Data Science 3 Project: Homework 3

Group 7: Aindrila, Suchandra, Chirag, Paritosh, Vanshika

14 March, 2024

1 Question 1

Infinite Dimensional Data:

"Infinite-dimensional data" typically refers to datasets or spaces where the number of dimensions is theoretically unbounded or infinitely large. In this field, data points are functions rather than vectors. Examples of infinite-dimensional data include functional data (where each data point is a function), spaces of functions such as Hilbert spaces, and stochastic processes. For example, we might have data consisting of curves, images, or signals. These datasets can be represented as functions in infinite-dimensional spaces. Dealing with infinite-dimensional data often requires advanced mathematical techniques from functional analysis, measure theory, and stochastic processes. Computational methods for infinite-dimensional data often involve approximations, discretizations, or other strategies to handle the infinite nature of the space.

Difference between Infinite Dimensional Data and High Dimensional Data:

The difference between high-dimensional and infinite-dimensional data lies primarily in the size and nature of the spaces in which the data resides.

- High-dimensional data refers to datasets where the number of dimensions (features or variables) is very large but still finite. Infinite-dimensional data refers to datasets or spaces where the number of dimensions is theoretically unbounded or infinitely large.
- In high-dimensional spaces, the number of dimensions is typically large enough to pose computational and analytical challenges, for example, hundreds or thousands of dimensions whereas in infinite-dimensional spaces, the number of dimensions is not finite, which presents unique mathematical and computational challenges.

Outliers in Functional Space:

Functional Data Analysis (FDA) is a statistical discipline focused on modeling and analyzing data recorded continuously over a range, such as trajectories or time courses. Unlike traditional time series analysis, the FDA considers repeated observations without assuming stationarity. It's well-suited for studying time-dependent and longitudinal data, common in biomedical research and other fields. FDA assumes that observed data come from independent identically distributed samples of a stochastic process, often presumed to be smooth and typically residing in spaces like L_2 or reproducing kernel Hilbert spaces. The primary goal of the FDA is to understand and model this underlying stochastic process.

A basic problem is that the smooth underlying process rarely is fully observed and the available discrete observations that are thought to be generated by the process are often noisy. In some cases the data are also sparsely observed, a frequently encountered scenario for longitudinal data.

Outliers in functional data analysis refer to observations that deviate substantially from the overall trend or pattern of the data. In functional spaces, such outliers can manifest as data points or curves that exhibit extreme behavior compared to the rest of the dataset.

For outlier detection in functional space, we need to know how to measure the distance between two observations where every observation has to be a function.

In $L_2[0,1] = \{x_n : \sum_{n=1}^{\infty} x_n^2 < \infty\}, n \in \mathbb{N}$, we know, $||f||_{L_2[0,1]} = [\int f^2(x)dx]^{1/2}$. One of the measures of distance in functional space can be defined as

$$||f - g||_{L_2[0,1]}.$$

Kosambi–Karhunen–Loève Expansion:

The Kosambi–Karhunen–Loève (KKL) expansion, also known as the Karhunen–Loève transform or Karhunen–Loève expansion, is a mathematical technique used in functional data analysis and signal processing. The KKL expansion is widely used for dimensionality reduction and feature extraction in functional data analysis.

The KKL expansion decomposes a stochastic process $X(t)$ as an infinite linear combination of orthogonal eigenfunctions, typically denoted by $\{e_k(t)\}$, which are determined by the covariance structure of the process, and associated uncorrelated random variables Z_k.

$$X(t) = \sum_{k=1}^{\infty} Z_k e_k(t).$$

where $X \in L_2[0,1]$. Here, Z_k are uncorrelated random variables with zero mean $E(Z(t)) = 0, \forall t \in [0,1]$. and unit variance, and $e_k(t)$ are the eigenfunctions of the covariance operator associated with the stochastic process. These eigenfunctions form an orthonormal basis for the space in which the process resides.

$$< e_k(t), e_{k'}(t) >_{L_2[0,1]} = 0; ||e_k(t)||_{L_2[0,1]} = 1$$

The eigenfunctions $\{e_k(t)\}$ are determined by solving the eigenvalue problem associated with the covariance operator of the stochastic process. This involves finding functions $e_k(t)$ and corresponding eigenvalues λ_k such that:

$$C(s,t) = \text{Cov}(X(s), X(t)) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s)\phi_k(t),$$

where $C(s,t)$ is the covariance function of $X(t)$.

This decomposition allows for the representation of the random process in terms of uncorrelated random variables called principal components. These principal components capture the most significant variability in the data and are ordered according to their importance.

A basic property of this KKL Expansion is,

$$\int_0^1 [X(t) - \sum_{k=1}^{N} Z_k e_k(t)]^2 dt \longrightarrow 0 \text{ as } N \longrightarrow \infty.$$

Brownian motion:

Brownian motion, named after the botanist Robert Brown who observed the erratic movement of pollen grains suspended in water, is a fundamental stochastic process in mathematics and physics. Mathematically, Brownian motion is often described using stochastic calculus, particularly utilizing the Wiener process. Here's a mathematical description:

In its continuous form, Brownian motion $(W_t)_{t \geq 0}$ is a stochastic process indexed by time $t \geq 0$. The following properties characterize it:

- **Increment Stationarity:** For any $s < t$, the random variable $W_t - W_s$ has the same distribution as $W_{t-s} - W_0$.

- **Gaussian Increments:** The increments $W_t - W_s$ are normally distributed with mean 0 and variance $t - s$, i.e., $W_t - W_s \sim \mathcal{N}(0, t - s)$.

- **Continuous Paths:** Brownian motion paths are continuous functions of time with probability 1 (almost surely). This means that with probability 1, the paths have no jumps or discontinuities.

- **Independent Increments:** Increments $W_t - W_s$ are independent of $W_u - W_v$ for disjoint time intervals $[s, t]$ and $[u, v]$.

Generating data from Brownian motion involves simulating the random movement of particles over time. Brownian motion is a stochastic process, meaning it involves randomness. Here's how you can generate data from Brownian motion:

Generate random numbers from a normal distribution (Gaussian distribution) with mean 0 and variance equal to the time step. These random numbers represent the increments of the Brownian motion. Add up these random increments cumulatively to obtain the positions of the particles at each time step.

This generates Brownian motion data over a time interval $[0, T]$ with N time steps, where T is the total time. You can adjust T and N according to your requirements. The increments are drawn from a normal distribution with a mean 0 and variance equal to the time step size. Then, the cumulative sum of these increments gives the Brownian motion W.

Data Generation:

To generate data from Brownian motion, we can use the fact that increments of Brownian motion are normally distributed. So, we can generate increments from a normal distribution and then cumulatively sum them to get the Brownian motion path. Let, $X(t)$ be a Gaussian Process, taking values $X(t_1), \ldots, X(t_L)$ where $t \in [0, 1]$, L is sufficiently large. We simulate data from $0.8 * X(t) + 0.2 * X(t)_{drifted}$ where $X(t) \sim N(0, 1)$ and $X(t)_{drifted} \sim N(100, 1)$. Here $L = 1000$. We generate 50 such Gaussian processes, shown in 1:

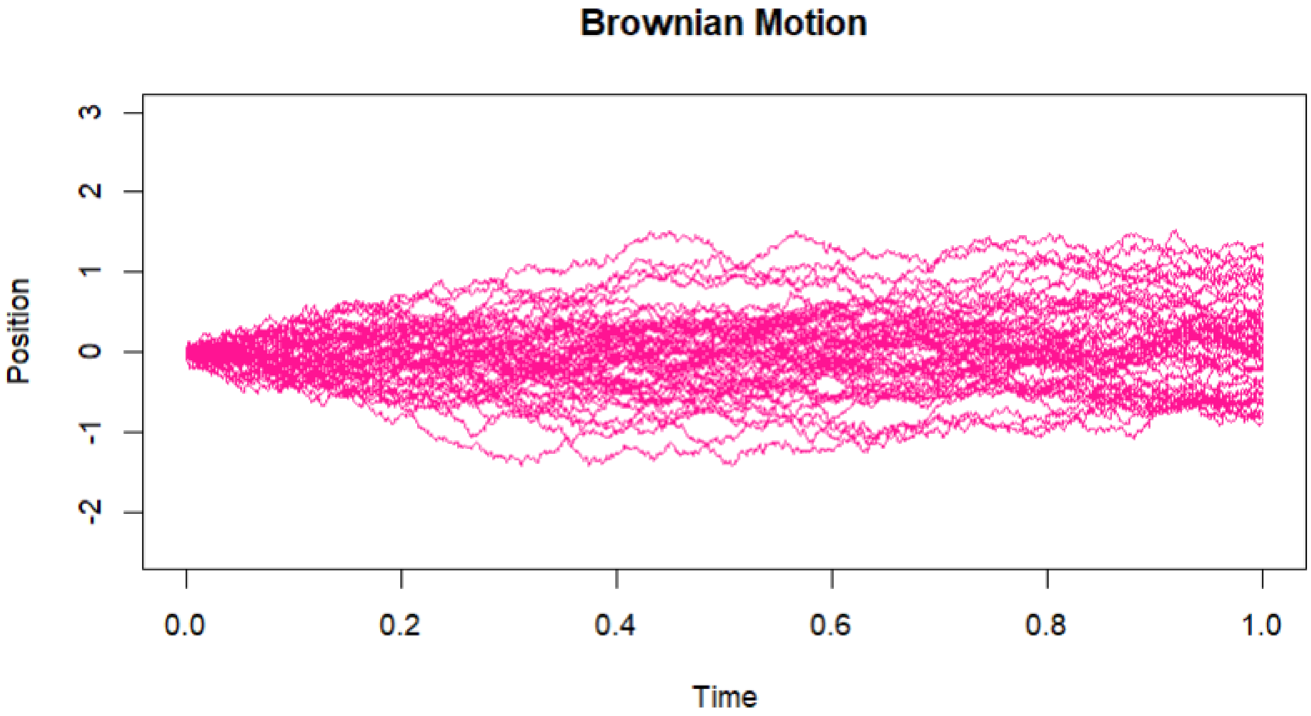

Figure 1

Methodology for Outlier Detection & Estimation of Proportion of Outliers:

1. **Functional Depth:** Functional depth measures, such as the band depth or modified band depth, can quantify the centrality of a functional observation relative to the entire dataset. Outliers typically have lower depth values, indicating their deviation from the bulk of the data.

   Here, our idea is based on pointwise maxima and minima. We calculate the pointwise local upper and lower bands considering all the Gaussian processes. Next, we classify a new Gaussian process as an outlier if 50 percent of its points lie beyond those bands.

   To find the lower and the upper bands, we compute pointwise 95% confidence interval and pointwise 90% percent confidence interval. For each of these intervals, we take one Gaussian process at a time and count the number of points that lie beyond the pointwise intervals. If the number of points exceeds half of the total number of points of that Gaussian process, we consider that Gaussian process to be outlier.

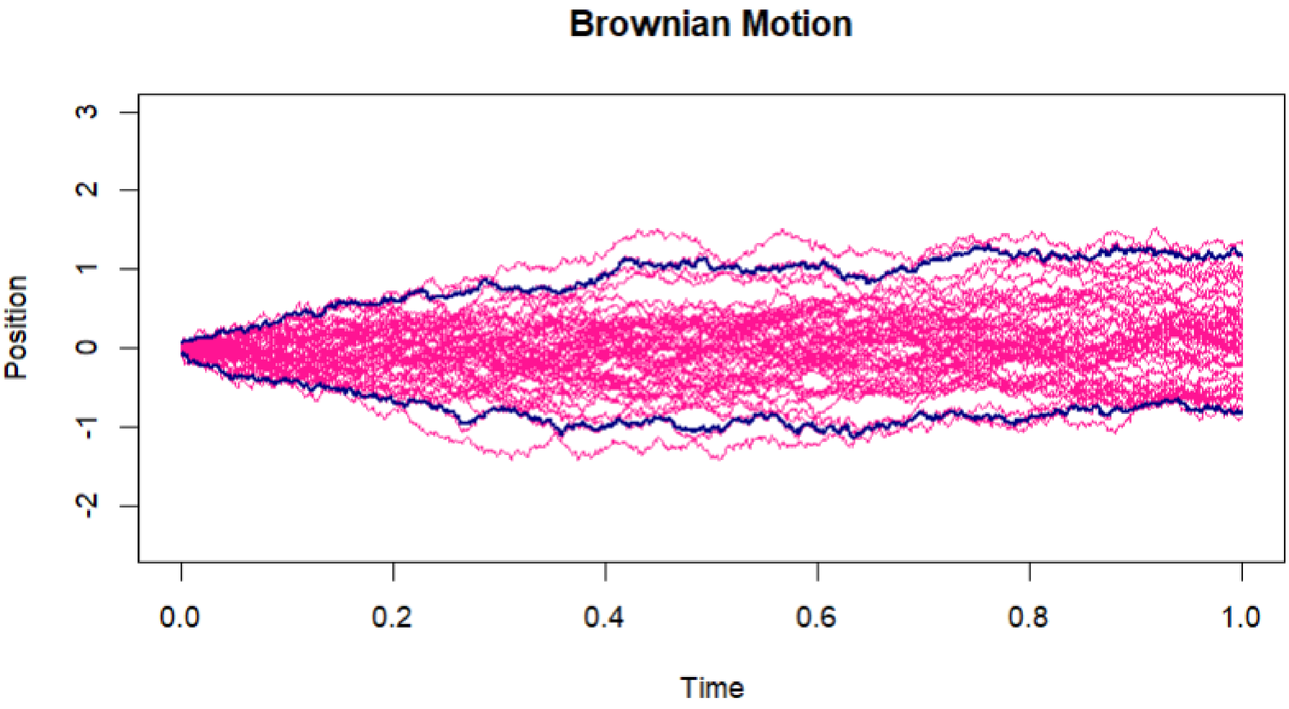

Figure 2: 95% Confidence Interval

Figure 2 shows the pointwise 95% confidence intervals. After computation, we find that 2 of the 50 Gaussian processes have half of the points lying outside the confidence intervals. Hence, the proportion of outliers = 2/50 = 0.04.

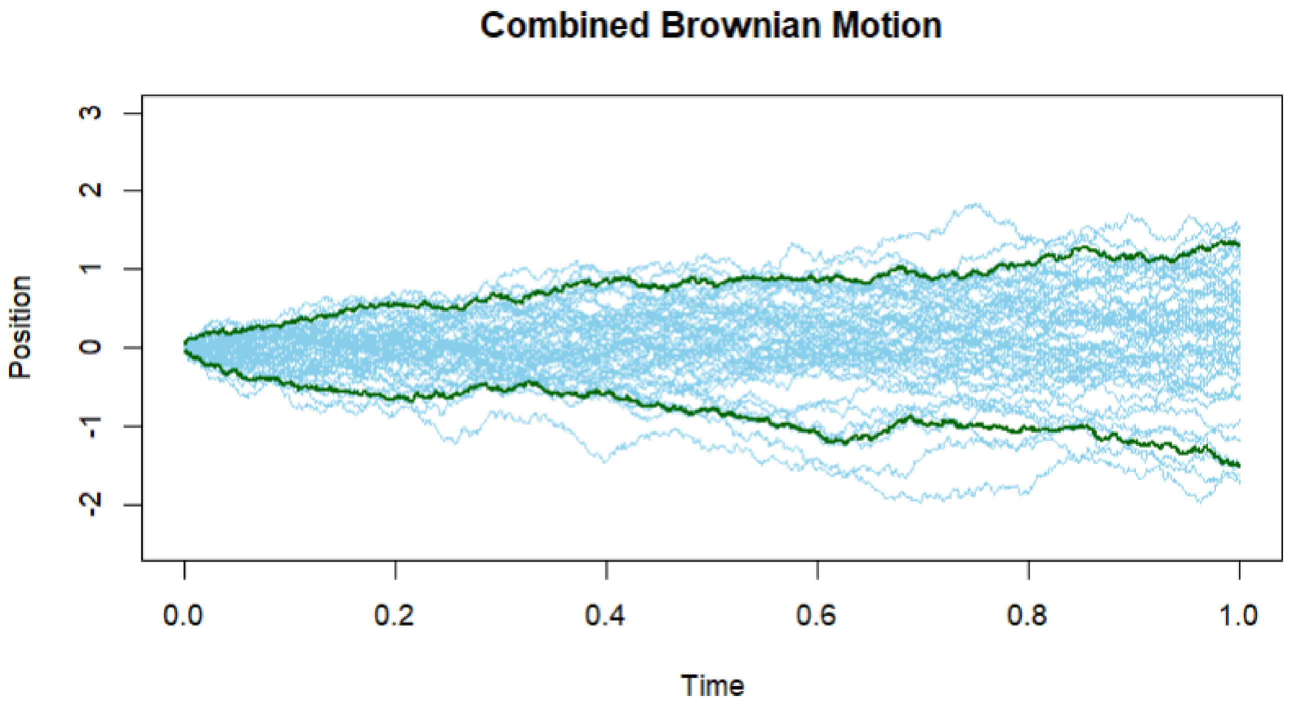

Figure 3: 90% Confidence Interval

Figure 3 shows the pointwise 90% confidence intervals. After computation, we find that 3 of the 50 Gaussian processes have half of the points lying outside the confidence intervals. Hence, the proportion of outliers = $3/50 = 0.06$.

2. **Based on Trimmed Mean:** Using robust estimators that are less sensitive to outliers can help mitigate their influence on statistical analyses. For example, robust measures of location and dispersion, such as the median and trimmed standard deviation, can provide more reliable estimates in the presence of outliers.

   Trimming or Winsorizing involves removing or downweighting extreme observations from the dataset. This approach can help reduce the impact of outliers on statistical estimates without discarding the entire observation.

   For the gaussian processs, $X_1(t), X_2(t), \ldots, X_n(t)$, the trimmed mean is

$$\bar{X}_\alpha = \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} X_i(t), \alpha \in (0, \frac{1}{2})$$

.

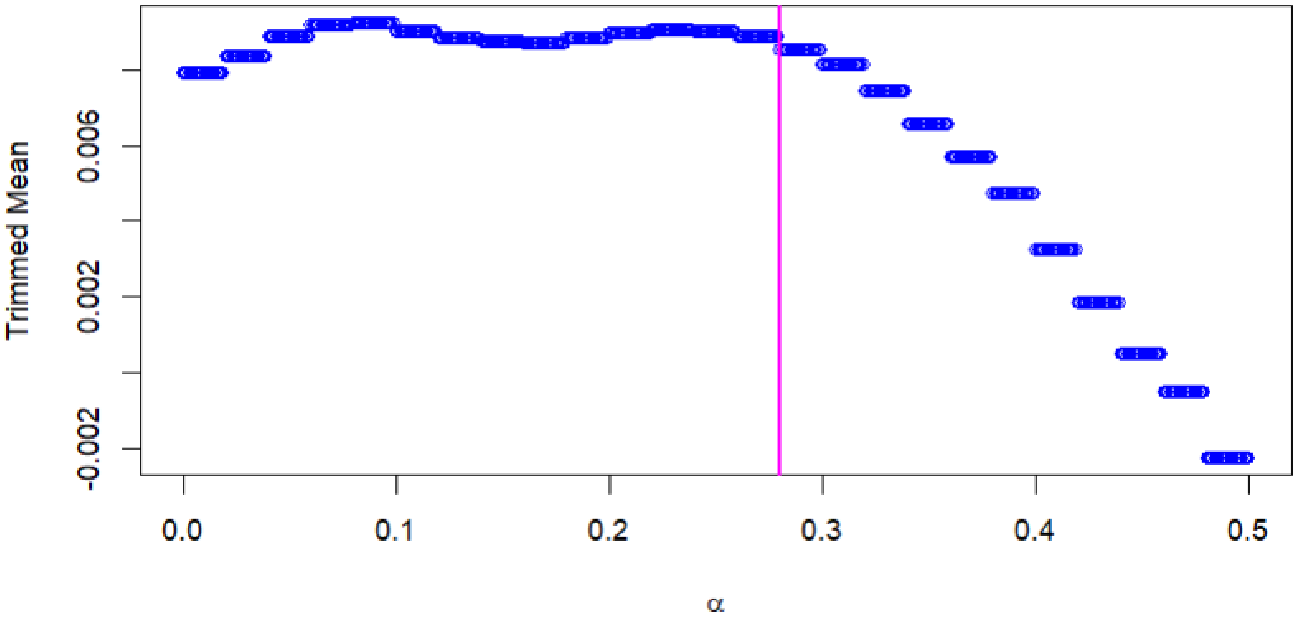

Figure 4: Figure for Trimmed Mean to compute Proportion of Outliers

Figure 4 shows the plot of $\bar{X}_\alpha$ against a sequence of values for α between $(0, 0.5)$. The pink vertical line in the plot at $\alpha = 0.28$ shows the cut-off line beyond which the values will be classified as outliers. The value of $\bar{X}_\alpha$ at $\alpha = 0.28$ is 0.008523449. Hence, we can understand from the plot that more than 5% outliers are present in the data.

2 Question 2

In general, the regression equation is $Y = m(X) + \epsilon$, $m : \mathbb{R} \longrightarrow \mathbb{R}$ but here we will work with $m : L_2[0,1] \longrightarrow \mathbb{R}$.

Proposed Estimator:
Nadaraya-Watson Estimator:

$$\hat{m_n}(x) = \frac{\sum_{i=1}^n y_i k(\frac{x - x_i}{h_n})}{\sum_{i=1}^n k(\frac{x - x_i}{h_n})}$$

where h is the bandwidth and k is the kernel density function for $m : \mathbb{R} \longrightarrow \mathbb{R}$.

$$\hat{m_n}(x) = \frac{\sum_{i=1}^n y_i k(\frac{||x - x_i||_{L_2[0,1]}}{h_n})}{\sum_{i=1}^n k(\frac{||x - x_i||_{L_2[0,1]}}{h_n})} \tag{1}$$

for $m : L_2[0,1] \longrightarrow \mathbb{R}$ at a particular function.

Data Generation:

Consider the model:

$$Y = m(X(t)) + \epsilon, m : L_2[0,1] \longrightarrow R \tag{2}$$

Let

$$m(X(t)) = \int_0^1 X^2(t)dt \tag{3}$$

Then,

$$Y = \int_0^1 X^2(t)dt + \epsilon \tag{4}$$

To generate data from Brownian motion, we can use the fact that increments of Brownian motion are normally distributed. So, we can generate increments from a normal distribution and then cumulatively sum them to get the Brownian motion path. Let, X(t) be a Gaussian Process, taking values $X(t_1), ..., X(t_L)$ where $t \in [0,1]$, L is sufficiently large. We simulate data from standard normal. Here L = 1000. We put these values in (4) to get Y. We generate 100 such Gaussian processes, shown in Figure 5.

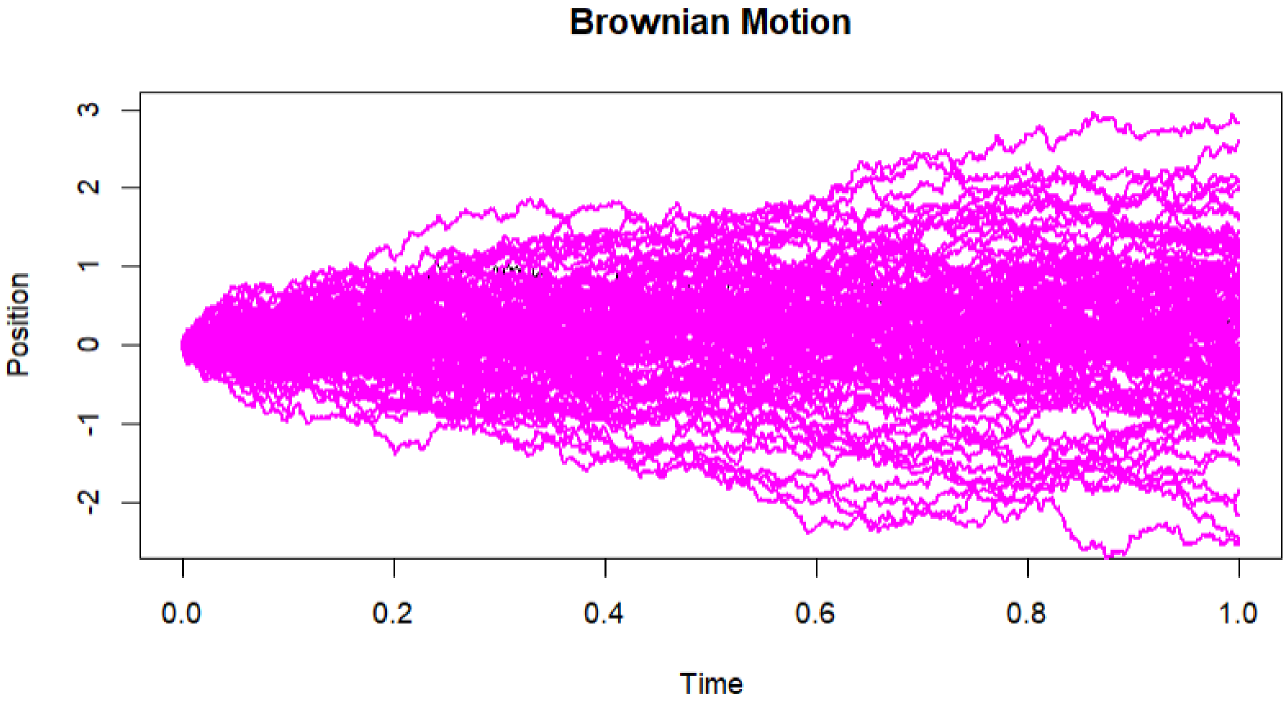

Figure 5

Comparison of Performance:

Comparing the performance of the Nadaraya-Watson estimator using data generated from Brownian motion can provide insights into how well the estimator captures the underlying structure of the data and its ability to make accurate predictions.

Consider $\hat{m_n}(x(t)) - \int_0^1 X^2(t)dt$. $\int_0^1 X^2(t)dt$ is a function independent of t. However, since the estimator $\hat{m_n}(x(t))$ is not independent of t, we can think that there is some bias in this estimator.

To study the performance of our proposed estimator, we will check if

$$sup_{t\in[0,1]}(\hat{m_n}(x(t)) - m(x(t))) \longrightarrow 0 \ as \ n \longrightarrow \infty, \tag{5}$$

i.e., $\hat{m_n}(x(t))$ converges to $m(x(t))$ almost surely.

Here, we will use "rmse" as a model evaluation criterion.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{n}}$$

Methodology:

70% of the generated data points are classified into a train set and the rest into a test set. We fix the bandwidth $h = 10$. Now for each of these sets, we calculate the value of $\hat{m_n}(x)$ using these points, from (1). Now using the obtained values from the proposed estimator, we calculate its RMSE from the value of those points, when put in (3). We repeat this for both the testing and the training sets.

The value of RMSE obtained is 1.03214 from the train set and 1.022297 from the test set.

If the RMSE is close to 1, it generally indicates good performance of the proposed estimator. An RMSE of close to 1 signifies that, on average, the proposed estimator's predictions are off by only 1 unit of whatever measurement scale is being used. This level of accuracy is often considered satisfactory, especially in many real-world applications where small errors are acceptable or even expected. Therefore, achieving an RMSE of 1 suggests that the proposed estimator is effectively capturing the underlying patterns in the data and making reliable predictions.

Real Life Examples where covariates are functions whereas the response is a scalar:

- Using historical data on disease outbreaks along with continuously evolving covariates to predict the future spread of diseases. The response, in this case, is the scalar count of reported cases within a defined time period.
  - Covariates: Time-varying factors such as temperature, humidity, population density, and vaccination rates.
  - Response: Number of reported cases of a contagious disease in a specific region.
- Using time-varying covariates to model and forecast inflation rates. The response is a scalar quantity representing the overall change in prices within a specific time frame.
  - Covariates: Economic indicators such as interest rates, money supply, and unemployment rates, which vary over time.
  - Response: Inflation rate, representing the percentage change in the general price level of goods and services over time.
- Using time series data of various covariates to predict the movement of stock prices. The response is a scalar representing the expected change in stock prices within a given time horizon.
  - Covariates: Historical stock prices, trading volumes, and market sentiment indicators, all of which change over time.
  - Response: Future stock price movement or return on investment.
- Using historical traffic data along with time-varying covariates to predict future traffic patterns. The response is a scalar representing the expected level of congestion or travel time at different times of the day.
  - Covariates: Time-dependent factors such as time of day, weather conditions, and special events affecting traffic flow.
  - Response: Traffic congestion levels or travel time between specific locations.

In each of these examples, the covariates are functions of time because they vary continuously over time. However, the response remains scalar, representing a single outcome or measurement of interest associated with the time-varying covariates.