Estimation of a probability density function

Laya Parkavousi *
laya_parkavosiyahoo.com
Student ID:99204032

Abstract

The problem we are trying to face with is the estimation of an unknown probability density p(x) for a sample of random numbers. In this homework, I am trying to estimate the probability distribution function in two ways. The first way is Binning for histogram and the second one is using Kernel Density Estimator. Furthermore, I am going to calculate the moments of a probability distribution and I will demonstrate that the higher the moment, the harder it is to estimate, in the sense that larger samples are required in order to obtain estimates of similar quality. This is due to the excess degrees of freedom consumed by the higher orders. Further, they can be subtle to interpret, often being most easily understood in terms of lower order moments.[1]

1 INTRODUCTION

Let $x_1, x_2, ..., x_N$ be a random sample from a normal distribution P with density $p(x)$. Here we consider $N \sim 10^6$. We want to recover the underlying probability density function generating our dataset.

A random variable is a variable whose value is determined by the outcome of a random experiment. A random variable is said discrete if its values are countable, or continuous otherwise. Even though, a computer can not generate a truly random number, we can change the seed values to make the pseudo-randoms less deterministic therefore the random sample would be accurate enough for our purpose.

In probability theory, a probability density function (PDF), or density of a continuous random variable, is a function whose value at any given sample (or point) in the sample space (the set of possible values taken by the random variable) can be interpreted as providing a relative likelihood that the value of the random variable would equal that sample.[2]

$$P(a < x < b) = \int_a^b p_X(x)dx \tag{1}$$

In other words, $p_X(x) = \frac{dP_X(x)}{dx}$.

When the probability law of x is unknown in explicit form but realizations of x can be drawn, a kernel estimation or binning may be used to reconstruct the probability density function of x. These are non-parametric methods that uses a finite data sample from x.[3]

Here we'll be concerned with estimators for the density function itself, hence $\hat{p}(x)$ is a random variable giving our estimate for density $p(x)$.[5]

2 Various Techniques

Some of the various techniques to recover the probability are discussed below.

2.1 Binning

The simplest way to estimate the density are histograms. A histogram is an approximate representation of the distribution of numerical data.[1] To construct the histogram, we should divide the range of values x into bins, of size h. Then count how many values fall into each interval. The bins are usually specified

*[1]Department of Physics, Sharif University of Technology, Tehran, Iran.

as consecutive, non-overlapping intervals of a variable. The bins (intervals) must be adjacent and are often (but not required to be) of equal size.[4]

Histograms give a rough sense of the density of the underlying distribution of the data, and often for density estimation: estimating the probability density function of the underlying variable. The total area of a histogram used for probability density is always normalized to 1. [1]

The number of bins, k can be assigned directly or can be calculated from a suggested bin width w as:[1]

$$k = \lceil \frac{\max x - \min x}{h} \rceil \tag{2}$$

Using wider bins where the density of the underlying data points is low reduces noise due to sampling randomness; using narrower bins where the density is high (so the signal drowns the noise) gives greater precision to the density estimation. Some theoreticians have attempted to determine an optimal number of bins. Scott [6, 7] derived formulas for the optimal bin width by minimizing the integrated mean squared error of the histogram model h(x) of the true underlying density $p(x)$,

$$L(h(x), p(x)) = \int dx (h(x) - f(x))^2. \tag{3}$$

Given an equally spaced mesh $\{t_k\}$ over the entire real line with $t_{k+1} - t_k = h$, where N denotes the sample size. For a fixed point x, the mean squared error of a histogram estimate, $\hat{f}(x)$, of the true density value, $f(x)$, is defined by

$$MSE(x) = \mathbf{E}\{\hat{f}(x) - f(x)\}^2 \tag{4}$$

The integrated mean squared error represents a global error measure of a histogram estimate and is defined by

$$IMSE = \int \mathbf{E}\{\hat{f}(x) - f(x)\}^2 dx. \tag{5}$$

2.1.1 Pointwise Mean Squared Error

Thus the pointwise variance of the histogram is $Np_k(1 - p_k)/(Nh)^2$, which is constant for all x in the k-th bin.[8]

$$\hat{f}(x) = \frac{\nu_k}{Nh} \tag{6}$$

We can calculate the integrated variance (IV) by integrating the pointwise variance over the k-th bin and summing over all bins,

$$IV = \sum_{k=-\infty}^{\infty} \frac{Np_k(1 - p_k)}{(Nh)^2} \times h \tag{7}$$

$$= \sum_{k=-\infty}^{\infty} \frac{p_k(1 - p_k)}{Nh} = \frac{1}{Nh} - \sum_k \frac{p_k^2}{Nh} \tag{8}$$

Next consider the bias of $\hat{p}$ at a fixed point, x, which is located in the k-th bin. Note that $\mathbf{E}\hat{p}(x) = Np_k/Nh = p_k/h$. A useful approximation to the bin probability is

$$p_k = \int_{t_k}^{t_{k+1}} p(y)dy \tag{9}$$

$$= hp(x) + h^2(\frac{1}{2} - \frac{x - t_k}{h})p'(x) + ..., \tag{10}$$

replacing the integrand $p(y)$ by $p(x) + (y - x)p'(x) + ...$ Thus the pointwise bias may be approximated by

$$\text{Bias} \quad \hat{p}(x) = \mathbf{E}\hat{p}(x) - p(x) = \frac{p_k}{h} - p(x) \tag{11}$$

$$= h(\frac{1}{2} - \frac{x - t_k}{h})p'(x) + ..., \tag{12}$$

Definition: A function is said to be Lipschitz continuous over an interval B_k if there exists a positive constant γ_k such that $|f(x) - f(y)| < \gamma_k |x - y|$ for all $x, y \in B_k$.

Then by the mean value theorem (MVT),

$$p_k = \int_{B_k} f(t)dt = hf(\xi_k) \quad \text{for} \quad \xi_k \in B_k. \tag{13}$$

It follows that

$$Var\hat{f}(x) \leqslant \frac{p_k}{Nh^2} \frac{f(\xi_k)}{Nh} \tag{14}$$

and

$$|Biasf(x)| = |f(\xi_k) - f(x)| \leqslant \gamma_k |\xi_k - x| \leqslant \gamma_k h; \tag{15}$$

Therefore,

$$MSE\hat{f}(x) \leqslant \frac{f(\xi_k)}{Nh} + \gamma_k^2 h^2 \tag{16}$$

The MSE(x) bound (16) is minimized when

$$h^*(x) = [\frac{f(\xi_k)}{2\gamma_k^2 N}]^{1/3}; \tag{17}$$

2.1.2 Global L_2 Histogram Error

Suppose that $x_1, ..., x_n$ is a random sample from a continuous probability density function $f(x)$ with two continuous and bounded derivatives. We shall need to identify the bin interval that contains a fixed point x as N varies. Let $I_N(x)$ be that interval and let $t_N(x)$ denote the left-hand endpoint of $I_N(x)$. Define the bin probability

$$p_N(x) = \int_{t_N(x)}^{t_N(x)+h(x)} f(y)dy \tag{18}$$

replacing the integrand $f(y)$ by $f(x) + (y - x)f'(x) + ...$ Thus the pointwise bias may be approximated by

$$\begin{aligned} p_N(x) &= \int_{t_N}^{t_N+h(x)} f(y)dy \\ &= h_N f(x) + h_N^2 (\frac{1}{2} - \frac{x - t_N(x)}{h}) f'(x) + ... \end{aligned} \tag{19}$$

Let ν_N be the number of data points falling in the $N - th$ bin. ν is a Binomial random variable with mean f_N and also, $\sum_N \nu_N = N$; hence,

$$\begin{cases} \mathbf{E}\nu_N = Nf_N \\ \text{Var}\nu_N = Nf_N(1 - f_N). \end{cases}$$

The histogram estimate is given by the random variable

$$\hat{f}(x) = \nu_N(x)/(Nh_N) \tag{20}$$

with expectation,

$$\begin{aligned} \mathbf{E}\{\hat{f}(x)\} &= p_N(x)/h_N(x) \\ &= f(x) + \frac{1}{2} h_N f'(x) - f'(x)\{x - t_N\} + ... \end{aligned} \tag{21}$$

Therefore the bias is

$$\frac{1}{2} h_N f'(x) - f'(x)\{x - t_N\} + ... \tag{22}$$

Now the variance of the histogram estimate at x is given by

$$\begin{aligned} var\{\hat{f}(x)\} &= p_N(x)\{1 - p_N(x)\}/(Nh_N^2) \\ &= \{h_N f(x) + O(h_N^2)\}\{1 - O(h_N)\}/(Nh_N^2) \\ &= f(x)/(Nh_N) + O(1/N) \end{aligned} \tag{23}$$

Combining, we have that

$$\begin{aligned} MSE(x) = &\frac{f(x)}{(Nh_N)} + \frac{1}{4}h_N^2 f'(x)^2 \\ &+ f'(x)^2\{x - t_N\}^2 - h_N f'(x)^2\{x - t_N\} + O(1/N + h_N^3) \end{aligned} \tag{24}$$

Integration of equation (1) over the real line implies that

$$\begin{aligned} IMSE = &1/(Nh_N) + \frac{1}{4}h_N^2 \int f'^2(x)dx + \int f'(x)^2\{x - t_N(x)\}^2 dx \\ &- h_N \int f'(x)^2\{x - t_N(x)\}dx + O(1/N + h_N^3) \end{aligned} \tag{25}$$

Then the third term in equation (25) may be written as

$$\begin{aligned} &\sum_{i=-\infty}^{\infty} \int_{t_N i}^{t_n i + h_N(x)} f'(x)^2\{x - t_N(x)\}^2 dx \\ &= \sum_{i=\infty}^{\infty} \int_0^{h_N} f'(t_{ni} + y)^2 y^2 dy \end{aligned} \tag{26}$$

Now $f'(t_{ni} + y) = f'(t_{ni}) + O(h_N)$, so that (26) becomes

$$\frac{1}{3}h_N^2 \int_{-\infty}^{\infty} f'(x)^2 dx + O(h_N^3) \tag{27}$$

A similar analysis for the fourth term in (25) yields

$$-\frac{1}{2}h_N^2 \int_{-\infty}^{\infty} f'(x)^2 dx + O(h_N^3) \tag{28}$$

Therefore

$$\begin{aligned} IMSE = &\frac{1}{Nh_N} + \frac{1}{12}h_N^2 \int_{-\infty}^{\infty} f'(x)^2 dx \\ &+ O(1/N + h_N^3) \end{aligned} \tag{29}$$

Minimizing the first two terms in (29), we obtain

$$h_N^* = \{\frac{6}{\int_{-\infty}^{\infty} f'(x)^2 dx}\}^{1/3} N^{-1/3} \tag{30}$$

All of the above calculation was re-obtain from the (Scott, 1979)[6].

With complete analysis of the bias [6, 8], For N data points, the optimal bin width h goes as $\alpha N^{-1/3}$, where α is a constant that depends on the form of the underlying distribution.

2.2 Kernel Density Estimator(KDE)

In this section I will talk about another approach density estimation, the kernel density estimator (KDE; also called kernel density estimation).

In statistics, kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable.

A one-dimensional smoothing kernel is any smooth function K such that, $\int K(x)dx = 1$, $\int xK(x)dx = 0$ and $\sigma_k^2 = \int x^2 K(x)dx > 0$.[5]

Some commonly used kernels are the following:

$$\textbf{Boxcar:} \quad K(x) = \frac{1}{2}I(x)$$
$$\textbf{Epanechnikov:} \quad K(x) = \frac{3}{4}(1 - x2)I(x)$$
$$\textbf{Gaussian:} \quad K(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$$
$$\textbf{Tricube:} \quad K(x) = \frac{70}{81}(1 - |x|^3)^3 I(x)$$

where $I(x) = 1 \quad$ if $|x| \leqslant 1$ and $I(x) = 0$ otherwise.

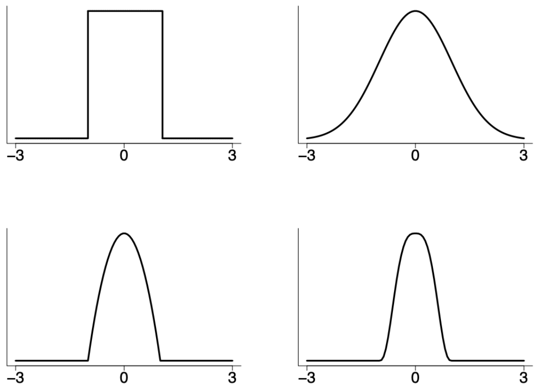

Figure 1: Examples of smoothing kernels: boxcar (top left), Gaussian (top right), Epanech- nikov (bottom left), and tricube (bottom right).[5]

Suppose that $x \in \mathbf{R}^d$. Given a kernel K and a positive number h, called the bandwidth, the kernel density estimator is defined to be

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h^d}K(\frac{\parallel x - x_i \parallel}{h}) \tag{31}$$

More generally, we define

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N} K_H(x - x_i) \tag{32}$$

where H is the positive definite bandwidth matrix and $K_H(x) = |H|^{-1/2}K(H^{-1/2}x)$.

2.2.1 Optimal Bandwidth Selection

Define, $U_1 = \int f^2(x)dx$ and $U_2 = \int xf^2(x)dx$. Also we can write the U_1 and U_2 in a more general density functionals $U = \int \gamma(x)f^2(x)$. We can obtain the kernel density functionals estimate of U by $\int \gamma(x)\hat{f}_n(x)dF_n(x)$, where $\hat{f}_n(x)$, where $\hat{f}_n(x)$ is the kernel density estimate of f and $F_n(x)$ is the empirical CDF. Thus, a kernel density functionals estimate of U is given by

$$\hat{U} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\gamma(x_i) + \gamma(x_j)}{2}K_h(x_i - x_j) \tag{33}$$

where $K_h(.) = (1/h)K(./h)$ and $K(.)$ is the kernel function.

For $\hat{U}$ in 33, the expected value and variance of $\hat{U}$ are. given by

$$\mathbf{E}(\hat{U}) = U + \frac{1}{2}\mathbf{E}[\gamma(y)f''(y)]\mu_2(K)h^2 + O(h^2), \tag{34}$$

$$\begin{aligned} Var(\hat{U}) = 4n^{-1}\{ \int \gamma^2(y)f^3(y)dy - \big(\int \gamma(y)f^2(y)dy)\} \\ 2n^{-2}h^{-1}R(K)\int \gamma^2(y)f^2(y)dy + \{O(n^{-1}) + O(n^{-2}h^{-1})\}, \end{aligned} \tag{35}$$

where $R(g) = \int g^2(y)dy$ and $\mu_2(g) = \int y^2 g(y)dy$. the proof for the theory above can be found on [9, 10]. Then the MSE of $\hat{U}$ can be written as follows:

$$\begin{aligned} MSU(\hat{U}) &= Var(\hat{U}) + Bias(\hat{U})^2 \\ &= 4n^{-1}\{\int \gamma^2(y)f^3(y)dy - (\int \gamma(y)f^2(y)dy)\} \\ & 2n^{-2}h^{-1}R(K)\int \gamma^2(y)f^2(y)dy \\ & \frac{1}{4}\mathbf{E^2}[\gamma(y)f''(y)]\mu_2^2(K)h^4 + \{O(n^{-2}h^{-1}) + O(h^4)\} \end{aligned} \tag{36}$$

Therefore, the optimal bandwidth selection for density functionals estimation of $U = \int \gamma(x)f^2(x)dx$ is h_{MSU}, the minimizer of $MSU(\hat{U})$. To obtain a closed form of optimal bandwidth for kernel functionals estimation of U, the min-imizer of the asymptotic mean square error (AMSE) of $\hat{U}$ is studied instead. The optimal bandwidth for estimation of U with respect to AMSE criterion is given by

$$h_{AMSE} = \left[\frac{2R(K)R(\gamma(x)f(x))}{\left(\int \gamma(x)f''(x)f(x)dx\right)^2 \mu_2^2(K)}\right]^{1/5} n^{-2/5} \tag{37}$$

However, h_{AMSE} in (37) is not computable since $R(\gamma(x)f(x)) = \int \gamma^2(x)f^2(x)dx$ and $\int \gamma(x)f''(x)f(x)dx$ depend on unknown function $f(x)$. A quick and simple guess of AMSE-optimal bandwidth is "Normal scale" bandwidth. Normal scale bandwidth selection will be studied for $\gamma = 1$ and x, respectively.

If f is Normal with mean 0 and variance σ^2, then $R(f)$ and $\mathbf{E}[f''(y)]$ can be calculated in terms of scale parameter σ. Note that

$$\begin{aligned} R(f) = \int f^2(y)dy &= \int \frac{1}{2\pi\sigma^2}e^{-y^2/\sigma^2}dy \\ &= \int \frac{1}{2\pi\sigma^2}e^{-z^2}\sigma dz = \frac{1}{2\sqrt{\pi}\sigma} \end{aligned} \tag{38}$$

$$\begin{aligned} \mathbf{E}[f''(y)] &= \int f''(y)f(y)dy \\ &= \int \left(\frac{y^2}{\sqrt{2\pi}\sigma^5}e^{-y^2/2\sigma^2} - \frac{1}{\sqrt{2\pi}\sigma^3}e^{-y^2/2\sigma^2}\right)\frac{1}{\sqrt{2\pi}\sigma}e^{-y^2/2\sigma^2}dy \\ &= \int \frac{1}{2\pi\sigma^4}(\frac{y^2}{\sigma^2} - 1)e^{-y^2/2\sigma^2}dy = -\frac{1}{4\sqrt{\pi}\sigma^3} \end{aligned} \tag{39}$$

Therefore

$$h_{AMSE;N(0,\sigma)} = \left[\frac{2R(K)(1/2\sqrt{\pi}\sigma}{(1/16\pi\sigma^6)\mu_2^2(K)}\right]^{1/5} n^{-2/5} \tag{40}$$

Then from the equation (38) and equation (39), equation (40) for the sample with the variance $\sigma = 1$ is simplified to

$$h \simeq 1.515717n^{-2/5} \tag{41}$$

From the equation above we can conclude that the bandwidth for kernel density estimator h goes as $\alpha N^{-2/5}$, where N is our sample size and is equal to n, in equation 40.

3 The Moments of a Probability Distribution

The n-th moment of a sample data x is represented as $\mathbf{E}[x^n]$, where,

$$\mathbf{E}[x^n] = \begin{cases} \sum_i x_i^n p(x_i), & \text{discrete,} \\ \int_{-\infty}^{\infty} x^n f(x)dx & \text{continues.} \end{cases}$$

First Moment is the average value or the mean of our sample data. It's the sum of the products of observations and their probabilities of occurrence aka.

Second Central Moment is the variance of the sample data. It demonstrates the spread of the observations from the average value.

Third Standardized Moment is called Skewness. Skewness gives an idea of the symmetry of the probability distribution around the mean.

Fourth Standardized Moment is called Kurtosis. Kurtosis gives an idea of how fat/heavy are the tails of a distribution, i.e how frequent extreme deviations (or outliers) are from the average value.

4 Results

4.1 Histogram

In this section the histograms of a set of random numbers with different bins are demonstrated.

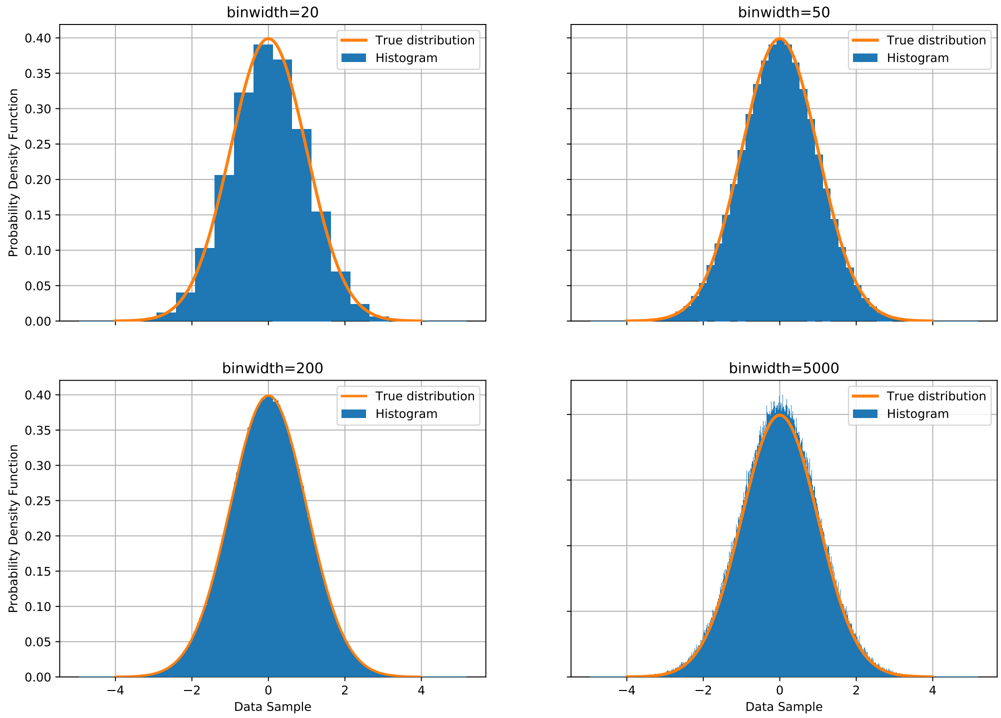

Figure 2: Histograms with different bins

The choice of the number of bins is important as it controls the coarseness of the distribution (number of bars) and, in turn, how well the density of the observations is plotted. It is a good idea to experiment with different bin sizes for a given data sample to get multiple perspectives or views on the same data.

As you can in the figure 2 increasing the number of bins does not really help out but we should use the optimal bins which was explained before. Also there is a try for calculating the optimal bins in the code (Jupyter Notebook code which is was sent as well). I found the optimal bins is equal to 199.

A histogram can be created using the Matplotlib library and the hist() function. The data is provided as the first argument, and the number of bins is specified via the "bins" argument either as an integer or as a sequence of the boundaries of each bin.

We can create a random sample drawn from a normal distribution and pretend we don't know the distribution, then create a histogram of the data. The normal() NumPy function will achieve this and we will generate $N \sim 10^6$ samples with a mean of 0 and a standard deviation of 1, e.g. a standard Gaussian.

4.2 KDE

There are several options available for computing kernel density estimates in Python. I used the Scikit-learn: KernelDensity library to perform KDE.

4.2.1 Bandwidth selection

The selection of bandwidth is an important piece of KDE, which was discussed earlier explicitly. For the same input data, different bandwidths can produce very different results.

There are two classes of approaches to this problem: in the statistics community, it is common to use reference rules, where the optimal bandwidth is estimated from theoretical forms based on assumptions about the data distribution. A common reference rule is Silverman's rule, which is derived for univariate KDE and included within both the Scipy and Statsmodels implementations. Other potential reference rules are ones based on Information Criteria, such as the well-known AIC and BIC.

In the Machine Learning world, the use of reference rules is less common. Instead, an empirical approach such as cross validation is often used. In cross validation, the model is fit to part of the data, and then a quantitative metric is computed to determine how well this model fits the remaining data. Such an empirical approach to model parameter selection is very flexible, and can be used regardless of the underlying data distribution.

Using cross validation within Scikit-learn is straightforward with the GridSearchCV meta-estimator. By using GridSearchCV I obtained $h = 0.181$ as an optimal bandwidth.

4.2.2 Kernels

In Fig.(1) four different kernels was demonstrated. Statsmodels contains seven kernels, while Scikit-learn contains six kernels, each of which can be used with one of about a dozen distance metrics, resulting in a very flexible range of effective kernel shapes. Again we can see all of the different kernels in Fig. 3.

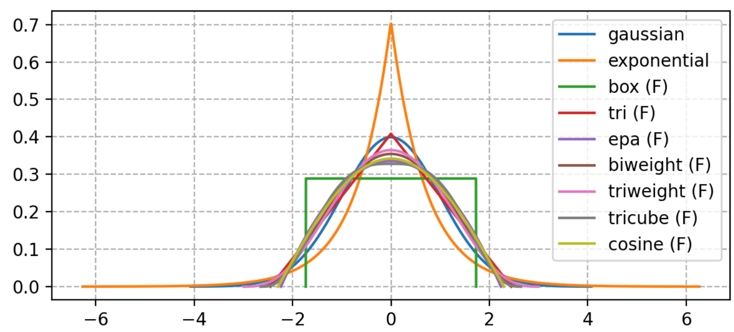

Figure 3: various kernel shapes [11].

Next page I included the plot which I obtained from the code with the sklearn library for the kernel density estimator with different bandwiths in Fig. (4). As it is obvious in the Fig. (4) too, the optimal bandwidth is approximately ~ 0.1 which it could be determined by the Eq.(40) as well too.

4.3 Moments

In Figure 5, I plotted first moment which is the mean to the 6-th moment. As we can see in the last plot, until the 5-th moment the tails in the end of the interval becomes zero, but in the 6-th moment and higher moments the moments diverge to infinity; therefor this higher moments are not reliable and should not be reported. So the 6-th moment should not be considered as a moment.

References

[1] `https://en.wikipedia.org/wiki/Histogram#Number_of_bins_and_width`

[2] Grinstead, Charles M.; Snell, J. Laurie (2009). Grinstead & Snell's Introduction to Probability. Orange Grove Texts. ISBN 161610046X. Retrieved 2019-07-25.

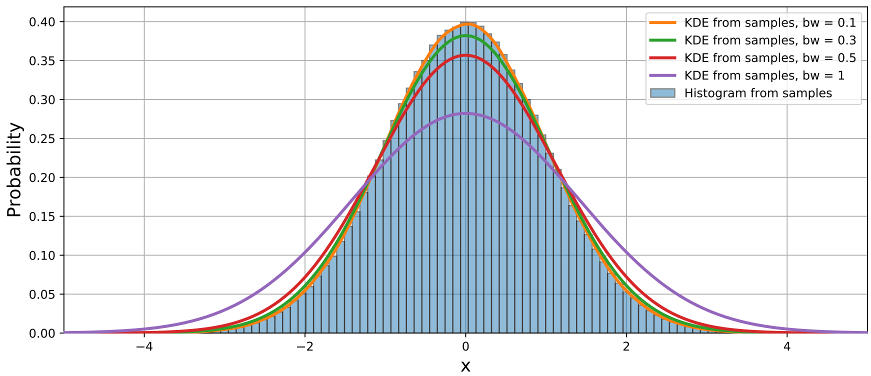

Figure 4: kernel density estimator with various bandwidths

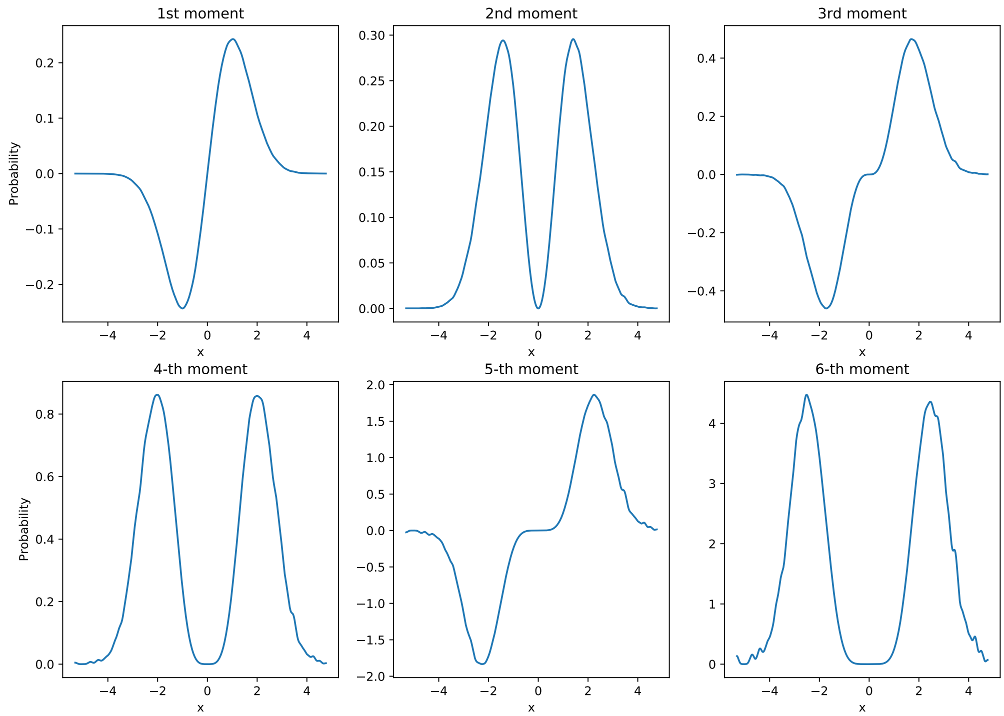

Figure 5: Moments

[3] Juan Carlos Cortés López & Marc Jornet Sanz; Math. Comput. Appl. 2020, 25, 33; doi:10.3390/mca25020033

[4] Howitt, D.; Cramer, D. (2008). Introduction to Statistics in Psychology (Fourth ed.). Prentice Hall. ISBN 978-0-13-205161-3.

[5] Frank Porter's Statistics Page. `http://www.hep.caltech.edu/~fcp/statistics/densityEst/densityEst.pdf`

[6] D. W. Scott, On optimal and data-based histograms, Biometrika 66 (1979), 605–610.

[7] D. W. Scott, Multivariate density estimation: Theory, practice, and visualization, John Wiley & Sons, 1992.

[8] Scott, David W. (2004) : Multivariate Density Estimation and Visualization, Papers, No. 2004,16, Humboldt-Universität zu Berlin, Center for Applied Statistics and Economics (CASE), Berlin

[9] A. J. Lee, U-Statistics: Theory and Practice, vol. 110, Marcel Dekker, New York, NY, USA, 1990.

[10] Su Chen, 2015. "Optimal Bandwidth Selection for Kernel Density Functionals Estimation," Journal of Probability and Statistics, Hindawi, vol. 2015, pages 1-21, August.DOI:10.1155/2015/242683

[11] `https://kdepy.readthedocs.io/en/latest/kernels.html`