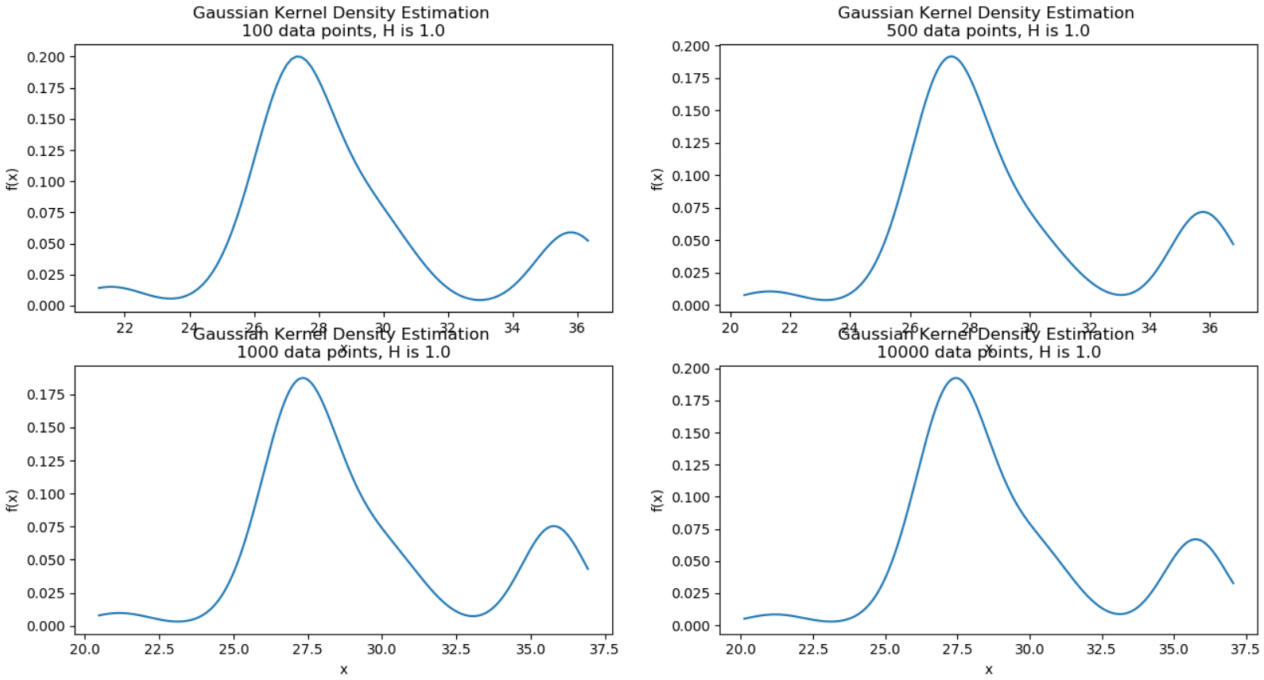
Assignment-1 Report

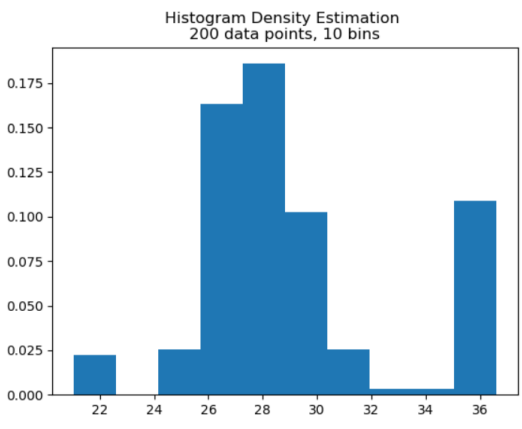
Requirement 1

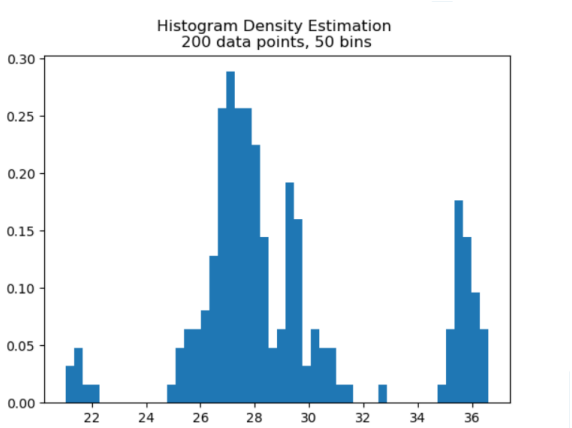
The estimation is gradually be more precise as the data grows within a certain range. (After reaching the threshold, the precision of the estimation is no longer significantly improved.)

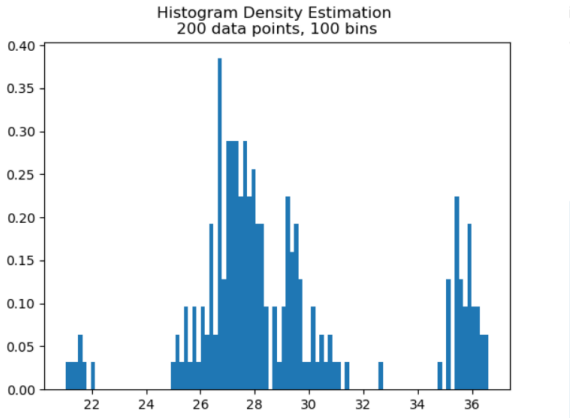
However, the improvement is not significant in some methods with good smooth parameters (as shown in the following figures) , in which the smooth parameters are of more importance.



Requirement 2







As shown in the figures above, the probability density model is very spiky, with a lot of structure that is not present in the underlying distribution, when the number of bins is very small (top figure with 10 bins). Conversely, if the number of bins is too large (bottom figure with 100 bins), then the density model is too smooth to capture the features of the underlying distribution.

The figure of the model with the best (or good) number of bins should have the following features:

- only a little protruded bars which are very thin.
- the edge is sufficiently curved.

Actually, there exists a way to choose number of bins called *Sturge's Rule*, the formula is **num_bins = 1 + 3.322*logN** where N is the number of data points. However, this method has been criticized for over-smoothing of histograms. Therefore, it should be considered a *rule-of-thumb* rather than an absolute formula with the perfect solution. A modified version of *Sturge's Rule* is *Doane's Rule*, in which the number of bins can be calculated by

$$\log_2(n) + 1 + \log_2(1 + \frac{\sqrt{b}}{\sigma\sqrt{b}})$$

Where

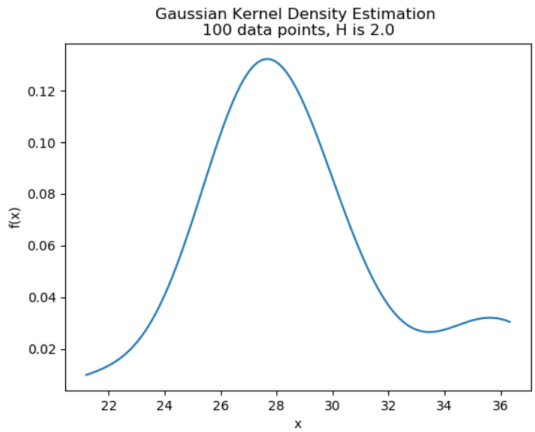
$$\sqrt{b} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{[\sum_{i=1}^{n}(X_i - \bar{X})^2]^{(3/2)}}$$

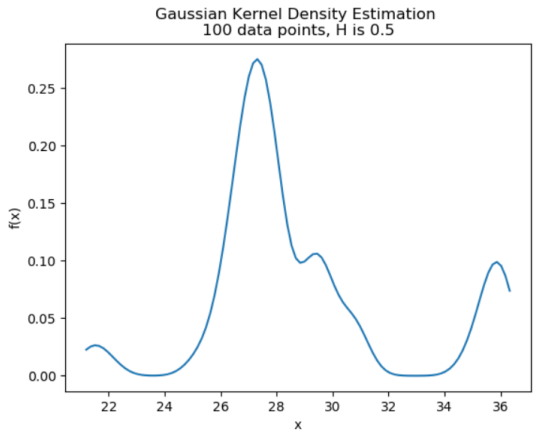
and

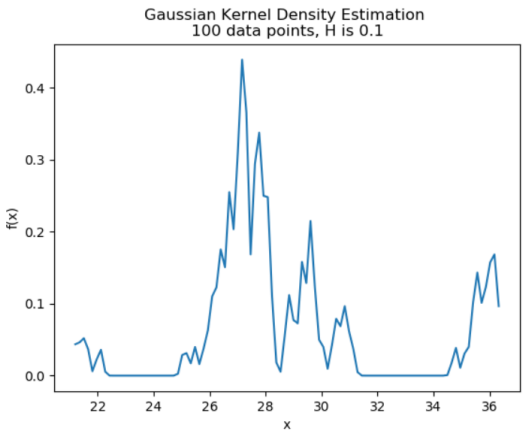
$$\sigma\sqrt{b} = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}}.$$

But it also has the problem of over-smoothing.

Requirement 3







The best density model with 100 data points is shown in the middle figure, in which the smooth parameter H is 0.5.

As shown in the figures above, there is a trade-off between sensitivity to noise at small H and over-smoothing at large H.

So as a conclusion, the figure of best (or good) density model should have the following features:

- only a little of spikes.

- enough inflection points.

These is a *rule-of-thumb* method to choose smooth parameter H (*h* in the following formula) when Gaussian density estimator is used to approximate univariate data. The formula of this method is

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-1/5},$$

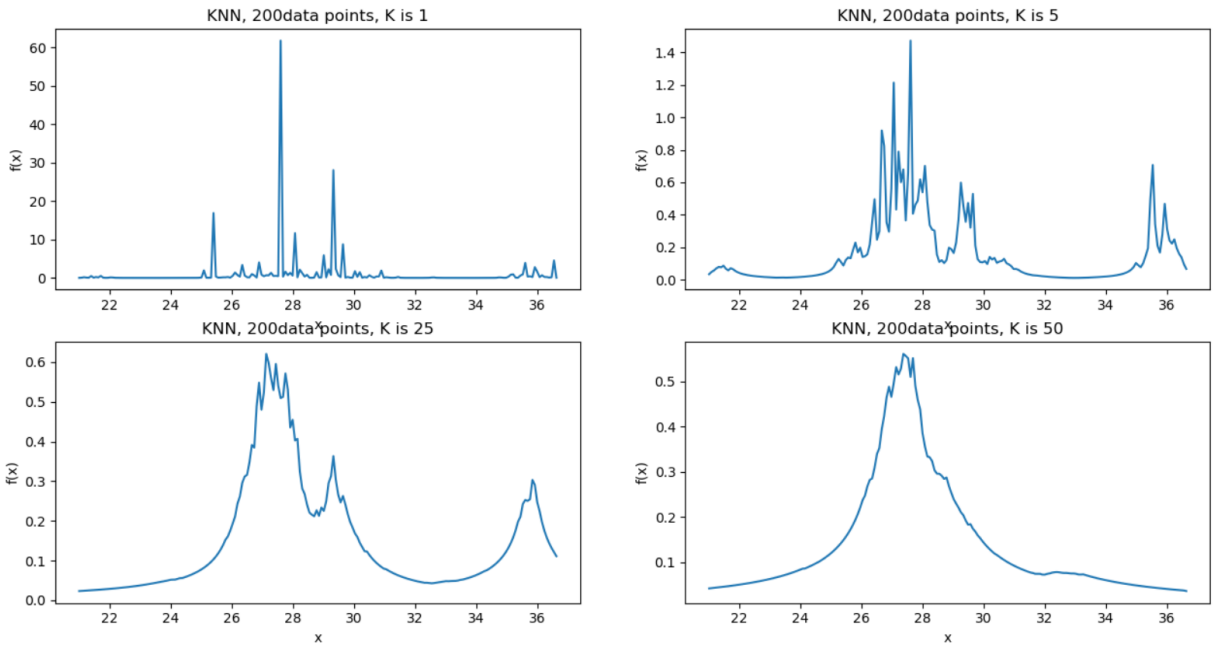
where

$$\hat{\sigma}$$

is the standard deviation of the sampled data points.

See more details in https://en.wikipedia.org/wiki/Kernel_density_estimation.

Requirement 4



As shown in the figures above, the density model is sensitive to noise when K is very small, which means the approximation error decreases but the estimation error increases, thus the model is likely to be overfitting. On the opposite, the approximation error increases but the estimation error decreases if the given K is very large, as the result, the model is likely to be underfitting.

Let's see the formula

$$p(x) = \frac{K}{NV}$$

It yields a valid distribution only when the sum of all regions' data points equals to N. In the K-Nearest Neighbour method, however, the union of all regions does not always cover the whole space, and the intersection of every two regions is not always empty. So we can draw the conclusion that the K-Nearest Neighbour method does not always yield a valid distribution.

We can also confirm this theoretically,

$$\int_{-\infty}^{+\infty} p(x)\,\mathrm{d}x = \sum \frac{K}{NV}\Delta V = \frac{K}{N}\sum \frac{1}{V}\Delta V,$$

where both K and V are constant.

The integral term

$$\int_{-\infty}^{+\infty} p(x)\,\mathrm{d}x$$

does not converge to 1 because series

$$\sum \frac{1}{V}$$

is divergent.