

Cuda Project: Kernel Density Estimation

Juan Salmeron Moya
Piotr Harmuszkiewicz

The goal of estimation is to approximate the probability density function of input data. The most important part is kernel. There are many types such as: triangular, espanechnikov, uniform and the one we are using Gaussian. It is a bit modified version to speed up the calculation:

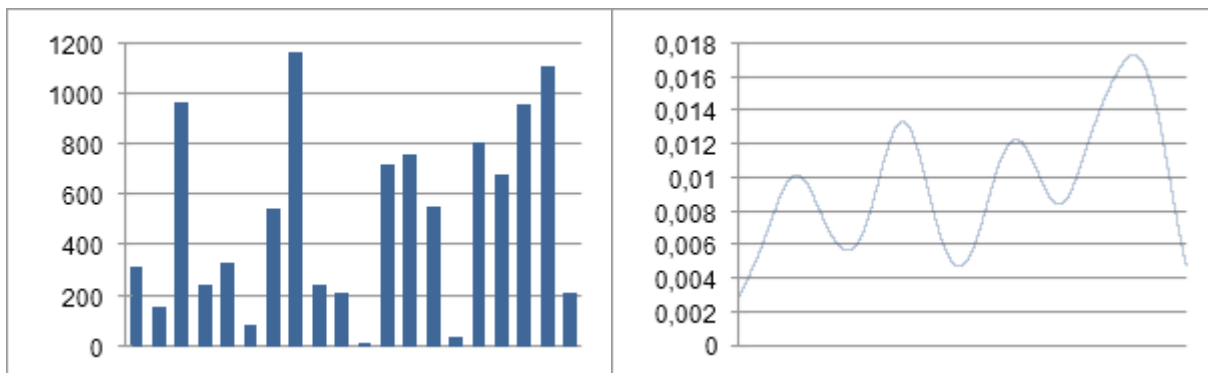
$$\frac{1}{n * h * \sqrt{2 * \pi}} \sum e^{-\frac{1}{2} \frac{(x-d_i)^2}{h^2}}$$

The first step is produce data. We want to compare our output data with histogram, so we make histogram as diverse as possible. That is why we draw the data in such a way.

The next step we calculate bandwidth. There is special formula as a rule of thumb.

Then we make estimation and as a result we get:

Data size: 10000
Estimation size: 10000



Expected value: 55.9845
Standard deviation: 28.8745
h = 4.85088
TIME: 14.6187 ms
Mean squared error: 0.000107274

The h value has big influence on how the plot looks like. The bigger h is the flatter plot will be.
To verify the results we can look at the plot or check the mean squared error.

The complexity of the kernel is dataSize * estimationSize and time changes according to this multiplication.