Reports-Tingyu Zhu

Research objectives

The goal of this project is to explore the Kernel Density Estimation method. The main idea of this method is to estimate an unknown density function by $\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)$. Here, $K_h(u)$ is the kernal function, and h is the bandwidth.

There're two important parts that may influence the estimation results: kernel functions $K_h()$ and the bandwidth h. Also, different sample sizes will lead to different results.

To assess the performance, 'Mean Absolute Deviation Errors' $MADE = \frac{1}{n}\sum_{k=1}^{n} |(\hat{f}(u_k) - f(u_k))|$ are computed for each estimation, where the $f(u_k)$ are the true densities.

So, my first task is to explore the influence of different choices of kernel functions and bandwidth with different sample sizes. The second task is to apply the KDE method to a real dataset.

Progress and results

Step 1 (week 5)

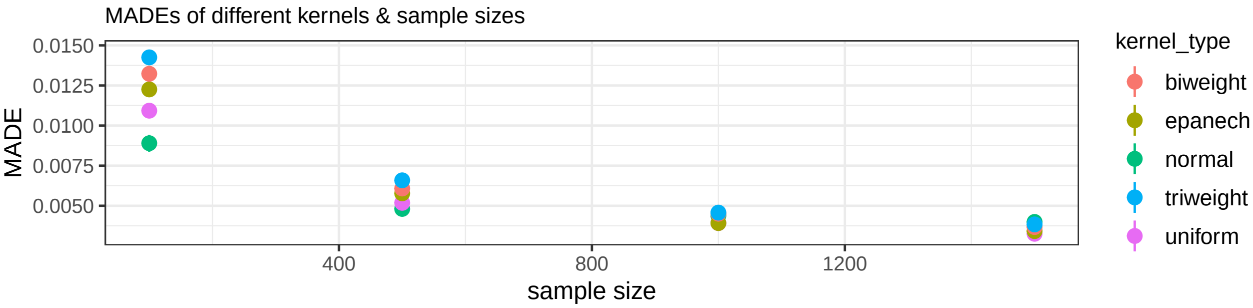
The key function `KDE_est()` was built. Using normally distributed samples, we found that our estimated densities are quite close to the true densities. Furthermore, by estimating a sample from the `exp(2)` distribution, we found that, as pointed out in the literature, the KDE method has trouble in estimating the boundary densities. (week 5) Details can be found in `simulation_week5.pdf` in the `simulation` folder.

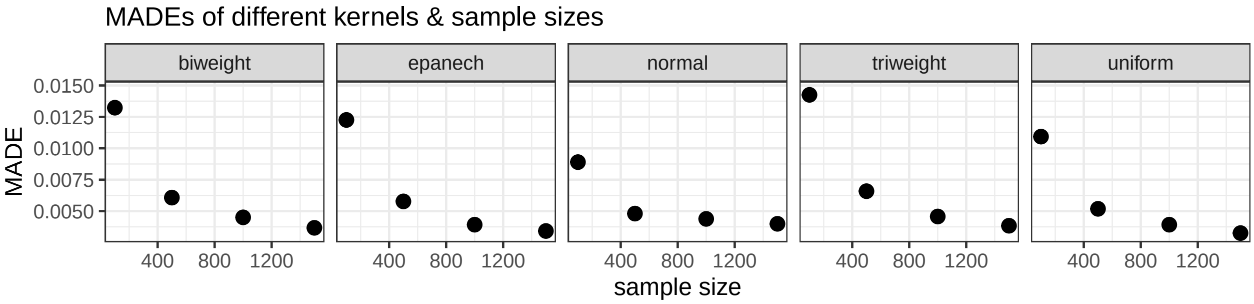
Step 2 (week 6)

Simulations are conducted under different combinations of kernel functions and sample sizes, with fixed bandwidth fixed at a given value. Under each combination of the parameter, simulations are repeated 100 times. For each simulation, the MADE is calculated. Then, compute the means and standard deviations of MADEs of all the 100 replicates.

Results show that, under the same sample size, the MADEs of different kernel functions are close to each other. And for all choice of kernels, larger sample sizes can generate more accurate results (smaller MADEs).

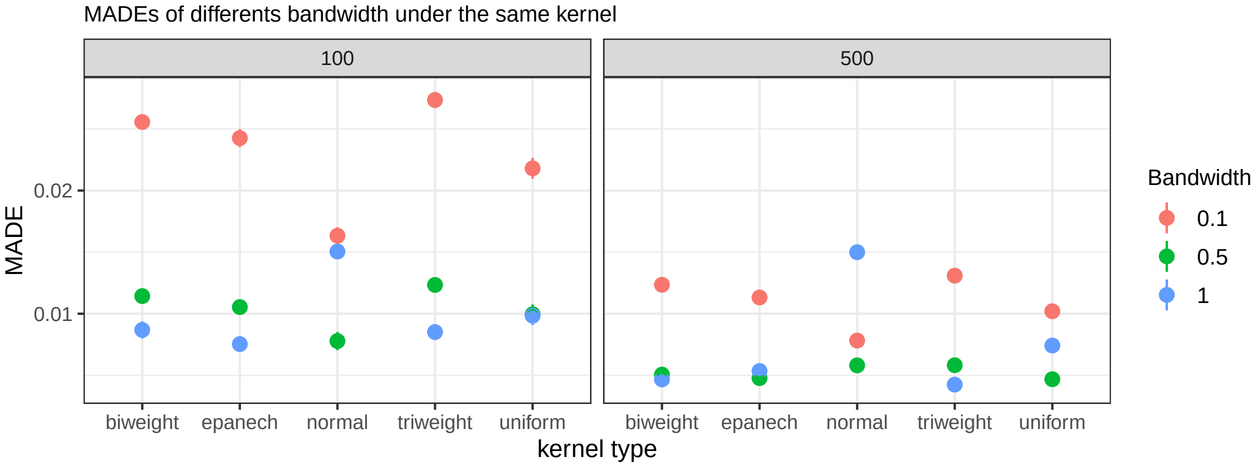
Thus, a large sample size can improve the KDE estimation and the choice of kernel functions is not very crucial. Details can be found in `simulation_week6.pdf` in the `simulation` folder and `Summary_week6` in the `analysis` folder.





Step 3 (week 7)

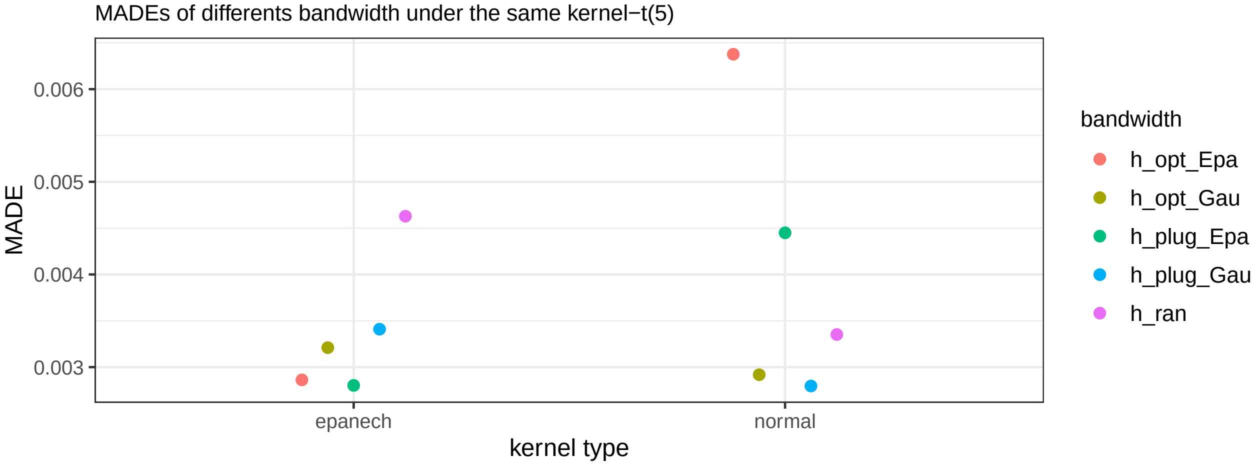
This step aims to explore the influence of the choice of bandwidth. First, simulations are conducted under different combinations of kernels, sample sizes, and bandwidths using the same way as in Step 2. Following plot shows that under the same sample size and kernel, the estimation results using different bandwidths are quite different. Thus, bandwidth is a crucial factor in the KDE method.



Then, we adopt two types of bandwidth selectors: the normal reference bandwidth selector and the plug-in bandwidth selector. Results in the literature stated that the former tends to have good performance when the distribution of data is close to normal. While the latter can be used for any data.

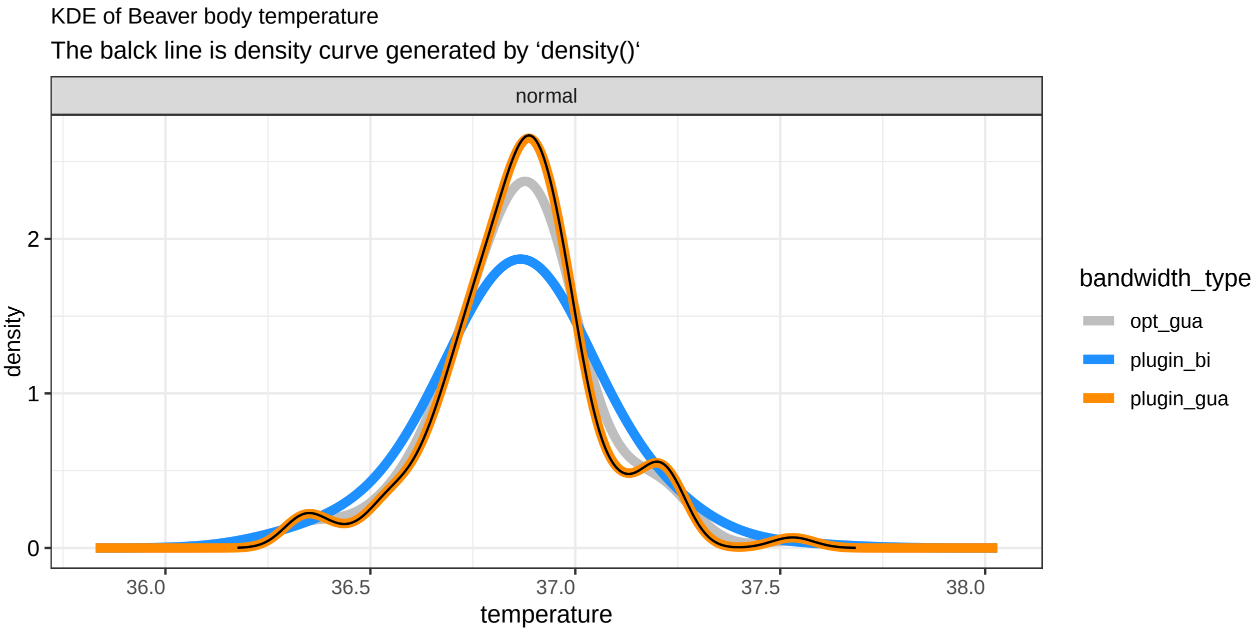
Here are the simulation results of the samples from the `t(5)` distribution under different combinations of the kernels and bandwidths with the same sample size. We can see that the plug-in bandwidths under the corresponding kernels (h_plug_Gau and h_plug_Epa) have the smallest MADEs.

However, other simulation results show that under some samples, the plug-in bandwidths may not have the smallest MADEs. The conclusion is that– if the data looks close to normal, the Normal reference bandwidth selector can be the first choice; if the data is far from normal, the plug-in bandwidth selector is a better choice. Details for this part can be found in `simulation_week7.pdf` in the `simulation` folder and `Summary_week7` in the `analysis` folder.



Step 4 (week8)

In the last step, we estimate the "beaver's Body Temperature" using the KDE method with different choices of bandwidth and kernels. Then, we compare the estimation results with those generated by the built-in `density()` function in R.



Plot shows our estimated density curve using the plug-in bandwidth is quite close to the curve generated by the `density()` function. Details can be found in `simulation_week8.pdf` in the `simulation` file and `pre_plots` in the `analysis` file.