Lecture 3

We have studied the kernel density estimator $\widehat{f}(x;h)$ at a fixed $x \in \mathbb{R}$. Next, we analyze a global risk of kernel density estimators. It is customary to 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 1.4. 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

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

In the proof of the next theorem, the following lemma will be used.

Lemma 1.1 (Generalised Minkowski inequality)**.** *For any Borel function g in $\mathbb{R}\times\mathbb{R}$ we have*

$$\int\left\{\int g(u,x)du\right\}^2 dx \le \left(\int\left[\int\{g(u,x)\}^2 dx\right]^{1/2} du\right)^2.$$

Proof: see Tsybakov (2009) *Introduction to nonparametric estimation*, pp. 191 – 192.

With these definitions, we now can prove the following theorem.

Theorem 1.2. *Let $X_1,\ldots,X_n \overset{i.i.d.}{\sim} F$, 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 $\int\{K(x)\}^2dx < \infty$. Then for any $h > 0$, $n \ge 1$ and any density f

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

(ii) Assume that $f \in \mathcal{F}_N = \{f : \ f \ge 0, \int f(x)dx = 1 \text{ and } f \in N_2^\beta(L)\}$ 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 $n \ge 1$

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

where

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

Proof

(i) As in the proof of Theorem 1.1 let

$$Z_i(x) = K\left(\frac{X_i - x}{h}\right) - \mathrm{E}\left\{K\left(\frac{X_i - x}{h}\right)\right\}.$$

and herewith

$$\begin{aligned}
\mathrm{var}\left\{\widehat{f}(x;h)\right\} &= \mathrm{E}\left[\left\{\frac{1}{nh}\sum_{i=1}^{n} Z_i(x)\right\}^2\right] = \frac{1}{nh^2}\mathrm{E}\left[\{Z_i(x)\}^2\right] \\
&\leq \frac{1}{nh^2}\mathrm{E}\left[\left\{K\left(\frac{X_i - x}{h}\right)\right\}^2\right] = \frac{1}{nh^2}\int\left\{K\left(\frac{u-x}{h}\right)\right\}^2 f(u)du
\end{aligned}$$

for all $x \in \mathbb{R}$. Therefore,

$$\begin{aligned}
\int \mathrm{var}\left\{\widehat{f}(x;h)\right\}dx &\leq \frac{1}{nh^2}\int\left[\int\left\{K\left(\frac{u-x}{h}\right)\right\}^2 f(u)du\right]dx \\
&= \frac{1}{nh^2}\int f(u)\left[\int\left\{K\left(\frac{x-u}{h}\right)\right\}^2 dx\right]du \\
&= \frac{1}{nh}\int\{K(u)\}^2 du.
\end{aligned}$$

(ii) Take any $x, u \in \mathbb{R}$, $h > 0$ and write Taylor expansion

$$f(x+uh) = f(x) + f^{'}(x)uh + \ldots + \frac{(uh)^{\ell}}{(l-1)!}\int_0^1 (1-\tau)^{\ell-1} f^{(\ell)}(x+\tau uh)d\tau.$$

Using the fact that the kernel K is of order ℓ we get

$$\begin{aligned}
\mathrm{bias}\left\{\widehat{f}(x;h)\right\} &= \int K(u)\frac{(uh)^{\ell}}{(l-1)!}\left\{\int_0^1 (1-\tau)^{\ell-1} f^{(\ell)}(x+\tau uh)d\tau\right\}du \\
&= \int K(u)\frac{(uh)^{\ell}}{(l-1)!}\left[\int_0^1 (1-\tau)^{\ell-1}\left\{f^{(\ell)}(x+\tau uh) - f^{(\ell)}(x)\right\}d\tau\right]du.
\end{aligned}$$

Applying the generalised Minkowski inequality (Lemma 1.1) and using the fact that

f belongs to $N_2^\beta(L)$ we get the result

$$
\begin{aligned}
&\int \left[\text{bias}\left\{\widehat{f}(x;h)\right\}\right]^2 dx \\
\leq\ & \int \left\{ \int |K(u)| \frac{|uh|^\ell}{(\ell-1)!} \int_0^1 (1-\tau)^{\ell-1} \left| f^{(\ell)}(x+\tau uh) - f^{(\ell)}(x)\right| d\tau du \right\}^2 dx \\
\leq\ & \left(\int |K(u)| \frac{|uh|^\ell}{(\ell-1)!} \left[\int \left\{ \int_0^1 (1-\tau)^{\ell-1} \left| f^{(\ell)}(x+\tau uh) - f^{(\ell)}(x)\right| d\tau \right\}^2 dx \right]^{1/2} du \right)^2 \\
\leq\ & \left\{ \int |K(u)| \frac{|uh|^\ell}{(\ell-1)!} \left(\int_0^1 (1-\tau)^{\ell-1} \left[\int \left\{ f^{(\ell)}(x+\tau uh) - f^{(\ell)}(x)\right\}^2 dx \right]^{1/2} d\tau \right) du \right\}^2 \\
\leq\ & \left[\int |K(u)| \frac{|uh|^\ell}{(\ell-1)!} \left\{ \int_0^1 (1-\tau)^{\ell-1} L |uh|^{\beta-\ell} d\tau \right\} du \right]^2 = C_2^2 h^{2\beta}.
\end{aligned}
$$

□

Hence, under assumptions of Theorem 1.2 we get

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

The minimizer 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 $MISE\left\{\widehat{f}(h_{MISE})\right\} = \mathcal{O}(n^{-2\beta/(2\beta+1)})$. Apparently, the behavior of the MISE is analogous to that of MSE at a fixed point x.

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 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 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,

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

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 1.3. *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).