STAT 535: Statistical Machine Learning **Autumn 2024**

Lecture 6: Density Estimation

Instructor: Yen-Chi Chen

Main reference: Section 6 of *All of Nonparametric Statistics* by Larry Wasserman.

A book about the methodologies of density estimation: *Multivariate Density Estimation: theory, practice, and visualization* by David Scott.

A more theoretical book (highly recommend if you want to learn more about the theory): *Introduction to Nonparametric Estimation* by A.B. Tsybakov.

Density estimation is the problem of reconstructing the probability density function using a set of given data points. Namely, we observe $X_1, \cdots, X_n$ and we want to recover the underlying probability density function generating our dataset.

6.1 Histogram

If the goal is to estimate the PDF, then this problem is called *density estimation*, which is a central topic in statistical research. Here we will focus on the perhaps simplest approach: histogram.

For simplicity, we assume that $X_i \in [0, 1]$ so $p(x)$ is non-zero only within $[0, 1]$. We also assume that $p(x)$ is smooth and $|p'(x)| \leq L$ for all x (i.e. the derivative is bounded). The histogram is to partition the set $[0, 1]$ (this region, the region with non-zero density, is called the support of a density function) into several bins and using the count of the bin as a density estimate. When we have M bins, this yields a partition:

$$B_1 = \left[0, \frac{1}{M}\right), B_2 = \left[\frac{1}{M}, \frac{2}{M}\right), \cdots, B_{M-1} = \left[\frac{M-2}{M}, \frac{M-1}{M}\right), B_M = \left[\frac{M-1}{M}, 1\right].$$

In such case, then for a given point $x \in B_\ell$, the density estimator from the histogram will be

$$\widehat{p}_M(x) = \frac{\text{number of observations within } B_\ell}{n} \times \frac{1}{\text{length of the bin}} = \frac{M}{n} \sum_{i=1}^{n} I(X_i \in B_\ell).$$

The intuition of this density estimator is that the histogram assign equal density value to every points within the bin. So for B_ℓ that contains x, the ratio of observations within this bin is $\frac{1}{n}\sum_{i=1}^{n} I(X_i \in B_\ell)$, which should be equal to the density estimate times the length of the bin.

Theorem 6.1 *Suppose that $p(x)$ has a uniformly bounded derivative, i.e., there exists L such that $\sup_x |p'(x)| \leq L$. Also, let $p_{\max} = \sup_x p(x) < \infty$. Then*

$$\textbf{bias}(\widehat{p}_M(x)) \leq \frac{L}{M}, \quad \mathsf{Var}(\widehat{p}_M(x)) \leq M\frac{p_{\max}}{n} + \frac{p_{\max}^2}{n}.$$

Proof:

Bias. Now we study the bias of the histogram density estimator.

$$\begin{aligned}
\mathbb{E}\left(\widehat{p}_M(x)\right) &= M \cdot P(X_i \in B_\ell) \\
&= M \int_{\frac{\ell-1}{M}}^{\frac{\ell}{M}} p(u)du \\
&= M\left(F\left(\frac{\ell}{M}\right) - F\left(\frac{\ell-1}{M}\right)\right) \\
&= \frac{F\left(\frac{\ell}{M}\right) - F\left(\frac{\ell-1}{M}\right)}{1/M} \\
&= \frac{F\left(\frac{\ell}{M}\right) - F\left(\frac{\ell-1}{M}\right)}{\frac{\ell}{M} - \frac{\ell-1}{M}} \\
&= p(x^*), \quad x^* \in \left[\frac{\ell-1}{M}, \frac{\ell}{M}\right].
\end{aligned}$$

The last equality is done by the mean value theorem with $F'(x) = p(x)$. By the mean value theorem again, there exists another point x^{**} between x^* and x such that

$$\frac{p(x^*) - p(x)}{x^* - x} = p'(x^{**}).$$

Thus, the bias

$$\begin{aligned}
\textbf{bias}(\widehat{p}_M(x)) &= \mathbb{E}\left(\widehat{p}_M(x)\right) - p(x) \\
&= p(x^*) - p(x) \\
&= p'(x^{**}) \cdot (x^* - x) \\
&\leq |p'(x^{**})| \cdot |x^* - x| \\
&\leq \frac{L}{M}.
\end{aligned} \tag{6.1}$$

Note that in the last inequality we use the fact that both x^* and x are within B_ℓ, whose total length is $1/M$, so the $|x^* - x| \leq 1/M$.

Variance. Now we turn to the analysis of variance.

$$\begin{aligned}
\mathsf{Var}(\widehat{p}_M(x)) &= M^2 \cdot \mathsf{Var}\left(\frac{1}{n}\sum_{i=1}^{n} I(X_i \in B_\ell)\right) \\
&= M^2 \cdot \frac{P(X_i \in B_\ell)(1 - P(X_i \in B_\ell))}{n}.
\end{aligned}$$

By the derivation in the bias, we know that $P(X_i \in B_\ell) = \frac{p(x^*)}{M}$, so the variance

$$\begin{aligned}
\mathsf{Var}(\widehat{p}_M(x)) &= M^2 \cdot \frac{\frac{p(x^*)}{M} \times \left(1 - \frac{p(x^*)}{M}\right)}{n} \\
&= M \cdot \frac{p(x^*)}{n} + \frac{p^2(x^*)}{n} \\
&\leq M \cdot \frac{p_{\max}}{n} + \frac{p_{\max}^2}{n}.
\end{aligned} \tag{6.2}$$

■

The analysis of the bias tells us that the more bins we are using, the less bias the histogram has. This makes sense because when we have many bins, we have a higher resolution so we can approximate the fine density

structure better. The analysis of the variance has an interesting result: the more bins we are using, the higher variance we are suffering.

In fact, to obtain the same convergence rate of the bias, we do not even need the existence of derivative of p. As long as p is *Lipschitz continuous*, we have the same convergence rate. Note that p is L-Lipschitz (continuous) if for every $x, y \in \mathbb{R}$ we have $|p(x) - p(y)| \leq L|x - y|$. With this, it is easy to see that

$$\begin{aligned} |\mathbf{bias}(\widehat{p}_M(x))| &= |\mathbb{E}(\widehat{p}_M(x)) - p(x)| \\ &= |p(x^*) - p(x)| \\ &\leq L|x^* - x| \leq \frac{L}{M}. \end{aligned}$$

Now if we consider the MSE, the pattern will be more inspiring. The MSE is

$$\mathsf{MSE}(\widehat{p}_M(x)) = \mathsf{bias}^2(\widehat{p}_M(x)) + \mathsf{Var}(\widehat{p}_M(x)) \leq \frac{L^2}{M^2} + M \cdot \frac{p_{\max}}{n} + \frac{p_{\max}^2}{n}. \tag{6.3}$$

An interesting feature of the histogram is that: *we can choose M, the number of bins.* When M is too large, the first quantity (bias) will be small while the second quantity (variance) will be large; this case is called *undersmoothing*. When M is too small, the first quantity (bias) is large but the second quantity (variance) is small; this case is called *oversmoothing*.

To balance the bias and variance, we choose M that minimizes the MSE, which leads to

$$M_{\mathsf{opt}} = \left(\frac{n \cdot L^2}{p_{\max}}\right)^{1/3}. \tag{6.4}$$

Although in practice the quantity L and $p_{\max}$ are unknown so we cannot chose the optimal M_{opt}, the rule in equation (6.4) tells us how we should change the number of bins when we have more and more sample size. Practical rule of selecting M is related to the problem of *bandwidth selection*, a research topic in statistics.

6.1.1 L_∞ analysis

We have derived the convergence rate for a histogram density estimator at a given point x. However, we have not yet analyzed the *uniform convergence rate* of the histogram. Using the Hoeffding's inequality, we are able to obtain such a uniform convergence rate.

Assume that $X_1, \cdots, X_n \sim F$ with a PDF p that has a non-zero density over $[0, 1]$. If we use a histogram with M bins:

$$B_1 = \left[0, \frac{1}{M}\right), B_2 = \left[\frac{1}{M}, \frac{2}{M}\right), \cdots, B_M = \left[\frac{M-1}{M}, 1\right].$$

Let $\widehat{p}_M(x)$ be the histogram density estimator:

$$\widehat{p}_M(x) = \frac{M}{n} \sum_{i=1}^{n} I(X_i \in B(x)),$$

where $B(x)$ is the bin that x belongs to.

The goal is to bound

$$\sup_x |\widehat{p}_M(x) - p(x)|.$$

Theorem 6.2 *Suppose that $p(x)$ has a uniformly bounded derivative and bounded density. Then*

$$\sup_x |\widehat{p}_M(x) - p(x)| = O\left(\frac{1}{M}\right) + O_P\left(\sqrt{\frac{M^2 \log M}{n}}\right).$$

Proof:

We know that the difference $\widehat{p}_M(x) - p(x)$ can be written as

$$\widehat{p}_M(x) - p(x) = \underbrace{\widehat{p}_M(x) - \mathbb{E}(\widehat{p}_M(x))}_{\text{stochastic variation}} + \underbrace{\mathbb{E}(\widehat{p}_M(x)) - p(x)}_{\text{bias}}.$$

The bias analysis we have done in Lecture 6 can be generalized to every point x, so we have

$$\sup_x |\mathbb{E}(\widehat{p}_M(x)) - p(x)| = O\left(\frac{1}{M}\right).$$

So we only need to bound the stochastic variation part $\sup_x |\widehat{p}_M(x) - \mathbb{E}(\widehat{p}_M(x))|$. Although we are taking supremum over every x, there are only M bins $B_1, \cdots, B_M$ so we can rewrite the stochastic part as

$$\begin{aligned}
\sup_x |\widehat{p}_M(x) - \mathbb{E}(\widehat{p}_M(x))| &= \max_{j=1,\cdots,M} \left| \frac{M}{n} \sum_{i=1}^n I(X_i \in B_j) - MP(X_i \in B_j) \right| \\
&= M \cdot \max_{j=1,\cdots,M} \left| \frac{1}{n} \sum_{i=1}^n I(X_i \in B_j) - P(X_i \in B_j) \right|.
\end{aligned}$$

Because the indicator function takes only two values: 0 and 1, we have

$$\begin{aligned}
P\left(\sup_x |\widehat{p}_M(x) - \mathbb{E}(\widehat{p}_N(x))| > \epsilon\right) &= P\left(M \cdot \max_{j=1,\cdots,M} \left| \frac{1}{n} \sum_{i=1}^n I(X_i \in B_j) - P(X_i \in B_j) \right| > \epsilon\right) \\
&= P\left(\max_{j=1,\cdots,M} \left| \frac{1}{n} \sum_{i=1}^n I(X_i \in B_j) - P(X_i \in B_j) \right| > \underbrace{\frac{\epsilon}{M}}_{=\epsilon'}\right) \\
&\le \sum_{j=1}^M P\left(\left| \frac{1}{n} \sum_{i=1}^n I(X_i \in B_j) - P(X_i \in B_j) \right| > \underbrace{\frac{\epsilon}{M}}_{=\epsilon'}\right) \\
&\le M \cdot 2e^{-2n\epsilon'^2} \qquad \text{(by Hoeffding's inequality)} \\
&= M \cdot 2e^{-2n\epsilon^2/M^2}.
\end{aligned}$$

Thus,

$$\sup_x |\widehat{p}_M(x) - \mathbb{E}(\widehat{p}_M(x))| = O_P\left(\sqrt{\frac{M^2 \log M}{n}}\right)$$

This, together with the uniform bias, implies

$$\begin{aligned}
\sup_x |\widehat{p}_M(x) - p(x)| &\le \sup_x |\mathbb{E}(\widehat{p}_N(x)) - p(x)| + \sup_x |\widehat{p}_M(x) - \mathbb{E}(\widehat{p}_N(x))| \\
&= O\left(\frac{1}{M}\right) + O_P\left(\sqrt{\frac{M^2 \log M}{n}}\right).
\end{aligned}$$

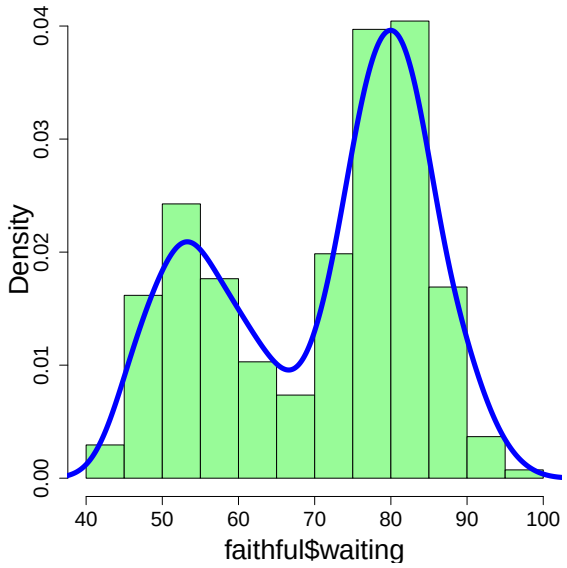
■

Note that this bound is not the tightest bound we can obtain. Using the Bernstein's inequality[1], you can obtain a faster convergence rate:

$$\sup_x |\widehat{p}_M(x) - p(x)| = O\left(\frac{1}{M}\right) + O_P\left(\sqrt{\frac{M \log M}{n}}\right).$$

6.2 Kernel Density Estimator

Here we will talk about another approach–the *kernel density estimator* (KDE; sometimes called kernel density estimation). The KDE is one of the most famous method for density estimation. The follow picture shows the KDE and the histogram of the `faithful` dataset in R. The blue curve is the density curve estimated by the KDE.

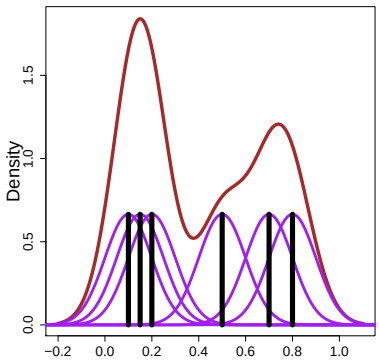


Here is the formal definition of the KDE. The KDE is a function

$$\widehat{p}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right), \tag{6.5}$$

where $K(x)$ is called the *kernel function* that is generally a smooth, symmetric function such as a Gaussian and $h > 0$ is called the smoothing bandwidth that controls the amount of smoothing. Basically, the KDE smoothes each data point X_i into a small density bumps and then sum all these small bumps together to obtain the final density estimate. The following is an example of the KDE and each small bump created by it:

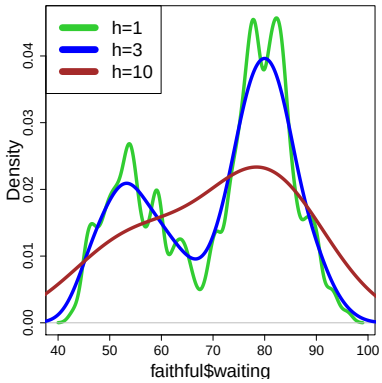
[1] https://en.wikipedia.org/wiki/Bernstein_inequalities_(probability_theory)



In the above picture, there are 6 data points located at where the black vertical segments indicate: $0.1, 0.2, 0.5, 0.7, 0.8, 0.15$. The KDE first smooth each data point into a purple density bump and then sum them up to obtain the final density estimate–the brown density curve.

6.2.1 Bandwidth and Kernel Functions

The smoothing bandwidth h plays a key role in the quality of KDE. Here is an example of applying different h to the `faithful` dataset:



Clearly, we see that when h is too small (the green curve), there are many wiggly structures on our density curve. This is a signature of *undersmoothing*–the amount of smoothing is too small so that some structures identified by our approach might be just caused by randomness. On the other hand, when h is too large (the brown curve), we see that the two bumps are smoothed out. This situation is called *oversmoothing*–some important structures are obscured by the huge amount of smoothing.

How about the choice of kernel function? A kernel function generally has two features:

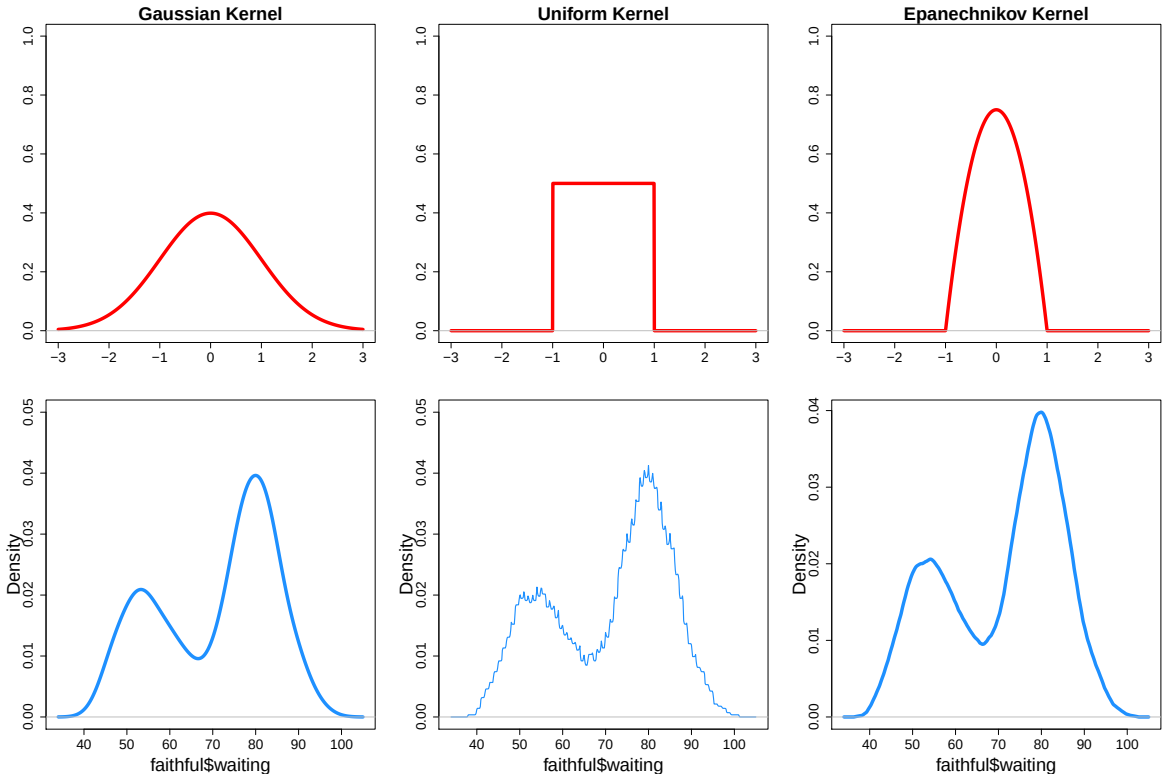
(K1) $K(x)$ is symmetric.

(K2) $\int K(x)dx = 1$.

(K3) $\lim_{x\to-\infty} K(x) = \lim_{x\to+\infty} K(x) = 0$.

In particular, the second requirement is needed to guarantee that the KDE $\widehat{p}_n(x)$ is a probability density function. Note that most kernel functions are positive; however, kernel functions could be negative [2].

In theory, the kernel function does not play a key role (later we will see this). But sometimes in practice, they do show some difference in the density estimator. In what follows, we consider three most common kernel functions and apply them to the `faithful` dataset:



The top row displays the three kernel functions and the bottom row shows the corresponding density estimators. Here is the form of the three kernels:

$$\begin{aligned}
\text{Gaussian} \quad & K(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}, \\
\text{Uniform} \quad & K(x) = \frac{1}{2} I(-1 \le x \le 1), \\
\text{Epanechnikov} \quad & K(x) = \frac{3}{4} \cdot \max\{1 - x^2, 0\}.
\end{aligned}$$

The *Epanechnikov* is a special kernel that has the lowest (asymptotic) mean square error.

Note that there are many many many other kernel functions such as triangular kernel, biweight kernel, cosine kernel, ...etc. If you are interested in other kernel functions, please see https://en.wikipedia.org/wiki/Kernel_(statistics).

[2] Some special types of kernel functions, known as the *higher order* kernel functions, will take negative value at some regions. These higher order kernel functions, though very counter intuitive, might have a smaller bias than the usual kernel functions.

6.2.2 Theory of the KDE

Now we will analyze the estimation error of the KDE. Assume that $X_1, \cdots, X_n$ are IID sample from an unknown density function p. In the density estimation problem, the parameter of interest is p, the true density function.

To simplify the problem, assume that we focus on a given point x_0 and we want to analyze the quality of our estimator $\widehat{p}_n(x_0)$.

Theorem 6.3 *Assume that $p(x)$ has bounded third derivatives and assumption (K1-3) holds for the kernel function. Then when $h \to 0$ as $n \to \infty$, we have*

$$\textbf{bias}(\widehat{p}_n(x_0)) = \frac{1}{2}h^2 p''(x_0)\mu_K + o(h^2), \quad \mathsf{Var}(\widehat{p}_n(x_0)) = \frac{1}{nh}p(x_0)\sigma_K^2 + o\left(\frac{1}{nh}\right).$$

Proof: Bias. The bias of KDE is

$$\begin{aligned}
\mathbb{E}(\widehat{p}_n(x_0)) - p(x_0) &= \mathbb{E}\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)\right) - p(x_0) \\
&= \frac{1}{h}\mathbb{E}\left(K\left(\frac{X_i - x_0}{h}\right)\right) - p(x_0) \\
&= \frac{1}{h}\int K\left(\frac{x - x_0}{h}\right) p(x)dx - p(x_0).
\end{aligned}$$

Now we do a change of variable $y = \frac{x-x_0}{h}$ so that $dy = dx/h$ and the above becomes

$$\begin{aligned}
\mathbb{E}(\widehat{p}_n(x_0)) - p(x_0) &= \int K\left(\frac{x - x_0}{h}\right) p(x)\frac{dx}{h} - p(x_0) \\
&= \int K(y)\, p(x_0 + hy)dy - p(x_0) \qquad \text{(using the fact that } x = x_0 + hy).
\end{aligned}$$

Now by Taylor expansion, when h is small,

$$p(x_0 + hy) = p(x_0) - hy \cdot p'(x_0) + \frac{1}{2}h^2y^2p''(x_0) + o(h^2).$$

Note that $o(h^2)$ means that it is a smaller order term compared to h^2 when $h \to 0$. Plugging this back to the bias, we obtain

$$\begin{aligned}
\mathbb{E}(\widehat{p}_n(x_0)) - p(x_0) &= \int K(y)\, p(x_0 - hy)dy - p(x_0) \\
&= \int K(y)\left[p(x_0) + hy \cdot p'(x_0) + \frac{1}{2}h^2y^2p''(x_0) + o(h^2)\right]dy - p(x_0) \\
&= \int K(y)\, p(x_0)dy + \int K(y)\, hy \cdot p'(x_0)dy + \int K(y)\,\frac{1}{2}h^2y^2p''(x_0)dy + o(h^2) - p(x_0) \\
&= p(x_0)\underbrace{\int K(y)\, dy}_{=1} + hp'(x_0)\underbrace{\int yK(y)\, dy}_{=0} + \frac{1}{2}h^2p''(x_0)\int y^2K(y)\, dy + o(h^2) - p(x_0) \\
&= p(x_0) + \frac{1}{2}h^2p''(x_0)\int y^2K(y)\, dy - p(x_0) + o(h^2) \\
&= \frac{1}{2}h^2p''(x_0)\int y^2K(y)\, dy + o(h^2) \\
&= \frac{1}{2}h^2p''(x_0)\mu_K + o(h^2),
\end{aligned}$$

where $\mu_K = \int y^2 K(y)\, dy$. Namely, the bias of the KDE is

$$\textbf{bias}(\widehat{p}_n(x_0)) = \frac{1}{2}h^2 p''(x_0)\mu_K + o(h^2). \tag{6.6}$$

Variance. For the analysis of variance, we can obtain an upper bound using a straight forward calculation:

$$\begin{aligned}
\mathsf{Var}(\widehat{p}_n(x_0)) &= \mathsf{Var}\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)\right) \\
&= \frac{1}{nh^2}\mathsf{Var}\left(K\left(\frac{X_i - x_0}{h}\right)\right) \\
&\le \frac{1}{nh^2}\mathbb{E}\left(K^2\left(\frac{X_i - x_0}{h}\right)\right) \\
&= \frac{1}{nh^2}\int K^2\left(\frac{x - x_0}{h}\right) p(x)dx \\
&= \frac{1}{nh}\int K^2(y)p(x_0 + hy)dy \quad (\text{using } y = \tfrac{x-x_0}{h} \text{ and } dy = dx/h \text{ again}) \\
&= \frac{1}{nh}\int K^2(y)\left[p(x_0) + hyp'(x_0) + o(h)\right] dy \\
&= \frac{1}{nh}\left(p(x_0)\cdot\int K^2(y)dy + o(h)\right) \\
&= \frac{1}{nh}p(x_0)\int K^2(y)dy + o\left(\frac{1}{nh}\right) \\
&= \frac{1}{nh}p(x_0)\sigma_K^2 + o\left(\frac{1}{nh}\right),
\end{aligned}$$

where $\sigma_K^2 = \int K^2(y)dy$. ■

This means that when we allow $h \to 0$, the bias is shrinking at a rate $O(h^2)$. Equation (6.15) reveals an interesting fact: the bias of KDE is caused by the *curvature* (second derivative) of the density function! Namely, the bias will be very large at a point where the density function curves a lot (e.g., a very peaked bump). This makes sense because for such a structure, KDE tends to smooth it too much, making the density function smoother (less curved) than it used to be.

Also, the variance shrinks at rate $O(\frac{1}{nh})$ when $n \to \infty$ and $h \to 0$. An interesting fact from the variance is that: at point where the density value is large, the variance is also large!

Now putting both bias and variance together, we obtain the MSE of the KDE:

$$\begin{aligned}
\textbf{MSE}(\widehat{p}_n(x_0)) &= \textbf{bias}^2(\widehat{p}_n(x_0)) + \mathsf{Var}(\widehat{p}_n(x_0)) \\
&= \frac{1}{4}h^4|p''(x_0)|^2\mu_K^2 + \frac{1}{nh}p(x_0)\sigma_K^2 + o(h^4) + o\left(\frac{1}{nh}\right) \\
&= O(h^4) + O\left(\frac{1}{nh}\right).
\end{aligned}$$

The first two term, $\frac{1}{4}h^4|p''(x_0)|^2\mu_K^2 + \frac{1}{nh}p(x_0)\sigma_K^2$, is called the asymptotic mean square error (AMSE). In the KDE, the smoothing bandwidth h is something we can choose. Thus, the bandwidth h minimizing the AMSE is

$$h_{\mathsf{opt}}(x_0) = \left(\frac{4}{n}\cdot\frac{p(x_0)}{|p''(x_0)|^2}\frac{\sigma_K^2}{\mu_K^2}\right)^{\frac{1}{5}} = C_1 \cdot n^{-\frac{1}{5}}.$$

And this choice of smoothing bandwidth leads to a MSE at rate

$$\mathbf{MSE}_{\mathsf{opt}}(\widehat{p}_n(x_0)) = O(n^{-\frac{4}{5}}).$$

The optimal MSE of the KDE is at rate $O(n^{-\frac{4}{5}})$, which is faster than the optimal MSE of the histogram $O(n^{-\frac{2}{3}})$! However, both are slower than the MSE of a MLE ($O(n^{-1})$). This reduction of error rate is the price we have to pay for a more flexible model (we do not assume the data is from any particular distribution but only assume the density function is smooth).

In the above analysis, we only consider a single point x_0. In general, we want to control the overall MSE of the *entire function*. In this case, a straight forward generalization is the *mean integrated square error (MISE)*:

$$\mathbf{MISE}(\widehat{p}_n) = \mathbb{E}\left(\int (\widehat{p}_n(x) - p(x))^2\right) = \int \mathbf{MSE}(\widehat{p}_n(x))dx.$$

Note that the second equality follows from the Fubini's theorem. Under a similar derivation, one can show that

$$\begin{aligned}\mathbf{MISE}(\widehat{p}_n) &= \frac{1}{4}h^4 \int |p''(x)|^2 dx \mu_K^2 + \frac{1}{nh}\underbrace{\int p(x)dx}_{=1}\sigma_K^2 + o(h^4) + o\left(\frac{1}{nh}\right) \\ &= \frac{\mu_K^2}{4}\cdot h^4 \cdot \underbrace{\int |p''(x)|^2 dx}_{\text{Overall curvature}} + \frac{\sigma_K^2}{nh} + o(h^4) + o\left(\frac{1}{nh}\right) \qquad (6.7) \\ &= O(h^4) + O\left(\frac{1}{nh}\right).\end{aligned}$$

The two dominating terms in equation (6.7), $\frac{\mu_K^2}{4}\cdot h^4 \cdot \underbrace{\int |p''(x)|^2 dx}_{\text{Overall curvature}} + \frac{\sigma_K^2}{nh}$, is called the *asymptotical mean integrated square error (AMISE)*. The optimal smoothing bandwidth is often chosen by minimizing this quantity. Namely,

$$h_{\mathsf{opt}} = \left(\frac{1}{n}\cdot\frac{4}{\int |p''(x)|^2 dx}\cdot\frac{\sigma_K^2}{\mu_K^2}\right)^{\frac{1}{5}} = C_2 \cdot n^{-\frac{1}{5}}. \qquad (6.8)$$

6.2.3 Bandwidth Selection

Although equation (6.8) provies an expression of the optimal bandwidth as h_{opt}, this choice is not applicable in practice because it involves the unknown quantity $\int |p''(x)|^2 dx$. Thus, how to choose h is an unsolved problem in statistics and is known as *bandwidth selection*[3]. Most bandwidth selection approaches are either proposing an estimate of AMISE and then minimizing the estimated AMISE or using an estimate of the curvature $\int |p''(x)|^2 dx$ and choose h_{opt} accordingly.

There are 4 common approaches for selecting the bandwidth.

- **Rule of thumb.** The rule of thumb is motivated by analyzing the problem of the optimal bandwidth for estimating the PDF of a univariate Gaussian with a Gaussian kernel. In this case,

$$h_{\mathsf{ROT},0} = 1.05 \cdot \widehat{\sigma} \cdot n^{-1/5} \quad \text{or} \quad h_{\mathsf{ROT},0} = 0.79 \cdot \widehat{\mathsf{IQR}} \cdot n^{-1/5}$$

[3] See https://en.wikipedia.org/wiki/Kernel_density_estimation#Bandwidth_selection for more details.

Later Silverman noticed that this bandwidth often oversmooth and miss the multi-modality so he recommended to revise it as

$$h_{\mathsf{ROT}} = 0.9 \cdot \min\{\widehat{\sigma}, \widehat{\mathsf{IQR}}/1.34\} \cdot n^{-1/5}.$$

The choice h_{ROT} is often known as the *Silverman's rule of thumb*.

- **Plug-in method.** The plug-in method is a very intuitive approach–we replace the unknown quantity with a consistent estimator. In this case, we estimate $p''(x)$ and then plug-in the estimated curvature into equation (6.8) and obtain the corresponding optimal bandwidth. However, a challenge is that we need to specify another bandwidth for the second derivative estimation.

- **Least squared cross-validation.** The least squared cross-validation (LSCV) is motivated by the following expansion of the MISE:

$$\mathbf{MISE}(\widehat{p}_n) = \int (\widehat{p}_n(x) - p(x))^2 dx = \int \widehat{p}_n^2(x)dx - \int \widehat{p}_n(x)p(x)dx + C_0.$$

  We may estimate the second integral using the idea of leave-one-out CV

$$\int \widehat{p}_n(x)p(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} \widehat{p}_{n,-i}(X_i),$$

  where $\widehat{p}_{n,-i}$ is the KDE computed without using X_i. This leads to a criterion

$$R_{\mathsf{LSCV}}(h) = \int \widehat{p}_n^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{p}_{n,-i}(X_i).$$

  We then choose h by minimizing $R_{\mathsf{LSCV}}(h)$. **WARNING:** this is not the cross-validation used in regression or classification problem although it does use the concept of leave-one out cross-validation.

- **Lepski's approach.** Recently, Goldenshluger and Lepski (2011) propose a method, known as the Lepskis approach, that treats the bandwidth selection problem as a model selection problem and proposes a new criterion for selecting the smoothing bandwidth. One feature of Lepskis approach is that the selected bandwidth enjoys many statistical optimalities. See the following two papers for more details:

  1. Goldenshluger, A., & Lepski, O. (2008). Universal pointwise selection rule in multivariate function estimation. *Bernoulli*, 14(4), 1150-1190.
  2. Goldenshluger, A., & Lepski, O. (2011). Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39(3), 1608-1632.

For more information on bandwidth selection, I would recommend

Sheather, Simon J. "Density Estimation." *Statistical Science* 19.4 (2004): 588-597.

6.2.4 Confidence Interval using the KDE

In this section, we will discuss an interesting topic–confidence interval of the KDE. For simplicity, we will focus on the CI of the density function at a given point x_0. Recall from equation (6.5),

$$\widehat{p}_n(x_0) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

where $Y_i = \frac{1}{h}K\left(\frac{X_i - x_0}{h}\right)$. Thus, the KDE evaluated at x_0 is actually a sample mean of $Y_1, \cdots, Y_n$. By CLT,

$$\sqrt{n}\left(\frac{\widehat{p}_n(x_0) - \mathbb{E}(\widehat{p}_n(x_0))}{\mathsf{Var}(Y_i)}\right) \xrightarrow{D} N(0,1).$$

However, one has to be very careful when using this result because from the analysis of variance,

$$\mathsf{Var}(Y_i) = \mathsf{Var}\left(\frac{1}{h}K\left(\frac{X_i - x_0}{h}\right)\right) = \frac{1}{h}p(x_0)\sigma_K^2 + o\left(\frac{1}{h}\right)$$

diverges when $h \to 0$. Thus, when $h \to 0$, the asymptotic distribution of $\widehat{p}_n(x_0)$ is

$$\sqrt{nh}\left(\widehat{p}_n(x_0) - \mathbb{E}(\widehat{p}_n(x_0))\right) \xrightarrow{D} N(0, p(x_0)\sigma_K^2).$$

Thus, a $1-\alpha$ CI can be constructed using

$$\widehat{p}_n(x_0) \pm z_{1-\alpha/2} \cdot \sqrt{p(x_0)\sigma_K^2}.$$

This CI cannot be used in practice because $p(x_0)$ is unknown to us. One solution to this problem is to replace it by the KDE, leading to

$$\widehat{p}_n(x_0) \pm z_{1-\alpha/2} \cdot \sqrt{\widehat{p}_n(x_0)\sigma_K^2}.$$

There is another approach for constructing CIs called the bootstrap approach. We will talk about this approach in the future lecture.

Remarks.

- A problem of these CI is that the theoretical guarantee of coverage is for the expectation of the KDE $\mathbb{E}(\widehat{p}_n(x_0))$ rather than the true density value $p(x_0)$! Recall from the analysis of bias, the bias is at the order of $O(h^2)$. Thus, if h is fixed or h converges to 0 slowly, the coverage of CI will be lower than the nominal coverage (this is called *undercoverage*). Namely, even if we construct a 99% CI, the chance that this CI covers the actual density value can be only 1% or even lower!

  In particular, when we choose $h = h_{\mathsf{opt}}$, we always suffers from the problem of undercoverage because the bias and stochastic variation is at a similar order.

- To handle the problem of undercoverage, a most straight forward approach is to choose $h \to 0$ faster than the optimal rate. This method is called *undersmoothing*. However, when we undersmooth the data, the MSE will be large (because the variance is going to get higher than the optimal case), meaning that the accuracy of estimation decreases.

- Aside from the problem of bias, the CIs we construct are only for a single point x_0 so the CI only has a pointwise coverage. Namely, if we use the same procedure to construct a $1-\alpha$ CI of every point, the probability that the *entire* density function is covered by the CI may be way less than the nominal confidence level $1-\alpha$.

- There are approaches of constructing a CI such that it simultaneously covers the entire function. In this case, the CI will be called a *confidence band* because it is like a band with the nominal probability of covering the entire function. In general, how people construct a confidence band is via the bootstrap and the band consists of two functions $L_\alpha(x), U_\alpha(x)$ that can be constructed using the sample $X_1, \cdots, X_n$ such that

$$P(L_\alpha(x) \le p(x) \le U_\alpha(x) \ \forall\ x) \approx 1 - \alpha.$$

6.2.5 L_∞ analysis

The L_∞ error, $\sup_x |\widehat{p}_n(x) - p(x)|$, is often used in many theoretical analysis. The L_∞ analysis of the KDE can be found in

> Giné, Evarist, and Armelle Guillou. "Rates of strong uniform consistency for multivariate kernel density estimators." *Annales de l'Institut Henri Poincare (B) Probability and Statistics.* Vol. 38. No. 6. No longer published by Elsevier, 2002.

The key result is the following theorem from Giné and Guillou (2002). This theorem is based on the Talagrand's inequality[4], one of the most important inequalities in statistical learning theory.

Theorem 6.4 (Giné and Guillou (2002)) *Assume the conditions in Giné and Guillou (2002). Then when $h \to 0$, there exists $c_1, c_2 > 0$ such that*

$$P(\sup_x |\widehat{p}_n(x) - p(x)| > \epsilon) \leq c_1 e^{-c_2 \cdot nh^d \cdot \epsilon^2} \tag{6.9}$$

for every

$$\epsilon \geq \sqrt{\frac{|\log h|}{nh^d}}. \tag{6.10}$$

The key assumption in Giné and Guillou (2002) is a *covering number* assumptions on the collection of kernel functions, which is satisfied by Gaussian kernel and any other compact support kernel function. Thus, this result holds for most commonly used kernel functions. With the above theorem, we can prove the following results.

Theorem 6.5 *Assume the conditions in Giné and Guillou (2002). Then when $h \to 0$ as $n \to \infty$,*

$$\sup_x |\widehat{p}_n(x) - p(x)| = O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh}}\right).$$

The extra $\sqrt{\log n}$ term has many interesting stories and it comes from the *empirical process theory*. We will learn this concept in the learning theory.

Proof: The quantity $O(h^2)$ is from the usual bias analysis so we omit the proof. Here we will focus on showing the O_P rate.

Let $\Delta_n = \sup_x |\widehat{p}_n(x) - p(x)|$. The restriction on ϵ actually constrains the rate to be $O_P\left(\sqrt{\frac{|\log h|}{nh^d}}\right)$. To see this, we first rewrite equation (6.9) using $t^2 = nh^d\epsilon^2$:

$$\begin{aligned} & P(\Delta_n > \epsilon) \leq c_1 e^{-c_2 \cdot nh^d \cdot \epsilon^2} \\ \Longrightarrow & P(\sqrt{nh^d}\Delta_n > \sqrt{nh^d}\epsilon) \leq c_1 e^{-c_2 \cdot nh^d \cdot \epsilon^2} \\ \Longrightarrow & P(\sqrt{nh^d}\Delta_n > t) \leq c_1 e^{-c_2 t^2}, \end{aligned}$$

[4] https://en.wikipedia.org/wiki/Talagrand%27s_concentration_inequality

when $t \geq \sqrt{|\log h|}$. Here you see that we cannot pick the right-hand-side arbitrarily small because of the lower bound on t. The above result directly leads to a bound on $\mathbb{E}(\sqrt{nh^d}\Delta_n)$:

$$\begin{aligned}
\mathbb{E}(\sqrt{nh^d}\Delta_n) &= \int_0^\infty P(\sqrt{nh^d}\Delta_n > t)dt \\
&= \int_{\sqrt{|\log h|}}^\infty P(\sqrt{nh^d}\Delta_n > t)dt + \int_0^{\sqrt{|\log h|}} P(\sqrt{nh^d}\Delta_n > t)dt \\
&\leq O(h^{-c_3}) + \int_0^{\sqrt{|\log h|}} 1dt \\
&= O(h^{-c_3}) + O(\sqrt{|\log h|}) = O(\sqrt{|\log h|}),
\end{aligned}$$

where c_3 is a positive constant. Thus, $\mathbb{E}(\Delta_n) = O\left(\sqrt{\frac{|\log h|}{nh^d}}\right)$ and by Markov's inequality

$$\Delta_n = O_P\left(\sqrt{\frac{|\log h|}{nh^d}}\right),$$

■

6.2.6 Remarks

- **Hölder class.** For any number β, let $\lfloor\beta\rfloor$ be the largest integer that is straightly less than β; for instance, $\lfloor 1.7\rfloor = 1$, $\lfloor 2.3\rfloor = 2$, and $\lfloor 2\rfloor = 1$. Let $f^{(m)}$ be the m-the derivative of f. A function f is called in the β-*Hölder* class if

  $$|f^{(\lfloor\beta\rfloor)}(x) - f^{(\lfloor\beta\rfloor)}(y)| \leq L|x-y|^{\beta-\lfloor\beta\rfloor}$$

  for all $x, y \in \mathbb{R}$. It is easy to see that 1-Hölder class is the collection of Lipschitz functions. And for any integer m, m-Hölder class is a larger class than the bounded m-th order derivative class, i.e, the Hölder continuity is weaker than the existence of derivative. Using the bias analysis, one can show that when p is in 2-Hölder class, the bias

  $$\textbf{bias}(\widehat{p}_n(x_0)) = O(h^2).$$

  In fact, different function classes are optimal for different density estimator. I would recommend the following book as a starting point if you are interested in this topic:

  Tsybakov, Alexandre B. Introduction to Nonparametric Estimation. Springer, 2009.

- **Derivative estimation.** The KDE can also be used to estimate the derivative of a density function. For example, when we use the Gaussian kernel, the first derivative $\widehat{p}'_n(x)$ is actually an estimator of the first derivative of true density $p'(x)$. Moreover, any higher order derivative can be estimated by the corresponding derivatives of the KDE. The difference is, however, the MSE error rate will be different. If we consider estimating the ℓ-th derivative, $p^{(\ell)}$, the MISE will be

  $$\textbf{MISE}(\widehat{p}^{(\ell)}) = \mathbb{E}\left(\int (\widehat{p}_n^{(\ell)}(x) - p^{(\ell)}(x))^2\right) = O(h^2) + O\left(\frac{1}{nh^{1+2\ell}}\right)$$

  under suitable conditions. The bias generally remains at a similar rate but the variance is now at rate $O\left(\frac{1}{nh^{1+2\ell}}\right)$. Namely, the variance converges at a slower rate. The optimal MISE for estimating the ℓ-th derivative of p will be

  $$\textbf{MISE}_{\mathsf{opt}}(\widehat{p}^{(\ell)}) = O\left(n^{-\frac{4}{5+2\ell}}\right), \quad h_{\mathsf{opt},\ell} = O(n^{-\frac{1}{5+2\ell}}).$$

♦ **Multivariate density estimation.** In addition to estimating the density function of a univariate random variable, the KDE can be applied to estimate the density function of a multivariate random variable. In this case, we need to use a multivariate kernel function. Generally, a multivariate kernel function can be constructed using a radial basis approach or a product kernel. Assume our data is d-dimensional. Let $\vec{x} = (x_1, \cdots, x_d)$ be the vector of each coordinate. The former uses $c_d \cdot K(\|\vec{x}\|)$ as the kernel function (c_d is a constant depending on d, the dimension of the data). The later uses $K(\vec{x}) = K(x_1)K(x_2)\cdots K(x_d)$ as the kernel function. In multivariate case, the KDE has a slower convergence rate:

$$\textbf{MISE}(\widehat{p}_n) = O(h^4) + O\left(\frac{1}{nh^d}\right) \Longrightarrow \textbf{MISE}_{\mathsf{opt}}(\widehat{p}_n) = O\left(n^{-\frac{4}{4+d}}\right), \quad h_{\mathsf{opt}} = O\left(n^{-\frac{1}{4+d}}\right).$$

Here you see that when d is large, the optimal convergence rate is very very slow. This means that we cannot estimate the density very well using the KDE when the dimension d of the data is large, a phenomenon known as the *curse of dimensionality*[5].

6.3 k-nearest neighbor

k-nearest neighbor (k-NN) is a cool and powerful idea for nonparametric estimation. Today we will talk about its application in density estimation. In the future, we will learn how to use it for regression analysis and classification.

Let $X_1, \cdots, X_n \sim p$ be our random sample. Assume each observation has d different variables; namely, $X_i \in \mathbb{R}^d$. For a given point x, we first rank every observation based on its distance to x. Let $R_k(x)$ denotes the distance from x to its k-th nearest neighbor point.

For a given point x, the kNN density estimator estimates the density by

$$\widehat{p}_{\mathsf{knn}}(x) = \frac{k}{n} \cdot \frac{1}{V_d \cdot R_k^d(x)} = \frac{k}{n} \cdot \frac{1}{\text{Volume of a } d\text{-dimensional ball with radius being } R_k(x)},$$

where $V_d = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$ is the volume of a unit d-dimensional ball and $\Gamma(x)$ is the Gamma function.

Here are the results when $d = 1, 2$, and 3 –

- $d = 1$, $V_1 = 2$: $\widehat{p}_{\mathsf{knn}}(x) = \frac{k}{n}\frac{1}{2R_k(x)}$.
- $d = 2$, $V_1 = \pi$: $\widehat{p}_{\mathsf{knn}}(x) = \frac{k}{n}\frac{1}{\pi R_k^2(x)}$.
- $d = 3$, $V_1 = \frac{4}{3}\pi$: $\widehat{p}_{\mathsf{knn}}(x) = \frac{k}{n}\frac{3}{4\pi R_k^3(x)}$.

What is the intuition of a kNN density estimator? By the definition of $R_k(x)$, the ball centered at x with radius $R_k(x)$

$$B(x, R_k(x)) = \{y : \|x - y\| \le R_k(x)\}$$

satisfies the fact that

$$\frac{k}{n} = \frac{1}{n}\sum_{i=1}^{n} I(X_i \in B(x, R_k(x))).$$

[5] There are many forms of curse of dimensionality; the KDE is just one instance. For other cases, see https://en.wikipedia.org/wiki/Curse_of_dimensionality

Namely, ratio of observations within $B(x, R_k(x))$ is k/n.

Recall from the relation between EDF and CDF, the quantity

$$\frac{1}{n}\sum_{i=1}^{n} I(X_i \in B(x, R_k(x)))$$

can be viewed as an estimator of the quantity

$$P(X_i \in B(x, R_k(x))) \approx \int_{B(x,R_k(x))} p(y)dy.$$

When n is large and k is relatively small compared to n, $R_k(x)$ will be small because the ratio $\frac{k}{n}$ is small. Thus, the density $p(y)$ within the region $B(x, R_k(x))$ will not change too much. Namely, $p(y) \approx p(x)$ for every $y \in B(x, R_k(x))$. Note that $p(x)$ is the center of the ball $B(x, R_k(x))$.

Therefore,

$$P(X_i \in B(x, R_k(x))) \approx \int_{B(x,R_k(x))} p(y)dy \approx p(x) \int_{B(x,R_k(x))} dy = p(x) \cdot V_d \cdot R_k^d(x).$$

This quantity will be the target of the estimator $\frac{1}{n}\sum_{i=1}^{n} I(X_i \in B(x, R_k(x)))$, which equals to $\frac{k}{n}$. As a result, we can say that

$$p(x) \cdot V_d \cdot R_k^d(x) \approx P(X_i \in B(x, R_k(x))) \approx \frac{1}{n}\sum_{i=1}^{n} I(X_i \in B(x, R_k(x))) \approx \frac{k}{n},$$

which leads to

$$p(x) \cdot V_d \cdot R_k^d(x) \approx \frac{k}{n} \Rightarrow p(x) \approx \frac{k}{n}\frac{1}{V_d \cdot R_k^d(x)}$$

This motivates us to use

$$\widehat{p}_{\mathsf{knn}}(x) = \frac{k}{n}\frac{1}{V_d \cdot R_k^d(x)}$$

as a density estimator.

Example. We consider a simple example in $d = 1$. Assume our data is $\mathcal{X} = \{1, 2, 6, 11, 13, 14, 20, 33\}$. What is the kNN density estimator at $x = 5$ with $k = 2$? First, we calculate $R_2(5)$. The distance from $x = 5$ to each data point in $\mathcal{X}$ is

$$\{4, 3, 1, 6, 8, 9, 15, 28\}.$$

Thus, $R_2(5) = 3$ and

$$\widehat{p}_{\mathsf{knn}}(5) = \frac{2}{8}\frac{1}{2 \cdot R_2(5)} = \frac{1}{24}.$$

What will the density estimator be when we choose $k = 5$? In this case, $R_5(5) = 8$ so

$$\widehat{p}_{\mathsf{knn}}(5) = \frac{5}{8}\frac{1}{2 \cdot R_5(5)} = \frac{5}{64}.$$

Now we see that different value of k gives a different density estimate even at the same x. How do we choose k? Well, just as the smoothing bandwidth in the KDE, it is a very difficult problem in practice. However, we can do some theoretical analysis to get a rough idea about how k should be changing with respect to the sample size n.

6.3.1 Asymptotic theory

The asymptotic analysis of a k-NN estimator is quiet complicated so here I only stated its result in $d = 1$. The bias of the k-NN estimator is

$$\textbf{bias}(\widehat{p}_{\mathsf{knn}}(x)) = \mathbb{E}(\widehat{p}_{\mathsf{knn}}(x)) - p(x) = b_1 \frac{p''(x)}{p^2(x)} \left(\frac{k}{n}\right)^2 + b_2 \frac{p(x)}{k} + o\left(\left(\frac{k}{n}\right)^2 + \frac{1}{k}\right),$$

where b_1 and b_2 are two constants. The variance of the k-NN estimator is

$$\mathsf{Var}(\widehat{p}_{\mathsf{knn}}(x)) = v_0 \cdot \frac{p^2(x)}{k} + o\left(\frac{1}{k}\right),$$

where v_0 is a constant. The quantity k is something we can choose. We need $k \to \infty$ when $n \to \infty$ to make sure both bias and variance converge to 0. However, how k diverges affects the quality of estimation. When k is large, the variance is small while the bias is large. When k is small, the variance is large and the bias tends to be small but it could also be large (the second component in the bias will be large). These results are from

> Mack, Y. P., and Murray Rosenblatt. "Multivariate k-nearest neighbor density estimates." *Journal of Multivariate Analysis* 9.1 (1979): 1-15.

To balance the bias and variance, we consider the mean square error, which is at the rate

$$\mathsf{MSE}(\widehat{p}_{\mathsf{knn}}(x)) = O\left(\frac{k^4}{n^4} + \frac{1}{k}\right).$$

This motivates us to choose

$$k = C_0 \cdot n^{\frac{4}{5}}$$

for some constant C_0. This leads to the optimal convergence rate

$$\mathsf{MSE}(\widehat{p}_{\mathsf{knn,opt}}(x)) = O(n^{-\frac{4}{5}})$$

for a k-NN density estimator.

Remark.

- When we consider a d-dimensional data, the bias will be of the rate

  $$\textbf{bias}(\widehat{p}_{\mathsf{knn}}(x)) = O\left(\left(\frac{k}{n}\right)^{\frac{2}{d}} + \frac{1}{k}\right)$$

  and the variance is still at rate

  $$\mathsf{Var}(\widehat{p}_{\mathsf{knn}}(x)) = O\left(\frac{1}{k}\right).$$

  This shows a very different phenomenon compared to the KDE. In KDE, the rate of variance depends on the dimension whereas the bias remains the same. In kNN, the rate of variance stays the same rate but the rate of bias changes with respect to the dimension. One intuition is that no matter what dimension is, the ball $B(x, R_k(x))$ always contain k observations. Thus, the variability of a kNN estimator is caused by k points, which is independent of the dimension. On the other hand, in KDE, the same h in different dimensions will cover different number of observations so the variability changes with respect to the dimension.

- The kNN approach has an advantage that it can be computed very efficiently using kd-tree algorithm[6]. This is a particularly useful feature when we have a huge amount of data and when the dimension of the data is large. However, a downside of the kNN is that the density often has a 'heavy-tail', which implies it may not work well when $|x|$ is very large. Moreover, when $d = 1$, the density estimator $\widehat{p}_{\mathsf{knn}}(x)$ is not even a density function (the integral is infinite!).

If you are interested in the theory of k-NN, I would recommend the following book:

> Devroye, Luc, László Györfi, and Gábor Lugosi. A probabilistic theory of pattern recognition. Vol. 31. Springer Science & Business Media, 2013.

6.4 Basis approach

In this section, we assume that the PDF $p(x)$ is supported on $[0, 1]$. Namely, $p(x) > 0$ only in $[0, 1]$. When the PDF $p(x)$ is smooth (in general, we need p to be squared integrable, i.e., $\int_0^1 p(x)^2 dx < \infty$), we can use an orthonormal basis to approximate this function. This approach has several other names: the basis estimator, projection estimator, and an orthogonal series estimator.

Let $\{\phi_1(x), \phi_2(x), \cdots, \phi_m(x), \cdots\}$ be a set of basis functions. Then we have

$$p(x) = \sum_{j=1}^{\infty} \theta_j \phi_j(x).$$

The quantity θ_j is the coefficient of each basis. In signal process, these quantities are refereed to as the signal.

The collection $\{\phi_1(x), \phi_2(x), \cdots, \phi_m(x), \cdots\}$ is called a basis if its elements have the following property:

- Unit 1:
$$\int_0^1 \phi_j^2(x) dx = 1 \tag{6.11}$$
for every $j = 1, \cdots$.
- Orthonormal:
$$\int_0^1 \phi_j(x)\phi_k(x) dx = 0 \tag{6.12}$$
for every $j \neq k = 1, \cdots$.

Here are some concrete examples of the basis:

- Cosine basis:
$$\phi_1(x) = 1, \quad \phi_j(x) = \sqrt{2}\cos(\pi(j-1)x), j = 2, 3, \cdots,$$
- Trigonometric basis:
$$\phi_1(x) = 1, \quad \phi_{2j}(x) = \sqrt{2}\cos(2\pi jx), \quad \phi_{2j+1}(x) = \sqrt{2}\sin(2\pi jx), j = 1, 2, \cdots$$

[6] https://en.wikipedia.org/wiki/K-d_tree

Often the basis is something we can choose so it is known to us. What is unknown to use is the coefficients $\theta_1, \theta_2, \cdots,$. Thus, the goal is to estimate these coefficients using the random sample $X_1, \cdots, X_n$.

How do we estimate these parameters? We start with some simple analysis. For any basis $\phi_j(x)$, consider the following integral:

$$\begin{aligned}
\mathbb{E}(\phi_j(X_1)) &= \int_0^1 \phi_j(x) dP(x) \\
&= \int_0^1 \phi_j(x) p(x) dx \\
&= \int_0^1 \phi_j(x) \sum_{k=1}^{\infty} \theta_k \phi_k(x) dx \\
&= \sum_{k=1}^{\infty} \theta_k \int_0^1 \underbrace{\phi_j(x)\phi_k(x)dx}_{=0 \text{ except } k=j} \\
&= \sum_{k=1}^{\infty} \theta_k I(k=j) \\
&= \theta_j.
\end{aligned}$$

Namely, the expectation of $\phi_j(X_1)$ is exactly the coefficient θ_j. This motivates us to use the sample average as an estimator:

$$\widehat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} \phi_j(X_i).$$

By construction, this estimator is unbiased, i.e., $\mathbb{E}(\widehat{\theta}_j) - \theta_j = 0$. The variance of this estimator is

$$\begin{aligned}
\mathsf{Var}(\widehat{\theta}_j) &= \frac{1}{n} \mathsf{Var}(\phi_j(X_1)) \\
&= \frac{1}{n} \left(\mathbb{E}(\phi_j^2(X_1)) - \mathbb{E}^2(\phi_j(X_1)) \right) \\
&= \frac{1}{n} (\mathbb{E}(\phi_j^2(X_1)) - \theta_j^2) \\
&= \frac{\sigma_j^2}{n},
\end{aligned}$$

where $\sigma_j^2 = \mathbb{E}(\phi_j^2(X_1)) - \theta_j^2$.

In practice, we cannot use all the basis because there will be infinite number of them being calculated. So we will use only M basis as our estimator. Namely, our estimator is

$$\widehat{p}_{n,M}(x) = \sum_{j=1}^{M} \widehat{\theta}_j \phi_j(x). \tag{6.13}$$

Later in the asymptotic analysis, we will show that we should not choose M to be too large because of the bias-variance tradeoff.

6.4.1 Asymptotic theory

To analyze the quality of our estimator, we use the mean integrated squared error (MISE). Namely, we want to analyze

$$\mathbf{MISE}(\widehat{p}_{n,M}) = \mathbb{E}(\int_0^1 (\widehat{p}_{n,M}(x) - p(x))^2 dx) = \int_0^1 \left[\mathbf{bias}^2(\widehat{p}_{n,M}(x)) + \mathsf{Var}(\widehat{p}_{n,M}(x))\right] dx$$

Theorem 6.6 *Suppose that we are using a cosine basis and $p(x)$ satisfies*

$$\int_0^1 |p''(x)|^2 dx \le L_0. \tag{6.14}$$

Then

$$\int_0^1 \mathbf{bias}^2(\widehat{p}_{n,M}(x))dx = O(M^{-4}), \quad \int \mathsf{Var}(\widehat{p}_{n,M}(x))dx = O\left(\frac{M}{n}\right).$$

Proof: Analysis of Bias. Because each $\widehat{\theta}_j$ is an unbiased estimator of θ_j, we have

$$\mathbb{E}(\widehat{p}_{n,M}(x)) = \mathbb{E}(\sum_{j=1}^M \widehat{\theta}_j \phi_j(x)) = \sum_{j=1}^M \mathbb{E}(\widehat{\theta}_j)\phi_j(x) = \sum_{j=1}^M \theta_j \phi_j(x).$$

Thus, the bias at point x is

$$\mathbf{bias}(\widehat{p}_{n,M}(x)) = \mathbb{E}(\widehat{p}_{n,M}(x)) - p(x) = \sum_{j=1}^M \theta_j\phi_j(x) - \sum_{j=1}^\infty \theta_j\phi_j(x) = -\sum_{j=M+1}^\infty \theta_j\phi_j(x).$$

Thus, the integrated squared bias is

$$\begin{aligned}
\int_0^1 \mathbf{bias}^2(\widehat{p}_{n,M}(x))dx &= \int_0^1 \left(-\sum_{j=M+1}^\infty \theta_j\phi_j(x)\right)^2 dx \\
&= \int_0^1 \left(\sum_{j=M+1}^\infty \theta_j\phi_j(x)\right)\left(\sum_{k=M+1}^\infty \theta_k\phi_k(x)\right) dx \\
&= \sum_{j=M+1}^\infty \sum_{k=M+1}^\infty \theta_j\theta_k \underbrace{\int_0^1 \phi_j(x)\phi_k(x)dx}_{=I(j=k)} \\
&= \sum_{j=M+1}^\infty \theta_j^2.
\end{aligned} \tag{6.15}$$

Namely, the bias is determined by the *signal strength* of the ignored basis, which makes sense because the bias should be reflecting the fact that we are not using all the basis and if there are some important basis (the ones with large $|\theta_j|$) being ignored, the bias ought to be large.

We know that in KDE and kNN, the bias is often associated with the smoothness of the density function. How does the smoothness comes into play in this case? It turns out that if the density is smooth, the remaining signals $\sum_{j=M+1}^\infty \theta_j^2$ will also be small. To see this, we consider a very simple model by assuming equation (6.14). This assumption implies that the overall curvature of the density function is bounded.

Using the fact that for cosine basis function $\phi_j(x)$,

$$\phi_j'(x) = -\sqrt{2}\pi(j-1)\sin(\pi(j-1)x),$$
$$\phi_j''(x) = -\sqrt{2}\pi^2(j-1)^2\cos(\pi(j-1)x) = -\pi^2(j-1)^2\phi_j(x).$$

Thus, equation (6.14) implies

$$\begin{aligned}
L_0 &\geq \int_0^1 |p''(x)|^2 dx \\
&= \int_0^1 |\sum_{j=1}^{\infty} \theta_j \phi_j''(x)|^2 dx \\
&= \int_0^1 |\sum_{j=1}^{\infty} \pi^2(j-1)^2\theta_j\phi_j(x)|^2 dx \\
&= \int_0^1 \left(\sum_{j=1}^{\infty} \pi^2(j-1)^2\theta_j\phi_j(x)\right)\left(\sum_{k=1}^{\infty} \pi^2(k-1)^2\theta_k\phi_k(x)\right) dx \\
&= \pi^4 \sum_{j=1}^{\infty}\sum_{k=1}^{\infty}(j-1)^2(k-1)^2\theta_j\theta_k \underbrace{\int_0^1 \phi_j(x)\phi_k(x)dx}_{=I(j=k)} \\
&= \pi^4 \sum_{j=1}^{\infty}(j-1)^4\theta_j^2.
\end{aligned}$$

Namely, equation (6.14) implies

$$\sum_{j=1}^{\infty}(j-1)^4\theta_j^2 \leq \frac{L_0}{\pi^4}. \tag{6.16}$$

This further implies that

$$\sum_{j=M+1}^{\infty} \theta_j^2 = O(M^{-4}). \tag{6.17}$$

An intuitive explanation is as follows. Equation (6.16) implies that when $j \to \infty$, the signal $\theta_j^2 = O(j^{-5})$; namely, the signal θ_j^2 cannot converge to 0 at a slower rate than j^{-5}. The reason is: if $\theta_j^2 \approx \frac{1}{j^{-5}}$, equation (6.16) becomes $\sum_{j=1}^{\infty}(j-1)^4\theta_j^2 \approx \sum_{j=1}^{\infty}(j-1)^4\frac{1}{j^{-5}} \approx \sum_{j=1}^{\infty}\frac{1}{j} \to \infty$! Thus, the *tail* signals have to be shrinking toward 0 at least of rate $O(j^{-5})$. As a result, $\sum_{j=M+1}^{\infty}\theta_j^2 \approx \sum_{j=M+1}^{\infty}O(j^{-5}) = O(M^{-4})$.

Equation (6.17) and (6.15) together imply that the bias is at rate

$$\int_0^1 \mathbf{bias}^2(\widehat{p}_{n,M}(x))dx = O(M^{-4}).$$

Analysis of Variance. Now we turn to the analysis of variance.

$$\begin{aligned}
\int_0^1 \mathsf{Var}(\widehat{p}_{n,M}(x))dx &= \int_0^1 \mathsf{Var}\left(\sum_{j=1}^M \widehat{\theta}_j \phi_j(x)\right) dx \\
&= \int_0^1 \left(\sum_{j=1}^M \phi_j^2(x)\mathsf{Var}(\widehat{\theta}_j) + \sum_{j\neq k=1}^M \phi_j(x)\phi_k(x)\mathsf{Cov}(\widehat{\theta}_j, \widehat{\theta}_k)\right) dx \\
&= \sum_{j=1}^M \mathsf{Var}(\widehat{\theta}_j) \underbrace{\int_0^1 \phi_j^2(x)dx}_{=1} + \sum_{j\neq k=1}^M \mathsf{Cov}(\widehat{\theta}_j, \widehat{\theta}_k) \underbrace{\int_0^1 \phi_j(x)\phi_k(x)dx}_{=0} \\
&= \sum_{j=1}^M \mathsf{Var}(\widehat{\theta}_j) \\
&= \sum_{j=1}^M \frac{\sigma_j^2}{n} \\
&= O\left(\frac{M}{n}\right).
\end{aligned}$$

■

Putting both bias and variance together, we obtain the rate of the MISE

$$\mathbf{MISE}(\widehat{p}_{n,M}) = \int_0^1 \left[\mathbf{bias}^2(\widehat{p}_{n,M}(x)) + \mathsf{Var}(\widehat{p}_{n,M}(x))\right] dx = O\left(\frac{1}{M^4}\right) + O\left(\frac{M}{n}\right).$$

Thus, the optimal choice is $M = M^* = C_0 n^{1/5}$ for some positive constant C). And this leads to

$$\mathbf{MISE}(\widehat{p}_{n,M^*}) = O(n^{-4/5}),$$

which is the same rate as the KDE and kNN.

Remark.

- **(Sobolev space)** Again, we see that the bias is dependent on the smoothness of the density function. Here, as long as the density function has an overall curvature (second derivative) being bounded, we have an optimal MISE at the rate of $O(n^{-4/5})$. In fact, if the density function has stronger smoothness, such as the overall third derivatives being bounded, we will have an even faster convergence rate $O(n^{-6/7})$. Now for any positive integer β and a positive number $L > 0$, we define

  $$W(\beta, L) = \left\{p : \int_0^1 |p^{(\beta)}(x)|^2 dx \leq L < \infty\right\}$$

  be a collection of smooth density functions. Then the bias between a basis estimator and any density $p \in W(\beta, L)$ is at the rate $O(M^{-2\beta})$ and the optimal convergence rate is $O(n^{-\frac{2\beta}{2\beta+1}})$. The collection $W(\beta, L)$ is a space of functions and is known as the Sobolev Space.

- **(Tuning parameter)** The number of basis M in a basis estimator, the number of neighborhood k in the kNN approach, and the smoothing bandwidth h in the KDE all play a very similar role in bias-variance tradeoff. These quantities are called *tuning parameters* in statistics and machine learning. Just like what we have seen in the analysis of the KDE, choosing these parameters is often a very

difficult task because the optimal choice depends on the actual density function, which is unknown to us. A principle approach to choosing the tuning parameters is based on minimizing an error estimator. For instance, in the basis estimator, we try to find an estimator for the MISE, $\widehat{\mathbf{MISE}}(\widehat{p}_{n,M})$. When we change value of M, this error estimator will also change. We then choose M by minimizing the error, i.e.,

$$\widehat{M}^* = \mathsf{argmin}_M \widehat{\mathbf{MISE}}(\widehat{p}_{n,M}).$$

Then we pick $\widehat{M}^*$ basis and construct our estimator.