Lecture 6

Kernel density estimation

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$, where F is an unknown c.d.f. with a Lebesgue p.d.f. f.
First, choose a starting point x_0, a binwidth $h > 0$ and define bins (or classes) $I_j = [x_0 + jh - h, x_0 + jh)$, $j \in \mathbb{Z}$. W.l.o.g. we set $x_0 = 0$. Since $f(x) = F'(x)$ *a.e.*, a simple estimator for $f(x)$ at $x \in I_j$ would be

$$f_n(x;h) = \frac{F_n(jh) - F_n(jh-h)}{h} = \frac{1}{nh}\sum_{i=1}^n \mathbb{I}(jh - h < X_i \leq jh) = \frac{1}{nh}\sum_{i=1}^n \mathbb{I}(X_i \in I_j).$$

This estimator is called the **regular histogram**.
In a histogram one fixes classes I_j and finds the number of observations that fall into each class, that is, $f_n(x;h) = h^{-1}\{F_n(jh) - F_n(jh-h)\}$. Recall again that

$$f(x) = F'(x) \approx \frac{F(x+h) - F(x-h)}{2h}$$

for some sufficiently small $h > 0$ and consider another approximation

$$\begin{aligned}
\widehat{f}(x;h) &= \frac{F_n(x+h) - F_n(x-h)}{2h} = \frac{1}{2hn}\sum_{i=1}^n \{\mathbb{I}(X_i \leq x+h) - \mathbb{I}(X_i \leq x-h)\} \\
&= \frac{1}{2hn}\sum_{i=1}^n \mathbb{I}(x-h < X_i \leq x+h) = \frac{1}{2hn}\sum_{i=1}^n \mathbb{I}\left(\frac{|X_i - x|}{h} \leq 1\right) \\
&=: \frac{1}{nh}\sum_{i=1}^n K_u\left(\frac{X_i - x}{h}\right),
\end{aligned}$$

where $h > 0$ and $h \to 0$ and

$$K_u(x) = \begin{cases} 1/2, & |x| \leq 1 \\ 0, & |x| > 1 \end{cases}$$

is the density of continuous uniform distribution on $[-1, 1]$.

Compared to a histogram, in $\widehat{f}(x;h)$ not the classes are fixed, but an interval around each X_i of length $2h$. Estimator $\widehat{f}(x;h)$ with K_u is known as the average shifted histogram or just the Rosenblatt estimator.

The Rosenblatt estimator, as well as a regular histogram, is piecewise constant (=not smooth), which is a clear drawback. A simple way out is to replace K_u by an appropriate

smooth function K. Such a more general estimator is known as Parzen-Rosenblatt kernel density estimator or just kernel density estimator.

Definition
Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} F$ with a given density $F' = f$. A **kernel density estimator** for f is defined via

$$\widehat{f}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),\ x \in \mathbb{R},\ h > 0.$$

Thereby $K : \mathbb{R} \to \mathbb{R}$, such that $\int_{-\infty}^{\infty} K(x)dx = 1$ is known as **kernel** and $h > 0$ is called **bandwidth**. A j**th moment** of a kernel K is defined as $\mu_j = \int_{-\infty}^{\infty} x^j K(x)dx$.

Some classical kernels:

1. $K(x) = 0.5\, \mathbb{I}(|x| \leq 1)$ (the rectangular or uniform kernel)
2. $K(x) = (1 - |x|)\, \mathbb{I}(|x| \leq 1)$ (the triangular kernel)
3. $K(x) = 0.75(1 - x^2)\, \mathbb{I}(|x| \leq 1)$ (the Epanechnikov kernel)
4. $K(x) = (2\pi)^{-1/2} \exp(-x^2/2)$ (the Gaussian kernel)

Let us consider the global risk of the kernel density estimator. Consider the mean integrated squared error

$$\begin{aligned} MISE\left\{\widehat{f}(h)\right\} &= \mathrm{E}\int\left\{\widehat{f}(x;h) - f(x)\right\}^2 dx = \int MSE\left\{\widehat{f}(x;h)\right\} dx \\ &= \int\left[\mathrm{bias}\left\{\widehat{f}(x;h)\right\}\right]^2 dx + \int \mathrm{var}\left\{\widehat{f}(x;h)\right\} dx, \end{aligned}$$

where in the second equality Tonelli-Fubini theorem was used.

Since MISE is a risk corresponding to the $L_2(\mathbb{R})$-norm, it is natural to assume that f is smooth w.r.t. this norm. For example, we may assume that f belongs to a Nikolsky class.

Definition
Let $\beta > 0$ and $L > 0$. The **Nikolsky** class $N_2^{\beta}(L)$ is defined as the set of functions $f : \mathbb{R} \to \mathbb{R}$ whose derivatives $f^{(\ell)}$ of order $\ell = \lfloor\beta\rfloor$ exist and satisfy

$$\left[\int\left\{f^{(\ell)}(x+u) - f^{(\ell)}(x)\right\}^2 dx\right]^{1/2} \leq L|u|^{\beta-\ell}, \quad \forall u \in \mathbb{R}.$$

Theorem
Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} F$, where c.d.f. F has a Lebesgue density f and $\widehat{f}(x;h) = (nh)^{-1}\sum_{i=1}^{n} K\{(X_i - x)/h\}$ be a kernel density estimator.

(i) Suppose that $K : \mathbb{R} \to \mathbb{R}$ is a function satisfying $\int \{K(x)\}^2 dx < \infty$. Then for any $h > 0$, $n \geq 1$ and any density f

$$\int var\left\{\widehat{f}(x;h)\right\} dx \leq \frac{1}{nh} \int \{K(x)\}^2 dx.$$

(ii) Assume that $f \in \mathcal{F}_N = \{f : \ f \geq 0, \int f(x)dx = 1$ and $f \in N_2^\beta(L)\}$ and let K be a kernel of order $\ell = \lfloor \beta \rfloor$ satisfying $\int |x|^\beta |K(x)| dx < \infty$. Then, for any $h > 0$ and $n \geq 1$

$$\int \left[bias\left\{\widehat{f}(x;h)\right\}\right]^2 dx \leq C_2^2 h^{2\beta},$$

where

$$C_2 = \frac{L}{\ell!} \int |x|^\beta |K(x)| dx.$$

Hence, under assumptions of this Theorem we get

$$MISE\left\{\widehat{f}(h)\right\} \leq C_2^2 h^{2\beta} + \frac{1}{nh} \int \{K(x)\}^2 dx.$$

The minimiser of the right hand side w.r.t. h is given by

$$h_{MISE} = \left[\frac{\int \{K(x)\}^2 dx}{2\beta C_2^2}\right]^{\frac{1}{2\beta+1}} n^{-\frac{1}{2\beta+1}},$$

so that $MISE\left\{\widehat{f}(h_{MISE})\right\} = \mathcal{O}(n^{-2\beta/(2\beta+1)})$.

So far we have not discussed the practical choice of K and h. First, we fix some kernel K and discuss how the bandwidth h can be chosen. One reasonable choice for h is a minimiser of $MISE\{\widehat{f}(h)\}$. However, $MISE\{\widehat{f}(h)\}$ (and hence the obtained h) depend on the unknown density f. Therefore, instead of minimising $MISE\{\widehat{f}(h)\}$, it is suggested to minimise an (approximately) unbiased estimator of $MISE\{\widehat{f}(h)\}$. First note that

$$\begin{aligned} MISE\left\{\widehat{f}(h)\right\} &= \mathrm{E} \int \left\{\widehat{f}(x;h) - f(x)\right\}^2 dx \\ &= \mathrm{E} \int \left\{\widehat{f}(x;h)\right\}^2 dx - 2\mathrm{E} \int \widehat{f}(x;h) f(x) dx + \int \{f(x)\}^2 dx. \end{aligned}$$

Since the last term is independent of h, minimisation of $MISE\{\widehat{f}(h)\}$ is equivalent to minimisation of

$$J(h) = \mathrm{E} \int \left\{\widehat{f}(x;h)\right\}^2 dx - 2\mathrm{E} \int \widehat{f}(x;h) f(x) dx.$$

Hence, it is sufficient to find an (approximately) unbiased estimator for each term of $J(h)$. Obviously, $\int\{\widehat{f}(x;h)\}^2dx$ is a trivial unbiased estimator for the first term. It remains to find an unbiased estimator for the second term. Let us show that

$$\frac{1}{n(n-1)h}\sum_{i=1}^{n}\sum_{j\neq i}K\left(\frac{X_j-X_i}{h}\right)$$

is an unbiased estimator for $\mathrm{E}\int\widehat{f}(x;h)f(x)dx$. Indeed, since $X_1,\ldots,X_n$ are i.i.d., we have

$$\begin{aligned}\mathrm{E}\left\{\frac{1}{n(n-1)h}\sum_{i=1}^{n}\sum_{j\neq i}K\left(\frac{X_j-X_i}{h}\right)\right\} &= \mathrm{E}\left\{\frac{1}{(n-1)h}\sum_{j\neq 1}K\left(\frac{X_j-X_1}{h}\right)\right\}\\ &= \mathrm{E}\left\{\frac{1}{(n-1)h}\sum_{j\neq 1}\int K\left(\frac{X_j-u}{h}\right)f(u)du\right\}\\ &= \frac{1}{h}\int f(x)\int K\left(\frac{x-u}{h}\right)f(u)dudx,\end{aligned}$$

provided that the last expression is finite. On the other hand,

$$\begin{aligned}\mathrm{E}\int\widehat{f}(x;h)f(x)dx &= \mathrm{E}\left\{\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{X_i-u}{h}\right)f(u)du\right\}\\ &= \frac{1}{h}\int f(x)\int K\left(\frac{x-u}{h}\right)f(u)dudx,\end{aligned}$$

proving the claim. Putting all together, an unbiased estimator for $J(h)$ results in

$$CV(h)=\int\left\{\widehat{f}(x;h)\right\}^2dx-2\frac{1}{n(n-1)h}\sum_{i=1}^{n}\sum_{j\neq i}K\left(\frac{X_j-X_i}{h}\right).$$

The function $CV(\cdot)$ is called the **(leave-one-out) cross-validation criterion**. We have proved the following result.

Theorem

Assume that for a function $K:\mathbb{R}\to\mathbb{R}$ and for a density f satisfying $\int\{f(x)\}^2dx<\infty$ and $h>0$ we have

$$\int\int f(x)\left|K\left(\frac{x-u}{h}\right)\right|f(u)dudx<\infty.$$

Then $E\{CV(h)\}=MISE\{\widehat{f}(h)\}-\int\{f(x)\}^2dx$.

Hence, functions $MISE\{\widehat{f}(h)\}$ and $\mathrm{E}\{CV(h)\}$ have the same minimisers. In turn, the minimizers of $\mathrm{E}\{CV(h)\}$ can be approximated by those of the function $CV(\cdot)$:

$$h_{CV}=\arg\min_{h>0}CV(h),$$

whenever the minimum is attained. Finally, we define the cross-validation kernel density estimator

$$\widehat{f}(x; h_{CV}) = \frac{1}{nh_{CV}} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h_{CV}}\right).$$

Note that h_{CV} also depends on the sample $X_1, \ldots, X_n$. It can be proved that under appropriate conditions the integrated squared error of $\widehat{f}(x; h_{CV})$ is asymptotically equivalent to that of $\widehat{f}(x; h_M)$, where $h_M = \arg\min_{h>0} MISE\{\widehat{f}(h)\}$ is unknown in practice (=oracle bandwidth).

The cross-validation is not the only way to obtain a bandwidth, which is optimal in some sense.

The choice of the kernel (among of the kernels of the same order) turns out to be less important than the choice of the bandwidth. First, let us state the theorem, which gives the asymptotic expression for $MISE\{\widehat{f}(h)\}$ for a *fixed density* f.

Theorem
Assume that K is a kernel of order 1 satisfying the conditions

$$\int \{K(x)\}^2 dx < \infty, \quad \int x^2 |K(x)| dx < \infty, \quad \int x^2 K(x) dx \neq 0$$

and the density f is differentiable on $\mathbb{R}$, such that f' is absolutely continuous on $\mathbb{R}$ and $\int \{f''(x)\}^2 dx < \infty$. Then for all $n \geq 1$

$$MISE\{\widehat{f}(h)\} = \left[\frac{1}{nh} \int \{K(x)\}^2 dx + \frac{h^4}{4} \int x^2 K(x) dx \int \{f''(x)\}^2 dx\right] \{1 + o(1)\},$$

where $o(1)$ is independent of n, but depends on f and tends to 0 as $h \to 0$.
Note that the $o(1)$ term depends on f, and hence this result holds for a fixed f, but not uniformly over a certain class of densities.

Obviously, one can scale the kernel K without violating assumptions on the kernel. We can take such a scaling parameter δ, that

$$\int \left\{\delta^{-1} K(x/\delta)\right\}^2 dx = \left\{\int x^2 \delta^{-1} K(x/\delta) dx\right\}^2.$$

It is easy to see that

$$\delta = \left[\frac{\int \{K(x)\}^2 dx}{\{\int x^2 K(x) dx\}^2}\right]^{1/5}$$

so that

$$\text{MISE}\{\widehat{f}(h)\} = C(K)\left[\frac{1}{nh} + \frac{h^4}{4}\int_{-\infty}^{\infty} f''(x)^2 dx\right]\{1 + o(1)\},$$

where

$$C(K) = \left[\int \{K(x)\}^2 dx\right]^{4/5}\left\{\int x^2 K(x) dx\right\}^{2/5}.$$

In particular, $C(K)$ is invariant to rescaling of K. A kernel scaled with δ is sometimes called *canonical kernel*. It permits "decoupling" of K and h (for a fixed f). For example, for a Gauss kernel $C(K) = (4\pi)^{-2/5}$. Canonical kernels are useful for (pictorial) comparison of density estimates based on different kernels of the same order, since they are defined in such a way that a particular single choice of bandwidth gives roughly the same amount of smoothing.

So far we considered kernel density estimators that are defined for densities on $\mathbb{R}$. However, there are many positive distributions with densities on $[0, \infty)$ or distributions with densities having a compact support.

Let $f(x) > 0$ for $x \in [0, \infty)$ and K is a kernel with the compact support $[-1, 1]$, so that $K\{(u - x)/h\}$ has support $[x - h, x + h]$. Then for $x < h$ $(x - h < 0)$

$$\begin{aligned}
\text{E}\left\{\widehat{f}(x; h)\right\} &= \int_0^{x+h} \frac{1}{h} K\left(\frac{u - x}{h}\right) f(u) du = \int_{-x/h}^{1} K(v) f(x + hv) dv \\
&= f(x)\int_{-x/h}^{1} K(v) dv + hf'(x)\int_{-x/h}^{1} vK(v) dv + \ldots \\
\text{var}\left\{\widehat{f}(x; h)\right\} &= \frac{1}{nh}\int_{-x/h}^{1} \{K(v)\}^2 dv + \ldots
\end{aligned}$$

Since $x < h$, then for $x/h < 1$ we have in the bias term that $\int_{-x/h}^{1} K(v) dv \neq 1$, as well as further possible terms in the bias $\int_{-x/h}^{1} v^j K(v) dv \neq 0$, $j = 1, 2, \ldots$. The variance term is little influenced. Hence, $\widehat{f}(x; h)$ is not consistent for $x < h$.

There are several approaches to correct the behaviour of the kernel density estimators at the boundary.