SDS8102: Advanced Mathematical Statistics

Week 12 – Lecture 1
November 17, 2026
Lecturer: Alexandre Tsybakov

---

5 Fourier analysis of kernel density estimators (contd.)

In the previous lecture we have proved the following result.

Theorem 2. *Assume that $p \in L_2(\mathbb{R})$, and λ is a symmetric real-valued function in $L_2(\mathbb{R})$, then the MISE of $\hat{p}_{n,\lambda}$ is*

$$\mathbb{E}\|\hat{p}_{n,\lambda} - p\|_2^2 = \left\{\frac{1}{2\pi}\int |\phi(w)|^2(1-\lambda(w))^2 dw + \frac{1}{2\pi n}\int \lambda^2(w)dw\right\} - \frac{1}{2\pi n}\int |\phi(w)|^2\lambda^2(w)dw.$$

We will refer to the first and the second terms in the curly brackets as the bias squared term and the variance term respectively. The last term will be call the "bonus" term because it is negative and thus makes the error smaller. However, it is usually negligible as compared to the bias and variance terms.

Remarks:

- Since $|\phi(w)| \le 1$, $\phi \in L_2(\mathbb{R})$ and $\lambda \in L_2(\mathbb{R})$, all the integrals in the theorem are finite.
- If $\lambda(w) = \hat{K}(hw)$ where K is a symmetric kernel, we can analyze the terms separately:
  1. *The variance term*:
  $$\frac{1}{2\pi n}\int \hat{K}^2(hw)dw = \frac{1}{2\pi nh}\int \hat{K}^2(w)dw = \frac{1}{nh}\int K^2(u)du.$$
  This is similar to the bound on the variance term obtained when analyzing the behavior of kernel estimator at a fixed point, but now it does not depend on the density. Its order of magnitude is $\frac{1}{nh}$, which is much larger than $\frac{1}{n}$ if h goes to 0.
  2. *The bonus term* is typically small. In fact, assume that $\hat{K} \in L_\infty(\mathbb{R})$ (i.e., $|\hat{K}(\cdot)|$ is uniformly bounded by a constant, which is satisfied for all the kernels we are interested in), then
  $$\begin{aligned}\frac{1}{2\pi n}\int |\phi(w)|^2\hat{K}^2(hw)dw &\le \frac{\|\hat{K}\|_\infty^2}{2\pi n}\int \phi^2(w)dw \\ &= \frac{\|\hat{K}\|_\infty^2}{n}\int p^2(x)dx = O(1/n).\end{aligned}$$
  Note that if $K \in L_1(\mathbb{R})$, then $\hat{K} \in L_\infty(\mathbb{R})$, because
  $$|\hat{K}(w)| = |\int e^{iwx}K(x)dx| \le \int |K| < \infty.$$
  Even if we do not have $K \in L_1(\mathbb{R})$, we can still have $\hat{K} \in L_\infty(\mathbb{R})$. For example, the sinc kernel has Fourier transform
  $$\hat{K}(w) = I(|w| \le 1).$$

3. *The bias squared term* $\frac{1}{2\pi}\int |\phi(w)|^2(1-\lambda(w))^2 dw$.

We have previously proved that the squared bias is of the order $h^{2\beta}$ if $p \in \Sigma(\beta, L)$. The squared error is computed either at a fixed point or integrated over a bounded interval. Now, on the real line, we will see that we also have this rate under some assumptions, and the rate is related to the smoothness of the density p.

Recall that if β is an integer, then $p \in \Sigma(\beta, L)$ "roughly" means that the βth derivative of p is uniformly bounded on the real line. We now consider the densities such that their derivative is bounded in $L_2(\mathbb{R})$. Let m be an integer. To characterize the set of densities that satisfy $\|p^{(m)}\|_2 \le L$ (the mth derivative is bounded in $L_2(\mathbb{R})$), introduce the Sobolev class of densities.

Assume first that $\phi \in L_1(\mathbb{R})$. Then, by the Fourier inversion formula we can write

$$p(x) = \frac{1}{2\pi}\int e^{-iwx}\phi(w)dw.$$

Formally differentiating m times under the sign of the intergal we obtain

$$\begin{aligned} p'(x) &= \frac{1}{2\pi}\int (-iw)e^{-iwx}\phi(w)dw, \\ &\cdots \\ p^{(m)}(x) &= \frac{1}{2\pi}\int (-iw)^m e^{-iwx}\phi(w)dw \end{aligned}$$

provided that the integrals on the right hand sides are finite. Thus, under appropriate conditions, the Fourier transform of $p^{(m)}$ is $(-iw)^m\phi(w)$, and by Plancherel's equality,

$$\|p^{(m)}\|_2^2 = \frac{1}{2\pi}\int |w|^{2m}|\phi(w)|^2 dw.$$

These are heuristic arguments, but it can be proved rigorously that $\|p^{(m)}\|_2 \le L$ if and only if

$$\int |w|^{2m}|\phi(w)|^2 dw \le 2\pi L^2.$$

Here is the formal definition of the Sobolev class.

Definition 3. *Let $\beta > 0$, $L > 0$. The Sobolev class $\mathcal{P}_S(\beta, L)$ is defined as*

$$\mathcal{P}_S(\beta, L) = \{p : p \text{ is a probability density on } \mathbb{R}, \int |w|^{2\beta}|\phi(w)|^2 dw \le 2\pi L^2\},$$

where ϕ denotes the characteristic function of p.

Now we can analyze the bias term.

Theorem 3. *Let $p \in L_2(\mathbb{R})$, $\beta > 0$. Let $K \in L_2(\mathbb{R})$ be a symmetric kernel such that there exists a version of its Fourier transform $\hat{K}$ satisfying*

$$\sup_{u\in\mathbb{R}\setminus\{0\}} \frac{|1-\hat{K}(u)|}{|u|^\beta} \le A < \infty. \tag{1}$$

Then the kernel estimator $\hat{p}_n$ with bandwidth $h = \alpha n^{-\frac{1}{2\beta+1}}$, $\alpha > 0$, satisfies

$$\sup_{p \in \mathcal{P}_S(\beta, L)} \mathbb{E}\|\hat{p}_n - p\|_2^2 \leq C n^{-\frac{2\beta}{2\beta+1}},$$

where C is a positive constant depending only on L, α, A and the kernel K.

Proof: The condition of the theorem allows us to bound the bias squared term.

$$\begin{aligned}\frac{1}{2\pi}\int |\phi(w)|^2(1-\hat{K}(hw))^2 dw \leq & \frac{1}{2\pi}\int |\phi(w)|^2 A^2 |hw|^{2\beta} dw \\ = & \frac{A^2}{2\pi} h^{2\beta} \int |\phi(w)|^2 |w|^{2\beta} dw.\end{aligned}$$

The integral is bounded by $2\pi L^2$ (definition of $\mathcal{P}_S(\beta, L)$), so the bias squared term is less than or equal to $A^2 L^2 h^{2\beta}$. Combining with the variance term we get

$$\mathbb{E}\|\hat{p}_n - p\|_2^2 \leq \frac{1}{nh}\int K^2(u) du + L^2 A^2 h^{2\beta}.$$

The minimum is obtained when h is of the order $n^{-\frac{1}{2\beta+1}}$. We have chosen $h = \alpha n^{-\frac{1}{2\beta+1}}$. The desired result follows. ■

Remark:

- When u approaches 0, condition of the theorem implies that $\hat{K}(0) = 1$. This can be viewed as extension of the condition $\int K(u)du = 1$. Indeed, if $K \in L_1(\mathbb{R})$, then $\hat{K}(0) = \int K(u)du$. If not, like in the case of the sinc kernel, this is some sort of approximation.

A natural question is: Which kernels K satisfy the condition of the theorem and what are the corresponding β? Consider some examples.

Kernel	$K(u)$	$\hat{K}(w)$
Gaussian	$\frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$	$e^{-\frac{w^2}{2}}$
Epanechnikov	$\frac{3}{4}(1-u^2) I(\lvert u\rvert \leq 1)$	$\frac{3}{w^3}(\sin w - w\cos w)$
Sinc	$\frac{\sin u}{\pi u}$	$I(\lvert w\rvert \leq 1)$
Spline	$*$	$\frac{1}{1+\lvert w\rvert^{\beta'}}$, $\beta' > 0$,
Pinsker	$*$	$(1-\lvert w\rvert^{\beta'})_+$, $\beta' > 0$.

$*$ explicit expression available for integer β'.

1. *Gaussian kernel.*
$$\frac{\left|1-e^{-\frac{w^2}{2}}\right|}{|w|^\beta} \le \begin{cases} |w|^{-\beta}, & |w|>1, \\ \frac{w^2/2}{|w|^\beta}, & |w|\le 1. \end{cases}$$
This is bounded by 1 if $\beta \le 2$. If $\beta > 2$, the supremum over $\mathbb{R}\backslash\{0\}$ is infinity.

2. *Spline kernel with parameter β'.*
$$\frac{\left|1-\frac{1}{1+|w|^{\beta'}}\right|}{|w|^\beta} = \frac{|w|^{\beta'-\beta}}{1+|w|^{\beta'}}.$$
Condition 1 of Theorem 3 holds if and only if $\beta \le \beta'$.

3. *Pinsker kernel with parameter β'.*
$$\frac{|1-(1-|w|^{\beta'})_+|}{|w|^\beta} = \begin{cases} |w|^{-\beta}, & |w|>1, \\ |w|^{\beta'-\beta}, & |w|\le 1. \end{cases}$$
Condition 1 of Theorem 3 holds if and only if $\beta \le \beta'$.

4. *Sinc kernel.*
$$\frac{|1-I(|w|\le 1)|}{|w|^\beta} = \begin{cases} |w|^{-\beta}, & |w|>1, \\ 0, & |w|\le 1. \end{cases}$$
The supremum over $w \in \mathbb{R}\backslash\{0\}$ is bounded for all β. Such a kernel is called infinite power kernel or super-kernel. There exist other examples of super-kernels. The Fourier transform of a super-kernel needs to take value 1 in some neighborhood of 0.

The sinc kernel is not in $L_1(\mathbb{R})$. But one can also obtain a super-kernel in $L_1(\mathbb{R})$ as shows the next exercise.

Exercise: Construct a kernel such that its Fourier transform is this piecewise continuous function:
$$\hat{K}(w) = \begin{cases} 0 & \text{if } |w|>a, \\ 1 & \text{if } |w|<b, \\ \text{linear} & \text{otherwise}, \end{cases}$$
where $0<b<a$. This "trapeze" Fourier trasform corresponds essentially to what is called the *de la Vallée Poussin* kernel. *Idea of the construction*: Consider first
$$\hat{K}_0 = \hat{K}_{sinc} * \hat{K}_{sinc},$$
where $*$ stands for the convolution and $\hat{K}_{sinc}$ is the Fourier transform of the sinc kernel. Since convolution of the Fourier transforms corresponds to multiplication of the corresponding kernels, we have, for a constant $C>0$,
$$K_0(u) = C\left(\frac{\sin x}{\pi x}\right)^2 \in L_1(\mathbb{R}).$$
Then, take a sum of shifts of the Fourier transform $\hat{K}_0$ to the left and to the right to obtain the desired result. The obtained kernel is a super-kernel. Its advantage over the sinc kernel is in the fact that it belongs to $L_1(\mathbb{R})$.

6 Unbiased risk estimation. Cross-validation

Let $\hat{p}_n$ be the kernel density estimator and let the kernel K be fixed. We already know that the bandwidth h is crucial to determine the behavior of the estimator. How to choose h in practice? Consider the risk

$$MISE(h) = \mathbb{E} \int (\hat{p}_n - p)^2(x)dx,$$

where $\hat{p}_n$ is the kernel density estimator with bandwidth h. The optimal value of h is the one that minimizes the $MISE$, that is,

$$h^* \in \operatorname*{argmin}_{h>0} MISE(h). \tag{2}$$

Assume here that the minimum is attained, so that such a minimizer h^* exists. This ideal bandwidth h^* depends on the true density p, so it is not available in practice. It is called the *oracle bandwidth*, and the estimator $\hat{p}_n$ with bandwidth $h = h^*$ is called the *oracle*. We would like to "mimic the oracle", i.e., to find a bandwidth $\hat{h}_n$ that only depends on the data $X_1, \ldots, X_n$, such that its risk is close to the risk of the oracle:

$$\mathbb{E}\big[||\tilde{p}_{n,\hat{h}_n} - p||_2^2\big] \approx \min_{h>0} MISE(h),$$

where $\tilde{p}_{n,h}$ denotes the kernel estimator $\hat{p}_n$ with bandwidth h. It turns out that this task can be achieved. The idea is to first estimate the $MISE(\cdot)$, and then to minimize in h the obtained estimator of $MISE(\cdot)$.

Note that the $MISE$ can be written as

$$MISE(h) = \mathbb{E} \int (\hat{p}_n - p)^2 = \mathbb{E}[\int \hat{p}_n^2 - 2 \int \hat{p}_n p] + \int p^2.$$

Only the expression in the square brackets depends on h. The last term is constant in h. Let

$$J(h) = \mathbb{E}[\int \hat{p}_n^2 - 2 \int \hat{p}_n p].$$

Since we are minimizing over h, minimizing $MISE(h)$ is equivalent to minimizing $J(h)$. Therefore, it is enough to look for an estimator of $J(h)$ denoted by $\hat{J}(h)$, because $MISE(h)$ and $J(h)$ have the same minimizers. A first idea is to take an unbiased estimator of $J(h)$. Write

$$\hat{J}(h) = \int \hat{p}_n^2 - \hat{A},$$

where $\hat{A}$ is a statistic to be determined. To make $\hat{J}(h)$ an unbiased estimator of $J(h)$, we need to find $\hat{A}$, such that

$$\mathbb{E}(\hat{A}) = 2\mathbb{E} \int \hat{p}_n p.$$

Rudemo(1982)[1] suggested the following estimator:

$$\hat{J}(h) = CV(h) \triangleq \int \hat{p}_n^2 - \frac{2}{n} \sum_{i=1}^{n} \hat{p}_{n,-i}(X_i),$$

[1] Rudemo, M. (1982) Empirical Choice of Histograms and Kernel Density Estimators. *Scandinavian Journal of Statistics*, 9, 65-78.

where CV stands for cross-validation, and $\hat{p}_{n,-i}$ is a *leave-one-out* estimator:

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

Define the cross-validated bandwidth and the cross-validated kernel density estimator:

$$\hat{h}_{CV} \in \operatorname*{argmin}_{h>0} CV(h),$$
$$\tilde{p}_n^{CV}(x) = \frac{1}{n\hat{h}_{CV}} \sum_{i=1}^{n} K\left(\frac{X_i - x}{\hat{h}_{CV}}\right).$$

Note that $\tilde{p}_n^{CV}$ is not a kernel density estimator as defined above because the bandwidth $\hat{h}_{CV}$ is a random variable depending on $X_1, \ldots, X_n$.

Now we prove that $CV(h)$ is an unbiased estimator of $J(h)$, i.e., we show that

$$\mathbb{E}[\int \hat{p}_n p] = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{p}_{n,-i}(X_i)\right). \tag{3}$$

Since $X_1, \ldots, X_n$ are i.i.d., the right hand side of (3) is equal to

$$\begin{aligned}
\mathbb{E}(\hat{p}_{n,-1}(X_1)) &= \mathbb{E}_{X_2,\ldots,X_n} \int \hat{p}_{n,-1}(x)p(x)dx \\
&= \mathbb{E}\left[\frac{1}{(n-1)h}\sum_{j=2}^{n} \int K\left(\frac{X_j - x}{h}\right)p(x)dx\right] \\
&= \frac{1}{h}\int\int K\left(\frac{z-x}{h}\right)p(z)p(x)dzdx.
\end{aligned}$$

This integral is finite if K is bounded in absolute value. The left hand side of (3) is equal to

$$\mathbb{E}\left[\int \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)p(x)dx\right],$$

which is the right hand side of (3). ∎