Kernel estimators

(notes by Pavel Chigansky)

Kernel estimator is one of the most common and practically useful methods of function estimation, introduced in the works of M.Rosenblatt [9] and E.Parzen [8]. It can be applied to a variety of statistical problems, including density estimation, nonparametric regression problem, etc.

1. Density estimation

Let $X_1,...,X_n$ be an i.i.d. sample from an unknown density p. It is natural to base construction of the density estimator on the empirical distribution function

$$\widehat{F}_n(x)=\frac{1}{n}\sum_{j=1}^n \mathbf{1}_{\{x\ge X_j\}}, \tag{1.1}$$

which is an unbiased, consistent and asymptotically minimax estimator for the distribution function

$$F(x)=\int_{-\infty}^x p(y)dy.$$

A naive attempt to produce a density estimator would be to differentiate $\widehat{F}_n(x)$ with respect to x. However this yields a meaningless result, since the derivative either equals zero or does not exist. A better idea is to smooth the indicators in (1.1), prior to taking the derivative. To this end let $I(x)$ be a differentiable function, increasing from 0 to 1. Then for any fixed $a\in\mathbb{R}$,

$$\lim_{h\searrow 0} I\Big(\frac{x-a}{h}\Big)=\mathbf{1}_{\{x\ge a\}}, \quad x\in\mathbb{R}\setminus\{a\}. \tag{1.2}$$

Now define "smoothed" empirical distribution function

$$\widehat{F}_{n,h}(x)=\frac{1}{n}\sum_{j=1}^n I\Big(\frac{x-X_j}{h}\Big),$$

where $h>0$ is a small constant, called the *bandwidth*. Clearly this is a legitimate estimator of the distribution function and it converges to the empirical distribution $\widehat{F}_n$ as $h\to 0$ almost everywhere. Taking the derivative, define the density estimator

$$\widehat{p}_n(x)=\frac{1}{nh}\sum_{j=1}^n K\Big(\frac{x-X_j}{h}\Big), \tag{1.3}$$

lecture notes for "Advanced Statistical Models B" course.

where $K(x)=\frac{d}{dx}I(x)$ is a nonnegative *kernel* function with $\int_{\mathbb{R}} K(x)dx = 1$. Here is a very partial list of kernels

$$\begin{aligned}
K(u) &= \frac{1}{2}\mathbf{1}_{\{|u|\le 1\}}, && \text{(uniform)},\\
K(u) &= (1-|u|)\mathbf{1}_{\{|u|\le 1\}}, && \text{(triangular)},\\
K(u) &= \frac{1}{\sqrt{2\pi}}e^{-u^2/2}, && \text{(Gaussian)},\\
K(u) &= \frac{15}{16}(1-u^2)^2\mathbf{1}_{\{|u|\le 1\}}, && \text{(biweight)}.
\end{aligned}\tag{1.4}$$

REMARK 1.1. Similar heuristics (Problem 1) leads to the following kernel estimator in the multivariate case with $X_j=(X_{j1},...,X_{jd})\in\mathbb{R}^d$

$$\widehat{p}_n(x)=\frac{1}{nh^d}\sum_{j=1}^n K\Big(\frac{x_1-X_{j1}}{h}\Big)...K\Big(\frac{x_d-X_{jd}}{h}\Big),\quad x\in\mathbb{R}^d. \tag{1.5}$$

Below we will often omit the subscripts n and h from the notations, whenever this does not cause confusion.

2. Risk upper bounds

As usually the quality of estimation can be measured by a risk

$$R(p,\widehat{p})=\mathbb{E}_p w\big(d(p-\widehat{p})\big),$$

where $d(\cdot)$ is a (semi-)norm [1] and w is a bowl shaped function. Here are some examples,

$$\begin{aligned}
&\mathbb{E}_p(\widehat{p}(x_0)-p(x_0))^2, && \text{(Mean Squared Error)},\\
&\mathbb{E}_p\|\widehat{p}-p\|_2^2=\mathbb{E}_p\int_{\mathbb{R}}(\widehat{p}(x)-p(x))^2dx, && \text{(Mean Integrated Squared Error)},\\
&\mathbb{E}_p\|\widehat{p}-p\|_\infty=\mathbb{E}_p\sup_{x\in\mathbb{R}}|\widehat{p}(x)-p(x)|, && \text{(Mean Uniform Error)}.
\end{aligned}$$

What kind of guarantees can we have on such risks?

2.1. The MSE risk at a fixed point.

The usual bias-variance decomposition

$$\mathrm{MSE}(p,\widehat{p})=\mathbb{E}_p(\widehat{p}(x_0)-p(x_0))^2=\mathrm{Var}_p(\widehat{p}(x_0))+(\mathbb{E}_p\widehat{p}(x_0)-p(x_0))^2$$

provides a useful insight. Let us calculate each term and see how the shape of the kernel K, the sample size n and the bandwidth parameter affect the risk. By the i.i.d. assumption, the variance term satisfies the bound

$$\mathrm{Var}_p(\widehat{p}(x_0))=\mathbb{E}_p\big(\widehat{p}(x_0)-\mathbb{E}_p\widehat{p}(x_0)\big)^2=\frac{1}{h^2}\frac{1}{n}\mathrm{Var}_p\Big(K\Big(\frac{x_0-X_1}{h}\Big)\Big)\le$$
$$\frac{1}{h^2}\frac{1}{n}\int_{\mathbb{R}}K\Big(\frac{x_0-u}{h}\Big)^2p(u)du\le\frac{1}{hn}\|p\|_\infty\|K\|_2^2=:\frac{C_1}{nh}, \tag{2.1}$$

[1] Note that the MSE at a fixed point x_0 is defined by the semi-norm $(|q(x_0)-p(x_0)|)^2$, since this quantity may vanish for distinct p and q.

with a constant C_1 independent of n and h. This reveals that the variance term decreases with the sample size and increases if the bandwidth is made smaller. To find a useful bound for the bias term we will need to restrict the class of densities under consideration.

DEFINITION 2.1. *A real function $f : \mathbb{R} \mapsto \mathbb{R}$ belongs to the Hölder class $\Sigma(\beta, L)$ with $\beta > 0$ and $L > 0$, if it has $\ell = \max\{k \in \mathbb{N} \cup \{0\} : k < \beta\}$ derivatives and the last derivative is $(\beta - \ell)$-Hölder continuous,*

$$|f^{(\ell)}(x) - f^{(\ell)}(y)| \le L|x-y|^{\beta-\ell} \quad \forall x, y \in \mathbb{R}.$$

When the true unknown density is assumed to belong to such a class, the bias term can be controlled efficiently by kernels with the following special properties.

DEFINITION 2.2. *A kernel K is said to be of order ℓ if*

$$\int_{\mathbb{R}} K(u)du = 1, \quad \int_{\mathbb{R}} u^j K(u)du = 0, \quad j = 1, ..., \ell,$$
$$\int_{\mathbb{R}} |K(u)||u|^{\ell+1} du < \infty.$$

Note that if $K(u) \ge 0$ is required, it cannot be of order greater than 1. However, we will argue later that this constraint can be dropped and show how to construct kernels of any desired order.

For a density $p \in \Sigma(\beta, L)$, the bias term satisfies

$$\begin{aligned}\mathbb{E}_p \widehat{p}(x_0) - p(x_0) = & \frac{1}{nh} \sum_{j=1}^{n} \int_{\mathbb{R}} K\Big(\frac{x_0 - u}{h}\Big) p(u)du - p(x_0) = \\ & \int_{\mathbb{R}} K(v) \Big(p(x_0 - vh) - p(x_0)\Big) dv.\end{aligned} \tag{2.2}$$

Expanding the density into Taylor's series aroudnd x_0, we obtain

$$\begin{aligned}& p(x_0 - vh) - p(x_0) = \\ & \sum_{j=1}^{\ell} \frac{p^{(j)}(x_0)}{j!}(-vh)^j + \frac{1}{\ell!}\big(p^{(\ell)}(x_0 - \xi vh) - p^{(\ell)}(x_0)\big)(-vh)^\ell\end{aligned} \tag{2.3}$$

with $0 \le \xi \le 1$. Therefore, if K is of order ℓ,

$$\begin{aligned}& \big|\mathbb{E}_p \widehat{p}(x_0) - p(x_0)\big| = \\ & \Big|\int_{\mathbb{R}} K(v) \frac{1}{\ell!}\big(p^{(\ell)}(x_0 - \xi vh) - p^{(\ell)}(x_0)\big)(-vh)^\ell\Big) dv\Big| \le \\ & \frac{1}{\ell!}\int_{\mathbb{R}} |K(v)| L |\xi vh|^{\beta-\ell} |vh|^\ell dv \le h^\beta \frac{L}{\ell!} \int_{\mathbb{R}} |K(v)||v|^\beta dv =: C_2 h^\beta.\end{aligned} \tag{2.4}$$

Thus the bias term does not depend on the sample size and decreases as $h \to 0$ at a rate higher for smoother functions.

Plugging the bounds for the bias and the variance, we obtain the following bound for the MSE risk

$$\mathbb{E}_p(\widehat{p}(x_0)-p(x_0))^2 \le \frac{C_1}{nh}+C_2^2h^{2\beta}. \tag{2.5}$$

The expression in the right hand side is minimized over $h>0$ by the choice

$$h_n^* := cn^{-\frac{1}{2\beta+1}}, \tag{2.6}$$

where c is a constant, which depends explicitly on C_1, C_2, that is, on the kernel K and the parameters β and L of the function class. [2] The exact dependence is not very important, since β and L are typically unknown in applications.

The bound (2.5) with the optimal bandwidth (2.6) becomes

$$\mathbb{E}_p(\widehat{p}(x_0)-p(x_0))^2 \le C(\beta,L,K)n^{-\frac{2\beta}{2\beta+1}},$$

where $C(\beta,L,K)$ is constant in n. Thus we obtained the upper bound on the minimax MSE risk

$$\sup_{p\in\Sigma(\beta,L)} \mathbb{E}_p\Big(n^{\frac{\beta}{2\beta+1}}\big(\widehat{p}(x_0)-p(x_0)\big)\Big)^2 \le C(\beta,L,K). \tag{2.7}$$

2.2. The MISE risk.

Let us now consider the risk

$$\mathrm{MISE}(p,\widehat{p}) = \int_{\mathbb{R}} \mathbb{E}_p(\widehat{p}(x)-p(x))^2dx.$$

As before, it can be decomposed into bias and variance terms. The variance satisfies a bound similar to (2.1),

$$\int_{\mathbb{R}} \mathrm{Var}_p(\widehat{p}(x))dx \le \frac{1}{h^2}\frac{1}{n}\int_{\mathbb{R}} p(u)\int_{\mathbb{R}} K\Big(\frac{x-u}{h}\Big)^2dxdu \le \frac{1}{hn}\|K\|_2^2 =: C_1\frac{1}{nh}.$$

The bias term in this case can be controlled on a different class of densities.

Definition 2.3. *A real function f belongs to Nikolski class $H(\beta,L)$ with $\beta>0$ and $L>0$ if it has $\ell=\min\{k\in\mathbb{N}\cup\{0\}: k<\beta\}$ derivatives and*

$$\big\|f^{(\ell)}(\cdot+t)-f^{(\ell)}(\cdot)\big\|_2 \le L|t|^{\beta-\ell}, \quad t\in\mathbb{R}.$$

To proceed we will need following classic inequality.

Lemma 2.4 (Generalized Minkowski inequality). *For any function $g:\mathbb{R}\times\mathbb{R}\mapsto\mathbb{R}$,*

$$\int\Big(\int g(u,x)du\Big)^2dx \le \bigg(\int\Big(\int g(u,x)^2dx\Big)^{1/2}du\bigg)^2.$$

[2] From (2.4) with $h=1$ it follows that

$$p(x_0)\le \mathbb{E}_p\widehat{p}(x_0)+C_2 \le \|K\|_\infty + C_2$$

and hence $\|p\|_\infty<\infty$ for $p\in\Sigma(\beta,L)$.

Now, for $p \in H(\beta, L)$, in view of (2.2) and (2.3),

$$\int_{\mathbb{R}} \left(\mathbb{E}_p \widehat{p}(x) - p(x)\right)^2 dx =$$
$$\int_{\mathbb{R}} \left(\int_{\mathbb{R}} K(v) \Big(\frac{1}{\ell!} \big(p^{(\ell)}(x-\xi v h) - p^{(\ell)}(x) \big) (-vh)^\ell \Big) dv \right)^2 dx \overset{\dagger}{\le}$$
$$\left(\frac{1}{\ell!} h^\ell\right)^2 \left(\int_{\mathbb{R}} K(v) |v|^\ell \Big(\int_{\mathbb{R}} \left(p^{(\ell)}(x-\xi v h) - p^{(\ell)}(x) \right)^2 dx \Big)^{1/2} dv \right)^2 \le$$
$$\left(\frac{L}{\ell!} h^\beta\right)^2 \left(\int_{\mathbb{R}} K(v)|v|^\beta dv \right)^2 =: C_2 h^{2\beta},$$

where in † we applied Minkowski inequality. Combining the bounds for the bias and variance implies that the MISE risk satisfies the same bound as in (2.5) and consequently attains the same minimax rate of $n^{-2\beta/(2\beta+1)}$, this time over the Nikolski class of densities.

3. Construction of kernels with any desired order

The statistic in (1.3) is a valid density, if the kernel K is a nonnegative function. However this has not been used in the derivation of the risk upper bound. Suppose we constructed $\widehat{p}(x_0)$ using a kernel, which is not constrained to be nonnegative everywhere (still keeping, of course, $\int K(u)du = 1$). Define $\widehat{p}^+(x_0) := \widehat{p}(x_0) \vee 0$, then since $p(x_0) > 0$,

$$\mathbb{E}_p (\widehat{p}^+(x_0) - p(x_0))^2 \le \mathbb{E}_p (\widehat{p}(x_0) - p(x_0))^2$$

and hence $\widehat{p}^+(x_0)$ is still an adequate estimator for $p(x_0)$, still being rate optimal.

How do we construct kernels of an arbitrary order ℓ? One approach is to use a family of polynomials φ_n with $\deg(\varphi_n) = n$, orthonormal with respect to some nonnegative weight function $\mu(\cdot)$,

$$\int_{\mathbb{R}} \varphi_m(u) \varphi_n(u) \mu(u) du = \begin{cases} 1, & m = n, \\ 0, & m \neq n. \end{cases}$$

For example, for $\mu(u) = \mathbf{1}_{\{|u| \le 1\}}$, these are the Legendre polynomials (Problem 5) and, for $\mu(u) = e^{-u^2/2}$, the Hermite polynomials, etc.

Define the kernel

$$K(u) = \sum_{m=0}^{\ell} \varphi_m(0) \varphi_m(u) \mu(u).$$

This kernel is of order ℓ, since $\varphi_0(u) \equiv \varphi_0(0) = 1/\sqrt{2}$ and, by orthonormality,

$$\int_{\mathbb{R}} K(u) du = \int_{\mathbb{R}} \varphi_0(0) \varphi_0(u) \mu(u) du +$$
$$\sum_{m=1}^{\ell} \frac{\varphi_m(0)}{\varphi_0(0)} \int_{\mathbb{R}} \varphi_0(u) \varphi_m(u) \mu(u) du = \|\varphi_0\|_{2,\mu}^2 = 1.$$

Since $\varphi_n(u)$ is a polynomial of degree n, any power u^m can be expanded as

$$u^m = \sum_{j=0}^m b_{jm}\varphi_j(u),$$

where b_{jm} are the scalar products of u^m with φ_j's. Then for any $1 \le m \le \ell$

$$\int_{\mathbb{R}} u^m K(u)du = \int_{\mathbb{R}} \sum_{j=0}^m b_{jm}\varphi_j(u) \sum_{i=0}^{\ell} \varphi_i(0)\varphi_i(u)\mu(u)du =$$

$$\sum_{j=0}^m b_{jm}\varphi_j(0) = u^m\Big|_{u=0} = 0.$$

4. Bandwidth tuning

An important practical issue is the choice of the bandwidth parameter h. For a given sample size, formula (2.6) can be viewed only as a rough guideline, rather than a concrete recipe. Various methodologies have been suggested to this end. One common practice is the Cross Validation technique. Let us sketch the ideas for estimation under the MISE risk.

Ideally we would like to choose the value of bandwidth h which minimises the risk, say $h_{\text{id}} = \text{argmin}_{h>0}\text{MISE}(h)$. This choice depends on the true density and is therefore cannot be used in estimation. It makes sense then to choose the bandwidth which minimises an unbiased estimator of MISE

$$\text{MISE}(h) = \mathbb{E}_p \int (\widehat{p}_n - p)^2 = \mathbb{E}_p\Big(\int \widehat{p}_n^2 - 2\int \widehat{p}_n p\Big) + \int p^2.$$

Since the last term does not depend on the estimator, the ideal choice h_{id} also minimizes the function

$$J(h) := \mathbb{E}_p\Big(\int \widehat{p}_n^2 - 2\int \widehat{p}_n p\Big).$$

We would like to find an unbiased estimator for $J(h)$. Trivially $\int \widehat{p}_n^2$ is an unbiased estimator for $\mathbb{E}_p \int \widehat{p}_n^2$. An unbiased estimator for $\mathbb{E}_p \int \widehat{p}_n p$ can be constructed by the Jackknife technique. Let

$$\widehat{p}_{n\backslash i}(x) = \frac{1}{(n-1)h}\sum_{j\ne i} K\Big(\frac{X_j - x}{h}\Big).$$

Then the statistic

$$\widehat{T}_n = \frac{1}{n}\sum_{i=1}^n \widehat{p}_{n\backslash i}(X_i),$$

is the desired unbiased estimator,

$$\mathbb{E}_p\widehat{T}_n = \mathbb{E}_p \frac{1}{n}\sum_{i=1}^n \frac{1}{(n-1)h}\sum_{j\ne i}\mathbb{E}_p\Big(K\Big(\frac{X_j - X_i}{h}\Big)\Big|X_i\Big) =$$

$$\mathbb{E}_p \int \frac{1}{nh}\sum_{i=1}^n K\Big(\frac{x - X_i}{h}\Big)p(x)dx = \mathbb{E}_p \int \widehat{p}_n p.$$

The function

$$\mathrm{CV}(h) = \int \widehat{p}_n^2 - \frac{2}{n}\sum_{i=1}^{n} \widehat{p}_{n/i}(X_i) \tag{4.1}$$

is called *leave-one-out cross validation* criterion and $h_{\mathrm{CV}} = \operatorname{argmin}_{h>0}\mathrm{CV}(h)$ can be a reasonable data-driven choice of the bandwidth. It can be shown that the kernel estimator with such a choice of the bandwidth retains its minimax rate optimality on the same classes of functions (Stone's theorem).

5. Nonparametric regression

The regression problem is to estimate the function $f : \mathbb{R} \mapsto \mathbb{R}$ given the sample of pairs $(X_1, Y_1), \dots, (X_n, Y_n)$, generated by the equation

$$Y_j = f(X_j) + \varepsilon_j,$$

where ε_j's are zero mean random errors.

Suppose that both the design points X_j's and the errors ε_j's are independent and are sampled independently from some densities. Then (X_j, Y_j) are i.i.d. random vectors from a non-degenerate unknown joint density, denote it by p_{XY}. Since ε_j's have zero means

$$f(x) = \mathbb{E}(Y_1|X_1 = x) = \frac{\int_{\mathbb{R}} y p_{XY}(x,y)dy}{p_X(x)}.$$

This formula suggests to construct an estimator of f by replacing the densities by their kernel estimators, cf. (1.5),

$$\widehat{f}_n(x) = \frac{\int_{\mathbb{R}} y \widehat{p}_{XY}(x,y)dy}{\widehat{p}_X(x)} = \frac{\int_{\mathbb{R}} y \frac{1}{nh^2}\sum_{j=1}^{n} K\big(\frac{x-X_j}{h}\big)K\big(\frac{y-Y_j}{h}\big)dy}{\frac{1}{nh}\sum_{j=1}^{n} K\big(\frac{x-X_j}{h}\big)} =$$

$$\frac{\frac{1}{nh^2}\sum_{j=1}^{n} K\big(\frac{x-X_j}{h}\big) h \int_{\mathbb{R}} (Y_j + vh)K(v)dv}{\frac{1}{nh}\sum_{j=1}^{n} K\big(\frac{x-X_j}{h}\big)} = \frac{\sum_{j=1}^{n} Y_j K\big(\frac{x-X_j}{h}\big)}{\sum_{j=1}^{n} K\big(\frac{x-X_j}{h}\big)},$$

where the last equality holds if K has of order 1 at least.

This is the Nadaraya-Watson kernel regressor. We will see that it can be viewed as a special case of a broader family of the *Local Polynomial* estimators, the main subject of the following chapter. In particular, we will see that its minimax MSE risk also decreases at the rate $n^{-2\beta/(2\beta+1)}$ on appropriate smoothness classes of functions.

6. More on theory and practice

6.1. Lower bounds and asymptotic minimax optimality. As we already saw, estimation of certain smooth, namely Hadamard differentiable, functionals of distribution can be done at the parametric rate of $\sqrt{n}$. Derivative operator however is not such a functional and hence it is not clear whether minimax rate $n^{-\beta/(2\beta+1)}$ is unimprovable on the above smoothness classes of functions. This turns out to be the case, as it is possible to derive lower bounds for the risk with the same rate. For example, one can prove that

$$\varliminf_n \inf_{\widehat{p}_n} \sup_{p\in\Sigma(\beta,L)} \mathbb{E}_p\Big(n^{\frac{\beta}{2\beta+1}}\big(\widehat{p}_n(x_0)-p(x_0)\big)\Big)^2 > 0, \tag{6.1}$$

where the infimum is taken over all estimators, based on the sample of size n. This shows that the above rate is in fact minimax optimal and thus suitably tuned kernel estimators are *rate* optimal, i.e. attain the best possible minimax rate. In some models, it is actually possible to find the exact value of the limit in (6.1) and to find estimators, whose minimax risk converges *exactly* to this limit.

Note that the optimal minimax rate $n^{\beta/(2\beta+1)}$ in the density estimation problem is slower than the usual parametric rate $\sqrt{n}$, which means that the former is a harder statistical problem. As β increases, that is, the functions in the class become smoother, the rate improves and approaches $\sqrt{n}$. There are, however, examples of infinite dimensional classes of functions, for which the rate of $\sqrt{n}$ is nevertheless attainable, see [10].

6.2. Other norms. While for many norms the minimax risk decreases at the same rate of $n^{-\beta/(2\beta+1)}$ as in (2.7), for some it does not. One example is the mean uniform risk

$$\mathbb{E}_p\big\|\widehat{p}_n(x)-p(x)\big\|_\infty = \mathbb{E}_p \sup_{x\in\mathbb{R}}|\widehat{p}_n(x)-p(x)|.$$

The optimal minimax rate on the same function classes as above turns out to be slower by a logarithmic factor, [4], [5].

6.3. Adaptivity to smoothness. Design of kernel estimators requires knowing the smoothness parameter β. An estimator which does not use β is called *adaptive* to smoothness and an interesting question then is whether it can attain the same optimal rate $n^{-\beta/(2\beta+1)}$ as in e.g. (2.7). Several such estimators were suggested and proved to be rate optimal in a number models, [6],[7]. On the other hand, in some models the optimal adaptive rate is known to be strictly slower, typically by a log factor.

6.4. Confidence bands. Constructing confidence bands for densities is a nontrivial matter as explained in [11, Section 5.7]. A natural quantity on which such a confidence band can be based is $p(x)-\widehat{p}_n(x)$ or its studentized version. However a close look at the bias-variance decomposition

$$p(x)-\widehat{p}_n(x) = \big(p(x)-\mathbb{E}_p\widehat{p}_n(x)\big)+\big(\mathbb{E}_p\widehat{p}_n(x)-\widehat{p}_n(x)\big),$$

reveals that under the scaling, which yields a non-degenerate limit distribution of the second term, the first (bias) term will not be asymptotically negligible. This is the result of choosing the bandwidth so that the bias and the variance contributions to the risk have the same order of magnitude. The limit bias term will depend on the unknown density and hence $p(x)-\widehat{p}_n(x)$ is not not a pivotal quantity, even asymptotically under the correct scaling. The classical paper on the subject is [1].

6.5. Curse of dimensionality. The optimal minimax rate in the estimation problem of multivariate densities from appropriate smoothness classes is $n^{-\beta/(2\beta+d)}$ (cf. Problem 7), where d is the dimension of the density support and β is the relevant smoothness parameter. Hence to achieve a small value ε of MSE risk one needs sample size of order $(1/\varepsilon)^{2+d/\beta}$, which grows exponentially with dimension d. This phenomenon, called the *curse of dimensionality*, is common in nonparametric problems.

6.6. Boundary effects. Note that our analysis of the kernel density estimators assumed that the densities are smooth functions on the whole real line. In many applications densities have bounded support and typically are less smooth at the boundaries. Consequently special adjustments must be introduced to keep rate optimality, see e.g. [2].

Computer experiment

Implement kernel density estimator with cross-validated bandwidth as in Section 4 using kernels of orders 1 and 2 from Problem 5.

(1) Apply your code with the kernel of order $\ell = 1$ to a sample of size $n = 100$ from some smooth density. Plot your estimate against the true density. Increase the sample size to $n = 1000$ and add the obtained estimate to the plot. Is the improvement in accuracy visible?

(2) Repeat with kernel of order $\ell = 2$ and add the estimates to the plot. Comment on your observations.

Exercises

PROBLEM 1. Following the heuristics of the one-dimensional case, derive the multivariate density estimator (1.5).

PROBLEM 2. Let $\mathcal{N}$ be the family of all normal densities

$$\mathcal{N} = \left\{ \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right) : \mu \in \mathbb{R}, \sigma \in \mathbb{R}_+ \right\},$$

where φ is $N(0,1)$ density.

(1) Show that $\mathcal{N} \cap \Sigma(\beta, L) \neq \emptyset$ for any $\beta > 0$ and $L > 0$.

(2) Show that $\mathcal{N} \setminus \Sigma(\beta, L) \neq \emptyset$ for any $\beta > 0$ and $L > 0$.

PROBLEM 3.

(1) Show that

$$q(x) = \frac{6}{\pi^2} \sum_{j=1}^{\infty} j \mathbf{1}_{\{j \le x \le j+1/j^3]\}}, \quad x \in \mathbb{R},$$

is a probability density function.

Hint: use the formula $\sum_{j=1}^{\infty} \frac{1}{j^2} = \frac{\pi^2}{6}$.

(2) Is the density q locally bounded, that is,

$$\sup_{|x|\le r} q(x) < \infty, \quad \forall r \in \mathbb{R}_+ \ ?$$

Bounded, $\|q\|_\infty < \infty$? Find $\underline{\lim}_{x\to\infty} q(x)$ and $\overline{\lim}_{x\to\infty} q(x)$.

(3) Show that for any $\beta, L \in \mathbb{R}_+$ and any $p \in \Sigma(\beta, L)$,

$$\lim_{x\to\pm\infty} p(x) = 0.$$

Hint: adapt the argument from the text which was used to show that

$$\sup_{p\in\Sigma(\beta,L)} \|p\|_\infty < \infty.$$

PROBLEM 4. Argue that the kernel estimator cannot be consistent, uniformly on any class which includes all densities with jump discontinuities.

Hint: inspect the bias term

PROBLEM 5. The Legendre polynomials

$$\phi_0(x) = \frac{1}{\sqrt{2}},$$
$$\phi_m(x) = \sqrt{\frac{2m+1}{2}}\frac{1}{2^m m!}\frac{d^m}{dx^m}\left(x^2-1\right)^m, \quad m = 1,2,..$$

form a complete orthogonal basis in $L_2([-1,1])$.

(1) Write the explicit formulas for the first three Legendre polynomials.

(2) Construct and plot kernels of all orders up to 3, using the Legendre basis.

(3) Specify the cross-validation criterion (4.1) for the kernels found above.

PROBLEM 6 (Exercise 1.2, [10]). Kernel estimator of the s-th derivative $p^{(s)}$ of a density $p \in \Sigma(\beta, L)$, $s < \beta$, can be defined as follows

$$\widehat{p}_n^{(s)}(x) = \frac{1}{nh^{s+1}}\sum_{j=1}^n K\Big(\frac{X_j - x}{h}\Big).$$

Here $h > 0$ is the bandwidth and $K : \mathbb{R} \mapsto \mathbb{R}$ is a bounded kernel with support $[-1,1]$ satisfying, for $\ell = \min\{k \in \mathbb{N}\cup\{0\} : k < \beta\}$,

$$\begin{aligned} &\int u^j K(u)du = 0, \quad j = 0,1,\dots,s-1,s+1,\dots,\ell \\ &\int u^s K(u)du = s! \end{aligned} \tag{6a}$$

(1) Prove that, uniformly over the class $\Sigma(\beta, L)$, the bias of $\widehat{p}_n^{(s)}(x_0)$ is bounded by $c_1 h^{\beta-s}$ and its variance is bounded by $c_2/(nh^{2s+1})$, where c_1 and c_2 are positive constants, and $x_0 \in \mathbb{R}$ is a given point.

(2) Prove that the maximum of the MSE of $\widehat{p}_n^{(s)}(x_0)$ over $\Sigma(\beta, L)$ is of order $O(n^{-\frac{2(\beta-s)}{2\beta+1}})$ as $n \to \infty$ if the bandwidth $h = h_n$ is chosen optimally.

(3) Let (ϕ_n) be the Legendre basis on $[-1,1]$. Show that the kernel

$$K(u) = \sum_{m=0}^{\ell} \phi_m^{(s)}(0)\phi_m(u)\mathbf{1}_{\{|u|\le 1\}}$$

satisfies conditions (6a).

PROBLEM 7 (Exercise 1.3, [10]). Consider the estimator $\widehat{p}_n$, cf. 1.5,

$$\widehat{p}_n(x,y) = \frac{1}{nh^2}\sum_{i=1}^{n} K\Big(\frac{X_i - x}{h}\Big)K\Big(\frac{Y_i - y}{h}\Big)$$

of the two dimensional probability density function p, which belongs to the Hölder class of densities on $\mathbb{R}^2$

$$|p(x,y) - p(x',y')| \le L(|x-x'|^\beta + |y-y'|^\beta), \quad \forall (x,y),(x',y') \in \mathbb{R}^2,$$

with given constants $0 < \beta \le 1$ and $L > 0$. Let (x_0, y_0) be a fixed point in $\mathbb{R}^2$. Derive upper bounds for the bias and the variance of $\widehat{p}(x_0,y_0)$ and an upper bound on the MSE risk at (x_0,y_0). Find the minimizer $h = h_n$ of the upper bound on the risk and the corresponding rate of convergence.

PROBLEM 8. The k Nearest Neighbours (k-NN) regression estimator is

$$\widehat{f}(x) = \frac{1}{k}\sum_{j \in N_k(x)} Y_j,$$

where $N_k(x)$ is the set of indices of the k design points, closest to x.

(1) Does the estimator $\widehat{f}(x)$ belong to $\Sigma(\beta, L)$ for some $\beta > 0$?

(2) Assuming uniform nonrandom design

$$X_j = j/n, \quad j = 1, \dots, n,$$

find a suitable choice of k, so that this estimator is minimax rate optimal with respect to the MSE risk on the class $\Sigma(1,L)$.

(3) For the same design as above, show that properly tuned k-NN estimator is minimax rate optimal with respect to the MSE risk on[3]$\Sigma(1,L) \cap \Sigma(2,L)$, if x_0 in the interior point of $[0,1]$.

(4) Is rate optimality retained on $\Sigma(1,L) \cap \dots \cap \Sigma(\beta, L)$ for $\beta > 2$?

(5) Explain how the k-nearest neighbours estimator is applied to functions of $d \ge 1$ variables.

[3] A function f belongs to both $\Sigma(1,L)$ and $\Sigma(2,L)$, if it's first derivative is bounded $\|f'\|_\infty \le L$ and also L-Lipschitz, i.e.

$$|f(y) - f(x)| \le L|y-x|, \quad y,x \in [0,1].$$

(6) Consider the uniform rectangular grid of design points $X_1,...,X_n \in [0,1]^d$, with $[n^{1/d}]$ points on each side of this cube. Find the value of k, which optimizes the minimax MSE error of the k-NN estimator on the class of Lipschitz functions

$$\Lambda(L) := \Big\{f : |f(x)-f(y)| \le L\|x-y\|,\ x,y \in [0,1]^d\Big\},$$

for the design as above.

PROBLEM 9. Let $X_1,...,X_n \overset{\text{i.i.d.}}{\sim} p$ where $p(\cdot)$ is an unknown density on $\mathbb{R}$. Let $T(p)$ be a functional of the form

$$T(p) = \int_{\mathbb{R}} \psi(x)p(x)dx,$$

where ψ is a fixed function. Consider the plug-in estimator of $T(p)$

$$T(\widehat{p}_n) = \int_{\mathbb{R}} \psi(x)\widehat{p}_n(x)dx,$$

where $\widehat{p}_n(x)$ is the kernel estimator (1.3). Assuming that ψ is L-Lipschitz function

$$|\psi(y)-\psi(x)| \le L|y-x|,$$

show that the kernel and the bandwidth can be chosen so that

$$\sup_{p\in\mathcal{P}(\psi,M)} \mathbb{E}_p\Big(\sqrt{n}\big(T(\widehat{p}_n)-T(p)\big)\Big)^2 \le C(K,\psi,M) < \infty,$$

where $C(K,\psi,M)$ is a constant, which depends only on the kernel and the class of densities under the supremum,

$$\mathcal{P}(\psi,M) = \Big\{p : \int \psi(v)^2 p(v)dv \le M\Big\}.$$

Note that the densities in this class are not required to be smooth, e.g. not even continuous.

References

[1] Bickel, Peter J., and Murray Rosenblatt. On some global measures of the deviations of density function estimates. The Annals of Statistics (1973): 1071-1095.

[2] Jones, M. Chris. ”Simple boundary correction for kernel density estimation.” Statistics and computing 3.3 (1993): 135-146.

[3] I.A. Ibragimov, R. Z. Hasminskii, Statistical estimation, asymptotic theory, Springer Verlag 1981

[4] Korostelev, A. P. Exact asymptotically minimax estimator for nonparametric regression in uniform norm. Theory Probab. Appl 38.4 (1993): 875-882.

[5] Korostelev, Alexander, and Michael Nussbaum. The asymptotic minimax constant for sup-norm loss in nonparametric density estimation. Bernoulli (1999): 1099-1118.

[6] Lepski, Oleg V. On problems of adaptive estimation in white Gaussian noise. Topics in nonparametric estimation 12 (1992): 87-106.

[7] Goldenshluger, Alexander, and Oleg Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. The Annals of Statistics 39.3 (2011): 1608-1632.

[8] Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. The Annals of Mathematical Statistics. 33 (3): 10651076.

[9] Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function, The Annals of Mathematical Statistics. 27 (3): 832837

[10] AB Tsybakov, Introduction to Nonparametric Estimation, Springer Science & Business Media, 22 Oct 2008

[11] Larry Wasserman, All of nonparametric statistics. Springer Science & Business Media, 2006.

DEPARTMENT OF STATISTICS, THE HEBREW UNIVERSITY, MOUNT SCOPUS, JERUSALEM 91905, ISRAEL

E-mail address: Pavel.Chigansky@mail.huji.ac.il