Indian Institute of Technology, Kanpur

Data Science Lab
(MTH312A)
HOME WORK-3
REPORT

Group Members:
Mrinmoy Saha (221352)
Sudesh Kumari (221440)
Vijay Soren (211163)
Manjeet Chaudhary (221346)
Pravin Raman Bharti (221375)

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

Answer :

$\mathbf{L}^2[0,1]space$

Informally, an L^2-function is a function $f : X \to \mathbb{R}$ that is square integrable, i.e.,

$$\|f\|^2 = \left(\int_X |f|^2 d\mu\right)^{1/2}$$

with respect to the measure μ, exists (and is finite), in which case $|f|$ is its L^2-norm. Here X is a measure space and the integral is the Lebesgue integral. The collection of L^2 functions on X is called $L^2(X)$ (ell-two) or L^2-space, which is a Hilbert space.

Approach

Brownian motion, a stochastic process, finds extensive applications across various fields, including finance, physics, and biology. This report presents an analysis of standard and drift Brownian motion, focusing on their trajectories and establishing bandwidth boundaries for the standard Brownian motion paths.

Generating Standard Brownian Motion Paths

We utilize a function to generate standard Brownian motion paths with irregular time intervals, ensuring realism in the simulated trajectories. These paths are representative of the typical behavior observed in the $L^2[0,1]$ space.

Generating Drift Brownian Motion Paths

To introduce outliers into the dataset, we generate drift Brownian motion paths by incorporating a drift term. These paths deviate systematically from the standard paths and act as outliers within the $L^2[0,1]$ space.

Magnitude of Drift

In the data generation process, the magnitude of drift is a critical parameter that determines the extent of deviation from standard Brownian motion. In our analysis, we set the drift term to a value of 4, indicating a significant deviation from the baseline behavior.

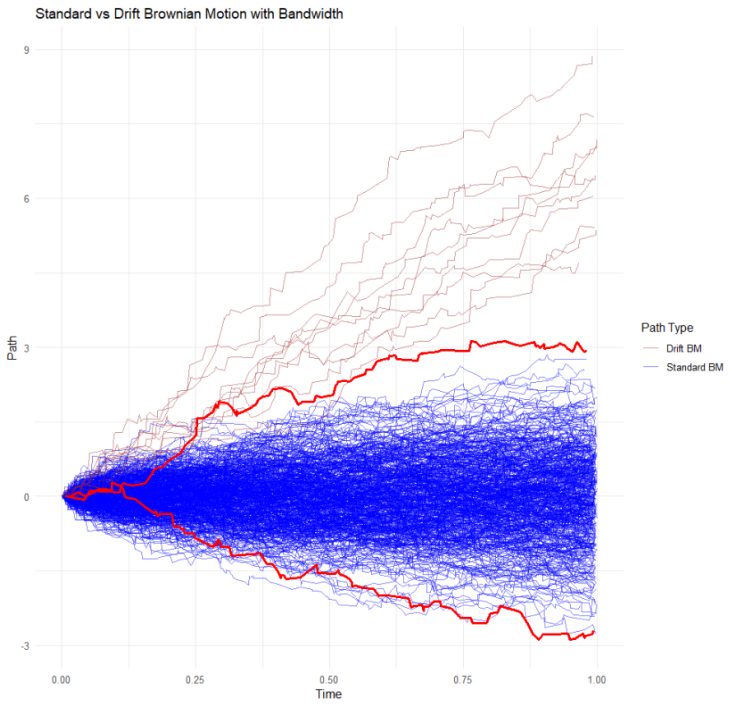

Figure 1: Standard vs drift Brownian Motion with Bandwidth

Outlier Detection

Collective maxima and minima for standard Brownian motion paths are computed. These values serve as bandwidth boundaries for the standard paths, allowing the identification of outliers.

Conclusion:

In conclusion, the report provides insights into how drift affects Brownian motion paths in the L2[0,1] space and how outlier detection techniques can be applied to identify and analyze these deviations. It underscores the importance of considering drift when analyzing stochastic processes and their trajectories

Question 2: Consider a regression model $Y = m(X) + \epsilon$, where $m : L_2[0,1] \to \mathbb{R}$. Propose an estimator of m for a given random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, and study the performance of your proposed estimator for simulated data.

Answer :

Theory

The Nadaraya-Watson estimator (often abbreviated as the N-W estimator) is a non-parametric technique used for estimating a regression function $E(Y|X = x)$ from a set of paired observations (X_i, Y_i), where X_i are the independent variables and Y_i are the dependent variables.

The N-W estimator is given by:

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} \left(K\left(\frac{||x - x_i||_{L_2[0,1]}}{h} \right) y_i \right)}{\sum_{i=1}^{n} \left(K\left(\frac{||x - x_i||_{L_2[0,1]}}{h} \right) \right)}$$

where $K(\cdot)$ is the kernel function with bandwidth h.

Overall, the N-W estimator is a flexible and widely used tool for non-parametric regression, particularly when the underlying relationship between variables is complex or unknown. However, its performance depends on the appropriate choice of bandwidth and kernel function.

Data Simulation

For data simulation, we use $X(t) = a \sin(2\pi t)$ where a is a coefficient. By changing the values of a, we obtain different sets of real-valued function data.

We define $Y = \int_0^1 X^2(t)dt + \epsilon$, where ϵ represents random noise.

Interpretation:

Here given $Y = m(X) + \epsilon$ where m is a function from L_2 to $\mathbb{R}$, and X is a set of real-valued functions.We use $X(t) = a\sin(2\pi t)$ where a is a coefficient which is represented in figure 2. Then we integrate it from 0 to 1 to get simulated Y(True value of Y). Here we simulate 30 functions, then we get 30 values which is Y, and then by using the extended version of the NW estimator, we estimate Y. Then we check for different values of h we get different estimates of y which is shown in Figure 3. in figure 4 we see for a table for true value and their estimates.

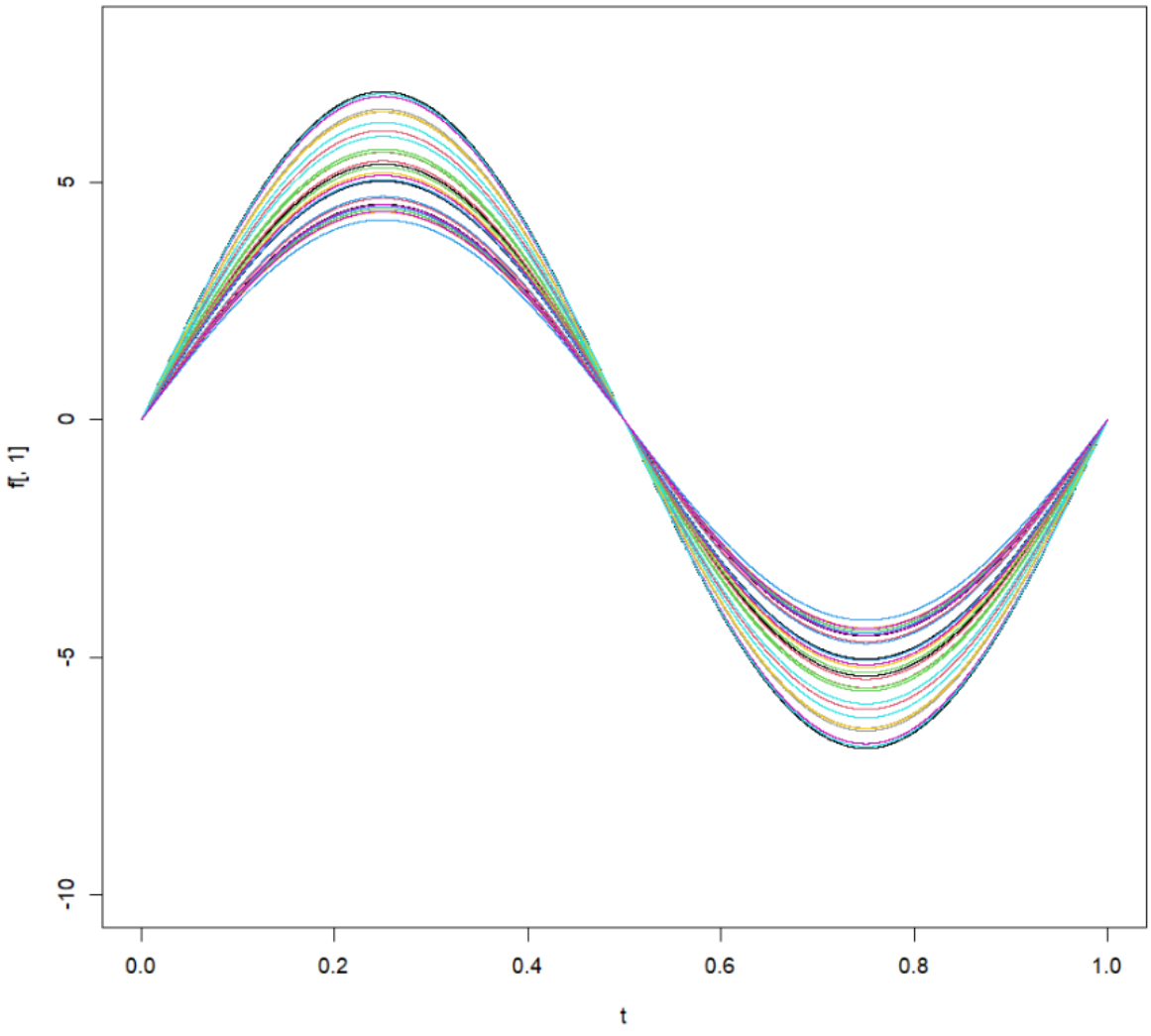

Figure 2: Above plot shows simulated X data

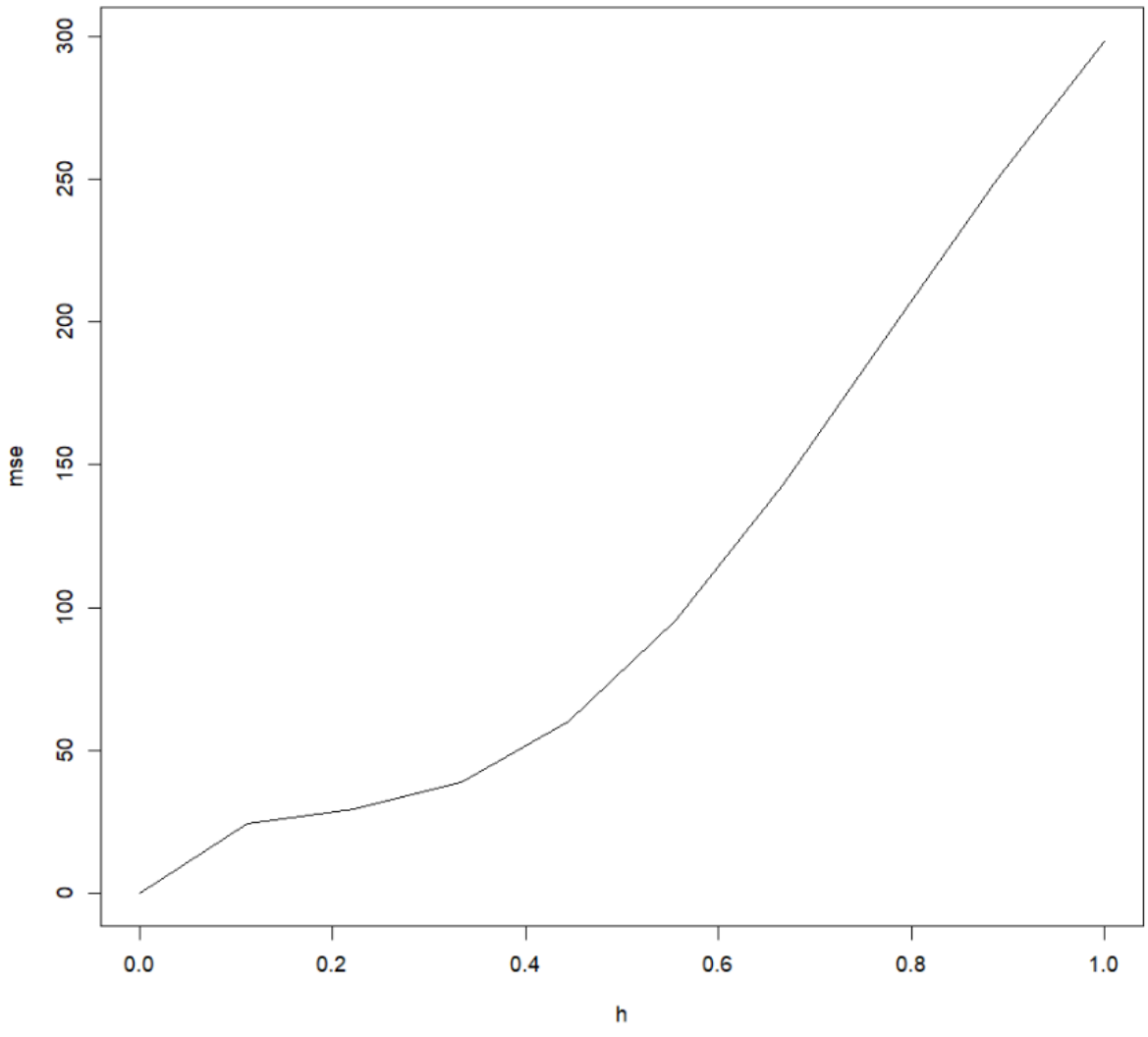

Figure 3: Bandwidth

: The plot shows how the MSE changes with different bandwidth values, helping to determine the optimal bandwidth for the NW estimator.

	true_y	est_y
1	8.061335	8.061335
2	19.527075	19.527075
3	16.394887	16.394887
4	11.156538	11.156538
5	23.767084	23.767084
6	25.418171	25.418171
7	8.425051	8.425051
8	22.716890	22.716890
9	16.556560	16.556560
10	15.965912	15.965912
11	13.554901	13.554901
12	11.600793	11.600793
13	19.132314	19.132314
14	11.109176	11.109176
15	13.892164	13.892164
16	22.259972	22.259972
17	24.325835	24.325835
18	12.015557	12.015557
19	13.943902	13.943902
20	8.148376	8.148376
21	17.318619	17.318619
22	11.600491	11.600491
23	20.291811	20.291811
24	9.348883	9.348883
25	12.463445	12.463445
26	14.556739	14.556739
27	7.932095	7.932095
28	12.016768	12.016768
29	25.625352	25.625352
30	10.289809	10.289809

Figure 4: True Value and their estimates

Conclusion :

As bandwidth h increases the Mean square error also increases for The Nadaraya-Watson estimator, so we get lesser the h values better the estimator (we can see that from the table above estimated Y values are very close to true Y values)