Imperial College London

Mathematics Second Year Project

Nonparametric Functional Data Analysis: A Powerful Prediction Method

Sanjay Patel
Lie XU
Richard Fu
Nikolai Rozanov

Supervisor:
Dr. Badr Missaoui

June, 2015

Contents

1 Introduction

This report focuses on the application and advantages of the combination of two statistical modelling techniques: nonparametric statistics and functional data analysis.

Examining the former of these two modelling techniques, we assert that nonparametric methods hold an important advantage over parametric functions. While parametric functions can yield some very accurate and precise estimates with our observed data, this is only possible if our assumptions on the parameters and on the distribution are correct; so if our assumptions are wrong, it may, in some cases, lead to very erroneous and misleading results. Hence, in comparison, nonparametric methods have the advantage of being more robust, as they make no unfounded inferences on the probability distribution of the variables, and thus are not limited by the accuracy of our beliefs. In addition, we see that nonparametric methods can still lead to very accurate predictions for given data.

As for the latter of the two modelling techniques, functional data analysis is a branch of statistics in which we analyze data that is presented in a continuum (often time), such as curves or surfaces. In other words, instead of studying the correlation between two variables as data points, in this case, we study the link between two variables that are functions, and predict new values of one of the variables given the other. Briefly phrased, functional data is created from piecing together consecutive data points from a continuous domain. We further elaborate functional data and nonparametric functions as well as other key terms in Chapter 2.

Thus, the aim of this report would be to show viable and applicable methods of using nonparametric functional data analysis, and identify commonly used nonparametric functional estimators, such as kernel distribution estimator (chapter 3). We will then provide specific examples, such as predicting electrical consumption using time-series, and analyze the results we obtained using R and MATLAB (in chapter 4). Finally, we conclude this report by elaborating the numerous advantages of nonparametric functional data analysis in comparison to traditional parametric functions and other statistical methods (in chapter 5).

2 Mathematical Framework

This section serves as a foundational chapter, outlining the major definitions, notations and notions of the Mathematics presented and elaborated on in this paper. In particular we will define what the functional and non-parametric settings are and later define the necessary technicalities, both for theoretical as well as practical methods.

2.1 The Functional and Non-parametric Framework

Before we define functional random variables, we should first examine when and why we would use them. From the continuous advancement in modern technology, especially progress in memory and computational capacity, we are becoming more and more capable of collecting and dealing with large sets of data. Thus, in many fields such as medicine, econometrics, biometrics, there are increasing situations where we collect data as curves. In other words, we can examine a large spread of values in a continuum, such as time or wavelength, and plot them in very thinly spaced out values across our domain. These consecutive points of data are so closely packed together that they can be connected with a function, from linear lines to splines. Thus, for instance, in a certain period of time (t_{min}, t_{max}), we can observe our data as a continuous family $\chi = X(t); t \in (t_{min}, t_{max})$. In order to work with this, we must first give a proper definition in a mathematical sense. From [1]:

Definition 2.1. *A random variable $\mathcal{X}$ is called functional variable (f.v.) if it takes values in an infinite dimensional space (or functional space). An observation χ of $\mathcal{X}$ is called a functional data.*

Definition 2.2. *Let Z be a random variable in some infinite dimensional space F, and let ϕ be a mapping defined on F and depending on the distribution of Z. A model for the estimation of ϕ consists in introducing some constraint of the form*

$$\phi \in C.$$

The model is called a functional parametric model for the estimation of ϕ if C is indexed by a finite number of elements of F. Otherwise, the model is called functional nonparametric model.

The infinite dimensional property of C gives way to the nonparametric aspect of our estimation of the data, while the infinite dimensional property of F gives way to the functional aspect of our data.

2.2 Distance Function

In a functional setting, the "distance" is not as in the natural sense as in the multivariate setting, where the "distance" is given by the Euclidean metric. A well-defined distance is a crucial to the estimation techniques we will use later in Chapter 3.

Definition 2.3. *$d : F \times F \to \mathbb{R}^+$ is a metric on some space F if*

1. $d(x, y) = d(y, x)\ , \forall\, x, y \in F$
2. $d(x, y) \geq 0$ *with equality if and only if* $x = y$
3. $d(x, z) \leq d(x, y) + d(y, z)\ \forall x, y, z \in F$

In the context of non-parametric functional Data Analysis it turns out that the use of Semi-metrics is advantageous, see discussion in Chapter 3 in [1].

Definition 2.4. *$d : F \times F \to \mathbb{R}^+$ is a semi-metric on some space F if*

1. $d(x, y) = d(y, x)\ , \forall\, x, y \in F$
2. $d(x, x) = 0\ \forall x \in F$
3. $d(x, z) \leq d(x, y) + d(y, z)\ \forall x, y, z \in F$

2.3 Kernel

In non-parametric statistics, the kernel is defined to be a weighting function used to estimate the probability density function of a random variable; this non-parametric method of estimation is called the kernel density estimation (**KDE**) [9]. While any density function can technically be considered a kernel, for the purposes of this report, we will only consider positive and symmetric kernels, as they are generally the most simplistic and commonly used ones.
Some commonly used Kernels are, graphs taken from [9]:

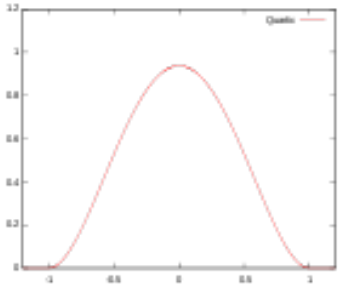

Figure 1: Quadratic Kernel

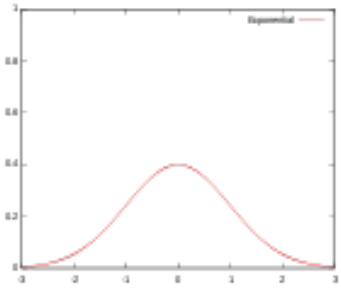

Figure 2: Gaussian Kernel

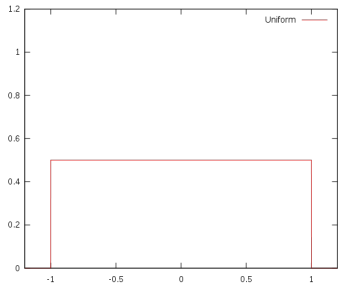

Figure 3: Box Kernel

The kernel function (which we will denote as K) often also has a smoothing parameter, h, which is called the bandwidth, which can alter the domain, and thus increase or decrease the weighting proportionally. For instance, the box kernel with bandwidth $h = 1$ as shown in the graph can be mathematically written in the form [1]:

$$Box\,kernel : K(u) = \frac{1}{2}1_{[-1,+1]}(u)$$

And for a general box kernel with bandwidth h centered around a real number x, we can express it as [1]:

$$K(u) = \frac{1}{h}1_{[x-h,x+h]}(u)$$

As we can see, the kernel will equate any value outside of the interval of length h around x (the bandwidth) to zero. Now let us consider another example: take a fixed real number x and n real random variables $X_1, X_2, ..., X_n$. Let us map these n r.r.v. using our kernel local weighting function into $P_1, P_2, ..., P_n$ such that [1]:

$$P_i = \frac{1}{h}K(\frac{x - X_i}{h})$$

As we can see, for each X_i, we attribute it a weight P_i based on its distance from x. Now generally, we prefer kernels that will give higher weighting to X_is closer to x. This makes intuitive sense when we are using weights for regression, median, and mode estimations as we will show in the next chapter, as we want the values closest to x to have a greater influence. Thus the choice of the kernel matters; in the case of the box kernel, it is very simple to calculate the weight of any value inside the bandwidth, but triangular, quadratic, or gaussian gives greater weights to values closer to x [1].

Multivariate Case

We can easily extend the previous kernel weighting to the multivariate case. One way to define our multivariate kernel $\mathbf{K}$ from $\mathbb{R}^p \to \mathbb{R}$ is the product of p real kernel functions $K_1, ..., K_p$ [1]:

$$\forall \mathbf{u} = (u_1, u_2, ..., u_p)^T \in \mathbb{R}^p, \mathbf{K}(u) = K_1(u_1) \times ... \times K_p(u_p)$$

Another way to define a multivariate kernel is with a norm (denoted $|| \cdot ||$) in $\mathbb{R}$ as follows[1]:

$$\forall \mathbf{u} \in \mathbb{R}^p, \mathbf{K}(\mathbf{u}) = K(||\mathbf{u}||)$$

Now, one thing to note that the norm of a vector is always a positive value, show our kernel K should always have a positive support, and thus we can use asymmetrical functions for K [1].

Functional Data

Now let us define $\chi_1, \chi_2, ..., \chi_n$ to be n f.r.v valued in E and let χ be a fixed element in E. Let d be a semi-metric on E, and let K be a real, asymmetrical kernel. We can extend our weighting technique to the functional case with [1]:

$$P_i = \frac{K(\frac{d(\chi,\chi_i)}{h})}{\mathbb{E}(K(\frac{d(\chi,\chi_i)}{h})}$$

3 Methodology - Nonparametric Prediction

In this section we aim to describe the main methods developed in [1] to predict a scalar valued random variable Y with compact support S from a functional explanatory random variable χ valued in the functional Space E. These methods will be used in a nonparametric setting, as according to definition 1.2. It is important to note that we can impose certain conditions on our explanatory f.r.v.; in [1] two types of restrictions are presented: the continuous type and the lipschitz type. These restrictions are often imposed partly because it is a necessity in order to satisfy the viability of our estimating methods. For instance, as we will show in Chapter 5, continuity is required to obtain convergence results for our estimators. On the other hand, lipschitz type conditions can allow us to obtain precise rates of convergence [1]. Thus while more restrictions are harder to satisfy, it can often give us more information.

3.1 Regression Method

This method refers to the regression nonlinear operator $r : E \to \mathbb{R}$ and we will consider the following model: $r \in C_E^0$ where [1]

$$C_E^0 = \left\{ f : E \to \mathbb{R}, \lim_{d(\chi,\chi') \to 0} f(\chi') = f(\chi) \right\}$$

or $\exists\, \beta > 0$ such that $r \in Lip_{E,\beta}$ where

$$Lip_{E,\beta} = \left\{ f : E \to \mathbb{R}, \exists C \in \mathbb{R}_*^+, \forall \chi' \in E, |f(\chi) - f(\chi')| < C d(\chi, \chi')^{\beta} \right\}.$$

Both the continuity framework and the lipschitz framework above are applicable. To condense the size of our report, we will only deal with continuous-type restrictions from now on. For more information on lipschitz-type restrictions, please refer to [1].

With the framework set, we can now estimate the regression, and here we use the following functional kernel regression estimator [1]:

$$\widehat{r}(\chi) = \frac{\sum_{i=1}^{n} Y_i K(h^{-1} d(\chi, \boldsymbol{\mathcal{X}}_i))}{\sum_{i=1}^{n} K(h^{-1} d(\chi, \boldsymbol{\mathcal{X}}_i))} \tag{3.1.1}$$

where K is an asymmetrical kernel and h depending on n is strictly positive real. To see how this estimator works, let us consider the following quantities:

$$w_{i,h}(\chi) = \frac{K(h^{-1} d(\chi, \boldsymbol{\mathcal{X}}_i))}{\sum_{i=1}^{n} K(h^{-1} d(\chi, \boldsymbol{\mathcal{X}}_i))} \tag{3.1.2}$$

Thus we can easily rewrite the kernel estimator $\widehat{r}(\chi)$ as follows:

$$\widehat{r}(\chi) = \sum_{i=1}^{n} w_{i,h}(\chi) Y_i \tag{3.1.3}$$

which is in fact a weighted average because

$$\sum_{i=1}^{n} w_{i,h}(\chi) = 1 \tag{3.1.4}$$

The behaviour of the $w_{i,h}(\chi)$'s can be deduced from the shape of the asymmetrical kernel function K.

Clarifying the idea behind this, if we consider positive kernels supported on [0,1], it is clear that the smaller $d(\chi, \boldsymbol{\mathcal{X}}_i)$, the larger $K(h^{-1}d(\chi, \boldsymbol{\mathcal{X}}_i))$. The closer $\mathcal{X}_i$ is to χ, the larger the weight assigned to Y_i. Also, as soon as $d(\chi, \boldsymbol{\mathcal{X}}_i) > h$, we have $w_{i,h}(\chi) = 0$. Therefore how we choose the parameter h is important. The larger h is, the larger the number of terms in the sum and the less sensitive $\widehat{r}(\chi)$ is with respect to small variations of the Y_i's. The smaller h is, the more sensitive $\widehat{r}(\chi)$ is to small variations of the Y_i's. In this sense h is called a smoothing parameter (or bandwidth) [1].

Other than the regression method, which in practice later we would see that it is more sensitive to outliers and extrema, we have other two methods to which we will compare and contrast in our analysis in later sections of this paper.

3.2 Conditional Median Method

This method uses the conditional median operator $F_Y^{\boldsymbol{\mathcal{X}}}$ which is a nonlinear operator from $E \times \mathbb{R}$ to $\mathbb{R}$. Without loss of generality we can assume that our operator is in the set. Compare to Chapter 5 in [1]

$$R_{cdf}^{\chi} = \{f : E \times \mathbb{R} \to \mathbb{R} \text{ s.t. } f(\chi, .) \text{ is a strictly increasing c.d.f.}\}.$$

Since the operator is in this set, the conditional median exists and it is the only one. The conditional median can be defined as

$$m(\chi) = {F_Y^{\chi}}^{-1}(1/2) \text{ where } F_Y^{\chi} = \begin{cases} \mathbb{R} \to [0,1] \\ y \mapsto F_Y^{\chi}(y) = F_Y^{\boldsymbol{\mathcal{X}}}(\chi, y). \end{cases}$$

Considering the constraint

$$C_{E\times\mathbb{R}}^{0} = \left\{ \begin{array}{c} f : E \times \mathbb{R} \to \mathbb{R}, \forall \chi' \in \aleph_{\chi} \text{ s.t.} \\ \lim\limits_{d(\chi,\chi')\to 0} f(\chi', y) = f(\chi, y) \\ \text{and } \forall y' \in \mathbb{R}, \lim\limits_{|y'-y|\to 0} f(\chi, y') = f(\chi, y) \end{array} \right\},$$

where $\aleph_{\chi} \subset E$ is the neighbour of χ, from this we can define a functional non parametric model for the conditional median operator like so

$$F_Y^{\boldsymbol{\mathcal{X}}} \in C_{E\times\mathbb{R}}^{0} \cap R_{cdf}^{\chi}.$$

Define K_0 to be a symmetric kernel and define H as

$$\forall u \in \mathbb{R}, \qquad H(u) = \int_{-\infty}^{u} K_0(v)\mathrm{d}v$$

We also define the kernel conditional c.d.f. estimator by

$$\widehat{F}_Y^{\boldsymbol{\mathcal{X}}}(\chi, y) = \frac{\sum_{i=1}^{n} K(h^{-1}d(\chi, \boldsymbol{\mathcal{X}}_i))H(g^{-1}(y - Y_i))}{\sum_{i=1}^{n} K(h^{-1}d(\chi, \boldsymbol{\mathcal{X}}_i))} \tag{3.2.1}$$

where g is a positive real number greater than zero.

From (3.2.1) we can define a kernel estimator of the functional conditional median $m(\chi)$ by

$$\widehat{m}(\chi) = \inf\left\{ y \in \mathbb{R}, \widehat{F}_Y^{\boldsymbol{\mathcal{X}}}(\chi, y) \geq \frac{1}{2} \right\}. \tag{3.2.2}$$

which helps us with the prediction problem.

3.3 Conditional Mode Method

The conditional mode predictor θ makes use of the nonlinear operator f_Y^{χ} and is naturally defined in the following way [1].

$$\theta(\chi) = arg\sup_{s \in S} f_Y^{\chi}(\chi, y)$$

The restriction, which we put on f_Y^{χ} will be such that it will guarantee uniqueness of the mode, hence we want a density function with a unique peak in its density. Therefore we will work with the following restrictive condition:

$$R_{dens}^{\chi} = \left\{ \begin{array}{l} f : E \times \mathbb{R} \to \mathbb{R} \; such\, that \; \exists \xi > 0, \exists !\, y_0 \in S \\ f(\chi, .) \text{is strictly increasing on} (y_0 - \xi, y_0) \text{and strictly decreasing on} (y_0, y_0 + \xi) \end{array} \right\}$$

We let $f_Y^{\chi} \in R_{dens}^{\chi} \times C_E^0$ this guarantees the unique solution of θ. In order to make practical use of this predictor we will again consider the Kernel Method. In fact we will use the CDF described in 3.2. to estimate f_Y^{χ}. In particular we will use $\frac{\partial}{\partial y}\widehat{F_Y^{\chi}}$. Thus our expression for $\widehat{f_Y^{\chi}}$ becomes

$$\frac{\partial}{\partial y}\widehat{F_Y^{\chi}} = \widehat{f_Y^{\chi}} = \frac{\sum_{i=1}^{n} K(h^{-1}d(\chi, \boldsymbol{\mathcal{X}}_i))\frac{\partial}{\partial y}H(g^{-1}(y - Y_i))}{\sum_{i=1}^{n} K(h^{-1}d(\chi, \boldsymbol{\mathcal{X}}_i))}$$

$$\implies \widehat{f_Y^{\chi}} = \frac{\sum_{i=1}^{n} K(h^{-1}d(\chi, \boldsymbol{\mathcal{X}}_i))K_0(g^{-1}(y - Y_i))}{\sum_{i=1}^{n} K(h^{-1}d(\chi, \boldsymbol{\mathcal{X}}_i))} \tag{3.3.1}$$

3.4 Summary

In this section we have outlined the main mathematical machinery we will use in order to predict r.r.v. from functional data. We will see in the next section how one can practically apply these methods. Another remark to make is about the metric or semi-metric and the bandwidth or

smoothness parameter h. As one can see these methodologies make extensive use of some metric d and the bandwidth h. The choice of such is a crucial step and will be briefly elaborated on in the next section. Further more we will also mention the asymptotic behavior of these approaches in the analysis chapter - in order to underpin this approach even more.

4 Application and Implementation

This chapter should serve as a a bridge between the theory presented up to this moment and real life problems. We will briefly touch upon the choice of the semi-metric and bandwidth parameter. Then we will focus on real data on US electricity consumption over the past 28 years we will test our methodology to predict the 28th year from the previous 27. We will introduce and explain R routines to handle data and to conduct non-parametric functional data analysis.

4.1 Preliminary mathematical discussion

So far we have only mentioned the use of a semi-metric d and the bandwidth(i.e. smoothness) parameter h. In order to make the methodology presented in this report applicable one has to make a choice of these parameters. In particular the choice should be made with some mathematical insight. This subsection should give a short account of this topic.

Smoothness Parameter

There are three main approaches in the kernel method approach in choosing the bandwidth parameter. See discussion in Chapter 7.1 in [1]. The first method will be about finding a global bandwidth from a "least squares problem". The other two are similar in the formulation, however, the main objective is not finding a fixed size of a neighborhood, induced by the semi-metric d, but to find the fixed optimal number of elements in a neighborhood of X.
In the case of the regression method one chooses the bandwidth parameter with the first approach by considering the regression sum (3.1.1).

$$R_{CV}^{kernel} = \frac{\sum_{i=1}^{n} y_i K(d(x_i, x)/h_{opt})}{\sum_{i=1}^{n} K(d(x_i, x)/h_{opt})} \tag{4.1.1}$$

Where h_{opt} is given by.

$$h_{opt} = arg\min_{h} CV(h)$$

and

$$CV(h) = \sum_{i=1}^{n} (y_i - R_{-i}^{kernel})^2$$

with

$$R_{-i}^{kernel} = \frac{\sum_{j=1, j\neq i}^{n} y_i K(d(x_i, x)/h)}{\sum_{j=1, j\neq i}^{n} K(d(x_i, x)/h)} \tag{4.1.2}$$

So one can see that this problem is similar to a least square problem, where finding the optimal h yields the least square solution. As mentioned before the other two approaches are about finding a optimal number of elements in the neighborhood of X. In this case we again use the sum (3.1.1), similar to (4.1.1) in this case h is a function of k and x, thus $h = h_k(x)$. Where k denotes the number of elements in the neighborhood of X. Hence we are optimising k.The global approach considers all x_i the local approach considers only the x_i closest to x.

$$R_{GCV}^{kernel} = \frac{\sum_{i=1}^{n} y_i K(d(x_i, x)/h_k(x_i))}{\sum_{i=1}^{n} K(d(x_i, x)/h_k(x_i))} \tag{4.1.3}$$

Where $k_{opt}(x_i)$ is given by.

$$k_{opt}(x_i) = arg\min_k GCV(k)$$

and

$$GCV(h) = \sum_{i=1}^{n}(y_i - R_{-i}^{kernel}(x_i))^2$$

with $R_{-i}^{kernel}(x_i)$ as in (4.1.2), with the only difference that h depends on k and x, and we are optimising k. For the local approach (LCV)we are optimising k by considering only the x_{i_0} closest to x. And the k_{opt} is given by the following.

$$k_{opt} = arg\min_k \left\| y_{i_0} - \frac{\sum_{j=1,j\neq i_0}^{n} y_{i_0} K(d(x_i,x)/h_k)}{\sum_{j=1,j\neq i_0}^{n} K(d(x_i,x)/h_k)} \right\| \tag{4.1.4}$$

For the other methods (i.e median and mode) very similar approaches are used and further discussed in Chapter 7.1 in [1].

Semi-Metric

The other crucial part for a practical approach is the choice of a semi-metric. It would be out of scope to discuss such a choice in depth in this report, however, a further discussion can be found in Chapter 13 in [1]. As part of this report we are going to present two typical classes of semi-metrics one can choose. One being the L^2 metric defined on the q-th derivative. i.e.:

$$d_q^{deriv}(\chi_i,\chi_j) = \sqrt{\int \left(\chi_i^{(q)}(t) - (X)^{(q)}(t)\right)^2 dt} \tag{4.1.5}$$

The other is based on the spectral-analysis decomposition of the Covariance Operator see [1] Chapter 3.4 and [2] Chapter 1.1. In particular it is based on a truncated sum of the FPCA(functional principal component analysis) terms. Hence it is given by.

$$d_q^{PCA}(\chi_i,\mathcal{X}) = \sqrt{\sum_{k=1}^{q}\left(\int [\chi_i(t) - \mathcal{X}(t)]v_k(t)dt\right)^2} \tag{4.1.6}$$

Where v_k is the eigenfunction of the expectation or covariance operator.[2]

$$\mathbb{E}(\mathcal{X}(.,t)f(.)) = \int_\Omega \mathcal{X}(\omega,t)f(\omega)dP(\omega)$$

4.2 Description of the Problem

The remainder of Chapter 4 will be about considering a concrete problem in order to show the reader how to apply the methods and practical approaches discussed earlier. The problem we will consider is predicting electricity consumption from data collected by the US. Parallel to this we will also test the effectiveness of the methods by predicting data we have already. The data is organised in the following way. Let $x_{i,j}$ be the electrical consumption in the j'th month of the i'th year.

	Jan	Feb	$\cdots$	Dec
Year 1	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,12}$
Year 2	$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,12}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Year 27	$x_{27,1}$	$x_{27,2}$	$\cdots$	$x_{27,12}$

Table 1: Data

Now, let $\mathbf{x}_i$ be the vector of electrical consumption for each month for the i'th year. We can interpolate these points to create the functional random variable $\boldsymbol{\mathcal{X}}_i$. To apply our methods we also need a scalar response variable $\mathbf{Y_i}$ and considering that we want to predict future electricity consumption a natural and smart choice for the scalar response variable is the s'th month of the $(i+1)$'th year, we will denote it by $\mathbf{Y_{i,s}}$. Setting up our model in this manner means we will have 12 separate prediction problems, each predicting the s'th month of the following year.

4.3 Implementation

We have obtained the raw data from [3], which consists of monthly electricity consumption in the United States from March 1973 until February 2001. In order for our methods to work we need to log the data so we can get rid of the heteroscedasticity, and we can get rid of the linearity by differencing the data. For a visual element you can see this graphically in [6] and for a further discussion of this see chapter 12 in [1]. This data can then be organised as shown in the matrix above (Table 1) where the functional data is given by the 28 yearly differenced log curves from March 1973 until February 2001. This data matrix will be logged into a file which is denoted by "npfda-electricity.dat".

We will use the first 27 years, to try and predict the 28th year with the methods outlined in chapter 4.2. Now that we have the data in the form that is usable, we can apply our methods using our matlab code.

```
Data = dlmread('npfda-electricity.dat');
%some omitted variables
for i = 1:27 %putting the data into functional form
    f{1,i} = @(x)interp1(months,Data(i,:),x);  end
for j=1:12 %
    for h= 1:2:30
        Y_Hat(j,h) = regression(Kernel_quadratic, ...);
    end  end
```

After organizing and some housekeeping of the electricity consumption data and putting it into functional form, we then predict the desired scalar response from our functional sample using the regression method.

```
function y_reg_hat = regression(Kernel, f_vec, y_vec, f_test, points,q,
    h,n);
if h<=0  h = 1 end
if n>size(f_vec) n= f_vec; end
for i = 1:n
    if abs(d_deriv(f_vec{1,i},f_test,points,q))<=h
                K(i) = Kernel(d_deriv(f_vec{1,i},f_test,points,q)/h);
    else K(i) = 0;
    end
end
y_reg_hat= K*y_vec'/sum(K);
```

We constructed a similar method for the median method.

```
function F_hat = median(Kernel, f_vec, y_vec, f_test, points,q,h,n);
f_temp = @H;
if h<=0
    h = 1;
end
if n>size(f_vec)
    n= f_vec;
end
for i = 1:n
    if abs(d_deriv(f_vec{1,i},f_test,points,q))<=h
                K(i) = Kernel(d_deriv(f_vec{1,i},f_test,points,q)/h);
    else K(i) = 0;
    end
end
F_hat = @(y) arrayfun(f_temp,(y-y_vec));
F_hat = @(y) K*F_hat(y)'/sum(K);
```

As the reader can see we have used the derivative based semi-metric. Concerning the Kernel in our actual analysis we have used a quadratic kernel, however, our methods allows the use of any Kernel. The routines above, given the Kernel and semi-metric, implement the methods presented in (3) but they do not find the optimal bandwidth parameter. Hence as one can see we run the regression for different h (the bandwidth) and determine the best fit by brute force calculations.

The results can be summarised in the following plots:

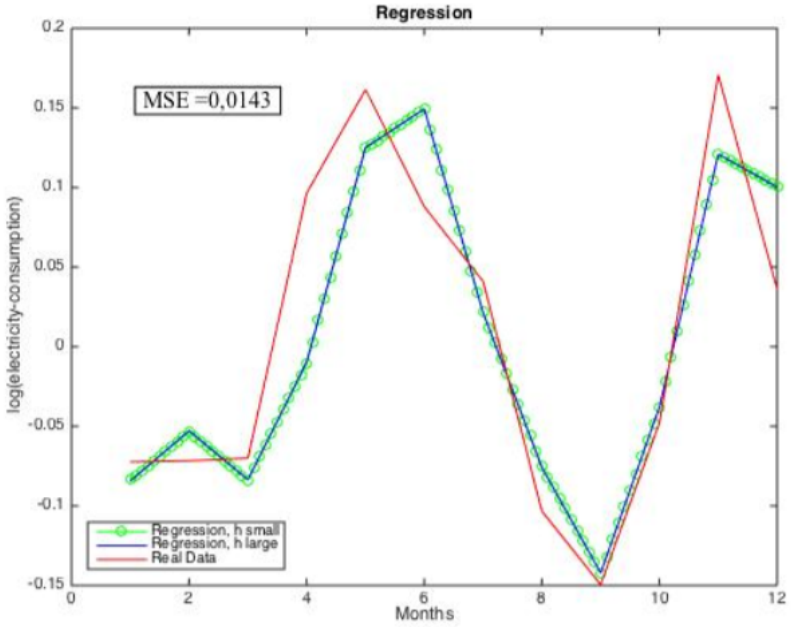

Figure 4: Regression

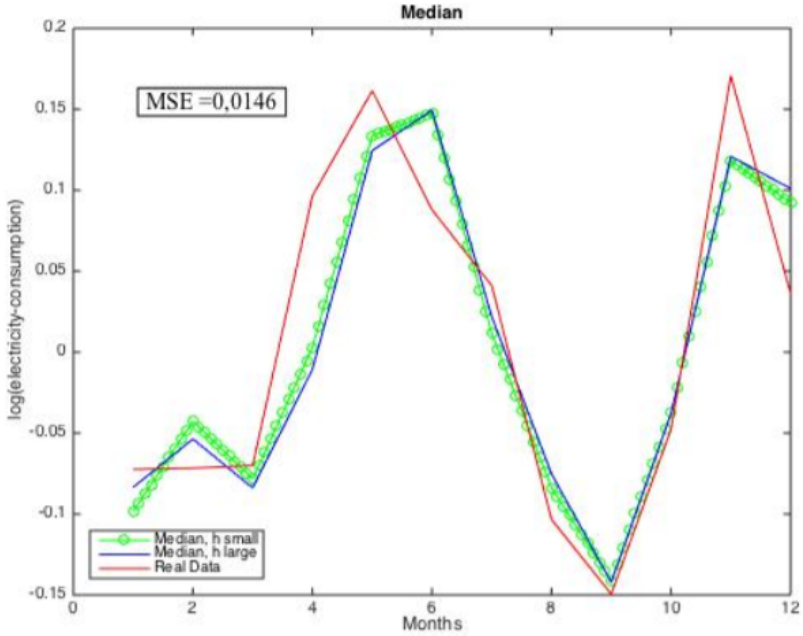

Figure 5: Median

To summarise some facts:

The regression method has a lower MSE, suggesting that with this Kernel and semi-metric the Regression method is a better estimator. Furthermore one can also observe that the bandwidth parameter does not strongly influence the regression method, not so with the median method. In general one sees that even a simple implementation of the methods yields powerful results.

We will also show an alternative implementation taken from [4]. This implementation in R makes extensive use of methodologies discussed in (4.1). I.e. calculating the optimal bandwidth and working with the PCA-based metric. The basic routines are the following:

```
35 result.pred.reg.step.s <- funopare.knn.lcv(elec.futur.s,elec.past.learn
       ,elec.past.testing,5,kind.of.kernel="quadratic",semimetric="pca")
                  # Kernel functional regression forecasting
36 result.pred.median.step.s <- funopare.quantile.lcv(elec.futur.s,elec.
       past.learn,elec.past.testing,2,alpha=0.5, Knearest=NULL, kind.of.
       kernel="quadratic",semimetric="pca")
                 # Kernel functional median forecasting
37 result.pred.mode.step.s <- funopare.mode.lcv(elec.futur.s,elec.past.
       learn,elec.past.testing,2,Knearest=NULL, kind.of.kernel="quadratic
       ",semimetric="pca")
                  # Kernel functional mode forecasting
```

The funopare.knn.lcv function refers to the local choice of k-nearest neighbours bandwidth selection method as mentioned in the previous chapter (4.1.3) (working with the regression method for kernel estimation). To see the code of this function, see reference [4]. The funopare.quantile.lcv and funopare.mode.lcv functions refer to the median and mode estimation methods respectively. The R code for these methods can also be found under the reference [4].The k mentioned is the same as in (4.1.4), also see appendix. Notice that the method is working with the PCA semi-metric and not the derivative one as explained in (4.1) as the PCA method captures the probability structure better see [2], however both are good families of metrics. Quadratic kernel is used for simplicity and it will be out of scope to address the best kernel to use in actual different practices.

Again to analyse the electricity consumption data, one would consider 12 separate models, calculating 12 different scalar estimates. We can summarise this approach in the following graphs.

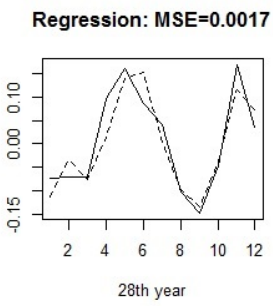

Figure 6: Regression

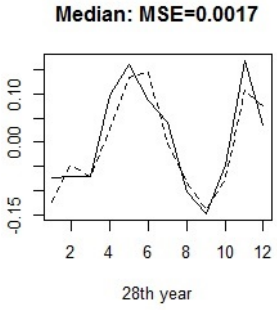

Figure 7: Median

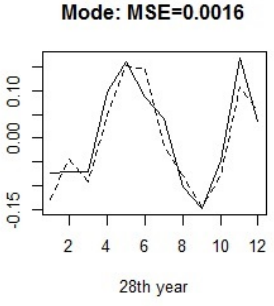

Figure 8: Mode

In the graphs above we see plots of different methods, namely the regression method, conditional median method and conditional mode method. The dotted line refers to the actual observed electrical consumption (for the 28th year) and the solid line refers to the forecasted electrical consumption. We see that each of methods give satisfactory results while the condition mode and median methods give slightly less MSE. This might be due to the regression method being more sensitive to extreme values which happened around the second month. It is also worth noting here that neither the dependence between the curves nor the small numbers of data obstructs the nice behaviour of our nonparametric functional methods.

4.4 Conclusion

In this Chapter we have seen how to apply the theory presented in (3) to real life data. For this purpose we have considered semi-metrics, bandwidth parameters and other subtleties in applying the theory. We saw that even our own matlab implementation works very well. Giving the reader an easily reproducible version for non-parametric function data analysis. The remaining questions are now: how to mathematically justify this approach and how to extend this methodology to non i.i.d cases.

5 Analysis and Comments

In this Chapter we want to present some mathematical justifications for the use of our methods. We will briefly discuss some very powerful theorems and thus justify this approach. In addition to this we will also point out some restrictions of this methodology and refer to further discussion in this field.

5.1 Convergence of Estimators

A very natural way to justify the effectiveness of an estimator is to show well-behaved asymptotics. In our case see Chapter 6 in [1] we have a proof of the following three theorems.

For the Regression Method:

Theorem 5.1. *Under the continuity-type model and under certain regularity conditions, refer to 6.2 in [1], then we have*

$$\lim_{n\to\infty} \widehat{r}(\chi) = r(\chi), \quad a.co.^1$$

For the Median Method:

Theorem 5.2. *Under the continuity-type model and under certain regularity conditions, refer to 6.2 in [1], then we have*

$$\lim_{n\to\infty} \widehat{m}(\chi) = m(\chi), \quad a.co.$$

For the Mode Method:

Theorem 5.3. *Under the continuity-type model and under certain regularity conditions, refer to 6.2 in [1], then we have*

$$\lim_{n\to\infty} \widehat{\theta}(\chi) = \theta(\chi), \quad a.co.$$

This means that our estimators also converge in probability to the true value, see [7]. This implies that we have a consistent sequence of estimators. Using standard results about estimators we can argue that having a consistent sequence of estimators justifies the use of this particular estimator.

[1] a.co. means almost sure convergence

5.2 Further Comments

Through our case study applications we realised that the model gives appealing results when the explanatory variables clearly follow a trend and therefore is in line with the assumption of the functional random variables being i.i.d.. This can be clearly seen in the electricity consumption data graph:

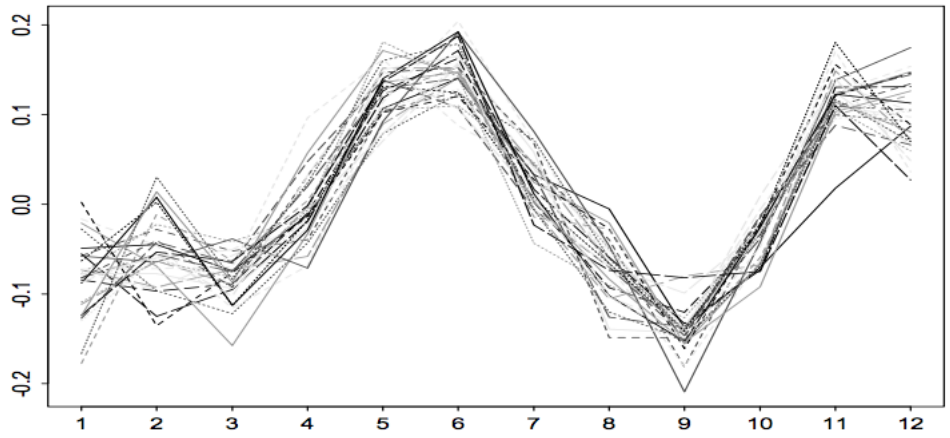

Figure 9: Transformed Electricity Consumption Data

This leads to our conclusion that the model might not be suitable in predicting sporadic or absolutely random processes, such as weather and stock market, where there are no visible trends. We hope further research can be made to extend the methods to predict models, where the functional random variables are not identical.

This concludes our analysis of the electricity consumption case study.

6 Endnote

As mentioned in the beginning of this report, we see that non-parametric functional data analysis yields very accurate results despite not having to hold any assumptions on parameters or distribution. In addition, implementing our estimator is rather simple, and we obtain good predictions even when we do not optimize our bandwidth or semi-metric. However, to conclude this report, we share and broadcast some open questions that we found to be interesting, from [1]. For instance, how can we find an optimal bandwidth automatically for any regression estimation, or conditional cdf? And how can we find the best semi-metric? While we have found certain methods of obtaining both the bandwidth and semi-metric, there is always room to develop even better strategies that can apply perhaps to all types of problems. Finally, in this report, we have only observed data as functions or curves, but we would like to consider the possibility of extending these data to even more complex shapes such as surfaces.

While non-parametric models are already quite widely used, we believe that with continuous advancements in technology, we will continue to gather data in a functional form. Recently, there has been many new publications on functional data. This shows a growing interest in this area of study, see Chapter 14 in [1]. All in all, while non-parametric functional data analysis is a relatively new type of modelling, we hope and can expect to see more breakthroughs in the future.

References

[1] F. Ferraty and P. Vieu. Nonparametric Functional Data Analysis. Theory and Practice. Springer series in statistics, Springer, 2006.

[2] Dauxois, J., Pousse, A., Romain, Y. Asymptotic theory for the princi- pal component analysis of a random vector function: some application to statistical inference. J. Multivariate Anal., 12, 136-154 (1982)

[3] http://www.math.univ-toulouse.fr/ ferraty/SOFTWARES/NPFDA/npfda-datasets.html

[4] http://www.math.univ-toulouse.fr/ ferraty/SOFTWARES/NPFDA/npfda-casestudies.html

[5] http://www.math.univ-toulouse.fr/ ferraty/SOFTWARES/NPFDA/npfda-elecfore.pdf

[6] http://www.math.univ-toulouse.fr/ ferraty/SOFTWARES/NPFDA/npfda-electricity-plot.pdf

[7] http://www.math.ist.utl.pt/ mjmorais/Chapter5incomplete.pdf

[8] http://www.psych.mcgill.ca/misc/fda/ex-weather-b1.html

[9] https://en.wikipedia.org/wiki/Kernel_(statistics)