Report (Group 9): Home Work 3

Course Code: Data Science Lab 3 (MTH312A)

Submitted by, Group 9

Anirban Ghosh(221271), Khyati Singh(221332)

Dasari Charithambika(210302)

Rajdeep Adhya(221385), Rohit Dutta(221396)

Instructor

Dr. Subhra Sankar Dhar

Associate Professor

Department of Mathematics and Statistics,

Indian Institute of Technology, Kanpur

Submission Date: 15 March, 2024

1 Question 1

Generate data with outliers, which can be embedded into $L_2[0,1]$ space. Propose a methodology for outlier detection/estimation of proportion of outliers in an infinite dimensional data and implement your methodology on the generated data.

- First let's begin with defining $L_2[0,1]$

  $L_2[0,1]$: **A Non-Euclidean Space** -

$$L_2[0,1] = \{f : \omega \mapsto \mathbb{R} | \int_0^1 f^2(x)dx < \infty\}$$

- **Wiener Process(Brownian Motion):**
  - It is an example of Elements from $L_2(0,1)$ space.
  - The Wiener process, also known as Brownian motion, is a continuous-time stochastic process characterized by real-valued increments. It is named after mathematician Norbert Wiener and is widely used in various fields such as finance, physics, and biology to model random movements or fluctuations over time. The characteristics of the process (W_t) are below:
    1. $W_0 = 0$ (almost surely)
    2. W has independent increments i.e for every $t \geq 0$, the future increments $W_{t+u} - W_u, u \geq 0$, are independent of $W_s, s < t$
    3. $W_{t+u} - W_u \sim \text{Normal}(0, u)$
    4. W_t is almost surely continuous in t
- **Drifted Brownian Motion:**

  Drifted Brownian motion is often denoted by X_t, where t represents time. This process incorporates a deterministic drift term, typically denoted by μ, which represents the average rate of change or trend in the process, and σ^2 infinitesimal variance. So, X_t can be expressed

as:

$$X_t = \mu t + \sigma W_t$$

Data Generation:

Our task is to provide data that contains outliers and can be embedded into the $L_2(0,1)$ space. Think about the combination of a Drifted Wiener Process ($W_2(t)$) and a Wiener Process ($W_1(t)$), where the drift is regarded as significant enough to include outliers. It is given by

$$W(t) = \alpha_1 W_1(t) + \alpha_2 W_2(t)$$

with $\alpha_1 = 0.8$ and $\alpha_1 = 0.2$

Algorithm:

- **Generation from Standard Wiener process:**

  Step 1: Define the time points at which data are collected: $t_0 = 0, t_1 = 0.1, t_2 = 0.2, ..., t_{100} = 10.0$.

  Step 2: Generate 100 random observations from the normal distribution $\mathbf{N}(0, 0.1)$.

  Step 3: Compute the cumulative sum of these random observations. This results in the observations corresponding to the Wiener process at each time point denoted as $W_{11}, W_{12}, ..., W_{1T}$. By definition, $W_{10} = 0$.

- **Generation from Drifted Wiener process:**

  Step 1: Generate Observations From Standard Wiener Process[from the above process] say, $W^*_{20}, ..., W^*_{2T}$.

  Step 2: Consider the values of $\mu = 2$ and $\sigma = 1$. $W_2(t) = \mu t + \sigma W^*_2(t)$

- **Generation from Mixture of Wiener processes:**

  Step 1: From the Bernoulli(0.85) distribution, produce an observation, such as u.

  Step 2: If $u = 1$, generate observation from Standard Wiener Process. Otherwise, generate it from the Drifted Wiener Process.

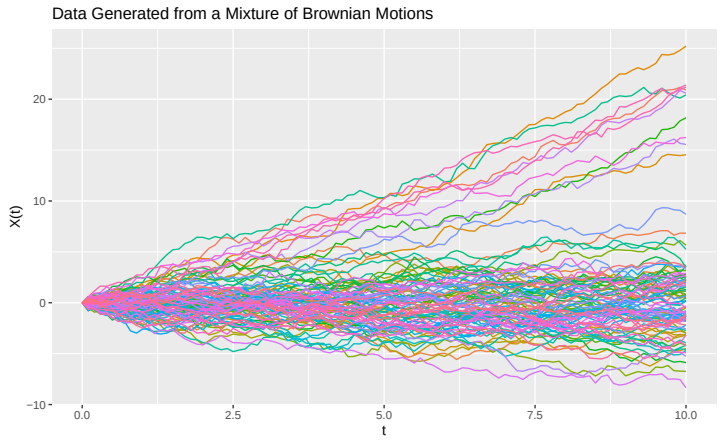

Figure 1: Data Generated from a Mixture of Brownian Motions

Using the above algorithm, we have generated 100 observations from the mentioned process. Figure 1 gives the plot of the data

Methods to Detect Outliers:

Method: Creating a Band of Extremes

- **Creating a Band :** Establish a pair of functions that connect the point-wise 10^{th} and 90^{th} percentiles. These two functions will be regarded as belonging to the lower and upper bands, respectively.
- **Outlier Detection :** If for a data point, values at more than 85% of the time points are above the upper band or below the lower band, it is suspected as an **Outlier**.

We have identified a few possible outliers in our simulated data using this technique. The indices are- 2, 8, 25, 43, 76, 79, 85, 92 and 93 (as shown in Figure 2).

2 Question 2

Consider a regression model $Y = m(X)+\epsilon$, where $m : L_2[0,1] \mapsto \mathbb{R}$.Propose an estimator of m for a given random sample $(X_1, Y_1), ..., (X_n, Y_n)$, and study the performance of your proposed estimator for a simulated data.

- **Regression Model:**

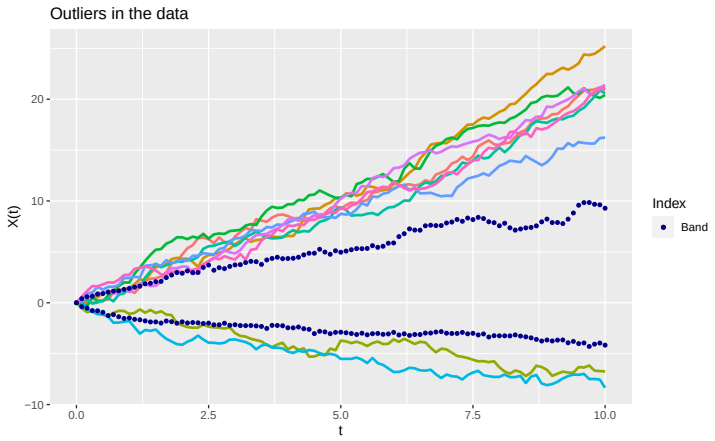

Figure 2: Figure of the Outliers and Bands

Consider a regression model $Y = m(X) + \epsilon$ where $m : L_2[0,1] \mapsto \mathbb{R}$.

- **Data Generation:**

  1. To generate observations X over a time interval ($[0, T]$ with 100 discrete time points, we use the Wiener process. This involves generating 100 observations from the Standard Brownian Motion (Wiener Process) algorithm outlined previously. Each observation corresponds to a time point within the interval.
  2. To generate observations Y, the true model is considered as $m(X) = \int_0^T X^2(t)dt$ and $\epsilon \sim \text{Normal}(0, 1)$. So finally

  $$Y_i = \int_0^T X_i^2(t)dt + \epsilon$$

  The integral for each data point was calculated during the simulation using the Riemann Sum.

- **An estimation of 'm':**

  We will be using **Nadaraya-Watson Estimator** in the case of non-parametric regression when $m : \mathbb{R} \mapsto \mathbb{R}$. We're trying to develop an estimator that is similar to this one for the

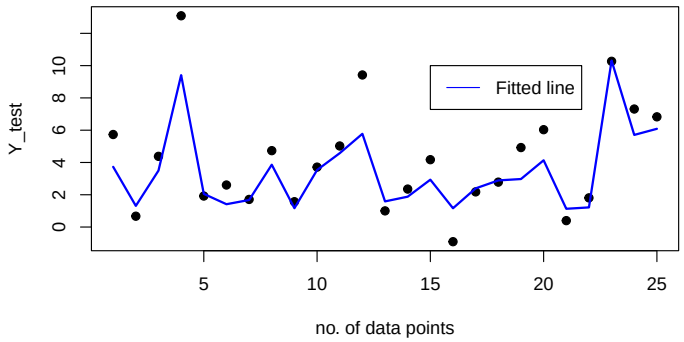

Figure 3: Figure of the Fitted Regression Curve on the Simulated Dataset

functional data here.

$$\hat{m}(X) = \frac{\sum_{i=1}^{n} K(\frac{\|X - X_i\|_{L_2[0,1]}}{h_n}) y_i}{\sum_{i=1}^{n} K(\frac{\|X - X_i\|_{L_2[0,1]}}{h_n})}$$

Here we have considered the Gaussian Kernel i.e. $K(z) = \frac{1}{\sqrt{2\pi}} \exp(\frac{-z^2}{2})$ and the bandwidth $h_n = 10$

- **Splitting Dataset:**

  Random sampling was the method used to divide the dataset. Of the entire data, 75% are part of the training set and the remaining 25% are part of the testing set. Using the training dataset, we fitted the suggested model, and using the test dataset, we verified the model's correctness.

- **Fitting:**

  Figure 3 gives us the fitting of the regression curve on the simulated dataset. In order to verify the model's correctness, we computed the RMSE of the fit on the test dataset, and the value is the 7.2437. Since, we have 25 data points on the test set which is low enough, we can conclude that, the fit on the data is quite good.