ISYE 8803 Homework 1 Problem 3

Nick DiNapoli, ndinapoli6@gatech.edu

May 29, 2022

1 Problem 3

In this problem, I aim to make accurate estimates of a dataset using a host of techniques. Specifically, I analyze data representing the amount of electrical energy produced by coal in the U.S. from 1950 to 2018 and implement cubic splines, B-splines, smoothing splines, and kernel regression. The data is shown in Figure 1 below.

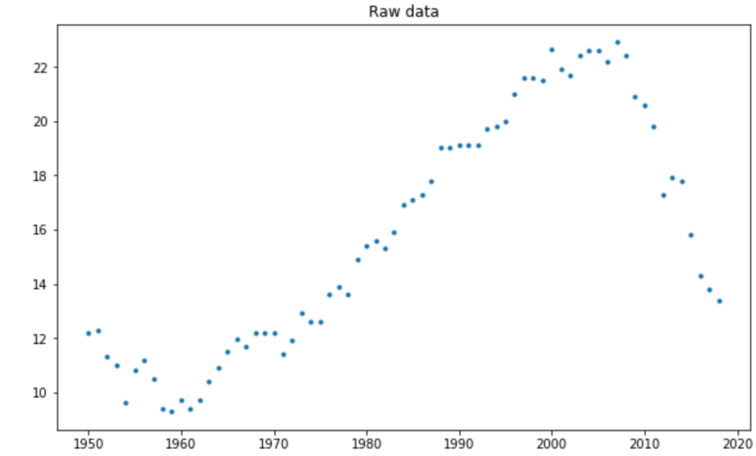

Figure 1: Electrical energy produced by coal in the U.S. from 1950 to 2018.

1.1 Cubic Splines

First, I use cubic splines to estimate the data and I vary the number of knots used from 6 to 15. For each iteration using a different number of knots, I visualize the estimation curve as well as calculate the mean squared error (MSE) for each model. Figure 2 shows the model estimates when varying the number of knots and Figure 3 show the MSE when varying the number of knots. It is clear that as the number of knots increases the MSE decreases as expected.

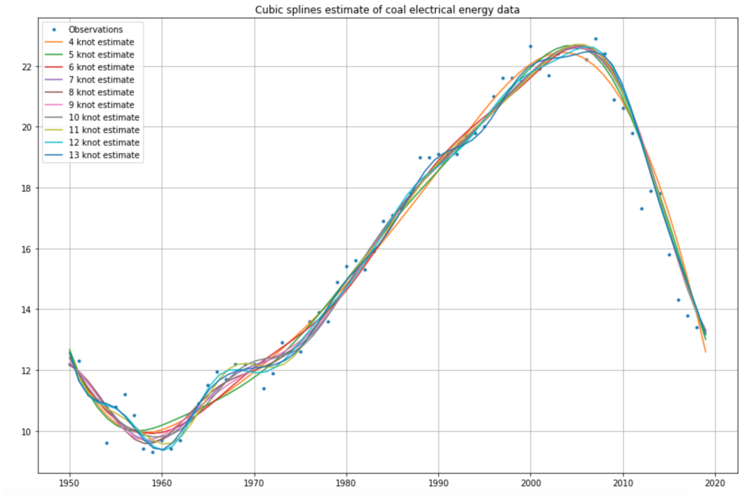

Figure 2: Cubic splines estimations as the number of knots is varied.

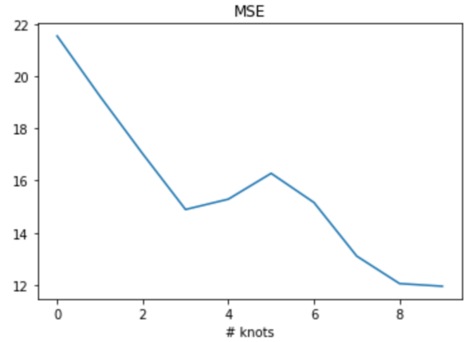

Figure 3: MSE as the number of knots is varied.

1.2 B-splines

I now complete the exact same task as the previous section but using the B-splines estimation. Figure 4 and 5 once again show the different models and the MSE as the number of knots increases. As it can be seen in in Figure 4, the models differ from one another slightly and they also differ from the cubic splines estimations.

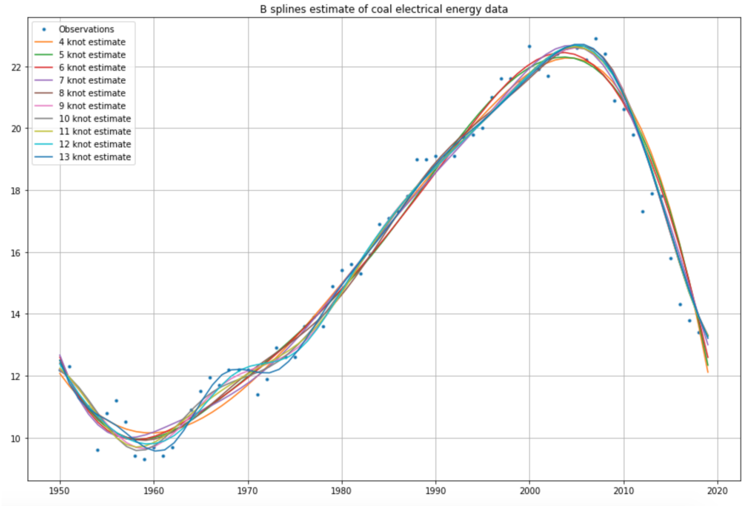

Figure 4: B-splines estimations as the number of knots is varied.

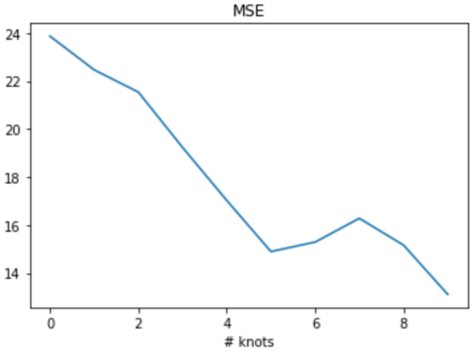

Figure 5: MSE as the number of knots is varied.

1.3 Smoothing Splines

I complete a very similar task as the previous two sections but this time using smoothing splines. For smoothing splines, I vary the smoothing parameter, λ, from 0 to .0001 and using generalized cross-validation I select the optimal λ to use for data estimation. I found the optimal λ to be 3.4e-5. Figure 6 shows the optimal smoothing splines estimation and Figure 7 shows the MSE as λ is varied.

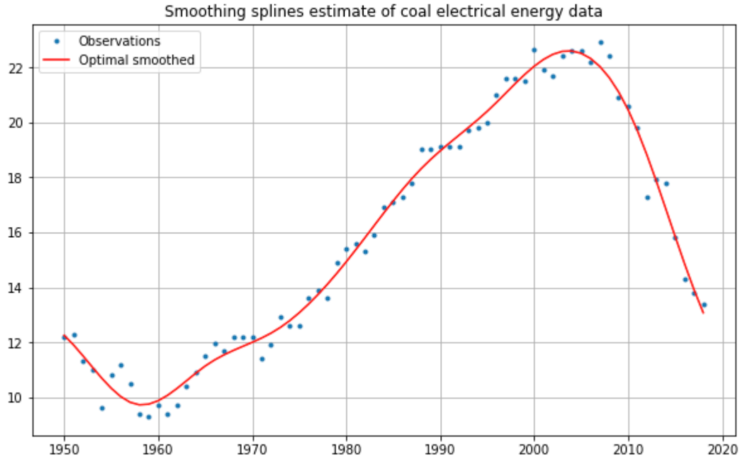

Figure 6: Smoothing splines estimation using the optimal smoothing parameter.

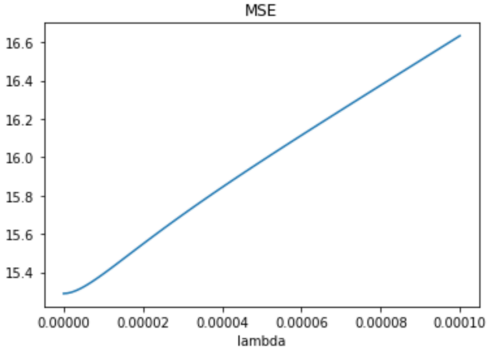

Figure 7: MSE λ is varied.

1.4 Kernel regression

Lastly, I estimate the data using kernel regression with a Gaussian kernel and once again vary λ from .001 to 1.001 and find the optimal value by computing the MSE using leave-one-out cross-validation (LOOCV) and finding the value that minimizes the MSE. I found the optimal λ to be .003. Figure 8 shows this estimation. and Figure 9 shows the MSE and λ is changed. It can be seen that the optimal λ causes the estimation to follow the data extremely closely so I decreased λ by a factor of 10 to get a better smoothes estimate.

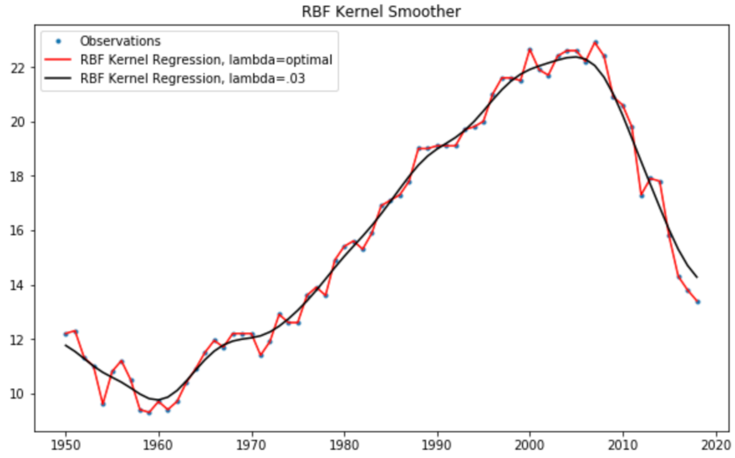

Figure 8: Kernel regression estimation using an RBF kernel and using the optimal bandwidth.

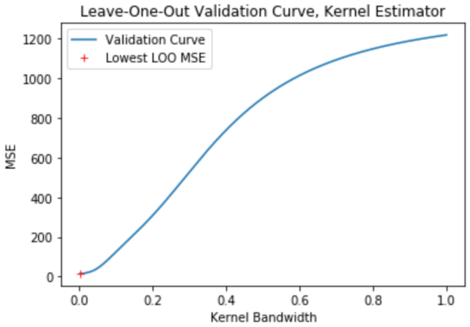

Figure 9: MSE λ is varied.

After computing the MSE using LOOCV for each method, it appears that smoothing splines produces the lowest MSE.

References

[1]