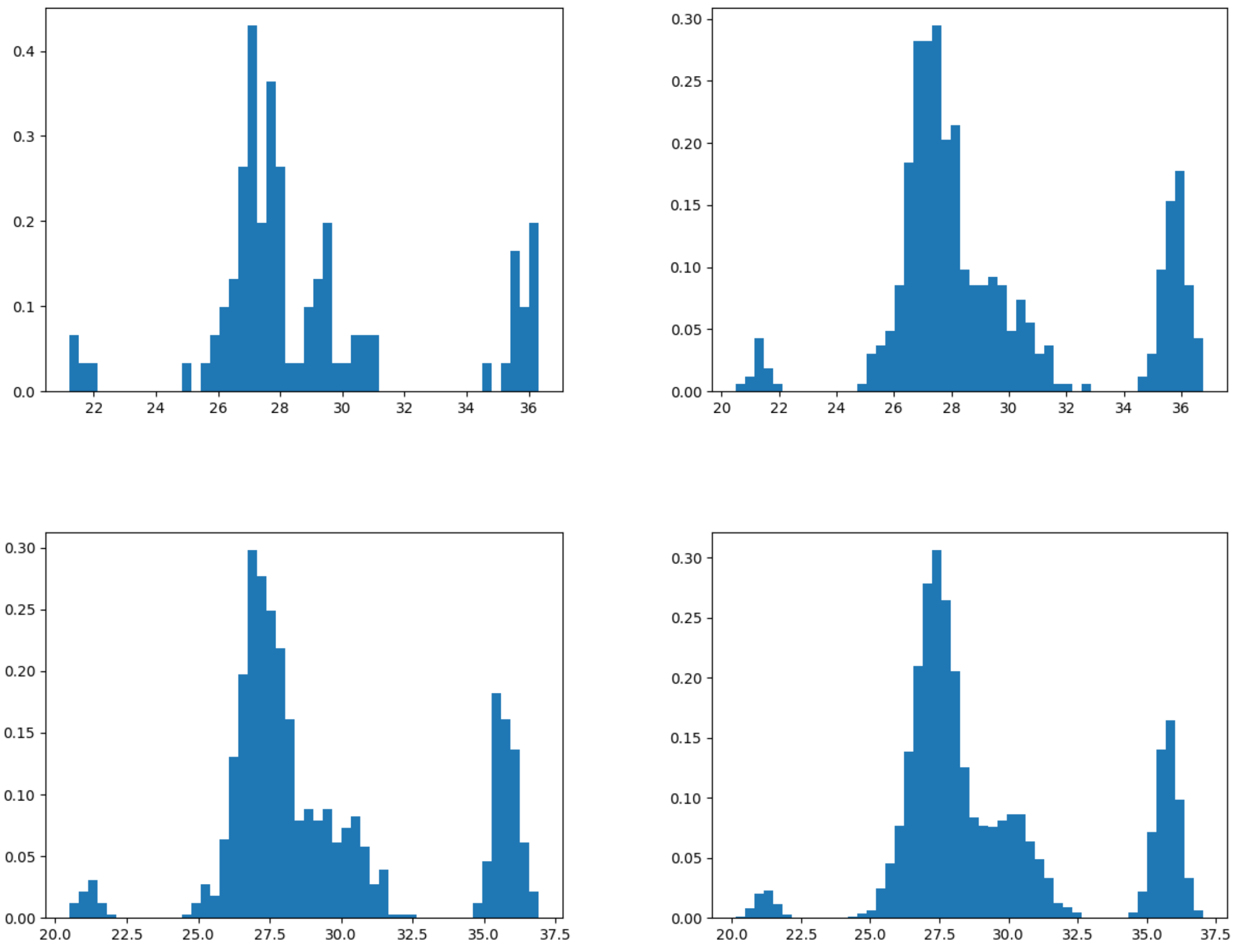
Non-Parametric Density Estimation

Anonymity
Department of Computer Science, Fudan University
March 19, 2019

1 Overview

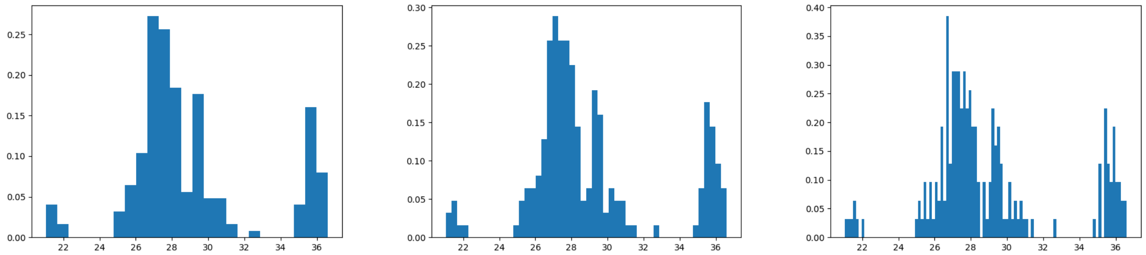
We've learned three kinds of non-parametric density estimation: **histogram, kernel and nearest neighbor** whose parameters are not determined. Intuitively, with more data distributed in the limited range to exclude contingency, we can get a smoother image and more reliable estimation, as the plots show.



2 Histogram estimation

Histogram estimation is a simple way to estimate the probability distribution of a continuous variable, which divide the entire range of values into a series of intervals and count how many values fall into each interval. Generally, the bins (intervals) are of equal size, with a rectangle erected over the bin, and the heights proportional to the frequency are always discrete. The plots below show when the bin is very small, the resulting density is very spiky, while it's too smooth to capture the valid estimation when the bin is large. So, to get better model, we should figure out the best choice for the number of bins.

There are many strategies for decision on the bins, such as Rule of Thumb, and Knuth's Rule[3]. But in a more practical way, we can also use Cross Validation just as what we do in Kernel Density Estimation to decide on h.



3 Kernel Density Estimation

3.1 Mean Squared Error

To determine the probability density on point **x**, we define a kernel function called Parzen window to count the number of points falling within a region of fixed volume centered on the point **x**, which infers the probability of x according to the continuity, like

$$p(x)=\frac{1}{n}\sum_{i=1}^{n}K_h(x-x_i)=\frac{1}{nh}\sum_{i=1}^{n}K(\frac{x-x_i}{h}) \tag{1}$$

And to get a smoother density model, we choose the Gaussian kernel function

$$p(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{(2\pi h^2)^{\frac{1}{2}}}exp\{-\frac{|x-x_i|^2}{2h^2}\} \tag{2}$$

To, optimize the estimation, we can introduce the mean squared error **MSE(x)** since the (absolute value of the) bias increases and the variance decreases as h increase.[1]

$$MSE(x)=E[(\widehat{f}(x)-f(x))^2]=(E[\widehat{f}(x)]-f(x))^2+Var(\widehat{f}(x)) \tag{3}$$

We can induce the bias and the variance

$$E[\widehat{f}(x)]=\int\frac{1}{h}K(\frac{x-y}{h})f(y)dy \tag{4}$$

$$\begin{aligned}&Var(\widehat{f}(x))=\frac{1}{nh^2}Var(K(\frac{x-X_i}{h}))\\&=\frac{1}{nh^2}E[K(\frac{x-X_i}{h})^2]-\frac{1}{nh^2}E[K(\frac{x-X_i}{h})]^2\\&=n^{-1}\int\frac{1}{h^2}K(\frac{x-y}{h})^2f(y)dy-n^{-1}(\int\frac{1}{h}K(\frac{x-y}{h})f(y)dy)^2\end{aligned} \tag{5}$$

$$\begin{aligned}&Bias(x)=\int\frac{1}{h}K(\frac{x-y}{h})f(y)dy-f(x)\\&=\int K(z)f(x-hz)dz-f(x)=\int K(z)(f(x-hz)-f(x))dz\end{aligned} \tag{6}$$

where

$$z=\frac{x-y}{h} \tag{7}$$

with Taylor expansion of f

$$f(x-hz)=f(x)-hz\dot{f}(x)+\frac{1}{2}h^2z^2\ddot{f}(x)+res \tag{8}$$

we can get

$$\begin{aligned}Bias(x)&=-h\dot{f}(x)\int zK(z)dz+\frac{1}{2}h^2\ddot{f}(x)\int z^2K(z)dz+res\\&=\frac{1}{2}h^2\ddot{f}(x)\int z^2K(z)dz+res\end{aligned} \tag{9}$$

$$\begin{aligned}&Var(\widehat{f}(x))=n^{-1}\int\frac{1}{h^2}K(\frac{x-y}{h})^2f(y)dy-n^{-1}(f(x)+Bias(x))^2\\&=n^{-1}h^{-1}\int f(x-hz)K(z)^2dz-(n^{-1}(f(x)+Bias(x))^2)_{=O(n^{-1})}\\&=n^{-1}h^{-1}\int f(x-hz)K(z)^2dz+O(n^{-1})=n^{-1}h^{-1}f(x)\int K(z)^2dz+o(n^{-1}h^{-1})\end{aligned} \tag{10}$$

to calculate the MSE(x), and $\frac{\partial}{\partial h}MSE(x) = 0$, we get

$$h_{opt} = n^{-1/5}(\frac{f(x)\int K^2(z)dz}{(\ddot{f}(x))^2(\int z^2K(z)dz)^2})^{1/5} \tag{11}$$

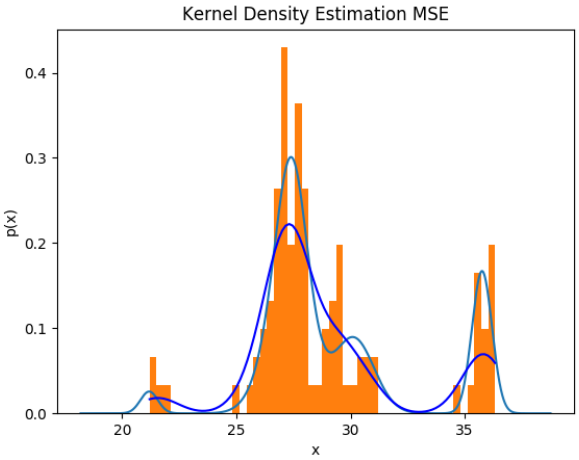
and since h(x) is dependent of x, so we can consider minimizing the MISE, i.e, $\int MSE(x)dx$, and the optimal bandwidth becomes

$$h_{opt} = n^{-1/5}(R(K)/\delta_K^4 * \frac{1}{R(\ddot{f})})^{1/5} \tag{12}$$

where $R(g) = \int g^2(x)dx$ and $\delta_k^2 = \int x^2K(x)dx$

Since R($\ddot{f}$) is unknown in this expression, the Rule of thumb replaces the unknown density function f in this functional by a reference distribution function, and finally estimate $\widehat{h}_{opt} = 1.06 min(\hat{\delta}, \frac{\widehat{R}}{1.34})n^{-1/5}$, where $\hat{\delta}$ is the standard deviation, and $\widehat{R}$ is the interquartile range.

As we can see, since the $\hat{\delta}$ is based on the sample, so the plot of estimation(blue) is kinda different from what the distribution(green) really is.



Inspired by Zhang Zuo Bai, we choose KL divergence to further estimate the KDE with h.

3.2 Cross Validation

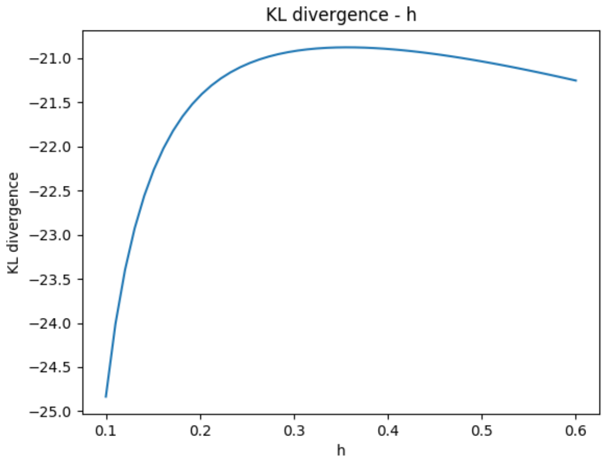
Consider the real distribution p(x), and we will modele this through an approximating distribution $p_h(x)$ using Gaussian kernel density estimation with bandwidth h, we have **relative entropy**,

$$\begin{aligned} KL(p_h||p) &= -\int p(x)\ln\frac{p_h(x)}{p(x)}dx \\ &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{x}^{Vi}p(x)\ln\frac{p_h(x)}{p(x)} \\ &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{x}^{Vi}p(x)(\ln p_h(x) - \ln p(x)) \end{aligned} \tag{13}$$

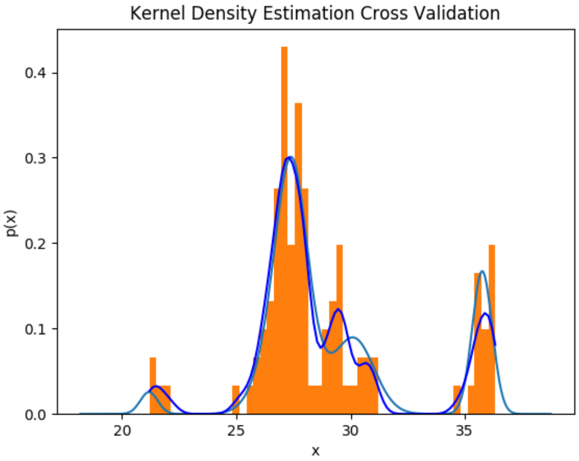
Since the probability of the sampled data is intuitively higher than that of the rest. So using Cross Validation, here we choose 10-Folds, we should minimize the relative entropy, i.e. the additional information, on the validation set(**Vi**). Hence We should optimize KL divergence by changing the bandwidth h. Here we know p(x) is the real distribution of sampled data, which is independent of h, in this case we can simplify the formula above,

$$\begin{gathered} \arg\min_h -\frac{1}{N}\sum_{i=1}^{N}\sum_{x}^{V_i}\ln p_h(x) \\ \Rightarrow \arg\max_h \sum_{x}\ln p_h(x) \end{gathered} \tag{14}$$

According to the plot, the KL divergence have the optimal solution on the validation set with h = 0.3551.



With the best bandwidth, we model(blue) the real distribution again and get a better result than the empirical formula with MSE.

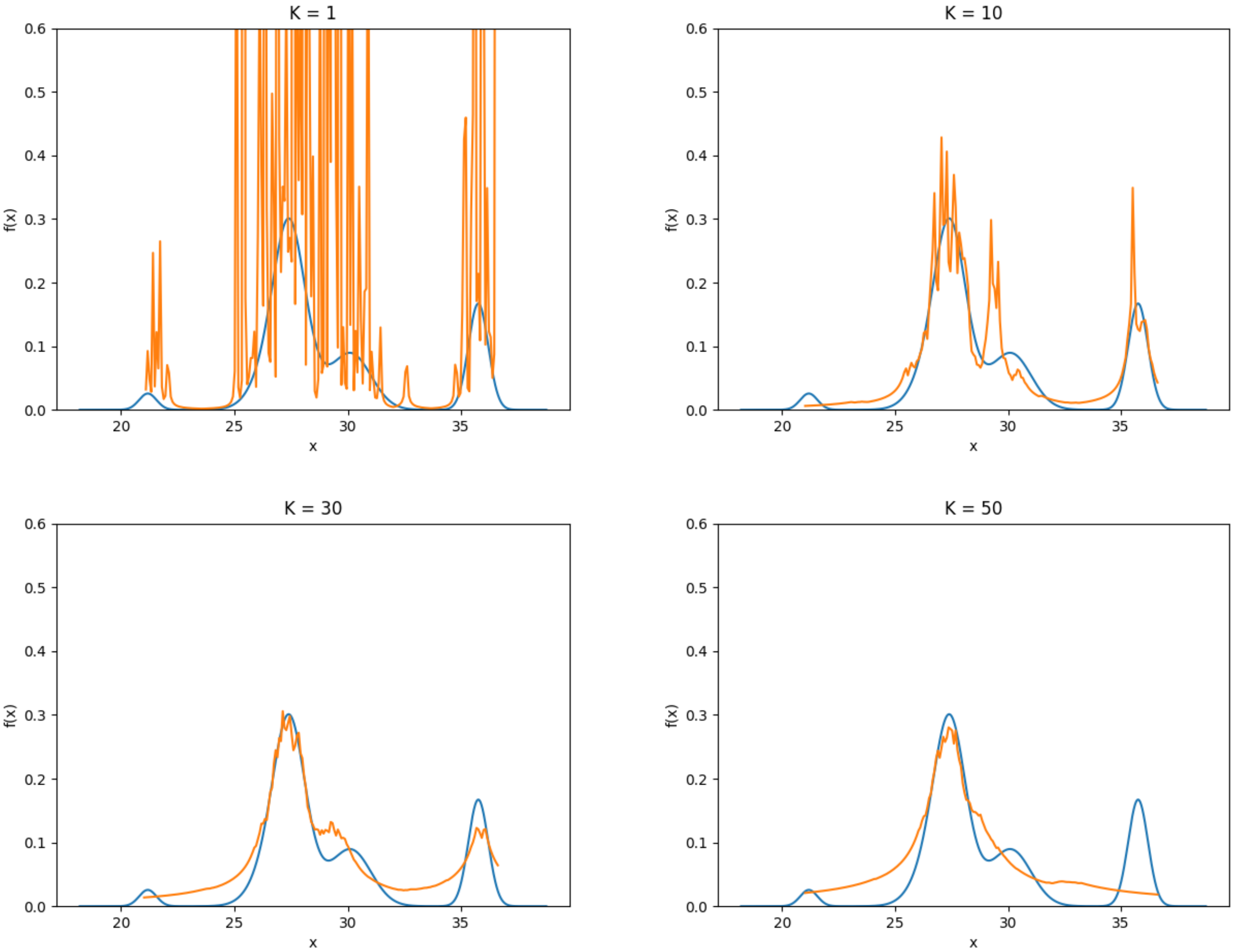


4 Nearest Neighbor Estimation

To address the problem that the optimal choice **h** of Parzen window may be dependent on location within the data space, we introduce the Nearest Neighbor Estimation. The KNN fix the size of data that we search around the **x**, to find the appropriate volume that need to cover all the K nearest data, like

$$p(x) = \frac{K}{NV} \tag{15}$$

We see that the parameter K governs the degree of smoothing, so that a small value of K(like K = 1) leads to a very noisy density model, whereas a large value(K = 50) smooths out the the bimodal nature of the true distribution. And when K = 30, we see that our model fit the the real distribution well.



Here we give a brief proof for that the 1NN does not always yield a valid distribution, consider that

$$\int_{-\infty}^{+\infty} p(x)dx = \int_{-\infty}^{x_{min}} p(x)dx + \int_{x_{min}}^{x_{max}} p(x)dx + \int_{x_{max}}^{+\infty} p(x)dx \geq \int_{-\infty}^{x_{min}} p(x)dx + \int_{x_{max}}^{+\infty} p(x)dx$$

$$= \frac{1}{2N}(\int_{-\infty}^{x_{min}} \frac{1}{x_{min}-x}dx + \int_{x_{max}}^{+\infty} \frac{1}{x-x_{max}}dx) = \frac{1}{2N}(-\ln(x_{min}-x)|_{-\infty}^{x_{min}} + \ln(x-x_{max})|_{x_{max}}^{+\infty}$$

x_{min}, x_{max} represents the boundary of x on the data space, and the right side of the equation above will diverge to positive infinity. Therefore, we draw a conclusion that $\int_{-\infty}^{+\infty} p(x)dx$ will never converge to 1.

5 References

[1] Park, B.U. and Marron, J.S., Comparison of Data-Driven Bandwidth Selectors, Journal of the American Statistical Association, 85 (1990) 66-72.

[2] Christopher M Bishop. Pattern recognition and machine learning. springer, 2006.

[3] Kevin H. Knuth. Optimal Data-Based Binning for Histograms, arXiv:physics/0605197.

[4] Thanks to Zhang Zuo Bai's help.