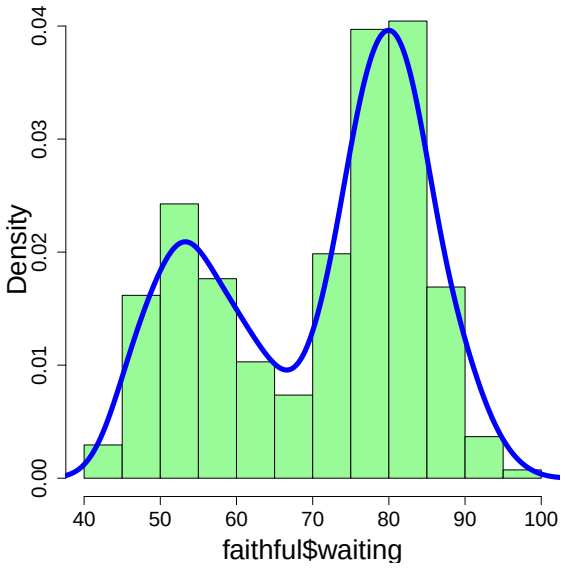
STAT/Q SCI 403: Introduction to Resampling Methods **Spring 2018**

Lecture 7: Density Estimation

Instructor: Yen-Chi Chen

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.

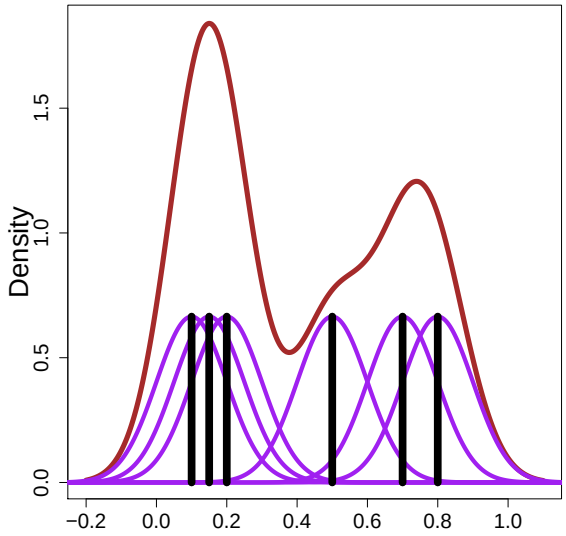
A classical approach of density estimation is the *histogram*. Here we will talk about another approach–the *kernel density estimator* (KDE; sometimes called kernel density estimation). The KDE is one of the most famous method for density estimation. The follow picture shows the KDE and the histogram 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

$$\widehat{p}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right), \tag{7.1}$$

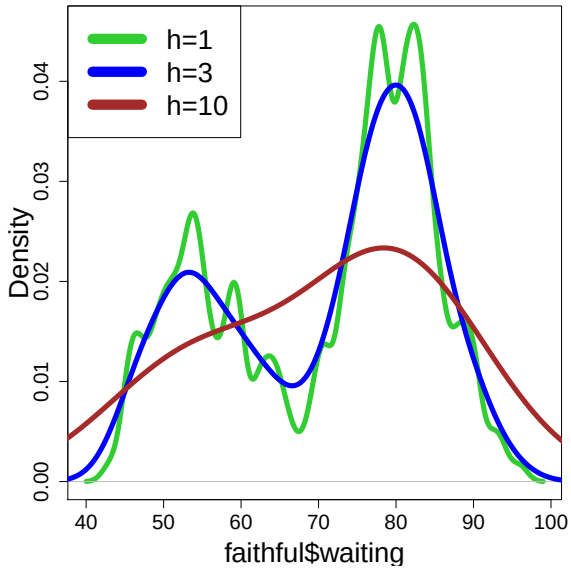
where $K(x)$ is called the *kernel function* that is generally a smooth, symmetric function such as a Gaussian and $h > 0$ is called the smoothing bandwidth that controls the amount of smoothing. Basically, the KDE smoothes each data point X_i into a small density bumps and then sum all these small bumps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



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.

7.1 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:

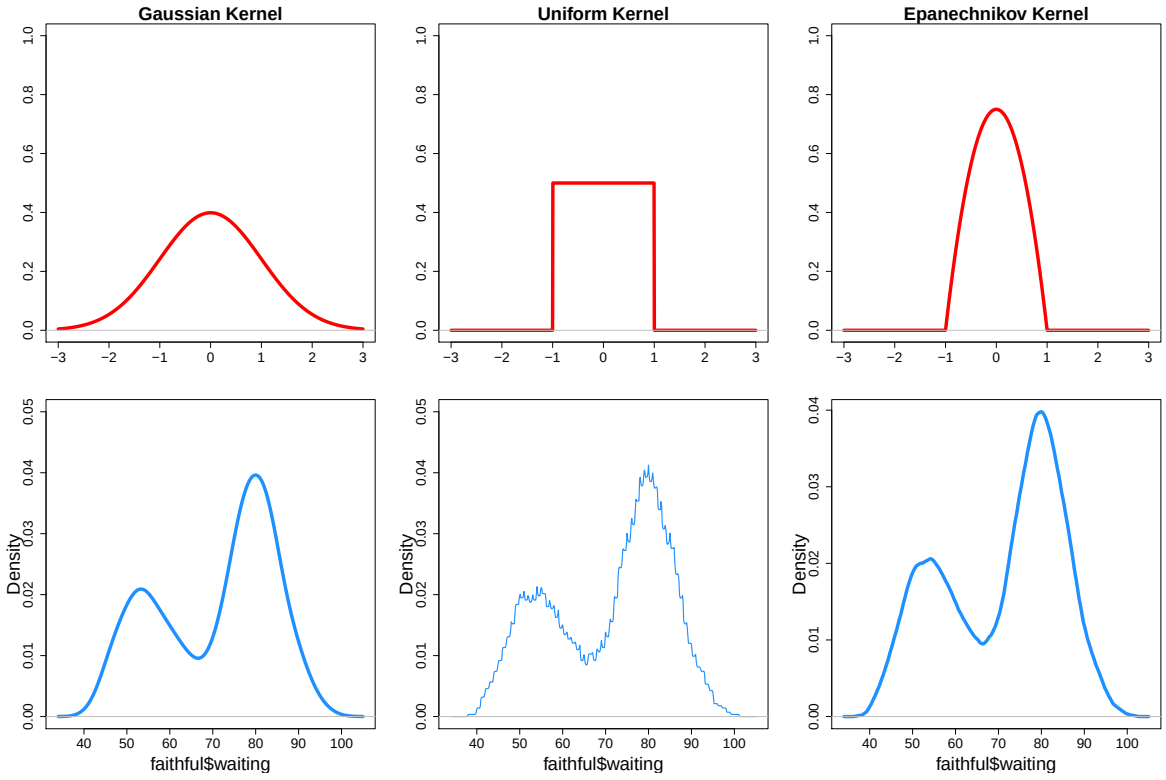
1. $K(x)$ is symmetric.

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

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

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 [1].

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 esti-

[1] 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.

mators. 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).

7.2 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 $\widehat{p}_n(x_0)$.

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

$$\textbf{bias}(\widehat{p}_n(x_0)) = \frac{1}{2}h^2p''(x_0)\mu_K + o(h^2). \tag{7.2}$$

This means that when we allow $h \to 0$, the bias is shrinking at a rate $O(h^2)$. Equation (7.2) 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.

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

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

where $\sigma_K^2 = \int K^2(y)dy$. Therefore, 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!

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

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

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

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

Remarks.

- 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)*:

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

  Under a similar derivation, one can show that

  $$\begin{aligned}\mathbf{MISE}(\widehat{p}_n) &= \frac{1}{4}h^4\int|p''(x)|^2dx\mu_K^2 + \frac{1}{nh}\underbrace{\int p(x)dx}_{=1}\sigma_K^2 + o(h^4) + o\left(\frac{1}{nh}\right) \\ &= \frac{\mu_K^2}{4}\cdot h^4\cdot\underbrace{\int|p''(x)|^2dx}_{\text{Overall curvature}} + \frac{\sigma_K^2}{nh} + o(h^4) + o\left(\frac{1}{nh}\right) \qquad (7.3) \\ &= O(h^4) + O\left(\frac{1}{nh}\right).\end{aligned}$$

- The two dominating terms in equation (7.3), $\frac{\mu_K^2}{4}\cdot h^4\cdot\underbrace{\int|p''(x)|^2dx}_{\text{Overall curvature}} + \frac{\sigma_K^2}{nh}$, is called the *asymptotical mean integrated square error (AMISE)*. The optimal smoothing bandwidth is often chosen by minimizing this quantity. Namely,

  $$h_{\mathsf{opt}} = \left(\frac{1}{n}\cdot\frac{4}{\int|p''(x)|^2dx}\cdot\frac{\sigma_K^2}{\mu_K^2}\right)^{\frac{1}{5}} = C_2\cdot n^{-\frac{1}{5}}. \qquad (7.4)$$

- However, the optimal smoothing bandwidth h_{opt} cannot be used 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*[2]. 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.

7.3 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 (7.1),

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

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

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

CI using the bootstrap variance. We can construct a confidence interval using the bootstrap as well. The procedure is simple. We sample with replacement from the original data points to obtain a new bootstrap sample. Using the new bootstrap sample, we construct a bootstrap KDE. Assume we repeat the bootstrap process B times, leading to B bootstrap KDE, denoted as

$$\widehat{p}_n^{*(1)}, \cdots, \widehat{p}_n^{*(B)}.$$

Then we can estimate the variance of $\widehat{p}_n(x_0)$ by the sample variance of the bootstrap KDE:

$$\widehat{\mathsf{Var}}_B(\widehat{p}_n(x_0)) = \frac{1}{B-1}\sum_{\ell=1}^{B}\left(\widehat{p}_n^{*(1)}(x_0) - \bar{\widehat{p}}_{n,B}^{*}(x_0)\right)^2, \quad \bar{\widehat{p}}_{n,B}^{*}(x_0) = \frac{1}{B}\sum_{\ell=1}^{B}\widehat{p}_n^{*(\ell)}(x_0).$$

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

And a $1-\alpha$ CI will be

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

CI using the bootstrap quantile. In addition to the above approach, we can construct a $1-\alpha$ CI of $p(x_0)$ using the quantile. Let

$$Q_\alpha = \alpha - \text{th quantile of } \left\{\widehat{p}_n^{*(1)}(x_0), \cdots, \widehat{p}_n^{*(B)}(x_0)\right\}.$$

A $1-\alpha$ CI of $p(x_0)$ is

$$[Q_{\alpha/2}, Q_{1-\alpha/2}].$$

This approach is generally called the bootstrap quantile approach. An advantage of this approach is that it does not requires asymptotic normality (i.e., it may work even if the CLT does not work). The bootstrap quantile approach works under a weaker assumption than the previous two methods.

Note that a more formal way of using the bootstrap quantile to construct a CI is as follows. Let

$$S_\alpha = \alpha - \text{th quantile of } \left\{|\widehat{p}_n^{*(1)}(x_0) - \widehat{p}_n(x_0)|, \cdots, |\widehat{p}_n^{*(B)}(x_0) - \widehat{p}_n(x_0)|\right\}.$$

Then an $1-\alpha$ CI of $p(x_0)$ is

$$\widehat{p}_n(x_0) \pm S_{1-\alpha}.$$

The idea of this CI is to use the bootstrap to estimate $|\widehat{p}_n(x_0) - p(x_0)|$ and construct a CI accordingly.

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

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

7.4 ♦ Advanced Topics

♦ **Other error measurements.** In the above analysis, we focus on the MISE, which is related to the L_2 error of estimating a function. In many theoretical analysis, researchers are often interested in L_q errors:

$$\left(\int |\widehat{p}_n(x) - p(x)|^q \, dx \right)^{1/q}.$$

In particular, the L_∞ error, $\sup_x |\widehat{p}_n(x) - p(x)|$, is often used in many theoretical analysis. And the L_∞ error has an interesting error rate:

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

where the O_P notation is a similar notation as O but is used for a random quantity. The extra $\sqrt{\log n}$ term has many interesting stories and it comes from the *empirical process theory*.

♦ **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

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

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

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

♦ **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:

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

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*[3].

[3] 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