STATS205 Autumn 2019 Homework 2

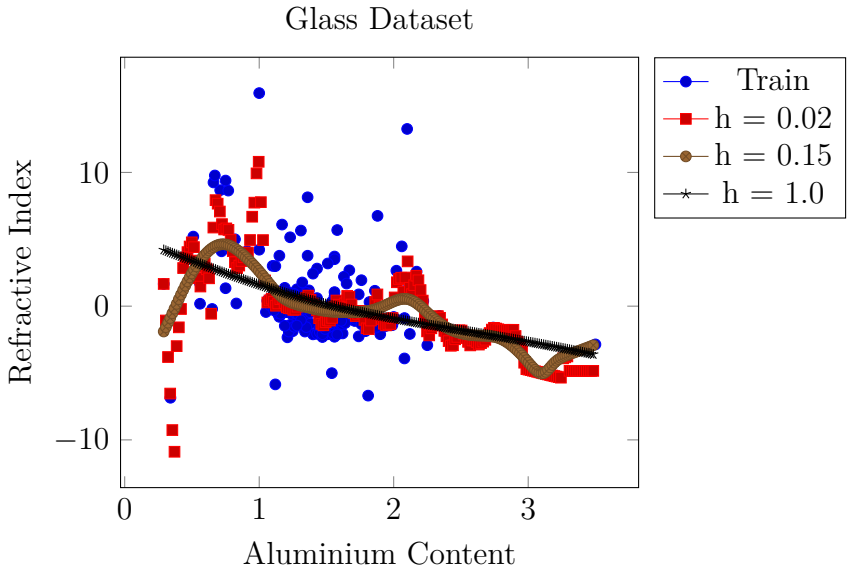
SUNet ID: raunakbh
Name: Raunak Bhattacharyya
Collaborators: [list all the people you worked with]

By turning in this assignment, I agree by the Stanford honor code and declare that all of this is my own work.

Problem a: Local Linear Regression

Part 1: Local linear regression on Glass dataset

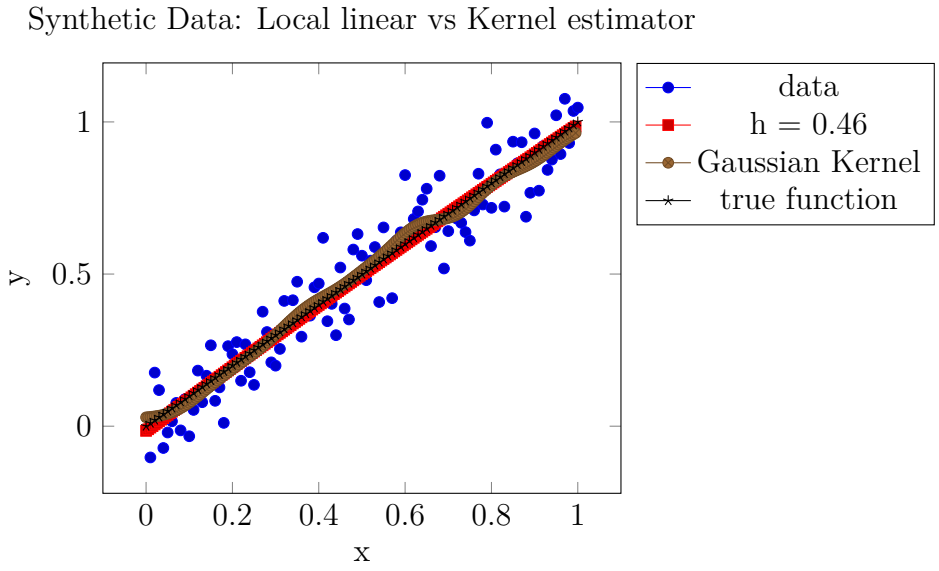
Here we see the impact of the bandwidth on the local linear regression performance.



From the figure $h = 0.02$ seems to be overfit to the data so looks like a case of high variance. $h = 1.0$ seems to miss capturing the trends in the data thereby displaying high bias. $h = 0.15$ seems to be better than both in terms of bias-variance tradeoff. We would need some empirical caluclation of MSE to make a more accurate statement.

Part 2: Local linear regression vs Kernel estimator

Here we compare local linear regression against the kernel estimator for the syntetic data generated by adding noise to $r(x) = x$. The Gaussian kernel is used for both the estimators. From the figure, both methods seem to be performing almost equally well (it is hard to

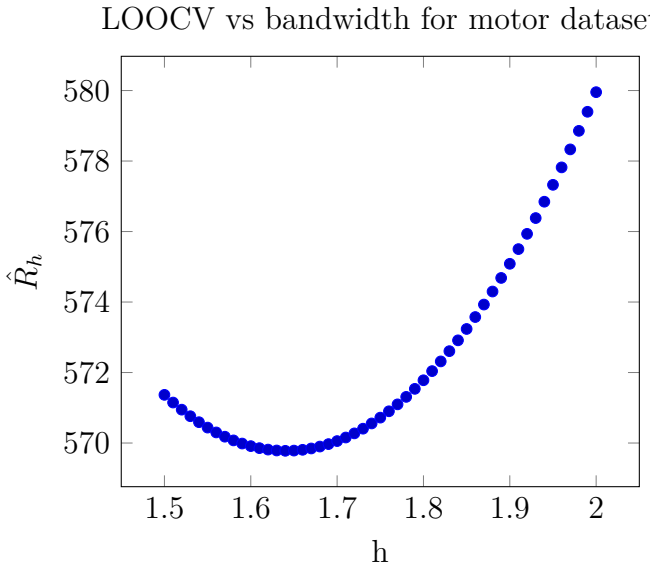


decide which is better just by looking at the plots). One possible reason is that the values of h used to generate these plots are the best for both methods, respectively as determined by leave one out cross validation.

Problem b: Cross Validation

Part 1: Motorcycle dataset

Here we see the variation of the LOOCV while the bandwidth is varied. The reason for

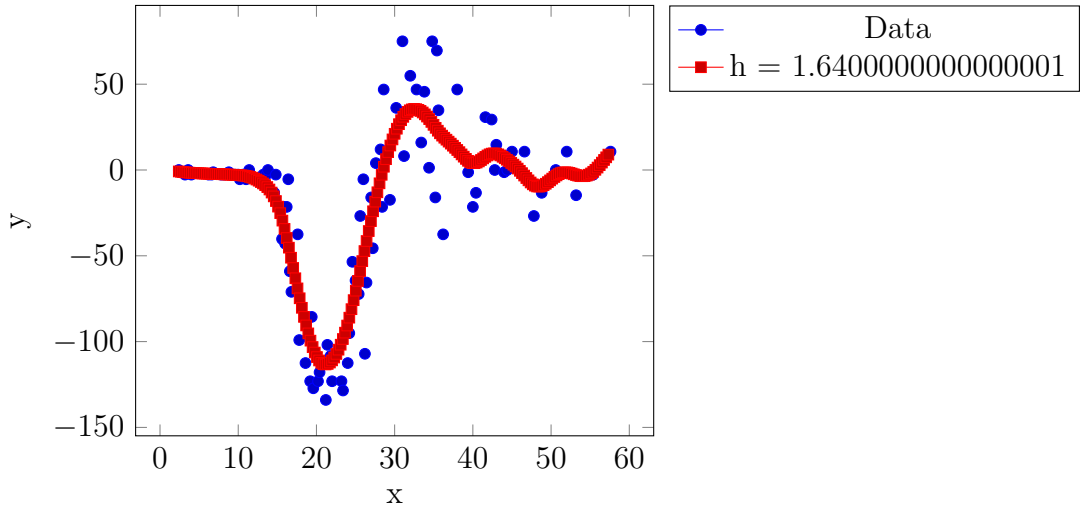


seeing a unimodal function is that there is a value of bandwidth that achieves the best tradeoff between bias and variance and hence the lowest empirical estimate of MSE.

Part 2: Optimal bandwidth

Here is the plot of the local linear regression estimate using the optimal bandwidth $h_{opt} = 1.64$.

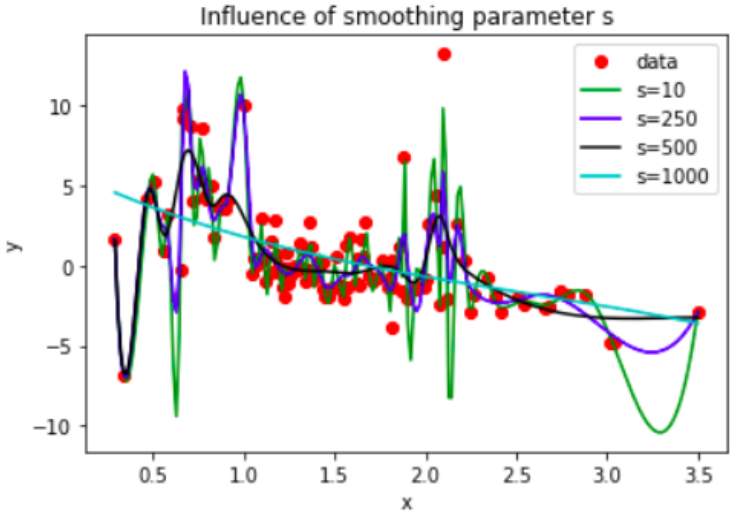
Local linear regression with optimal bandwidth for motor dataset



Problem c: Splines

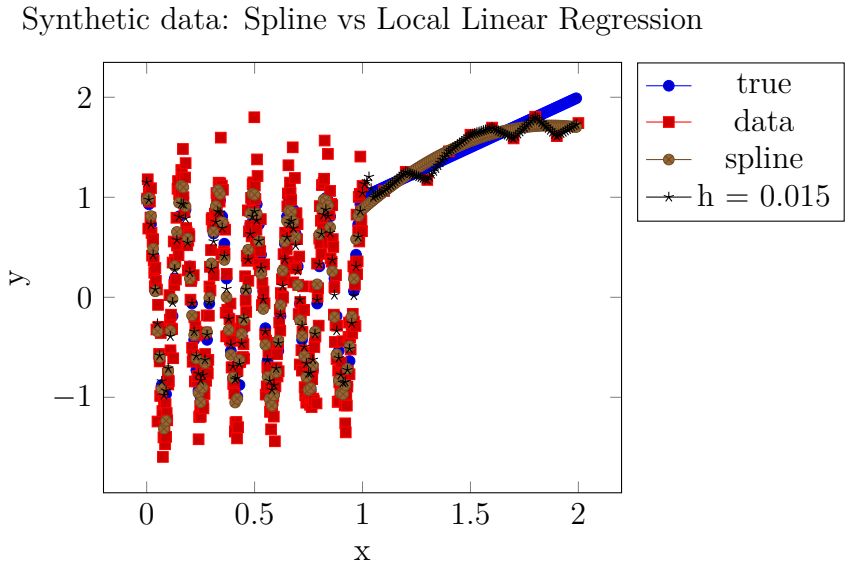
Part 1: Glass dataset

Here we see the bias-variance tradeoff depending on the choice of the smoothing parameter in case of the glass dataset.



Part 2: Cubic spline vs local linear regression

Figure shows the comparison of splines against local linear regression in case of the synthetic data which is cosine when $x \in [0, 1)$ and linear when $x \in [1, 2]$



It is difficult to compare the two methods because the provided smoothing parameter for splines, and bandwidth for local linear regression are good choices (maybe even optimal) thus both methods perform quite well in estimating ground truth.

However, the splines does output a smoother estimate as compared to locally linear regression. This is as expected because splines solve the objective function which aims to optimize smoothness of the fit in addition to MSE.