Functional Data Regression

Madhur Bansal (210572)

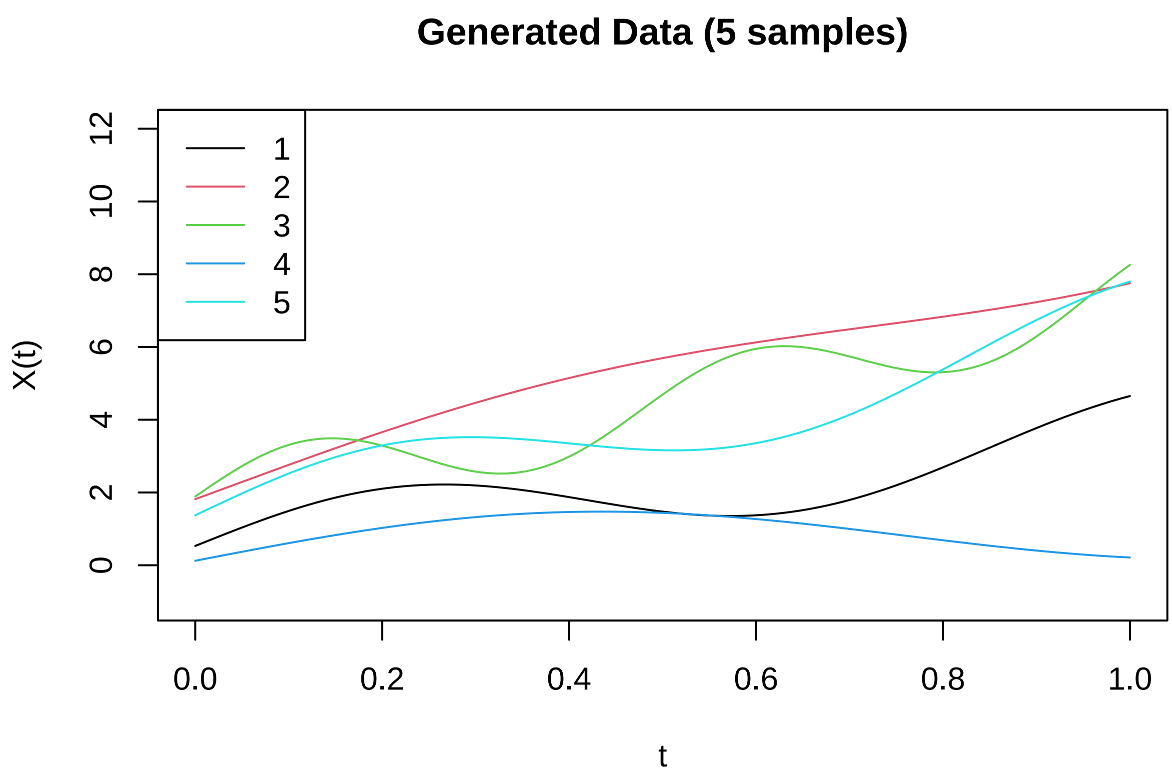
2024-03-14

Simulating the data

We construct a random variable $X_{(t)}$ from $L^2_{[0,1]}$ space using the following:

$$X_{(t)} = (c_1 * e^t) + sin(c_2 * 10t) + (c_3 * 2t)$$

where c_1, c_2, c_3 are random variables drawn from $uniform(0,2)$. We generate (n = 100) samples of $X_{(t)}$ as our given data. Here is the plot of 5 samples from the generated data.



We simulate real Y using the following:

$$Y = m(X_{(t)}) + \epsilon$$

where $m(X_{(t)}) = \int_0^1 X^2_{(t)}(sin(t) + cos(t))dt$ and $\epsilon \sim N(0,1)$

Here are the real values of y_i for above $X^i_{(t)}$ in respective order:

```
## [1]  8.204814 41.079967 33.139107  1.021508 26.661513 25.103260
```

Proposed Estimator (Based on Nadaraya-Watson Estimator)

I am using something similar to Nadaraya-Watson estimator for estimating m. The proposed estimator is as follows:

$$\hat{m}(X_{(t)}) = \frac{\Sigma_{i=1}^{n} K(||X - X^i_{(t)}||_{L_2}/h) y_i}{\Sigma_{i=1}^{n} K(||X - X^i_{(t)}||_{L_2}/h)}$$

where:

$K(.)$ is a valid kernel. I have used standard Normal $N(0,1)$

h is the band-width

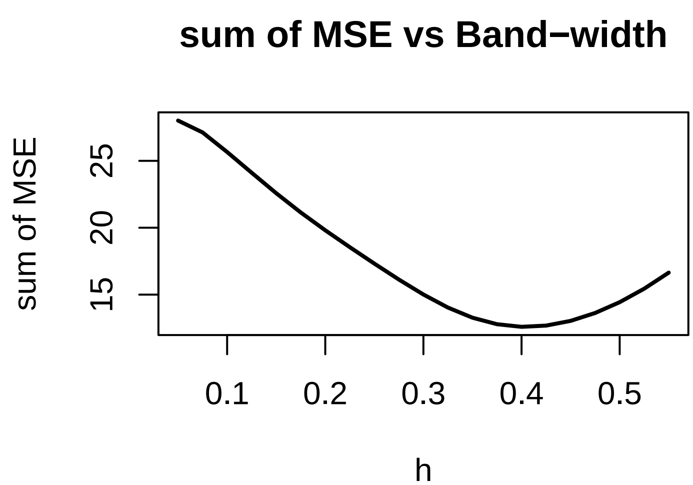
$||.||$ is the norm on $L_{2[0,1]}$ space

Choosing the Bandwidth (h)

In order to choose the appropriate band-width for the simulated sample data, I am using K-fold cross validation.

Here are the steps I followed:

1. I generate a grid of band-width containing 21 values - $h = \{0.05, 0.075, ...0.55\}$
2. Then I perform K-fold cross validation (with $K = 4$), and calculate the sum of the MSE obtained from cross-validation, for every h in the grid.
3. Chose the h which gives the lowest sum of MSE from the K cross validation results.




```
## [1] "Chosen Band-width: 0.4"
```

Estimation for new data

Now, we generate 100 new $(X_{(t)}, y)$ in order to evaluate the performance of our estimator.

Here are some of the estimates made for the newly generated data:

```
##          Real y Estimated y       Error
## [1,] 25.039944   25.519573 -0.47962918
## [2,] 21.278169   20.465545  0.81262407
## [3,]  1.686397    2.611027 -0.92463012
## [4,]  8.952435    9.773870 -0.82143463
## [5,] 18.974028   18.906697  0.06733104
## [6,] 14.837879   14.011128  0.82675184
```

```
## [1] "MSE =  1.08949042918219"
```

```
## [1] "R2 =  0.989799998756462"
```

Conclusion

The implementation of the Nadaraya-Watson estimator for functional regression was successful, with the model achieving a Mean Squared Error (MSE) of 1.09 and an R-squared (R^2) score of 0.98. These results show that the model performs very well in terms of accuracy and explains almost all of the variability in the data.