Computer Intensive Statistics and Applications
Chapter 5: Density Estimation

Shaobo Jin

Department of Mathematics

Problem

Suppose that we have observed an iid sample X_1, ..., X_n. We want to the density of X, denoted by $f(x)$, using the observed data.

1. Parametric approach: the distribution belongs to a known distribution family with unknown parameters.
2. Nonparametric approach: the unknown density satisfies

$$f \in \mathcal{F} = \left\{ f : \mathbb{R} \to \mathbb{R}, f(x) \geq 0, \int f(x)\, dx = 1, f \text{ is continuous} \right\}.$$

Estimating the density allows us to visualize the data and also provides a starting point for regression analysis.

Nonexistence of Unbiased Estimator

Theorem (Theorem 5.1)

Suppose that $(X_1, \cdots, X_n)$ *is an iid sample of the random variable* X *with unknown density* f*. Consider* $\hat{f}(x) = \hat{f}(x, X_1, ..., X_n)$*. Then, there exists no* $\hat{f} \in \mathcal{G}$ *such that* $E\left[\hat{f}(x)\right] = f(x)$*, for all* $x \in \mathbb{R}$ *and for all* $f \in \mathcal{F}$*, where*

$$\begin{aligned} \mathcal{G} &= \Big\{\hat{f} : \mathbb{R} \to \mathbb{R}, \hat{f}(x) \geq 0, \int \hat{f}(x)\, dx = 1, \hat{f}(x) \text{ is continuous in } x \\ &\qquad \text{and measurable in } X_1, ..., X_n\Big\}. \end{aligned}$$

Mean Squared Error

For a fixed x, the mean squared error between our estimator $\hat{f}(x)$ and the true density $f(x)$ is

$$
\begin{aligned}
\operatorname{MSE}\left(\hat{f}(x)\right) &= \mathrm{E}\left[\left(\hat{f}(x) - f(x)\right)^2\right] \\
&= \operatorname{Var}\left[\hat{f}(x)\right] + \left\{\mathrm{E}\left[\hat{f}(x)\right] - f(x)\right\}^2.
\end{aligned}
$$

Theorem 5.1 implies that the bias term cannot be zero. Hence, we want to find a balance between the variance and the bias.

1. When the estimator has a large variance, it is called undersmoothing.
2. When the bias becomes too large, it is called oversmoothing.

Under- and Oversmoothing: Example

Example

Suppose that, for a fixed x, we approximate the density by

$$f(x) \approx \frac{\mathrm{P}(x-\epsilon \leq X \leq x+\epsilon)}{2\epsilon} \equiv \frac{p}{2\epsilon}.$$

We can estimate p by $n^{-1}\sum_{i=1}^{n} 1(x-\epsilon \leq x_i \leq x+\epsilon)$. Then,

$$\begin{aligned} \mathrm{E}\left[\hat{f}(x)\right] &= \mathrm{E}\left[\frac{\sum_{i=1}^{n} 1(x-\epsilon \leq x_i \leq x+\epsilon)}{2\epsilon n}\right] = \frac{p}{2\epsilon}, \\ \mathrm{Var}\left[\hat{f}(x)\right] &= \mathrm{Var}\left[\frac{\sum_{i=1}^{n} 1(x-\epsilon \leq x_i \leq x+\epsilon)}{2\epsilon n}\right] = \frac{p(1-p)}{4\epsilon^2 n}. \end{aligned}$$

- If ϵ is small, the bias is small and variance is large.
- If ϵ is large, the bias is large and variance is small.

Histogram

Histogram is a standard method to approximate/estimate density.

- Consider a bin with bandwidth $h > 0$ as

$$B_j = [x_0 + (j-1)\,h,\ x_0 + jh],$$

where j is an integer (positive or negative), and x_0 is the origin of the histogram.

- For $x \in B_j$, we approximate $f(x)$ by $h^{-1}\mathrm{P}(X \in B_j)$, and estimate $\mathrm{P}(X \in B_j)$ by $n^{-1}\sum_{i=1}^{n} 1(X_i \in B_j)$.
- The histogram is given by

$$\begin{aligned}\hat{f}_h(x) &= \sum_j 1(x \in B_j)\frac{1}{h}\left[\sum_{i=1}^{n}\frac{1}{n}1(X_i \in B_j)\right].\\ &= \frac{1}{nh}\sum_{i=1}^{n}\sum_j 1(X_i \in B_j)\,1(x \in B_j).\end{aligned}$$

Histogram: Bias and Variance

It is easy to show $\hat{f}_h(x) \geq 0$ and

$$\int \hat{f}_h(x)\,dx = 1.$$

Consider a fixed x. Then, if $x \in B_j$,

$$\begin{aligned} \mathrm{E}\left[\hat{f}(x)\right] &= \frac{\mathrm{P}(X \in B_j)}{h}, \\ \mathrm{Var}\left[\hat{f}(x)\right] &= \frac{\mathrm{P}(X \in B_j)\left[1 - \mathrm{P}(X \in B_j)\right]}{nh^2}. \end{aligned}$$

Hence, histogram is a biased estimator unless $f(x)$ is constant over B_j. The bias depends on h.

Histogram: Taylor Expansion

Suppose that the true density $f(x)$ is smooth enough. Let b_j be the midpoint of B_j. Taylor expansion for $x \in B_j$ yields

$$f(x) = f(b_j) + f'(b_j)(x - b_j) + o(h), \qquad \text{as } h \to 0.$$

Hence, skipping proof,

$$\begin{aligned} \mathrm{E}\left[\hat{f}(x)\right] - f(x) &= f'(b_j)(b_j - x) + o(h), \\ \mathrm{Var}\left[\hat{f}(x)\right] &= \frac{1}{nh} f(x) + o\left(\frac{1}{nh}\right), \end{aligned}$$

as $h \to 0$ and $nh \to \infty$.

Mean Squared Error

For a fixed x, the mean squared error satisfies

$$\begin{aligned} \mathrm{MSE}\left(\hat{f}(x)\right) &= \mathrm{Var}\left[\hat{f}(x)\right] + \left\{\mathrm{E}\left[\hat{f}(x)\right] - f(x)\right\}^2 \\ &= \frac{1}{nh} f(x) + \left[f'(b_j)\right]^2 (b_j - x)^2 + o\left(\frac{1}{nh}\right) + o\left(h^2\right). \end{aligned}$$

As $h \to 0$ and $nh \to \infty$, $\mathrm{MSE}\left(\hat{f}(x)\right) \to 0$. Hence, $\hat{f}(x)$ is a consistent estimator of $f(x)$.

Mean Integrated Squared Error

The mean integrated squared error (MISE) for the goodness of estimation:

$$\mathrm{MISE}\left(\hat{f}\right) = \mathrm{E}\left[\int \left[\hat{f}(x) - f(x)\right]^2 dx\right].$$

The MISE of the histogram satisfies

$$\mathrm{MISE}\left(\hat{f}\right) = \frac{1}{nh} + \frac{h^2}{12}\left\|f'(x)\right\|_2^2 + o\left(\frac{1}{nh}\right) + o\left(h^2\right),$$

where

$$\left\|f'(x)\right\|_2 = \sqrt{\int \left[f'(x)\right]^2 dx}.$$

Optimal Binwidth

The leading term of MISE is called the asymptotic MISE:

$$\mathrm{AMISE}\left(\hat{f}\right) = \frac{1}{nh} + \frac{h^2}{12}\int\left[f'(x)\right]^2 dx.$$

We can minimize AMISE to obtain the optimal binwidth.

- Minimizing $\mathrm{AMISE}\left(\hat{f}\right)$ as a function in h yields

$$h_0 = \left[\frac{6}{n\int\left[f'(x)\right]^2 dx}\right]^{1/3},$$

  the same order as $n^{-1/3}$.
- With the optimal h_0,

$$\mathrm{AMISE}\left(\hat{f}\right) = \frac{1}{nh} + \frac{h^2}{12}\int\left[f'(x)\right]^2 dx.$$

- However, a dilemma is that we do not know $f'(x)$.

Specify Break Points

Suppose that the jth bin is B_j with width h_j. The bandwidth can vary across bins.

- For $x \in B_j$, the density can be estimated by

$$\frac{\sum_{i=1}^{n} 1\left(X_i \in B_j\right)}{n h_j}.$$

- Hence,

$$\hat{f}\left(x\right) = \sum_{j} 1\left(x \in B_j\right) \frac{\sum_{i=1}^{n} 1\left(X_i \in B_j\right)}{n h_j}.$$

Modifying Histogram

The bins in the histogram

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n}\sum_{j} 1(X_i \in B_j)\, 1(x \in B_j)$$

are not adapted to x.

- Choose bins first and then check which bin x belongs to.

One alternative is to define bins according to the x of interest.

Alternative

Let X be a random variable with density f. For a fixed x, define $B_h(x) = \left[x - 2^{-1}h, x + 2^{-1}h\right]$. We can approximate the density by

$$
\begin{aligned}
f(x) &\approx \frac{P(X \in B_h(x))}{h} = \frac{1}{h} \int_{x-2^{-1}h}^{x+2^{-1}h} f(u)\, du \\
&= \int \frac{1}{h} f(u) K\left(\frac{x-u}{h}\right) du = \mathrm{E}\left[\frac{1}{h} K\left(\frac{x-X}{h}\right)\right],
\end{aligned}
$$

where $K(\cdot)$ is the density of Uniform $[-0.5, 0.5]$.

- The corresponding density estimator is

$$
\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).
$$

Kernel Density Estimation

In general, we can follow the above idea and estimate the density by

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where $K()$ is a suitably chosen function.

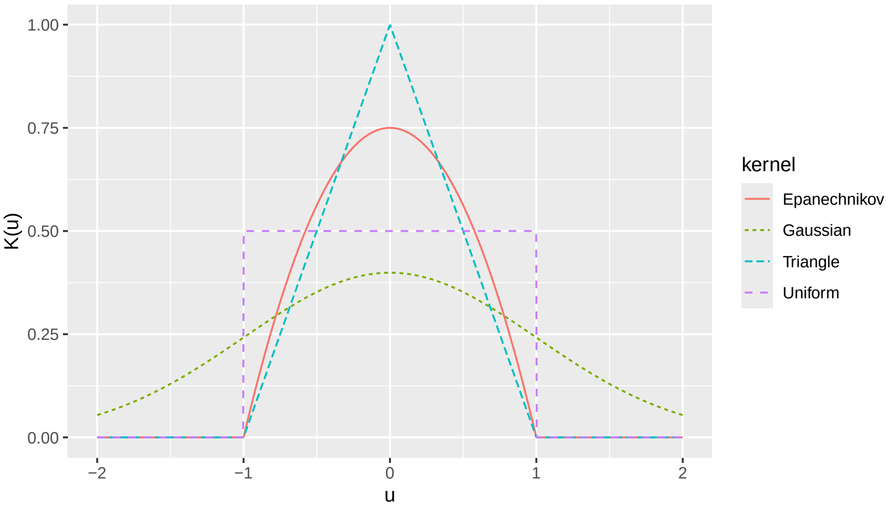
Definition

A non-negative function $K(x)$ is a kernel function if $\int K(x)\,dx = 1$ and $K(x) = K(-x)$. Such density estimator using a kernel function is called a kernel density estimator and such h is called bandwidth.

Kernel Function: Examples

$$
\begin{aligned}
\text{Uniform}: &\quad K(u) = \frac{1}{2} I(|u| \leq 1). \\
\text{Triangle}: &\quad K(u) = (1 - |u|) I(|u| \leq 1). \\
\text{Epanechnikov}: &\quad K(u) = \frac{3}{4}\left(1 - u^2\right) I(|u| \leq 1). \\
\text{Quartic (Biweight)}: &\quad K(u) = \frac{15}{16}\left(1 - u^2\right)^2 I(|u| \leq 1). \\
\text{Triweight}: &\quad K(u) = \frac{35}{32}\left(1 - u^2\right)^3 I(|u| \leq 1). \\
\text{Gaussian}: &\quad K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right). \\
\text{Cosine}: &\quad K(u) = \frac{\pi}{4} \cos\left(\frac{\pi u}{2}\right) I(|u| \leq 1).
\end{aligned}
$$

Kernel Function: Example



Kernel Density: Bias and Variance

Suppose that the true density $f(x)$ is smooth enough. Taylor expansion yields

$$f(x+ht) = f(x) + f'(x)\,ht + \frac{1}{2}f''(x)\,h^2t^2 + o\left(h^2\right), \quad \text{as } h \to 0.$$

Hence, we can show that

$$\begin{aligned} \mathrm{E}\left[\hat{f}_h(x)\right] - f(x) &= \frac{1}{2}h^2 f''(x) \int t^2 K(t)\,dt + o(h), \\ \mathrm{Var}\left[\hat{f}_h(x)\right] &= \frac{1}{nh} f(x) \int K^2(t)\,dt + o\left(\frac{1}{nh}\right), \end{aligned}$$

as $h \to 0$ and $nh \to \infty$.

- A small h yields a small bias but a large variance.
- A large h yields a large bias but a small variance.

MSE

For a fixed x, the mean squared error satisfies

$$\begin{aligned}\operatorname{MSE}\left(\hat{f}_h(x)\right) &= \frac{1}{nh}\|K\|_2^2 f(x) + \frac{1}{4}h^4\mu_2^2(K)\left[f''(x)\right]^2 \\ &\quad + o\left(\frac{1}{nh}\right) + o\left(h^4\right),\end{aligned}$$

where

$$\|K\|_2^2 = \int K^2(t)\,dt \qquad \mu_2(K) = \int t^2 K(t)\,dt.$$

As $h \to 0$ and $nh \to \infty$, $\operatorname{MSE}\left(\hat{f}(x)\right) \to 0$. Hence, $\hat{f}(x)$ is a consistent estimator of $f(x)$.

Optimal Binwidth

The MISE satisfies

$$\text{MISE}\left(\hat{f}_h\right) = \underbrace{\frac{1}{nh}\|K\|_2^2 + \frac{1}{4}h^4\mu_2^2(K)\|f''(x)\|_2^2}_{=\text{AMISE}\left(\hat{f}_h\right)} + o\left(\frac{1}{nh}\right) + o\left(h^4\right).$$

We can minimize AMISE to obtain the optimal binwidth.

- Minimizing AMISE $\left(\hat{f}\right)$ as a function in h yields

$$h_0 = \left[\frac{\|K\|_2^2}{n\mu_2^2(K)\|f''(x)\|_2^2}\right]^{1/5},$$

  the same order as $n^{-1/5}$.
- However, a dilemma is that we do not know $f''(x)$.

Rate of Convergence

- With the optimal h_0, the AMISE of histogram satisfies

$$\begin{aligned}\text{AMISE}\left(\hat{f}_h\right) &= \frac{1}{nh_0} + \frac{h_0^2}{12}\left\|f'(x)\right\|_2^2 \\ &= O\left(n^{-2/3}\right).\end{aligned}$$

- With the optimal h_0, the AMISE of kernel density estimation satisfies

$$\begin{aligned}\text{AMISE}\left(\hat{f}_h\right) &= \frac{1}{nh}\left\|K\right\|_2^2 + \frac{1}{4}h^4\mu_2^2(K)\left\|f''(x)\right\|_2^2 \\ &= \frac{5\left\|f''(x)\right\|_2^{2/5}}{4n^{4/5}}\mu_2^{2/5}(K)\left\|K\right\|_2^{8/5} = O\left(n^{-4/5}\right),\end{aligned}$$

  converging faster than histogram.

Choice of Kernel and Bandwidth

- In theory, the optimal nonnegative kernel that minimizes AMISE $\text{AMISE}\left(\hat{f}_{h_0}\right)$ is the Epanechnikov kernel.
  - However, it is not practical, since the optimal bandwidth

$$h_0 = \left[\frac{\|K\|_2^2}{n\mu_2^2(K)\|f''(x)\|_2^2}\right]^{1/5},$$

    depends on unknown $f''(x)$.
- In practice, the effect of kernel is often minor, relative to the choice of the bandwidth.

Choice of Bandwidth: Rule-of-Thumb

Silverman's rule-of-thumb (plug-in method): Regardless of the true f, we use a Gaussian kernel and the density of $N\left(0, \sigma^2\right)$ as the true f. Then,

$$h_0 = \left[\frac{\|K\|_2^2}{n\mu_2^2(K)\|f''(x)\|_2^2}\right]^{1/5} = \left[\frac{4}{3}\right]^{1/5}\hat{\sigma}n^{-1/5},$$

where $\hat{\sigma}^2$ is the sample variance.

- It works well if the true density is not too far from normal.

But estimating σ^2 by sample variance is sensitive to outliers. A more robust estimator is based on the interquartile range. For $X \sim N\left(\mu, \sigma^2\right)$, the difference between the 75%-quantile and the 25%-quantile is 1.34σ. The bandwidth becomes

$$h_0 = \left[\frac{4}{3}\right]^{1/5}\min\left\{\hat{\sigma}, \frac{R}{1.34}\right\}n^{-1/5},$$

where R is the interquartile range.

Choice of Bandwidth: Cross Validation

Consider the integrated squared error (ISE) for the goodness of estimation:

$$\begin{aligned} \text{ISE}\left(\hat{f}\right) &= \int \left[\hat{f}\left(x\right) - f\left(x\right)\right]^2 dx \\ &= \int \hat{f}_h^2\left(x\right) dx - 2\int \hat{f}_h\left(x\right) f\left(x\right) dx + \int f^2\left(x\right) dx. \end{aligned}$$

We can estimate the second integral by $n^{-1}\sum_{i=1}^{n} \hat{f}_{h,-i}\left(X_i\right)$, where

$$\hat{f}_{h,-i}\left(x\right) = \frac{1}{\left(n-1\right)h}\sum_{j\neq i} K\left(\frac{x - X_j}{h}\right).$$

The leave-one-out cross validation criterion minimizes

$$\text{CV}\left(h\right) = \int \hat{f}_h^2\left(x\right) dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{h,-i}\left(X_i\right).$$

Limiting Distribution

Recall that the leading terms of bias and variance are

$$\begin{aligned} \mathrm{E}\left[\hat{f}_h(x)\right] - f(x) &\approx \frac{1}{2}h^2 f''(x)\,\mu_2(K), \\ \mathrm{Var}\left[\hat{f}_h(x)\right] &\approx \frac{1}{nh} f(x)\,\|K\|_2^2. \end{aligned}$$

For a fixed x, if the bandwidth satisfies $h = cn^{-1/5}$ for a constant c, we would expect

$$n^{2/5}\left\{\hat{f}_h(x) - f(x)\right\} \xrightarrow{d} N\left(b(x),\ v^2(x)\right),$$

where

$$\begin{aligned} b(x) &= \frac{c^2}{2} f''(x)\,\mu_2(K), \\ v^2(x) &= \frac{1}{c} f(x)\,\|K\|_2^2. \end{aligned}$$

Pointwise Confidence Band

Let λ_α be the α quantile of $N(0,1)$. Then,

$$1-\alpha \approx \mathrm{P}\left(-\lambda_{1-\alpha/2} \leq \frac{n^{2/5}\left\{\hat{f}_h(x) - f(x)\right\} - b(x)}{v(x)} \leq \lambda_{1-\alpha/2}\right).$$

An asymptotic interval for $f(x)$ is

$$\hat{f}_h(x) - \frac{h^2}{2} f''(x)\mu_2(K) \pm \lambda_{1-\alpha/2}\sqrt{\frac{1}{hn} f(x)\|K\|_2^2},$$

which depends on unknown $f''(x)$.

- One ad-hoc way is to ignore the bias term $\frac{h^2}{2} f''(x)\mu_2(K)$.
- An alternative is to estimate $f''(x)$, e.g., use the derivative of the kernel density estimator.

Pointwise Confidence Band

However, the central limit theorem implies that

$$\sqrt{n}\left\{\hat{f}_h(x) - \mathrm{E}\left[\hat{f}_h(x)\right]\right\} \overset{d}{\to} N\left(0,\ \mathrm{Var}\left[\frac{1}{h}K\left(\frac{x-X}{h}\right)\right]\right).$$

We can show that

$$\mathrm{Var}\left[\frac{1}{h}K\left(\frac{x-X}{h}\right)\right] \approx \frac{1}{h}f(x)\|K\|_2^2.$$

Hence, an asymptotic interval for $\mathrm{E}\left[\hat{f}_h(x)\right]$ is

$$\hat{f}_h(x) \pm \lambda_{1-\alpha/2}\sqrt{\frac{1}{nh}\hat{f}_h(x)\|K\|_2^2},$$

the same interval as the ad-hoc solution of the asymptotic interval for $f(x)$.

Bootstrap Confidence Band

We can also bootstrap to construct confidence band if we don't want to rely on asymptotic normality.

- Suppose that nonparametric bootstrap is used to create B bootstrap samples.
- For each bootstrap sample, we apply kernel density estimation and obtain $\hat{f}^{(b)}(x)$.
- The bootstrap confidence interval methods can be used to construct a bootstrap confidence interval using the bootstrap replicates $\hat{f}^{(1)}(x)$, ..., $\hat{f}^{(B)}(x)$.

However, the bootstrap confidence interval is still a confidence interval for $\mathrm{E}\left[\hat{f}(x)\right]$.

Bins

It is easy to generalize the histogram for uni-dimension to a histogram for multi-dimension.

- The bin for the univariate case

$$B_j(x_0, h) = [x_0 + (j-1)h,\ x_0 + jh],$$

  is simply an interval.
- The bin for the bivariate case can be a rectangle such as

$$B_j(x_{10}, h_1) \times B_k(x_{20}, h_2).$$

- In general, we have a "rectangular grid" with length h_t in the tth coordinate.

Histogram

Consider $x \in \mathbb{R}^d$. Suppose that each bin B_j is of the form

$$B_{j_1}(x_{10}, h_1) \times \cdots \times B_{j_d}(x_{d0}, h_d).$$

For $x \in B_j$, we still approximate $f(x)$ by

$$\frac{1}{\prod_{k=1}^d h_k} \mathrm{P}(X \in B_j).$$

Then, the histogram becomes

$$\begin{aligned} \hat{f}_h(x) &= \sum_j 1(x \in B_j) \frac{1}{\prod_{k=1}^d h_k} \left[\sum_{i=1}^n \frac{1}{n} 1(X_i \in B_j) \right]. \\ &= \frac{1}{n \prod_{k=1}^d h_k} \sum_{i=1}^n \sum_j 1(X_i \in B_j) 1(x \in B_j). \end{aligned}$$

Introducing Kernel Function

An alternative to the bins in histogram is to make the interval along each coordinate centered around the corresponding element as

$$\left[x_1 - \frac{h_1}{2}, x_1 + \frac{h_1}{2}\right] \times \cdots \times \left[x_d - \frac{h_d}{2}, x_d + \frac{h_d}{2}\right].$$

Then, we can approximate the density by

$$\begin{aligned} f(x) &\approx \frac{\mathrm{P}(X \in B(x))}{h_1 \cdots h_d} \\ &= \frac{1}{h_1 \cdots h_d} \int_{x_1 - \frac{h_1}{2}}^{x_1 + \frac{h_1}{2}} \cdots \int_{x_d - \frac{h_d}{2}}^{x_d + \frac{h_d}{2}} f(u)\, du_1 \cdots du_d \\ &= \int \frac{1}{h_1 \cdots h_d} f(u)\, K\left(\frac{x_1 - u_1}{h_1}, \cdots, \frac{x_d - X_d}{h_d}\right) du_1 \cdots du_d, \end{aligned}$$

where $K(\cdot)$ is the density of a uniform distribution on $B(x)$.

Multivariate Kernel density

It means that we can generalize the kernel density estimation technique to the multidimensional case. But now the kernel function operates on d arguments such as

$$K\left(\frac{x_1 - X_{i1}}{h_1}, \cdots, \frac{x_d - X_{id}}{h_d}\right).$$

In such a case, the kernel density estimator is

$$\hat{f}_h(x) = \frac{1}{n\prod_{k=1}^{d} h_k} \sum_{i=1}^{n} K\left(\frac{x_1 - X_{i1}}{h_1}, \cdots, \frac{x_d - X_{id}}{h_d}\right),$$

where $h = \begin{bmatrix} h_1 & \cdots & h_d \end{bmatrix}$ is the bandwidth vector.

Specify Kernel Function

1. The easiest way to specify a multidimensional kernel is

$$K(u) = \prod_{j=1}^{d} K_1(u_j),$$

where each $K_1(u_j)$ is the unidimensional kernel.

2. We can also use a multivariate kernel such that it is not multiplicative:

$$K(u) \propto K_1\left(u^T u\right),$$

where $\int K(u)\,du = 1$. For example,

$$\text{Epanechnikov}: \quad K(u) \propto \left(1 - u^T u\right) I\left(u^T u \leq 1\right).$$

$$\text{Gaussian}: \quad K(u) \propto \exp\left(-\frac{u^T u}{2}\right).$$

Bandwidth Matrix

A general approach is to use a bandwidth matrix H. The multivariate density estimator is

$$\hat{f}_H(x) = \frac{1}{n \det(H)} \sum_{i=1}^{n} K\left[H^{-1}(x - X_i)\right],$$

where H is a symmetric and positive definite matrix, and the kernel function $K()$ is spherical symmetric and satisfy $\int K(u)\, du = 1$.

- Suppose that $H = \operatorname{diag}\{h_j\}$. Then, we obtain the multiplicative kernel.
- Another example, the multivariate Gaussian kernel yields

$$K\left[H^{-1}(x - X_i)\right] = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{(x - X_i)^T H^{-2} (x - X_i)}{2}\right).$$

Bias and Variance

The derivations of bias and variance are similar to the univariate case. Let $\mu_2(K) = \int u_j^2 K(u)\, du$ for all $j = 1, ..., d$, because of spherical symmetry. Then

$$\begin{aligned} \mathrm{E}\left[\hat{f}_H(x)\right] - f(x) &= \frac{1}{2}\mu_2(K)\,\mathrm{tr}\left[H \frac{\partial^2 f(x)}{\partial x \partial x^T} H\right] + o\left(\mathrm{tr}\left[H^2\right]\right), \\ \mathrm{Var}\left[\hat{f}_H(x)\right] &= \frac{1}{n \det(H)} f(x) \|K\|_2^2 + o\left(\frac{1}{n \det(H)}\right), \end{aligned}$$

as $H \to 0$ and $n \det(H) \to \infty$, where $\|K\|_2^2 = \int K^2(u)\, du$.

MSE and AMISE

Hence, the leading term in MSE is

$$\mathrm{MSE}\left(\hat{f}_H(x)\right) \approx \frac{f(x)\|K\|_2^2}{n\det(H)} + \frac{1}{4}\mu_2^2(K)\,\mathrm{tr}^2\left[H\frac{\partial^2 f(x)}{\partial x\partial x^T}H\right].$$

The AMISE is

$$\mathrm{AIMSE}\left(\hat{f}_H(x)\right) = \frac{\|K\|_2^2}{n\det(H)} + \frac{1}{4}\mu_2^2(K)\int \mathrm{tr}^2\left[H\frac{\partial^2 f(x)}{\partial x\partial x^T}H\right]dx.$$

The AMISE optimal bandwidth matrix minimizes AIMSE $\left(\hat{f}_H(x)\right)$.

Curse of Dimensionality

For simplicity, suppose that $H = hI_d$, where I_d is a $d \times d$ identity matrix. Then, the AMISE optimal bandwidth matrix is

$$h_0 = \left[\frac{d \left\|K\right\|_2^2}{n\mu_2^2\left(K\right)\int \mathrm{tr}^2\left[\frac{\partial^2 f(x)}{\partial x \partial x^T}\right] dx}\right]^{1/(d+4)} = O\left(n^{-1/(d+4)}\right).$$

The convergences rate is much slower than the rate in the unidimensional case $h_0 = O\left(n^{-1/5}\right)$, especially for large d. Hence, multidimensional density estimation is not reliable for large d.

k-Nearest Neighbor Estimator

Suppose that we have observed a random sample X_1, ..., $X_n \in \mathbb{R}^d$. Instead of considering a bin, we consider a ball $B(x,\rho)$ with the center x and radius $\rho(x)$.

- For $x \in B(x,\rho)$, we still approximate $f(x)$ by

$$\frac{1}{\mathrm{Vol}\left(B(x,\rho)\right)}\mathrm{P}\left(X \in B(x,\rho)\right),$$

  where $\mathrm{Vol}\left(B(x,\rho)\right) = \frac{\pi^{d/2}}{\Gamma(1+2^{-1}d)}\rho^d$ is the volume of $B(x,\rho)$.
- The k-nearest neighbor (kNN) estimator is

$$\begin{aligned}\hat{f}_k(x) &= \frac{1}{n\mathrm{Vol}\left(B(x,R_k(x))\right)}\left[\sum_{i=1}^{n}\frac{1}{n}1\left(X_i \in B(x,R_k(x))\right)\right],\\ &= \frac{k}{n\mathrm{Vol}\left(B(x,R_k(x))\right)},\end{aligned}$$

  where $R_k(x)$ be the distance of x to its kth nearest point.

Nearest Neighbor Estimator

Using the kernel function, the k-nearest neighbor (kNN) estimator becomes

$$\hat{f}_k(x) = \frac{1}{nR_k^d(x)} \sum_{i=1}^{n} K\left(\frac{x - X_i}{R_k(x)}\right),$$

where $K()$ is the kernel function.

In contrast, the histogram is

$$\hat{f}_h(x) = \frac{1}{n\prod_{k=1}^{d} h_k} \sum_{i=1}^{n} K\left(\frac{x_1 - X_{i1}}{h_1}, \cdots, \frac{x_d - X_{id}}{h_d}\right),$$

where $K(\cdot)$ is the density of a uniform distribution on a d-dimensional bin, and the kernel density estimator with $H = hI_d$ is

$$\hat{f}_H(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{d}\right).$$