Bandwidth Selector for KDE

Laor Spitz 208649707, Tamar Yanetz 207139940

March 2022

1 Abstract

Kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable. It is a fundamental data smoothing problem where inferences about the population are made, based on a finite data sample. In this paper, we offer an automated solution that tries different methods for bandwidth selection, compares the results and returns the optimal bandwidth.

We test our solution on synthetic data created from known distributions and real world data sets. The results show that in some cases using our offered solution gives a good estimation of the data distribution.

2 Problem Description

Bandwidth selection in KDE is an important part of the data preprocessing phase in the Data Science pipeline. Choosing a correct value of bandwidth has a great effect on KDE plot smoothness. If the bandwidth is too low it will lead to under-smoothing: the plot is a combination of peaks, one peak for each sample element. If the bandwidth is too high there will be over-smoothing: the plot won't show the non-unimodal properties of the data distribution (as shown in Figure 1 and Figure 2). A poorly chosen bandwidth value may lead to undesired transformations of the density plot.

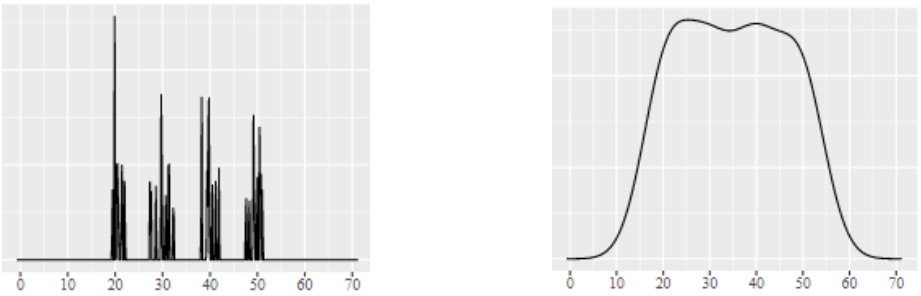

Figure 1: under smoothing vs. over smoothing

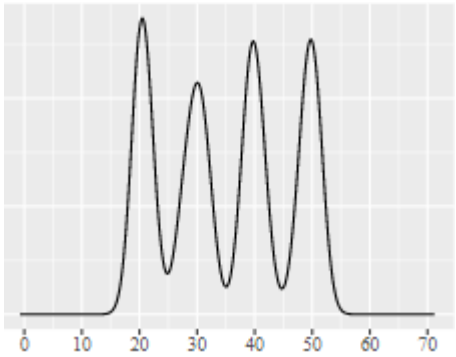

Figure 2: what we actually want

3 Solution Overview

To automate this complicated process we offer a solution that splits each data set to train and test sets. Then we run different bandwidth selection methods on the train set and compare their performance on the test set; the method with the lowest Kolmogorov-Smirnov statistic (KS) will determine the bandwidth value for the KDE. The KS statistic is defined by $D_n = sup_x|F_n(x) - F(x)|$, where F_n is the empirical distribution function and F is the cumulative distribution. The methods we use are: Silverman's rule of thumb, Scott, Sheather and Jones and Maximum likelihood cross validation (MLCV). Silverman's rule of thumb works well only on normally distributed data. Therefore, we take it into consideration if the data distribution is normal (KS test doesn't reject the null hypothesis, which is the data distribution is normal). After computing the bandwidth values on the train set we test them on the test set with KS test and choose the one with the lowest statistic that does not reject the null hypothesis, if all the bandwidth values reject the null hypothesis we return the bandwidth calculated by Improved Sheather and Jones method.

4 Experimental Evaluation

We tested our solution both on synthetic data, generated from normal and exponential distributions and on real world data-set: The Boston House-Price Data (taken from Kaggle). As mentioned above the calculated bandwidth is computed by the train set and tested on the test set, in addition to that we added graphs to visually compare the final KDE and the data distribution. We noticed that on some data the graph with the selected bandwidth is similar to the distribution, but in other cases we get either over smoothed plot or under smoothed plot.. In addition,we noticed that there are other factors that probably affected the smoothness of the graph such as the data-set size, the quality of the data etc.

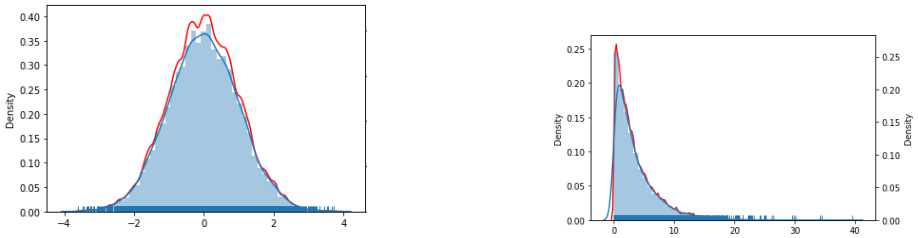

Figure 3: Results on synthetic data: red plot indicates the selected bandwidth plot

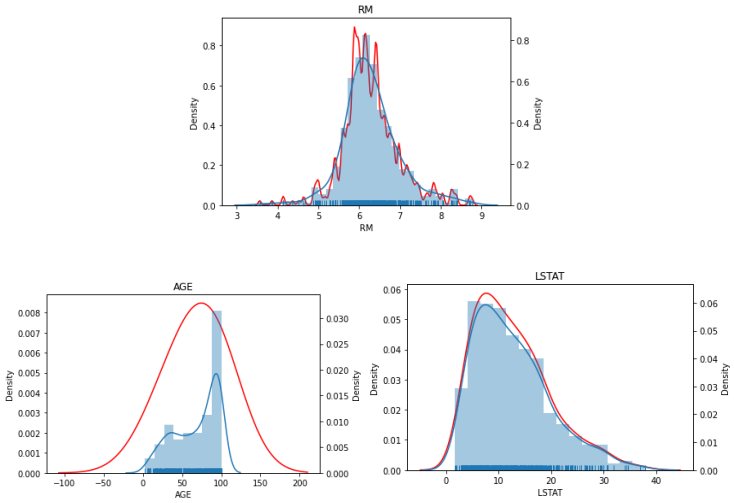

Figure 4: Results on different categories from Boston House Prices Data Set: red plot indicates the selected bandwidth plot

5 Related Work

Bandwidth selection is crucial when you are trying to visualize your distributions. The existing tools and libraries that work with various bandwidth selection methods specified by the user. We used these tools in our project. Unfortunately, there is no universal bandwidth selector that fits all the situations and that is what motivated us to study this subject. While working on the project we researched for different bandwidth selections methods and encountered famous papers such as: 'DENSITY ESTIMATION FOR STATISTICS AND DATA ANALYSIS' by B.W. Silverman, 'A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation' by S. J. Sheather and M. C. Jones and On optimal databased bandwidth selection in kernel density estimation by Hall, P., Sheater, S.J., Jones, M.C., Marron, J.S. We ended up comparing the bandwidths calculates by Silverman Rule of thumb, Scott, Sheather and Jones and Maximum likelihood cross validation.

6 Conclusion

Our results were inconclusive, in some cases the bandwidth selected was good and in other cases it led to over/under smoothing. The advantage of using this automation method is that it simplifies the data reprocessing phase and makes it possible for users who are less knowledgeable in this subject, a "black box" that gives the desired estimation saves time and energy. The system need improvements, because we encountered cases where it didn't work well. It's important to explore different aspects and factors that affect the results in order to attain a higher degree of success.