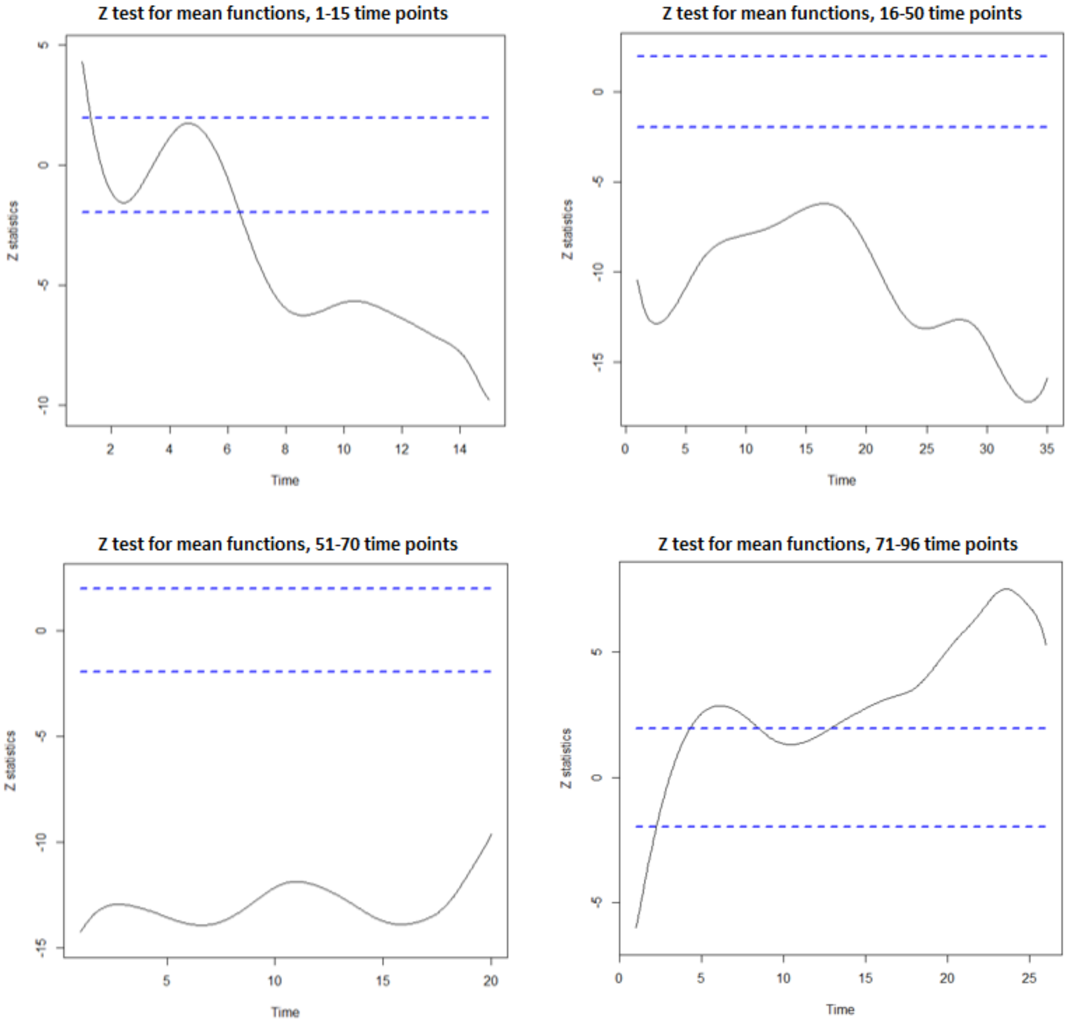
After checking the hypothesis on the whole interval of range, further interest is in checking how the previously mentioned tests work on separate parts of the interval. We can clearly see from the graphs (Figures 3, 5) that the functions of both samples at the beginning and at the end of the range are close to zero and overlap with each other. Based on viewership patterns, we split the time range into four smaller intervals: Early morning time (1-15 time points), Morning-midday (16-50 time points), Peak time (51-70 time points), and Late evening-night time (71-96 time points). The results of Z-tests can be seen in Figure 8. It is actually the same as we got during testing on the whole time interval, but this time we can see better how the test reacts at different time points. The Z-statistics of the majority of the points do not fall into the critical values region, and the null hypothesis, that means are equal, must be rejected. However, now we can clearly notice that the Z-statistics of approximately the first seven time points fall into the critical values region. Also, some Z-statistics from the Late evening-night interval also fall into the critical values region.

The results of L^2-Norm-Based and MMD-Based tests can be seen in Table 2. We can see that all the p-values of L^2-Norm-Based test are not even close to the chosen $\alpha = 0.05$. However, as already discussed, L^2-Norm-Based test does not take into account the pooled sample covariance function $\hat{\gamma}(s, t)$ and works well with large sample sizes, so this might be the problem with this test. However, if we choose 1-7 time points, L^2-Norm-Based test does not reject the null hypothesis. Overall, the MMD-Based test gives us quite different results. The p-value of the Early morning interval (1-15 time points) is equal to 0.146, which is > 0.05, so we do not reject the null hypothesis that two samples are equal. Also, the p-value of the Late evening-night interval (71-96 time points) is equal to 0.568, and we once again do not reject the null hypothesis.

Period	L^2-Norm	MMD
All time range [1-96]	0	0.015
Early morning [1-15]	0	0.146
Morning-midday [16-50]	0	0
Peak [51-70]	0	0
Late evening-night [71-96]	0	0.568

Table 2: P-values of two sample tests for specific periods.

Figure 8: Pointwise Z-test for men and women samples for different intervals.


Z test for mean functions, 1-15 time points

Z test for mean functions, 16-50 time points

Z test for mean functions, 51-70 time points

Z test for mean functions, 71-96 time points

To sum everything up, at the majority of the time points, the pointwise Z-test rejects the null hypothesis that the mean functions of two functional samples are equal, except at some time points early in the morning, where viewership is usually close to zero, and late in the evening-at night, where viewership is also usually close to zero or drops sharply. L^2-Norm-Based test rejects the null hypothesis in all cases, except at 1-7 time points. However, the MMD-Based test tells us that we should not reject the null hypothesis early in the morning or late at night. This seems quite reasonable since, during these periods, viewership is usually close to zero.

3.4 Cross-covariance Analysis

While analyzing primary data and its origin, it was noticed that since each function represents one day, Mondays can be highly correlated with Mondays, Tuesdays with Tuesdays, and so on. This comes from the television program grid: the same days of the week have very similar programs.

Also, people's viewing habits may be different on different days of the week. In this case, there is a need to transform data in such a way that this relationship would be diminished. To prove the need for transformation, we can look at cross-covariance graphs with different lags. You can see in Figure 9 what the cross-covariance contour graph of men data looks like. It is easily noticeable that the contour of lag 0 is very similar to the contour of lag 7, the contour of lag 1 is very similar to the contour of lag 8, and so on. Also, the same relationship can be noticed with women data (Figure 10). This gives the idea that we should take weekly differences and a newly reconstructed set of functions should be analyzed.

Figure 9: Cross-covariance of men data.

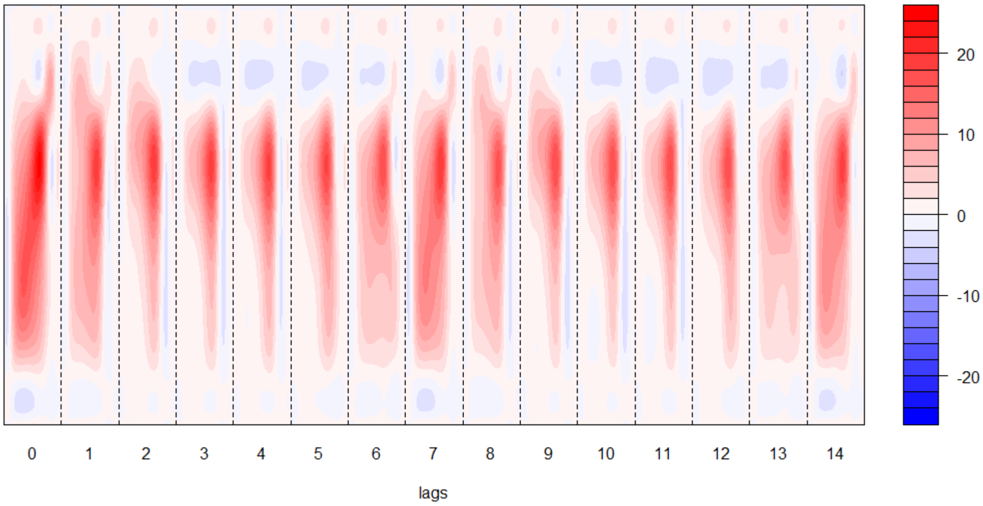
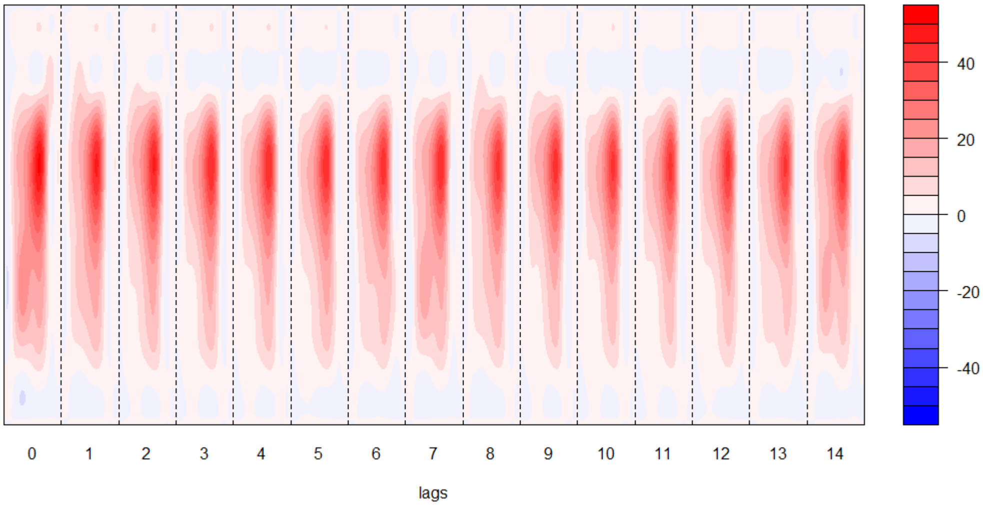


Figure 10: Cross-covariance of women data.



4 New constructed data of weekly differences

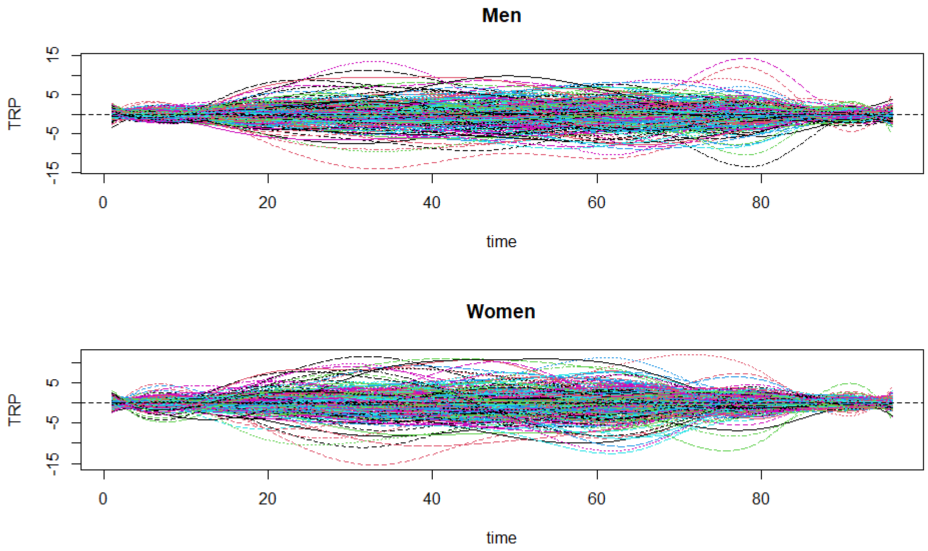
4.1 Exploratory analysis

Now both functional samples are reconstructed in such a way that new functions are obtained:

$$x_i^* = x_i - x_{i-7},$$

where x_i is the primary function of day i. How the newly constructed functions of men and women look can be seen in Figure 11. Now we can see that there is no relationship visible in cross-covariance plots (Figure 12, Figure 13) - contours are fading when lags are increasing.

Figure 11: New constructed functions of weekly differences.



As datasets have already been transformed and there is no longer cross-correlation visible, let's look at the descriptive statistics of functional data. First of all, the interest is to plot boxplots and see if there are any outliers in the newly constructed data (Figure 14). According to boxplots, men's data have two outlying functions, and women's data have one outlying function. One outlier is common for both datasets. However, these functions are not constructed from some special or exceptional days. Outliers are dealt with by replacing those curves with ones constructed from the same days but a week before. After making this transformation, no outliers are detected. How the functions look after dealing with outliers can be seen in Figure 15. Grey dashed curves represent smoothed functions of difference data after dealing with outliers.

In Figure 15, mean and bounds of +− 2 standard deviations from the mean can be seen. Since we have differences, as expected, the mean is around zero for both samples. The majority of the functions fall into +− 2 standard deviations from mean bounds. For both datasets, deviations' patterns look very similar. Deviation is smaller in the morning (from 0 to 15 time points) and during

Figure 12: Cross-covariance of new reconstructed men functions.

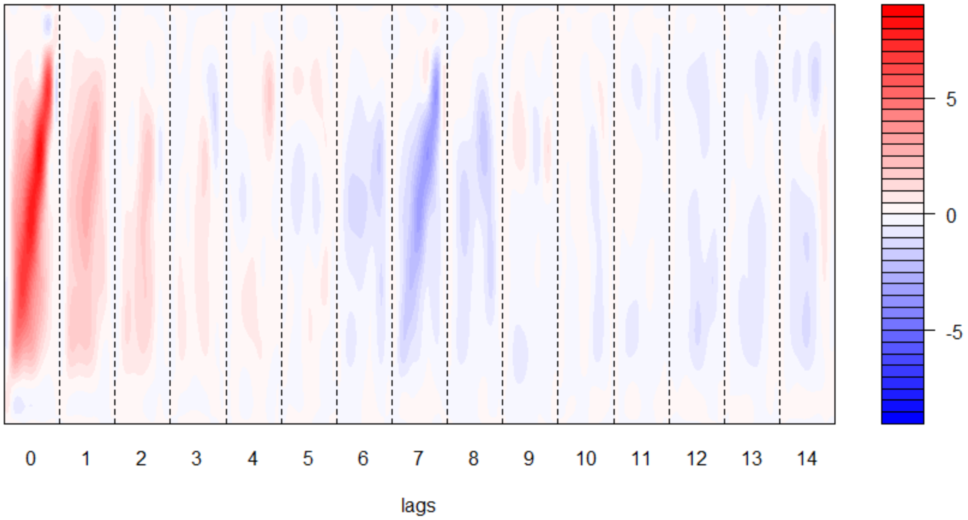
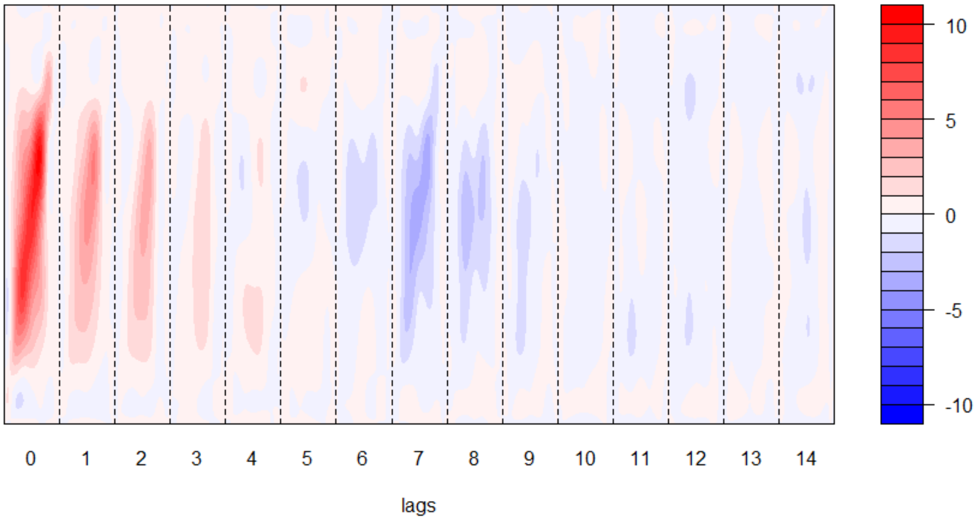


Figure 13: Cross-covariance of new reconstructed women functions.



the night (from 85 to 96 time points), when there are fewer people watching TV. As a reminder, in television terms, the day begins at 5 a.m. Deviation is higher during the day and highest around the evening, from 50 to 75 time points.

4.2 Two sample test using autoregression coefficients

As a reminder, the main goal of this work is to investigate whether or not men and women have the same TV viewing habits. Another method we chose to examine this is hard to find in the existing literature. If the person has concrete (existing) viewing habits at one point, we expect him

Figure 14: Boxplots of newly constructed data.

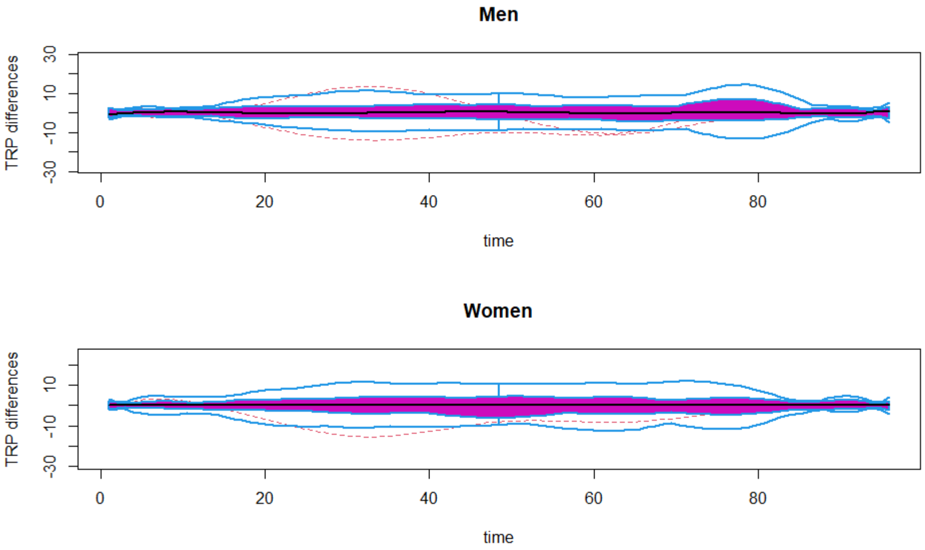
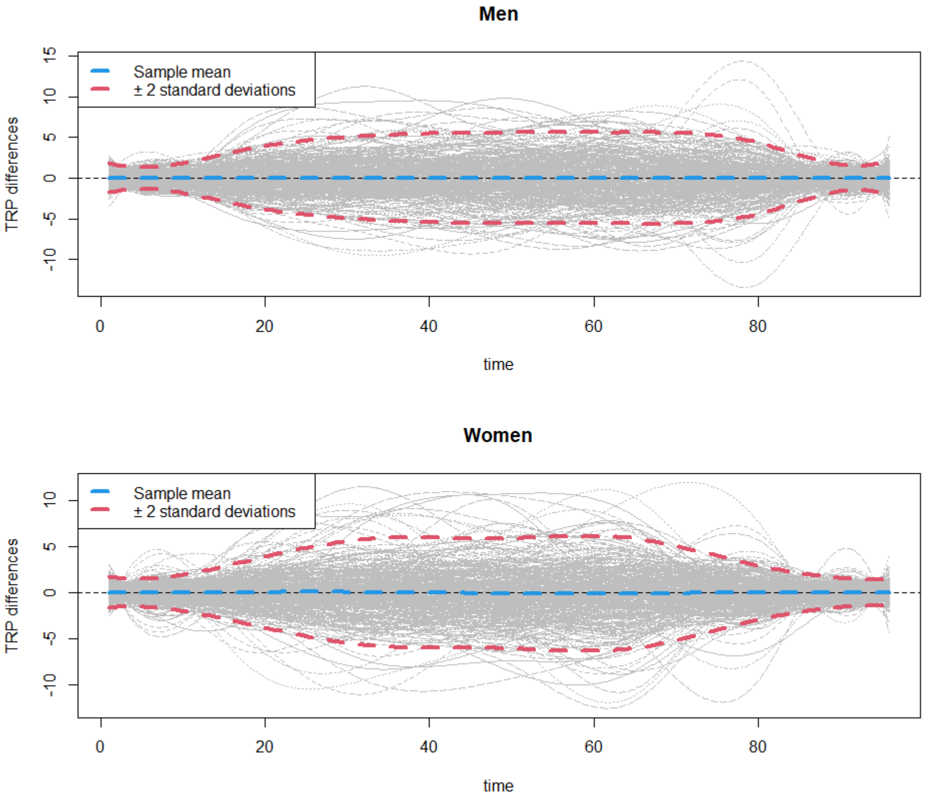


Figure 15: Samples means and standard deviations of difference data.



to act the same at the same time the next day. What we want to do is fit many autoregression models of order 1 (AR(1)) at concretely chosen time points. Autoregression of order 1 can be expressed in such a way [19]:

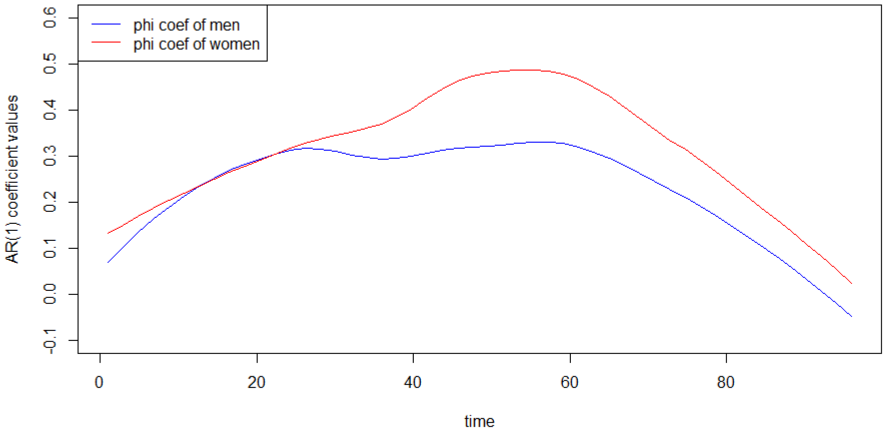
$$z_t = \delta + \phi_1 z_{t-1} + \omega_t, \quad \omega_t \sim i.i.d.\, N(0, \sigma_\omega^2), \quad t = 1, \ldots, N,$$

where z_t is a function at time t, δ and ϕ_1 are parameters of the model, and ω_t is white noise.

Since we have functions (which are continuous), we can take function values at any chosen time point in function existing interval. We will divide our time interval into 50 equal time points, perform 50 AR(1) models for both men and women's data, and extract evaluated coefficients ϕ_1 from each model at the time point t, say $\hat{\phi_w}(t)$ and $\hat{\phi_m}(t)$. Then we will be able to make two functions from the extracted coefficients and compare them.

After performing the procedure described above, the plotted AR(1) coefficients for men and women data as smoothed functions can be seen in Figure 16.

Figure 16: AR(1) coefficients of men and women data



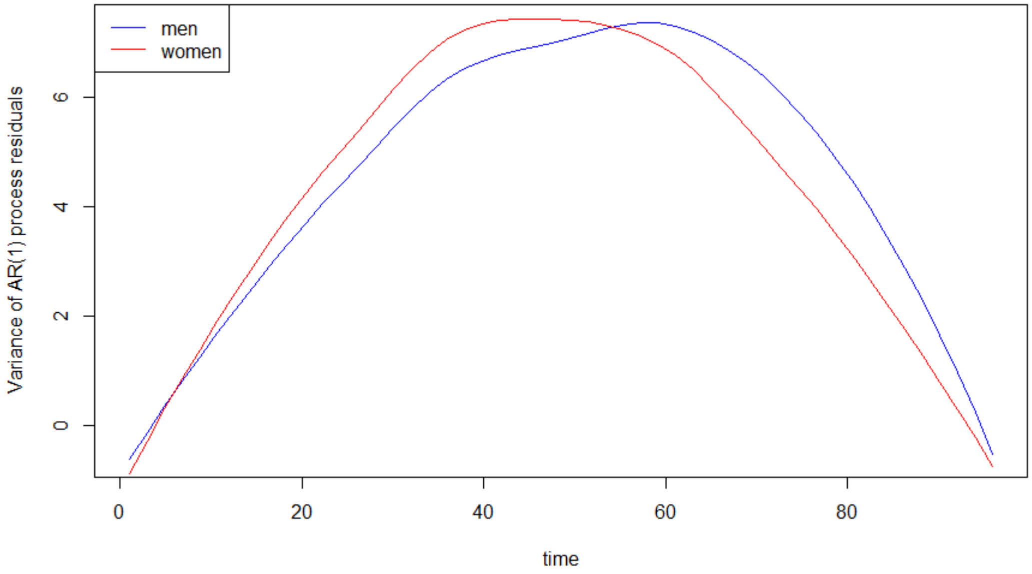
Both curves are in the interval between -1 and 1, which means that AR(1) processes are stationary. Also, both curves follow a very similar shape. We can also compute the maximum difference, which is:

$$\left\| \hat{\phi_m} - \hat{\phi_w} \right\| = \max_t \left| \hat{\phi_m}(t) - \hat{\phi_w}(t) \right| = 0.1601145.$$

However, you can notice that AR(1) coefficients are really small values, and we can look at residuals. More precisely, we want to look at the variance of residuals. As a reminder, we performed the AR(1) process for each of 50 equally spaced points for men and women data. For the residuals of each process, we calculate variance and get two functions of variance. How variance functions look can be seen in Figure 17.

As we can see, the variance of residuals functions look very similar and close to each other. However, men's function is more shifted to the right. However, both variances are quite large. We can look at the difference between those variances. We still want to test if the variances of men and

Figure 17: Variance of AR(1) processes for men and women data.



women AR(1) processes residuals are equal. If variances of residuals are very similar and close, we might conclude that AR(1) processes are very similar as well. We wish to test:

$$\begin{aligned} H_0: &\quad \sigma_m(t) = \sigma_w(t), \quad t \in \tau, \quad \text{versus} \\ H_1: &\quad \sigma_m(t) \neq \sigma_w(t), \quad \text{for some} \quad t \in \tau, \end{aligned}$$

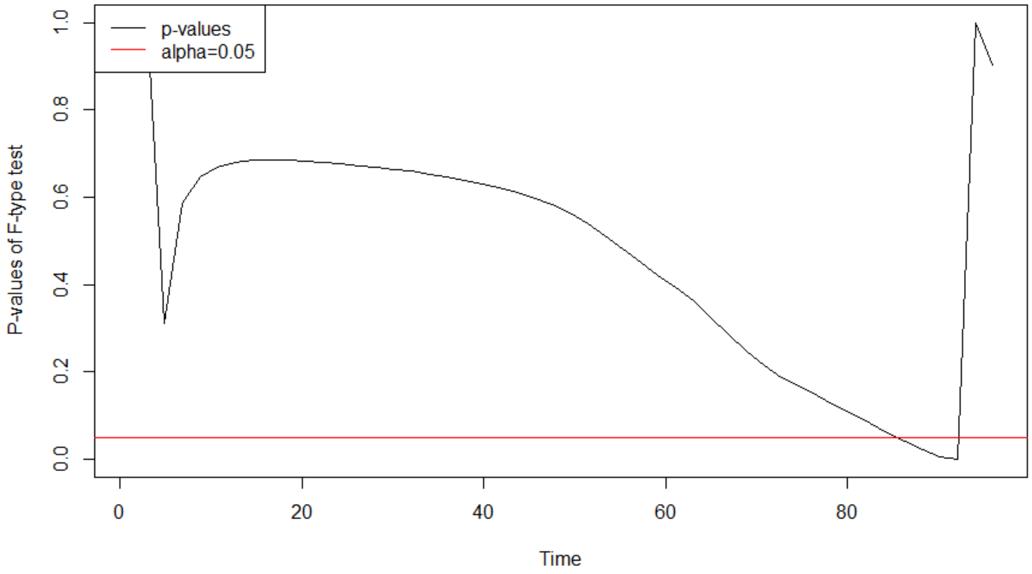
where $\sigma_m(t)$ and $\sigma_w(t)$ are variances of AR(1) processes at time t of men and women data, respectively. To do that, we use the F-type test for testing the variance of two functional samples pointwise. The F statistic at point t is simply given by [20]:

$$F = \frac{Var[\epsilon_m(t)]}{Var[\epsilon_w(t)]},$$

where $Var[h_m(t)]$ and $Var[h_w(t)]$ are the variances at time t of men and women AR(1) processes residuals.

After calculating the F statistic, we compare it with the critical value of the F-distribution $F_{\alpha,n,m}$, where n and m are sample sizes of $\epsilon_m(t)$ and $\epsilon_w(t)$, and α is the significance level, which traditionally is 0.05. If $F_{\alpha,n,m} <$ F, we reject the null hypothesis of equal variances [20]. Also, p-values can be computed from that. P-values can be seen in Figure 18. The majority of the p-values are greater than the chosen significance level of 0.05, so for the majority of the points, the F-type test does not reject the null hypothesis that variances are equal. The only interval in which the null hypothesis is rejected is at the end of the time interval, which is the night time, approximately around 85-90 time points.

Figure 18: P-values of F-type test for variances of AR(1) processes.



5 Conclusions and Recommendations

In this work, we performed two different sample tests for functional data in order to analyze whether or not men's and women's television viewing habits differ. We tried different tests for sample mean functions: the simple pointwise Z test, the L^2-Norm-Based and the MMD-Based tests. When tested for the whole interval of the time scale, all the tests rejected the null hypothesis (for the pointwise Z test, the null hypothesis was rejected for the majority of the time points).

However, even though the null hypothesis was rejected for the whole time period, maybe the viewing is similar at some specific, smaller intervals. So we divided our time scale into smaller intervals and tested the same hypotheses. Results were a bit different. Of course, the Z-test showed the same results as performed on the whole interval: there were some points where the null hypothesis was not rejected, but at the majority of the time points, the test rejected the null hypothesis. More surprising results were obtained from L^2-Norm-Based and MMD-Based tests. L^2-Norm-Based test rejected the null hypothesis at all intervals with a p-value of 0. However, the MMD-Based test did not reject the null hypothesis for the Early morning time period (1-15 time points) and Late evening-night time period (71-96 time points). This proves that we should be very careful when choosing tests to perform; assumptions should always be considered wisely. Different tests performed might give us quite different results.

Furthermore, it was noticed that the same days of week have very similar television programs, and viewing on those days is very similar. Cross-covariance analysis proved that we should also take a look into newly constructed functions of weekly differences.

For another approach, we divided our time scale into 50 equally spaced time points and performed AR(1) models at those time points for both men and women newly made samples. If men and women have the same viewing habits, we expect them to act the same as the day before (one function before), and then AR(1) processes for men and women should be very similar. We took coefficients, smoothed them, and obtained coefficient functions that were compared by shape and maximum difference. Even though the shapes of the coefficient functions were quite similar, it is not enough to prove similar viewing habits. The problem occured because that we do not have the means to say whether or not two functions of AR(1) coefficients differ statistically significantly, and to prove that would be a very difficult task. We decided not to tackle that task this time. We took a look into the variances of residuals functions as well. Even though functions seemed very close to each other and similar in shape, we also tested the hypothesis of whether or not variances of AR(1) processes residuals for men and women data differ statistically significantly. For this task, we used a pointwise F-type test for the variances of two functional samples. For the majority of the time points, the test did not reject the null hypothesis, and that means that the variances of residuals are equal for the majority of the points.

From all the results we got, we can draw the conclusion that at some intervals of the day, men and women have very similar viewing habits - in the early morning and late at night. The analysis of autoregressions also showed, that there might be overall similarity in viewing habits, even though all of the tests we tried rejected the null hypothesis that the means of two samples are equal. Also, we analyzed data on overall viewing habits, not excluding weekends, when people tend to have a different lifestyle than during the weekdays. However, when we checked outliers, there were no specific patterns of outlying weekend days. For better analysis, it would be good to cluster the data into weekdays and weekends and then test the same hypotheses.

References

[1] Pros & Cons of Media Types. *https://acrobatant.com/pros-cons-of-media-types/*

[2] Katie J. Damratoski, April R. Field, Katie N. Mizell, Michael C. Budden. An Investigation Into Alternative Television Viewership Habits of College Students. *The Journal of Applied Business Research.* 2011 January/February, Volume 27, No. 1.

[3] Dilek Altas, Hakan Oztunc. An Empirical Analysis of Television Commercial Ratings in Alternative Competitive Environments Using Multinomial Logit Model. *Eurasian Journal of Business and Economics.* 2013, Volume 6(11), 39-51.

[4] Jin-Ting Zhang. *Analysis of Variance for Functional Data.* CRC Press, U.S., 2014.

[5] Jong Soo Lee, Dennis D. Cox, Michele Follen. A Two Sample Test for Functional Data. *Communications for Statistical Applications and Methods.* 2015, Volume 22, No. 2, 121-135.

[6] Miroslaw Krzysko, Lukasz Smaga. Two-sample Tests for Functional Data Using Characteristic Functions. *Austrian Journal of Statistics.* 2021, Volume 50, 53-64.

[7] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, Alexander J. Smola. A Kernel Method for the Two-Sample-Problem. *Journal of Machine Learning Research 1.* 2008, 1-10.

[8] TV Auditorijos Tyrimas, TNS LT.
https://www.kantar.lt/lt/top/paslaugos/media-auditoriju-tyrimai/tv-auditorijos-tyrimas/

[9] James O. Ramsay, Giles Hooker, Spencer Graves. *Functional Data Analysis with R and MATLAB.* Springer, New York, 2009.

[10] Peter Hall, Ingrid Van Keilegom. Two-sample Tests in Functional Data Analysis Starting From Discrete Data. *Statistica Sinica.* 2007. Volume 17, 1511-1531.

[11] Oluwasegun Ojo, Rosa E. Lillo, Antonio Fernández Anta. *Outlier Detection for Functional Data with R Package fdaotlier.*
https://arxiv.org/pdf/2105.05213.pdf

[12] Functional Data Analysis lecture notes by Jurgita Markevičiūtė. *Exploratory Data Analysis.* Faculty of Mathematics and Informatics, Vilnius University. 2023.

[13] Ying Sun, Marc G. Genton. Functional Boxplots. *Journal of Computational and Graphical Statistics*, Volume 2, Number 2, American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of North America, 2011, 316-334.

[14] Jonas M. Kübler, Wittawat Jitkrittum, Bernard Schölkopf, Krikamol Muandet. *A Witness Two-Sample Test. https://arxiv.org/abs/2102.05573*

[15] George Wynne, Andrew B. Duncan. *A Kernel Two-Sample Test for Functional Data.* Imperial College London, Department of Mathematics, The Alan Turing Institute, 2020, *https://arxiv.org/abs/2008.11095*

[16] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, Alexander Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research 13.* 2012, 723-773.

[17] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, Alexander Smola. A Kernel Two-Sample Test. Max Planck Institute for Biological Cybernetics. Technical report no 157, 2008.

[18] Antoin Schrab, Ilmun Kim, Mélisande Albert, Béatrice Laurent, Benjamin Guedj, Arthur Gretton. MMD Aggregated Two-Sample Test. *Journal of Machine Learning Research 24.* 2023, 1-81.

[19] Sample ACF and properties of AR(1) model. *https://online.stat.psu.edu/stat510/lesson/1/1.2*

[20] Jose Fernandez. *Econometrics for Business Analysis. https://bookdown.org/cuborican/RE_STAT/*

[21] Kernel Two-sample Test with Maximum Mean Discrepancy. *https://rdrr.io/cran/maotai/man/mmd2test.html.*

[22] Denis Bosq. *Linear Processes in Functional Spaces: Theory and Applications.* Vol. 149, Springer-Verlag New York, 2011.

[23] Johan Strandberg. *Non-parametric methods for functional data.* Doctoral disertation, Department of Mathematics and Mathematical Statistics, Umea University, 2020.

[24] Jane-Ling Wang, Jeng-Min Chiou, Hans-Georg Müller. Review of functional data analysis. *Annual Reviews.* 2015, 1-41.

[25] Germán Aneiros, Ricardo Cao, Ricardo Fraiman, Christian Genest, Philippe Vieu. Recent advances in functional data analysis and high-dimensional statistics. *Journal of Multivariate Analysis.* 2019, Volume 170, 3-9.

[26] Algimantas Birbilas, Alfredas Račkauskas. Functional Modelling of Telecommunications Data. *Mathematical Modelling and Analysis.* 2022, Volume 27, No. 1, 117-133.

[27] Algirdas Laukaitis, Alfredas Račkauskas. Functional Data Analysis of Payment Systems. *Nonlinear Analysis: Modelling and Control.* 2002, Volume 7, No. 2, 53-68.

[28] Neda Mohammadi, Victor M. Panaretos. Functional data analysis with rough sample paths? *Journal of Nonparametric Statistics. https://doi.org/10.1080/10485252.2023.2215347*

[29] Zhiping Qiu, Jianwei Chen, Jin-Ting Zhang. Two-sample tests for multivariate functional data with applications. *Computational Statistics & Data Analysis.* 2021, Volume 157, 107160.

[30] Qing Jiang, Simos G. Meintanis, Lixing Zhu. Two-sample tests for multivariate funstional data. *Functional Statistics and Related Fiels.* 2017, 145-154.

[31] Katherine S. King. Functional Data Analysis With Application to United States Weather Data. *UVM Honors College Senior Theses. 10.* 2014.

6 Appendix A

Program text can be given by using `\verbatim` command:

```
library(readxl)
library(fda)
library(fda.usc)
library(fds)
library(rainbow)
library(lattice)
library(freqdom.fda)
library(fdaACF)
library(far)
library(ftsa)
library(maotai)
library(fdANOVA)
library(dplyr)
library(fdatest)
library(refund)

set.seed(123) #for some of the functions that use randomness

#uploading data
data_15min <- read_excel("C:/Users/.../data_15min.xlsx")
#View(data_15min)

#raw data plot
matplot(cbind(data_15min$'M4+'[1:2000], data_15min$'W4+'[1:2000]), type="l",
        col=c("blue","red"),lty=c(1, 1), ylab="TRP", xlab="time",
        main="TV ratings, men and women")
legend("topleft", legend=c("men", "women"), col=c("blue", "red"), lty=1:1)

#transforming by adding 10 to each observation to avoid functions being below 0
#1 function - 1 day
men <- matrix(data_15min$'M4+', 96, byrow=FALSE) + 10
women <- matrix(data_15min$'W4+', 96, byrow=FALSE) + 10

#plot with daily curves
matplot(cbind(men, women), type="l", col=c("blue","red"), lty=c(1, 1),
        ylab="TRP", xlab="time", main="TV ratings, men and women")
legend("topleft", legend=c("men", "women"), col=c("blue", "red") , lty=1:1)


######### Primary Data. Smothing #########
```

```
#Trial how smoothing looks like
#B-spline basis functions expansions method

f_men=fdata(t(men))
f_women=fdata(t(women))

argval <- argvals(f_men)
nbasis <- 4
basisobj <- create.bspline.basis(c(1, 96), nbasis)
plot(basisobj)
men_s <- smooth.basis(argvals=argval, y=men, fdParobj=basisobj)
menfd <- men_s$fd
#plotfit.fd(men, argval, menfd)

#RMSE
rmse <- sqrt(mean((eval.fd(argval, menfd) - men)^2))
print(rmse)

# We want the best GCV (or RMSE), but we also want the smallest
#  number of basis functions, which in this case is the degrees of
#  freedom for error (df).  Small df implies a stable estimate.

#Deciding on number of basis functions. Using cubic polynomials.
#(basically deciding on number of interior knots (4 - no knots, 30 - 26 knots))
#Knots equally spaced in the time range

for (nbasis in 4:30){
  basisobj = create.bspline.basis(c(1, 96), nbasis)
  men_s = smooth.basis(argvals=argval, y=men, fdParobj=basisobj)
  menfd = men_s$fd
  gcv = mean(men_s$gcv)
  RMSE = mean(sqrt(mean((eval.fd(argval, menfd) - men)^2)))
  cat(paste(nbasis, round(RMSE, 3), round(gcv, 3), "\n"))
  }
#by Elbow rule and GCV, it would be 9 basis functions.

for (nbasis in 4:30){
  basisobj = create.bspline.basis(c(1, 96), nbasis)
  women_s = smooth.basis(argvals=argval, y=women, fdParobj=basisobj)
  womenfd = women_s$fd
  gcv = mean(women_s$gcv)
  RMSE = mean(sqrt(mean((eval.fd(argval, womenfd) - women)^2)))
  cat(paste(nbasis, round(RMSE, 3), round(gcv, 3), "\n"))
}
```

```
#for women, by Elbow rule 8 basis functions, but we will use 9 for both

#Final smoothing:
basisobj <- create.bspline.basis(c(1, 96), nbasis=9)
#plot(basisobj)

men_smoothed = smooth.basis(argvals=argval, y=men, fdParobj=basisobj)
men_smoothed$fd$fdnames = list("Time", "Day", "Ratings (TRP)")
women_smoothed = smooth.basis(argvals=argval, y=women, fdParobj=basisobj)
women_smoothed$fd$fdnames = list("Time", "Day", "Ratings (TRP)")

matplot(cbind(men, women), type="l", col="grey", xlab="Time",
        ylab="Ratings (TRP)",
        main="Smoothed functions of men and women viewership" )
lines(men_smoothed$fd, col="blue")
lines(women_smoothed$fd, col="red")
legend("topleft", legend=c("raw data", "men", "women"),
       col=c("grey", "blue", "red"), lty=c(1, 2, 2))


######### Primary Data. Exploratory Analysis #########

par(mfrow=c(2, 1))
#boxplot with outliers and median
boxplot(men_smoothed$fd, main="Boxplot of men viewing", xlab="time", ylab="TRP")
#only 1 outlier
boxplot(women_smoothed$fd, main="Boxplot of women viewing", xlab="time",
        ylab="TRP")
#only 2 outliers
#pink plot includes 50% of our deepest curves
#dotted red line denotes outliers
#black line - deepest curve

par(mfrow=c(1, 1))

#removing outliers (changing to the previous weeks same day):
men_smoothed$fd$coefs[,362] <- men_smoothed$fd$coefs[,355]
women_smoothed$fd$coefs[,362] <- women_smoothed$fd$coefs[,355]
women_smoothed$fd$coefs[,139] <- women_smoothed$fd$coefs[,132]

#after removing outliers:
par(mfrow=c(1,1))
matplot(cbind(men, women), type="l", col="grey", xlab="time", ylab="TRP",
        main="Smoothed data, men and women" )
```

```
lines(men_smoothed$fd, col="blue")
lines(women_smoothed$fd, col="red")
legend("topleft", legend=c("raw data", "men", "women"),
       col=c("grey", "blue", "red"), lty=c(1, 2, 2))

#boxplots:
par(mfrow=c(2, 1))
boxplot(men_smoothed$fd, main="Boxplot of men viewing", xlab="time", ylab="TRP")
boxplot(women_smoothed$fd, main="Boxplot of women viewing", xlab="time",
        ylab="TRP")

#sample mean and standard deviation:
par(mfrow=c(1, 2))
fmen_mean <- mean.fd(men_smoothed$fd)
fmen_std <- std.fd(men_smoothed$fd)

plot(men_smoothed$fd, col="grey", ylab="TRP",
     main="Men sample mean and standard deviations")
lines(fmen_mean, lwd=4, lty=2, col=4)
lines(fmen_mean + 2*fmen_std, lwd=4, lty=2, col=2)
lines(fmen_mean - 2*fmen_std, lwd=4, lty=2, col=2)
legend("topleft", legend=c("Sample mean", "± 2 standard deviations"),
       col=c(4, 2),lwd=4:4, lty=2:2)

fwomen_mean <- mean.fd(women_smoothed$fd)
fwomen_std <- std.fd(women_smoothed$fd)

plot(women_smoothed$fd, col="grey", ylab="TRP",
     main="Women sample mean and standard deviations")
lines(fwomen_mean, lwd=4, lty=2, col=4)
lines(fwomen_mean - 2*fmen_std, lwd=4, lty=2, col=2)
lines(fwomen_mean + 2*fmen_std, lwd=4, lty=2, col=2)
legend("topleft", legend=c("Sample mean", "± 2 standard deviations"),
       col=c(4, 2),lwd=4:4, lty=2:2)

#covariance:
par(mfrow=c(1, 2))

f_men.cov <- var.fd(men_smoothed$fd)
grid <- (1:96)
f_men.cov.mat <- eval.bifd (grid, grid, f_men.cov)
levelplot(row.values=grid, column.values=grid, x=f_men.cov.mat,
          contour=TRUE, labels=FALSE, xlab="time", ylab="time",
          main="Levelplot of covariance of men data")
```

```
f_women.cov <- var.fd(women_smoothed$fd)
f_women.cov.mat <- eval.bifd (grid, grid, f_women.cov)
levelplot(row.values=grid, column.values=grid, x=f_women.cov.mat,
          contour=TRUE, labels=FALSE, xlab="time", ylab="time",
          main="Levelplot of covariance of women data")

par(mfrow=c(1, 1))


######### Tests for mean functions

#MMD:
t <- (1:96)
smoothed_matrix_men <- as.matrix(eval.fd(t, men_smoothed$fd))
smoothed_matrix_women <- as.matrix(eval.fd(t, women_smoothed$fd))

MMD.test(smoothed_matrix_men, smoothed_matrix_women) #this one used

#Z two sample test
t.sq <- seq(1,96, length=500)
stat <- Ztwosample(x=men_smoothed$fd, y=women_smoothed$fd, t.seq=t.sq)
stat$params
stat$statistics.pointwise

#L2-Norm-Based Test
t.sq <- seq(1,96, length=96)
stat <- L2.stat.twosample(x=men_smoothed$fd, y=women_smoothed$fd, t.seq=t.sq,
                          method=1)
stat

smoothed_matrix_men <- as.matrix(eval.fd(t, men_smoothed$fd))
smoothed_matrix_women <- as.matrix(eval.fd(t, women_smoothed$fd))

data_for_fanova <- cbind(smoothed_matrix_men, smoothed_matrix_women)
data_for_fanova <- as.data.frame(data_for_fanova)

labels <- rep(0, 728)
labels[365:728] <- 1 #women

fanova <- fanova.tests(x=data_for_fanova, group.label=as.character(labels))
summary(fanova)
```

```
####### ------Tests for different intervals-------- #########

#Dividing range into four smaller ones:
#Early morning time - 1-15 time points
#Morning-mid-day - 15-50 time points
#Peak time - 50-70 time points
#Late-evening-night time - 70-96 time points

#using the same number of basis functions as in all data.
#also checked by GCV and Elbow methods


### Morning [1, 15]
morning_men <- men[1:15, ]
morning_women <- women[1:15, ]

basisobj <- create.bspline.basis(c(1, 15), nbasis=9)
argval <- (1:15)

morning_men_smoothed = smooth.basis(argvals=argval, y=morning_men,
                                    fdParobj=basisobj)
morning_men_smoothed$fd$fdnames = list("Time", "Day", "Ratings (TRP)")
morning_women_smoothed = smooth.basis(argvals=argval, y=morning_women,
                                      fdParobj=basisobj)
morning_women_smoothed$fd$fdnames = list("Time", "Day", "Ratings (TRP)")

boxplot(morning_men_smoothed$fd, main="Men", xlab="time", ylab="TRP")
#no outliers
boxplot(morning_women_smoothed$fd, main="Women", xlab="time", ylab="TRP")
#no outliers

matplot(cbind(morning_men, morning_women), type="l", col="grey", xlab="Time",
        ylab="Ratings (TRP)",
        main="Smoothed functions of men and women viewership, 1-15 time points")
lines(morning_men_smoothed$fd, col="blue")
lines(morning_women_smoothed$fd, col="red")
legend("topleft", legend=c("raw data", "men", "women"),
       col=c("grey", "blue", "red"), lty=c(1, 2, 2))

t.sq <- seq(1, 15, length=370)
stat <- Ztwosample(x=morning_men_smoothed$fd, y=morning_women_smoothed$fd,
                   t.seq=t.sq)

t.sq <- seq(1, 15, length=100)
```

```
stat <- L2.stat.twosample(x=morning_men_smoothed$fd, y=morning_women_smoothed$fd,
                          t.seq = t.sq, method=1)
stat

#MMD
MMD.test(smoothed_matrix_men[1:15,], smoothed_matrix_women[1:15,])

#fANOVA
fanova_morning <- fanova.tests(x=data_for_fanova[1:15,],
                               group.label=as.character(labels))
summary(fanova_morning)


### Midday [16, 50]
mid_men <- men[16:50, ]
mid_women <- women[16:50, ]

basisobj <- create.bspline.basis(c(1, 35), nbasis=9)

mid_men_smoothed = smooth.basis(argvals=1:35, y=mid_men, fdParobj=basisobj)
mid_men_smoothed$fd$fdnames = list("Time", "Day", "Ratings (TRP)")
mid_women_smoothed = smooth.basis(argvals=1:35, y=mid_women, fdParobj=basisobj)
mid_women_smoothed$fd$fdnames = list("Time", "Day", "Ratings (TRP)")

boxplot(mid_men_smoothed$fd, main="Men", xlab="time", ylab="TRP")
#no outliers
boxplot(mid_women_smoothed$fd, main="Women", xlab="time", ylab="TRP")
#21, 63 and 364 functions outliers

mid_women_smoothed$fd$coefs[,21] <- mid_women_smoothed$fd$coefs[,14]
mid_women_smoothed$fd$coefs[,63] <- mid_women_smoothed$fd$coefs[,56]
mid_women_smoothed$fd$coefs[,364] <- mid_women_smoothed$fd$coefs[,357]

matplot(cbind(mid_men, mid_women), type="l", col="grey", xlab="Time",
        ylab="Ratings (TRP)",
        main="Smoothed functions of men and women viewership" )
lines(mid_men_smoothed$fd, col="blue")
lines(mid_women_smoothed$fd, col="red")
legend("topleft", legend=c("raw data", "men", "women"),
       col=c("grey", "blue", "red"), lty=c(1, 2, 2))

t.sq <- seq(1, 35, length=370)
stat <- Ztwosample(x=mid_men_smoothed$fd, y=mid_women_smoothed$fd, t.seq=t.sq)
#stat$params
```

```
#stat$statistics.pointwise

t.sq <- seq(1, 35, length=35)
stat <- L2.stat.twosample(x=mid_men_smoothed$fd, y=mid_women_smoothed$fd,
                          t.seq=t.sq, method=1)
stat

#MMD
MMD.test(smoothed_matrix_men[16:50,], smoothed_matrix_women[16:50,])

fanova_mid <- fanova.tests(x=data_for_fanova[16:50,],
                           group.label=as.character(labels))
summary(fanova_mid)



### Peak [51, 70]
peak_men <- men[51:70,]
peak_women <- women[51:70,]

basisobj <- create.bspline.basis(c(1, 20), nbasis=9)

peak_men_smoothed = smooth.basis(argvals=1:20, y=peak_men, fdParobj=basisobj)
peak_men_smoothed$fd$fdnames = list("Time", "Day", "Ratings (TRP)")
peak_women_smoothed = smooth.basis(argvals=1:20, y=peak_women, fdParobj=basisobj)
peak_women_smoothed$fd$fdnames = list("Time", "Day", "Ratings (TRP)")

boxplot(peak_men_smoothed$fd, main="Men", xlab="time", ylab="TRP")
#no outliers
boxplot(peak_women_smoothed$fd, main="Women", xlab="time", ylab="TRP")
#no outliers

matplot(cbind(peak_men, peak_women), type="l", col="grey", xlab="Time",
        ylab="Ratings (TRP)",
        main="Smoothed functions of men and women viewership")
lines(peak_men_smoothed$fd, col="blue")
lines(peak_women_smoothed$fd, col="red")
legend("topleft", legend=c("raw data", "men", "women"),
       col=c("grey", "blue", "red"), lty=c(1, 2, 2))

t.sq <- seq(1, 20, length=370)
stat <- Ztwosample(x=peak_men_smoothed$fd, y=peak_women_smoothed$fd, t.seq=t.sq)

t.sq <- seq(1, 20, length=20)
stat <- L2.stat.twosample(x=peak_men_smoothed$fd, y=peak_women_smoothed$fd,
```

```
                          t.seq=t.sq, method=1)
stat

#MMD
MMD.test(smoothed_matrix_men[51:70,], smoothed_matrix_women[51:70,])

fanova_peak <- fanova.tests(x=data_for_fanova[51:70,], group.label=as.character(labels))
summary(fanova_peak)


### Late night [71, 96]
late_men <- men[71:96,]

#checking by GCV for interest
f_late_men=fdata(t(late_men))
argval <- argvals(f_late_men)
for (nbasis in 4:15){
  basisobj = create.bspline.basis(c(1, 26), nbasis)
  men_s = smooth.basis(argvals=argval, y=late_men, fdParobj=basisobj)
  menfd = men_s$fd
  gcv = mean(men_s$gcv)
  RMSE = mean(sqrt(mean((eval.fd(argval, menfd) - late_men)^2)))
  cat(paste(nbasis, round(RMSE, 3), round(gcv, 3), "\n"))
}
#9 functions

late_women <- women[71:96,]
f_late_women=fdata(t(late_women))
argval <- argvals(f_late_women)
for (nbasis in 4:15){
  basisobj = create.bspline.basis(c(1,26), nbasis)
  women_s = smooth.basis(argvals=argval, y=late_women, fdParobj=basisobj)
  womenfd = women_s$fd
  gcv = mean(women_s$gcv)
  RMSE = mean(sqrt(mean((eval.fd(argval, womenfd) - late_women)^2)))
  cat(paste(nbasis, round(RMSE, 3), round(gcv, 3), "\n"))
}
#also 9 functions

basisobj <- create.bspline.basis(c(1, 26), nbasis=9)
late_men_smoothed = smooth.basis(argvals=1:26, y=late_men, fdParobj=basisobj)
late_men_smoothed$fd$fdnames = list("Time", "Day", "Ratings (TRP)")
late_women_smoothed = smooth.basis(argvals=1:26, y=late_women, fdParobj=basisobj)
late_women_smoothed$fd$fdnames = list("Time", "Day", "Ratings (TRP)")
```

```
boxplot(late_men_smoothed$fd, main="Men", xlab="time", ylab="TRP")
#139, 189 and 362 functions
boxplot(late_women_smoothed$fd, main="Women", xlab="time", ylab="TRP")
#139 and 362 functions

late_men_smoothed$fd$coefs[,139] <- late_men_smoothed$fd$coefs[,132]
late_men_smoothed$fd$coefs[,189] <- late_men_smoothed$fd$coefs[,182]
late_men_smoothed$fd$coefs[,362] <- late_men_smoothed$fd$coefs[,355]

late_women_smoothed$fd$coefs[,139] <- late_women_smoothed$fd$coefs[,132]
late_women_smoothed$fd$coefs[,362] <- late_women_smoothed$fd$coefs[,355]

matplot(cbind(late_men, late_women), type="l", col="grey", xlab="Time",
        ylab="Ratings (TRP)",
        main="Smoothed functions of men and women viewership")
lines(late_men_smoothed$fd, col="blue")
lines(late_women_smoothed$fd, col="red")
legend("topleft", legend=c("raw data", "men", "women"),
       col=c("grey", "blue", "red"), lty=c(1, 2, 2))

t.sq <- seq(1,26, length=370)
stat <- Ztwosample(x=late_men_smoothed$fd, y=late_women_smoothed$fd, t.seq = t.sq)

t.sq <- seq(1,26, length=20)
stat <- L2.stat.twosample(x=late_men_smoothed$fd, y=late_women_smoothed$fd,
                          t.seq=t.sq, method=1)
stat

#MMD
MMD.test(smoothed_matrix_men[71:96,], smoothed_matrix_women[71:96,])

fanova_late <- fanova.tests(x=data_for_fanova[71:96,],
                            group.label=as.character(labels))
summary(fanova_late)


###### cross-covariance #######
fts.plot.covariance(men_smoothed$fd, men_smoothed$fd, lags=0:14)
fts.plot.covariance(women_smoothed$fd, women_smoothed$fd, lags=0:14)


#---------------------------------------------------------------------------
# new datasets constructed: y_t = x_t - x_(t-7)
```

```
#since cross-covariance analysis showed us, that there are similarities among
#weekdays and there is logic that program grid is the same during same
#weekdays, it allows to make assumption that we need to analyze new datasets of
#differences: y_t = x_t - x_(t-7)

basisobj <- create.bspline.basis(c(1, 96), nbasis=9)
argval <- argvals(f_men1)

men_smoothed_1 = smooth.basis(argvals=argval, y=men[,8:364], fdParobj=basisobj)
men_smoothed_1$fd$fdnames = list("Time", "Day", "Ratings (TRP)")
women_smoothed_1 = smooth.basis(argvals=argval, y=women[,8:364],
                                fdParobj=basisobj)
women_smoothed_1$fd$fdnames = list("Time", "Day", "Ratings (TRP)")

men_smoothed_2 = smooth.basis(argvals=argval, y=men[,1:357], fdParobj=basisobj)
men_smoothed_2$fd$fdnames = list("Time", "Day", "Ratings (TRP)")
women_smoothed_2 = smooth.basis(argvals=argval, y=women[,1:357],
                                fdParobj=basisobj)
women_smoothed_2$fd$fdnames = list("Time", "Day", "Ratings (TRP)")

diff_men_fd <- minus.fd(men_smoothed_1$fd, men_smoothed_2$fd)
diff_women_fd <- minus.fd(women_smoothed_1$fd, women_smoothed_2$fd)

#graph of both differentiated datasets:
par(mfrow=c(2, 1))
plot(diff_men_fd, main="Men", ylab="TRP")
plot(diff_women_fd, main="Women", ylab="TRP")

#cross-covariance to see if exists some relationship or not anymore:
par(mfrow=c(1, 1))
grid <- (1:96)
f_new_men.cov <- var.fd(diff_men_fd)
f_new_men.cov.mat <- eval.bifd (grid, grid, f_new_men.cov)

contour(grid, grid, f_new_men.cov.mat, lwd=2)
persp(grid, grid, f_new_men.cov.mat, xlab="s", ylab="t", zlab="c(s,t)")

f_new_women.cov <- var.fd(diff_women_fd)
f_new_women.cov.mat <- eval.bifd (grid, grid, f_new_women.cov)

contour(grid, grid, f_new_women.cov.mat, lwd=2)
persp(grid, grid, f_new_women.cov.mat, xlab="s", ylab="t", zlab="c(s,t)")
```

```
fts.plot.covariance(diff_men_fd, diff_men_fd, lags=0:14)
fts.plot.covariance(diff_women_fd, diff_women_fd, lags=0:14)
#bright contours of colors are fading away so no relationship visible among
#different days now


###### Exploratory analysis ######
par(mfrow=c(2,1))
#dealing with outliers (changing to the difference of the days that was week ago):
boxplot(diff_men_fd, main="Men",xlab="time", ylab="TRP differences")
#3 outliers - 92 and 348 and 355 functions
boxplot(diff_women_fd, main="Women",xlab="time", ylab="TRP differences")
#4 outliers - 92, 132, 139 and 355
#pink plot includes 50% of our deepest curves
#dotted red line denotes outliers
#black line - deepest curve

diff_men_fd$coefs[,92] <- diff_men_fd$coefs[,85]
diff_men_fd$coefs[,348] <- new_men_smoothed$fd$coefs[,341]
diff_men_fd$coefs[,355] <- new_women_smoothed$fd$coefs[,348]

diff_women_fd$coefs[,92] <- diff_men_fd$coefs[,85]
diff_women_fd$coefs[,132] <- new_men_smoothed$fd$coefs[,125]
diff_women_fd$coefs[,139] <- new_men_smoothed$fd$coefs[,125]
diff_women_fd$coefs[,355] <- new_women_smoothed$fd$coefs[,348]

#Boxplots after dealing with outliers:
boxplot(new_men_smoothed$fd, main="Men",xlab="time", ylab="TRP differences")
#no outliers
boxplot(new_women_smoothed$fd, main="Women",xlab="time", ylab="TRP differences")
#no outliers
#pink plot includes 50% of our deepest curves
#dotted red line denotes outliers
#black line - deepest curve

#How the data look like after dealing with outliers:
par(mfrow=c(2, 1))
plot(diff_men_fd, main="Men", ylab="TRP")
plot(diff_women_fd, main="Women", ylab="TRP")

#Sample mean and standard deviation:
par(mfrow=c(2, 1))
fmen_mean_new <- mean.fd(diff_men_fd)
fmen_std_new <- std.fd(diff_women_fd)
```

```
plot(diff_men_fd, col="grey", main="Men", ylab="TRP differences")
lines(fmen_mean_new, lwd=4, lty=2, col=4)
lines(fmen_mean_new + 2*fmen_std_new, lwd=4, lty=2, col=2)
lines(fmen_mean_new - 2*fmen_std_new, lwd=4, lty=2, col=2)
legend("topleft", legend=c("Sample mean", "± 2 standard deviations"),
       col=c(4, 2), lwd=4:4, lty=2:2)

fwomen_mean_new <- mean.fd(diff_women_fd)
fwomen_std_new <- std.fd(diff_women_fd)

plot(diff_women_fd, col="grey", main="Women", ylab="TRP differences")
lines(fwomen_mean_new, lwd=4, lty=2, col=4)
lines(fwomen_mean_new + 2*fwomen_std_new, lwd=4, lty=2, col=2)
lines(fwomen_mean_new - 2*fwomen_std_new, lwd=4, lty=2, col=2)
legend("topleft", legend=c("Sample mean", "± 2 standard deviations"),
       col=c(4, 2), lwd=4:4, lty=2:2)


#AR(1) model at 50 equally spaced time points:

#men:
men_int <- matrix(NA, nrow=50, ncol=2)  #empty matrixes for coefficients
men_slope <- matrix(NA, nrow=50, ncol=2)
men_res_var <- matrix(NA, nrow=50, ncol=2)  #empty matrix for residuals

points <- seq(1, 96, len=50)
row=1

for(i in points) {
  data <- t(eval.fd(i, diff_men_fd))
  model <- arima(data, order=c(1, 0, 0))
  men_int[row, 1] <- i
  men_slope[row, 1] <- i
  men_res_var[row, 1] <- i
  men_int[row, 2] <- model$coef["intercept"]
  men_slope[row, 2] <- model$coef["ar1"]
  men_res_var[row, 2] <- var(residuals(model))
  row = row+1
}

colnames(men_int) <- c("point", "intercept")
colnames(men_slope) <- c("point", "coef")
colnames(men_res_var) <- c("point", "var")
```

```
#women:
women_int <- matrix(NA, nrow=50, ncol=2)  #empty matrixes for coefficients
women_slope <- matrix(NA, nrow=50, ncol=2)
women_res_var <- matrix(NA, nrow=50, ncol=2)  #empty matrix for residuals

points <- seq(1, 96, len=50)
row=1

for(i in points) {
  data <- t(eval.fd(i, diff_women_fd))
  model <- arima(data, order=c(1, 0, 0))
  women_int[row, 1] <- i
  women_slope[row, 1] <- i
  women_res_var[row, 1] <- i
  women_int[row, 2] <- model$coef["intercept"]
  women_slope[row, 2] <- model$coef["ar1"]
  women_res_var[row, 2] <- var(residuals(model))
  row = row+1
}

colnames(women_int) <- c("point", "intercept")
colnames(women_slope) <- c("point", "coef")
colnames(women_res_var) <- c("point", "var")

women_slope <- as.data.frame(women_slope)
men_slope <- as.data.frame(men_slope)
men_res_var <- as.data.frame(men_res_var)
women_res_var <- as.data.frame(women_res_var)

par(mfrow=c(1, 1))
#smooth coefficient functions:
val_m <- loess(men_slope$coef˜men_slope$point)
val_w <- loess(women_slope$coef˜women_slope$point)

matplot(men_slope$point, predict(val_m), type="l", xlab="time",
        ylim=c(-0.1, 0.6), ylab="AR(1) coefficient values", col="blue")
lines(women_slope$point, predict(val_w),col="red")
legend("topleft", legend=c("phi coef of men", "phi coef of women"),
       col=c("blue", "red") , lty=1:1)

#without smoothing:
plot(men_slope$point, men_slope$coef, xlab="time", ylim=c(-0.1, 0.6),
     ylab="AR(1) coefficient values", type="l", col="blue")
```

```
lines(x=women_slope$point, y=women_slope$coef, type="l", col="red")
legend("topleft", legend=c("phi coef of men", "phi coef of women"),
       col=c("blue", "red") , lty=1:1)
#values from -1 to 1, that means this is a stationary process

#maximum difference without smoothing
differences_slope <-abs(men_slope$coef - women_slope$coef)
max(differences_slope)

#maximum difference after smoothing
differences_slope_s <-abs(predict(val_m) - predict(val_w))
max(differences_slope_s)

#### Variance of residuals:
res_m <- loess(men_res_var$var~men_res_var$point)
res_w <- loess(women_res_var$var~women_res_var$point)

#plotting residuals smoothed:
matplot(men_res_var$point, predict(res_m), type="l", xlab="time",
        ylab="Variance of AR(1) process residuals", col="blue")
lines(women_res_var$point, predict(res_w), col="red")
legend("topleft", legend=c("men", "women"), col=c("blue", "red"), lty=1:1)

#difference:
diff_var_res <- men_res_var
diff_var_res$var <- men_res_var$var - women_res_var$var

diff_var_res_s <-abs(predict(res_m) - predict(res_w))

matplot(diff_var_res$point, diff_var_res_s, type="l", xlab="time",
        ylab="Diff. of variance of AR(1) processes residuals")
abline(h=0, col="red")

#Checking if variances of residuals are equal:
variance_1 = predict(res_m)
variance_2 = predict(res_w)
f_values = variance_1 / variance_2
p_values = pf(f_values, df1 = 50 - 1, df2 = 50 - 1, lower.tail = FALSE)
matplot(men_res_var$point, p_values, xlab="Time",
        ylab="P-values of F-type test", type="l")
abline(h=0.05, col="red")
legend("topleft", legend=c("p-values", "alpha=0.05"),
       col=c("black", "red"), lty=1:1)
```

Pointwise Z two sample test code:

```
Ztwosample <- function(x, y, t.seq, alpha=0.05) {
  if(class(x) != "fd") stop("X must be fd object")
  if(class(y) != "fd") stop("Y must be fd object")
  k <- length(t.seq)

  mu.x <- mean.fd(x)
  mu.y <- mean.fd(y)

  n <- dim(x$coef)[2]
  m <- dim(x$coef)[2]

  delta <- (mu.x - mu.y)
  delta.t <- eval.fd(t.seq, delta)

  z.x <- center.fd(x)
  z.y <- center.fd(y)

  z.x.t <- eval.fd(t.seq, z.x)
  z.y.t <- eval.fd(t.seq, z.y)
  z.t <- cbind(z.x.t, z.y.t)

  if(n > k) {
    Sigma <- (t(z.t) %*% z.t)/(n-2)
  } else {
    Sigma <- (z.t %*% t(z.t))/(n-2)
  }

  gamma.t <- diag(Sigma)
  Zpointwise <- sqrt((n*m)/(n+m)) * delta.t/sqrt(gamma.t)

    crit <- qt(1-alpha/2, n-2)
    crit.val <- rep(crit, k)
    params <- list(critical.value = crit)



  mx <- max(cbind(Zpointwise, crit.val))
  mn <- min(cbind(Zpointwise, -crit.val))

  dev.new()
  plot(t.seq, Zpointwise, type="l", xlab = 'Time', ylab = "Z statistics",
       main = "Two samples t-test", ylim=c(mn-0.5, mx+0.5))
  lines(t.seq, crit.val, lty=2, lwd=2, col="blue")
  lines(t.seq, -crit.val, lty=2, lwd=2, col="blue")
```

```
  return(list(statistics.pointwise = Zpointwise,
              params = params))
}
```

L^2-Norm-Based test code:

```
L2.stat.twosample <- function(x, y, t.seq, alpha=0.05, method=1:2, replications=100) {
  # method = 1: T stat
  # method = 2: Bootsrap
  if(class(x) != "fd") stop("X must be fd object")
  if(class(y) != "fd") stop("Y must be fd object")

  mu.x <- mean.fd(x)
  mu.y <- mean.fd(y)

  n <- dim(x$coefs)[2]
  m <- dim(y$coefs)[2]

  k <- length(t.seq)

  cn <- (n*m)/(n+m)
  delta <- (mu.x - mu.y)
  delta.t <- eval.fd(t.seq, delta)

  z.x <- center.fd(x)
  z.y <- center.fd(y)

  z.x.t <- eval.fd(t.seq, z.x)
  z.y.t <- eval.fd(t.seq, z.y)
  z.t <- cbind(z.x.t, z.y.t)

  if(n > k | m > k) {
    Sigma <- (t(z.t) %*% z.t)/(n-2)
  } else {
    Sigma <- (z.t %*% t(z.t))/(n-2)
  }

  trace <- function(data) sum(diag(data))

  A <- trace(Sigma)
  B <- trace(Sigma^2)

  L2stat <- cn * t(delta.t) %*% delta.t
```

```
  L2stat <- L2stat[1]

    A2 <- A^2
    B2 <- B
    alp <- B2/A
    df <- A2/B2
    pvalue <- 1-pchisq(L2stat/alp, df)
    params <- list(alpha = alp, df = df)

 list(statistics = L2stat, pvalue = pvalue, params=params)
}
```