Lecture 4

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 1.4. *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 $n \geq 1$

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

where $o(1)$ is independent of n, but depends on f and tends to 0 as $h \to 0$.

Proof
See Tsybakov (2009), pp. 192 – 195.

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 $\delta \in (0, \infty)$ 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 the 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 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 $x < h$ $(x - h < 0)$

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

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 $f(x)$ if $x < h$.

There are several approaches to correct the behavior of the kernel density estimators at the boundary. A straightforward solution seems to rescale $\widehat{f}(x;h)$ as follows

$$\widehat{f}_S(x;h) = \frac{\widehat{f}(x;h)}{\int_{-x/h}^{1} K(v)dv}.$$

Note that $\int_{-x/h}^{1} K(v)dv = 1$ for $x \geq h$, so that this method works for all $x \in [0, \infty)$. Then, the bias becomes

$$\mathrm{E}\left\{\widehat{f}_S(x;h)\right\} = f(x) + hf^{'}(x)\frac{\int_{-x/h}^{1} vK(v)dv}{\int_{-x/h}^{1} K(v)dv} + \ldots.$$

Hence, if we make the assumption $f \in \mathcal{F}(\beta, L)$ and take the kernel K to be of order $\lfloor\beta\rfloor$, then the order of the bias at the boundary (for $x < h$) is h, while the bias at the interior (for $x \geq h$) is of order h^β, which is smaller than h for $\beta > 2$. Such an effect when the bias at the boundaries is larger than that at the interior is referred to as **boundary effects**.

A more appealing solution is to use kernels that depend on x/h, such that $\int_{-x/h}^{1} K_b(x)dx = 1$ and $\int_{-x/h}^{1} v^j K_b(x)dx = 0$, $j = 1, 2, \ldots$. Note that such kernels can be found for general order, helping to avoid boundary effects. Also, this approach is more general and is applicable to the estimation of densities with compact support, as well as in the regression context.

A simple ad-hoc solution for the **kernel of first order** is given in Gasser, T. and Müller, H.G. (1979) *Kernel estimation of regression functions*.

$$K_b(u; x/h) = \frac{\mu_{2,x/h}(K) - \mu_{1,x/h}(K)u}{\mu_{0,x/h}(K)\mu_{2,x/h}(K) - \{\mu_{1,x/h}(K)\}^2} K(u)\mathbb{I}\{u \in [-1, x/h)\},$$

where $\mu_{\ell,x/h}(K) = \int_{-1}^{x/h} u^\ell K(u)du$.

A more general solution for **kernels of arbitrary order** is given in Müller, H.G. (1991) *Smooth optimum kernel estimators near endpoints*. The idea is to find a kernel that is a solution to the following variational problem:

$$\begin{aligned}
&\text{Minimise } \int_{-1}^{x/h} \left\{K_b^{(\mu)}(u; x/h)\right\}^2 du \\
&\text{subject to } K_b(\cdot; x/h) \in \mathcal{C}^{\mu-1}(\mathbb{R}), \;\; K_b^{(j)}(-1; x/h) = K_b^{(j)}(x/h; x/h), \;\; 0 \le j < \mu \\
&\text{as well as } K_b(\cdot; x/h) \text{ has support } [-1, x/h] \text{ and is of order } \ell \text{ on } [-1, x/h].
\end{aligned}$$

This gives the boundary kernel for the left boundary, and a right boundary kernel is obtained as $K_b(-u; x/h)$. The motivation behind is connected to variance-minimizing kernels. A general solution is given in terms of normalized ultraspherical polynomials of order μ, see Müller (1991) for an exact expression.

2 Nonparametric regression

2.1 Regression and L_2-risk

Let (X, Y) be a random vector; X is $\mathbb{R}^d$-valued ($d \ge 1$) and Y is $\mathbb{R}$-valued.
Let both X and Y be integrable and let q denote the density of X.
In **regression analysis** one is interested how the value of the **response variable** Y depends on the value of the **covariate vector** X.

That is, our goal is to find some (measurable) $f : \mathbb{R}^d \to \mathbb{R}$, such that $f(X)$ is a "good" approximation of Y, that is $|f(X) - Y|$ should be "small" in some sense.

One possibility: consider the L_p-risk $\mathrm{E}|f(X) - Y|^p$, $1 \leq p \leq \infty$.

In the following, we will consider $p = 2$.
That is, we are looking for $f : \mathbb{R}^d \to \mathbb{R}$ such that $\mathrm{E}|f(X)-Y|^2 = \min_{f:\mathbb{R}^d\to\mathbb{R}} \mathrm{E}|f(X)-Y|^2$.
It can be shown (exercise) that L_2-risk $\mathrm{E}|f(X) - Y|^2$ is minimal if $f(x) = \mathrm{E}(Y|X = x)$.

Hence, $f(x) = \mathrm{E}(Y|X = x)$ is the **optimal approximation of Y by a function of X with respect to L_2-risk**. Function $f(x)$ is called the **regression function**. A regression model is typically written as

$$Y = f(X) + \epsilon, \ \text{ for } \epsilon = Y - f(X), \ \mathrm{E}(\epsilon|X) = 0.$$

Since the distribution of (X, Y) is usually unknown, one can not predict Y using $f(x) = E(Y|X = x)$. In practice, one observes data $(x_1, y_1), \ldots, (x_n, y_n), \ \ x_i \in \mathbb{R}^d, \ y_i \in \mathbb{R}$, which are seen as realizations of

$$D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}, \text{ with } (X_i, Y_i) \text{ i.i.d as } (X, Y).$$

Typical (minimal) assumption $\mathrm{E}(Y^2) < \infty$.

We aim to find $\widehat{f}_n(x) = \widehat{f}_n(x; D_n)$ (a random function) as an estimator for f based on D_n, where for (X_i, Y_i) holds

$$\begin{aligned} Y_i &= f(X_i) + \epsilon_i, \\ \epsilon_i &= Y_i - f(X_i) : \ \mathrm{E}(\epsilon_i|X_i) = 0, \ \ i = 1, \ldots, n. \end{aligned}$$

This model is called a **random design regression model**.

Sometimes the data are sampled at some fixed pre-specified points $x_i \in \mathbb{R}^d$. Usually this is possible for $d = 1$ or $d = 2$ only and the data are taken to be equispaced.

$$\begin{aligned} Y_i &= f(x_i) + \epsilon_i, \ \ i = 1, \ldots, n \\ \epsilon &= (\epsilon_i, \ldots, \epsilon_n)^t : \ \mathrm{E}(\epsilon) = 0, \ \ \mathrm{cov}(\epsilon) = \sigma^2 I_n, \ \ \sigma^2 \in (0, \infty) \end{aligned}$$

This model is referred to as **fixed design regression model**.

We are looking for such $\widehat{f}_n$ that is a "good" approximation of f. Since f is the optimal approximation of Y by a function of X with respect to L_2-risk, it is natural to measure the performance of $\widehat{f}_n$ also with the L_2-risk.

One can show (exercise) that the L_2-risk of $\widehat{f}_n$ can be decomposed as

$$\mathrm{E}\left\{ |Y - \widehat{f}_n(X)|^2 \middle| D_n \right\} = \int_{\mathbb{R}^d} |\widehat{f}_n(x) - f(x)|^2 q(x) dx + \mathrm{E}|Y - f(X)|^2.$$

Hence, $\mathrm{E}\{|Y - \widehat{f}_n(X)|^2 | D_n\}$ is close to the optimal value (which is $\mathrm{E}|Y - f(X)|^2$) if and only if $\int_{\mathbb{R}^d} |\widehat{f}_n(x) - f(x)|^2 q(x) dx$ is close to zero. This expression (a random variable!) is called the **L_2-error of $\widehat{f}_n$**.

2.2 Parametric regression

How to find $\widehat{f}_n$, that estimates f and has minimal L_2-risk?
In some applications one can assume a simple parametric structure for f, for example $f(x) = g(x, \beta)$, where $g(\cdot, \beta) : \mathbb{R}^d \to \mathbb{R}$, is a known parametric function of the unknown parameters $\beta \in \mathbb{R}^p$, $p \leq n$. Then the estimator can be found by minimizing the empirical L_2-risk

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} |Y_i - g(X_i, \beta)|^2 \right\}.$$

Advantages of parametric regression:

1. Low parameter dimension and hence suitable even for small n
2. Fast convergence rates
3. Straightforward interpretation

Disadvantage: Leads to a large error, if f is wrongly specified.