NONPARAMETRIC KERNEL DENSITY ESTIMATION

In this lecture, we discuss kernel estimation of probability density functions (PDF). Nonparametric density estimation is one of the central problems in statistics. In economics, nonparametric density estimation plays important roles in various areas such as, for example, industrial organization (Guerre et al., 2000), empirical finance (Ait-Sahalia, 1996), and etc. These notes borrow from the following sources: Li and Racine (2007), Pagan and Ullah (1999), and Härdle and Linton (1994).

Kernel density estimator

Assumption 1 (a) Suppose \( \{X_i : i = 1, \ldots, n\} \) is a collection of iid random variables drawn from a distribution with the CDF \( F \) and PDF \( f \).

(b) In the neighborhood \( N_x \) of \( x \), \( f \) is bounded and twice continuously differentiable with bounded derivatives.

(When discussing \( f(x) \), we will implicitly assume that \( f(x) \) exists at \( x \).) The econometrician’s objective is to estimate \( f \) without imposing any functional form (parametric) assumptions on the PDF.

First, consider estimation of \( F \). Since \( F(x) = \mathbb{E}1\{X_i \leq x\} \), an estimator of \( F \) can be constructed as

\[
\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq x\}.
\]

The function \( \hat{F}_n(x) \) is called the empirical CDF of \( X_i \). The WLLN implies that for all \( x \),

\[
\hat{F}_n(x) \to_p F(x).
\]

As a matter of fact, a stronger results can be established (Glivenko-Gantelli Theorem, see Chapter 19.1 of van der Vaart (1998)):

\[
\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| \to_{a.s.} 0.
\]

Next, by the CLT,

\[
n^{1/2} \left( \hat{F}_n(x) - F(x) \right) \to_d N(0,F(x)(1-F(x))).
\]

Furthermore, for any \( x_1, x_2 \in \mathbb{R} \), \( n^{1/2} \left( \hat{F}_n(x_1) - F(x_1) \right) \) and \( n^{1/2} \left( \hat{F}_n(x_2) - F(x_2) \right) \) are jointly asymptotically normal with mean zero and the covariance

\[
F(x_1 \wedge x_2) - F(x_1) F(x_2),
\]

where \( x_1 \wedge x_2 \) denotes the minimum between \( x_1 \) and \( x_2 \).

Since

\[
f(x) = \frac{dF(x)}{dx} = \lim_{h \to 0} \frac{F(x + h) - F(x - h)}{2h},
\]

from (1), one can consider the following estimator for the PDF \( f \):

\[
\hat{f}_n(x) = \frac{\hat{F}_n(x + h_n) - \hat{F}_n(x - h_n)}{2h_n} = \frac{1}{n} \sum_{i=1}^{n} 1\{x-h_n \leq X_i \leq x+h_n\}.
\]
where \( h_n \) is a small number (note that we consider continuously distributed random variables, so that \( P(X_i = x - h_n) = 0 \)). We write \( h_n \) instead of just \( h \) because, typically, it will be a function of the sample size \( n \) such that \( \lim_{n \to \infty} h_n = 0 \). Now, define the following kernel function:

\[
K(u) = \frac{1}{2} \cdot 1 \{|u| \leq 1\}.
\] (2)

Then, the kernel PDF estimator is given by

\[
\hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n}\right).
\] (3)

Thus, with the kernel function \( K \) defined according to (2), the kernel density estimator is an average number of observations in the small neighborhood of \( x \) as defined by the smoothing parameter or bandwidth (also kernel window).

The kernel function in (2) is called uniform, because it corresponds to the uniform distribution (we have that \( \int K(u) \, du = 1 \)). It has a disadvantage of giving equal weights to all observations inside the \( h_n \)-window with the center at \( x \), regardless of how close they are to the center. Also, if one considers \( \hat{f}_n(x) \) as a function of \( x \), it is rough having jumps at the points \( X_i \pm h \), and has a zero derivative everywhere else. Those problems can be resolved if one considers alternative kernel functions, for example, the quadratic kernel:

\[
K(u) = \frac{15}{16} (1 - u^2) \cdot 1 \{|u| \leq 1\}.
\]

The class of estimators (3) with a kernel satisfying \( \int K(u) \, du = 1 \) is referred to as Rosenblatt-Parzen Kernel Estimator.

**Small sample properties of the kernel density estimator**

We will make the following assumption concerning \( K \):

**Assumption 2** (a) \( \int K(u) \, du = 1 \).

(b) \( K(u) = K(-u) \).

(c) \( K \) is compactly supported on \([-1, 1] \) and bounded.

(d) \( \int u^2K(u) \, du \neq 0 \).

The kernel density estimator is biased:

**Lemma 1** Under Assumptions 1(a) and 2(a), \( E\hat{f}_n(x) - f(x) = \int K(u) (f(x + uh_n) - f(x)) \, du \).

**Proof.**

\[
E\hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} EK\left(\frac{X_i - x}{h_n}\right)
= h_n^{-1}EK\left(\frac{X_i - x}{h_n}\right)
= h_n^{-1} \int K\left(\frac{u - x}{h_n}\right) f(u) \, du.
\]

Next, using change of variable \( y = (u - x) / h_n, u = x + yh_n, \) and \( du = h_n \, dy \), we obtain

\[
E\hat{f}_n(x) = \int K(u) f(x + uh_n) \, du,
\]

and the result follows since \( f(x) \int K(u) \, du = f(x) \) by Assumption 2(a).
Lemma 2 Under Assumptions 1(a) and 2(a), the variance of \( \hat{f}_n(x) \) is given by

\[
\text{Var} \left( \hat{f}_n(x) \right) = \frac{1}{nh_n} \int K^2(u) f(x + uh_n) \, du - \frac{1}{n} \left( \int K(u) f(x + uh_n) \, du \right)^2.
\]

Proof. Since the data are iid,

\[
\text{Var} \left( \hat{f}_n(x) \right) = \frac{1}{n} \text{Var} \left( \frac{1}{h_n} K \left( \frac{X_i - x}{h_n} \right) \right)
\]

\[
= \frac{1}{n} \left( \frac{1}{h_n^2} \text{EK}^2 \left( \frac{X_i - x}{h_n} \right) - \left( \frac{1}{h_n} \text{EK} \left( \frac{X_i - x}{h_n} \right) \right)^2 \right).
\]

By the same change of variable argument as in the proof of Lemma 1, we obtain

\[
\text{EK}^2 \left( \frac{X_i - x}{h_n} \right) = \int K^2 \left( \frac{u - x}{h_n} \right) f(u) \, du
\]

\[
= h_n \int K^2(u) f(x + uh_n) \, du.
\]

From Lemma 1, one can expect that the bias increases with \( h_n \); a bigger bandwidth implies that more observations away from \( x \) have non-zero weights which contributes to the bias. On the other hand, the variance decreases with \( h_n \), as the estimator averages over more observations. The theorem below establishes more formally the bias-variance trade-off for the kernel estimator.

Let \( f^{(s)} \) denote the \( s \)-order derivative of \( f \):

\[
f^{(s)}(x) = \frac{d^s f(x)}{dx^s}.
\]

Theorem 1 Suppose that \( h_n \to 0 \) and \( nh_n \to \infty \) as \( n \to \infty \). Then, under Assumptions 1 and 2,

(a) \( E\hat{f}_n(x) - f(x) = c_1(x) h_n^2 + O(h_n^2) \), where \( c_1(x) = f^{(2)}(x) \int u^2 K(u) \, du/2 \).

(b) \( \text{Var} \left( \hat{f}_n(x) \right) = c_2(x) / (nh_n) + O(1/n) \), where \( c_2(x) = f(x) \int K^2(u) \, du \).

Proof. Since the first two derivative of \( f \) exist by Assumption 1(b), consider the following expansion for \( f(x + uh_n) \):

\[
f(x + uh_n) = f(x) + f^{(1)}(x) uh_n + \frac{1}{2} f^{(2)}(x_n(u)) u^2 h_n^2,
\]

where \( x_n(u) \) lies between \( x \) and \( x + uh_n \). From Lemma 1 we have

\[
E\hat{f}_n(x) - f(x) = \int K(u) \left( f^{(1)}(x) uh_n + \frac{1}{2} f^{(2)}(x_n(u)) u^2 h_n^2 \right) \, du
\]

\[
= \frac{h_n^2}{2} \int K(u) f^{(2)}(x_n(u)) u^2 \, du
\]

\[
= c_1 h_n^2 + \frac{h_n^2}{2} \int K(u) \left( f^{(2)}(x_n(u)) - f^{(2)}(x) \right) u^2 \, du.
\]

The second equality follows because by Assumption 2(b), \( \int K(u) \, du = 0 \).

We will show next that

\[
\int K(u) \left( f^{(2)}(x_n(u)) - f^{(2)}(x) \right) u^2 \, du = o(1),
\]

(5)
and therefore the second summand in (4) is \( o \left( h_n^2 \right) \). By Assumption 2(c), we only need to consider \( |u| \leq 1 \); by Assumption 1(b) and since \( x_n (u) \) lies between \( x \) and \( x + uh_n \),
\[
\left| f^{(2)} (x_n (u)) - f^{(2)} (x) \right| 
\leq 2 \sup_{z \in \mathcal{N}_x} f^{(2)} (z) < \infty.
\]
Next, since \( h_n \to 0 \),
\[
\lim_{n \to \infty} x_n (u) = x.
\]
Now, by the dominated convergence theorem,
\[
\lim_{n \to \infty} \int K (u) \left( f^{(2)} (x_n (u)) - f^{(2)} (x) \right) u^2 \, du 
= \int K (u) \left( \lim_{n \to \infty} f^{(2)} (x_n (u)) - f^{(2)} (x) \right) u^2 \, du = 0,
\]
which establishes (5) and concludes the proof of part (a) of the theorem.

For part (b),
\[
\frac{1}{nh_n} \int K^2 (u) f (x + uh_n) \, du = \frac{1}{nh_n} f (x) \int K^2 (u) \, du + \frac{1}{n} f^{(1)} (x) \int K^2 (u) \, du
+ \frac{h_n}{2n} \int K^2 (u) f^{(2)} (x_n (u)) u^2 \, du
= \frac{c_2}{nh_n} + O \left( \frac{h_n}{n} \right), \quad (6)
\]
since \( \int K^2 (u) \, du = 0 \) by symmetry (Assumption 2(b)), and \( \int K^2 (u) f^{(2)} (x_n (u)) u^2 \, du = O (1) \) as in the proof of part (a). The result of part (b) follows from (6) and Lemma 2. □

Again, Theorem 1 shows the bias-variance trade-off. The optimal choice of bandwidth can be found by minimizing some function that combines bias and variance, for example, the mean squared error (MSE):
\[
MSE \left( \hat{f}_n (x) \right) = E \left( \hat{f}_n (x) - f (x) \right)^2 
= Var \left( \hat{f}_n (x) \right) + \left( E \hat{f}_n (x) - f (x) \right)^2
= \frac{c_2 (x)}{nh_n} + c_1^2 (x) h_n^4 + O \left( \frac{1}{n} \right) + o \left( h_n^4 \right)
= \frac{c_2 (x)}{nh_n} + c_1^2 (x) h_n^4 + o \left( \frac{1}{nh_n} + h_n^2 \right). \quad (7)
\]
Minimization of the leading term of MSE gives
\[
4c_1^2 (x) h_n^3 = \frac{c_2 (x)}{nh_n^2}, \quad \text{or}
\]
\[
h_n = \left( \frac{c_2 (x)}{4c_1^2 (x)} \right)^{1/5} n^{-1/5} = \frac{\left( f (x) \int K^2 (u) \, du \right)^{1/5}}{\left( f^{(2)} (x) \int u^2 K (u) \, du \right)^{2/5}} n^{-1/5}. \]
When the optimal (in the MSE sense) bandwidth is selected, neither bias nor variance components of the MSE dominate each other asymptotically as \( \text{Var} = \text{Bias}^2 = O(n^{-4/5}) \). When the Integrated MSE criterion is employed, \( \int \text{MSE} \left( \hat{f}_n (x) \right) dx \), the optimal bandwidth becomes

\[
h_n = \frac{\left( \int K^2 (u) \, du \right)^{1/5}}{\left( \int (f^{(2)} (x))^2 \, dx \right)^{1/5} \left( \int u^2 K (u) \, du \right)^{2/5}} n^{-1/5}.
\]

Let \( \hat{\sigma}_n^2 \) denote the sample variance of the data. The following rules of thumb often used in practice:

\[
h_n = 1.364 \hat{\sigma}_n \left( \frac{\int K^2 (u) \, du}{\int u^2 K (u) \, du} \right)^{1/5} n^{-1/5},
\]

which is optimal for \( f(x) \sim N(\mu, \sigma^2) \), and

\[
h_n = 1.06 \hat{\sigma}_n n^{-1/5},
\]

which is optimal for \( f(x) \sim N(\mu, \sigma^2) \) and when \( K \) is the standard normal density.

**Consistency of the kernel density estimator**

Consistency of \( \hat{f}_n (x) \) follows immediately from Theorem 1 by Chebychev’s inequality.

**Corollary 1** Suppose that \( h_n \to 0 \) and \( nh_n \to \infty \) as \( n \to \infty \). Then, under Assumptions 1 and 2, \( \hat{f}_n (x) \to_p f(x) \).

**Proof.** By Chebychev’s inequality,

\[
P \left( \left| \hat{f}_n (x) - f(x) \right| > \varepsilon \right) \leq \frac{E \left( \hat{f}_n (x) - f(x) \right)^2}{\varepsilon^2} = \frac{c_2(x)}{\varepsilon^2 nh_n} + \frac{c_1(x) h_n^4}{\varepsilon^2} + o \left( \frac{1}{nh_n} + h_n^4 \right)
\]

\[
\to 0,
\]

where the second line is by (7). \( \blacksquare \)

A stronger result can be given, see Newey (1994). Suppose that \( f \) admits at least \( m \) continuous derivatives on some interval \([x_1, x_2] \); \( K \) has at least \( m \) continuous derivatives, is compactly supported and of order \( m \):

\[\int u^j K (u) \, du = 0 \text{ for all } j = 1, \ldots, m - 1; \int u^m K (u) \, du \neq 0, \text{ and } \int K (u) \, du = 1.\]

Then

\[
\sup_{x \in [x_1, x_2]} \left| \hat{f}_n (x) - f(x) \right| = O_p \left( \frac{nh_n}{\log n} \right)^{-1/2} + h^m.
\]

The derivatives of \( f \) can be estimated by the derivative of \( \hat{f}_n \), however, with a slower rates of convergence. Newey (1994) shows that

\[
\sup_{x \in [x_1, x_2]} \left| \hat{f}_n^{(k)} (x) - f^{(k)} (x) \right| = O_p \left( \frac{nh_n^{2k}}{\log n} \right)^{-1/2} + h^m.
\]
Asymptotic normality of the kernel density estimator

Write

\[ \hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K \left( \frac{X_i - x}{h_n} \right) \]

\[ = n^{-1} \sum_{i=1}^{n} v_{ni}, \text{ where} \]

\[ v_{ni} = \frac{1}{h_n} K \left( \frac{X_i - x}{h_n} \right). \]

Note that \( h_n \) and consequently \( v_{ni} \) depend on \( n \). The collection \( \{ v_{ni} : i = 1, \ldots, n \} : n \in \mathbb{N} \) is called a triangular array. In our case, under Assumption 1, \( v_{ni} \)'s are iid. The following CLT is available for independent triangular arrays (Lehmann and Romano, 2005, Corollary 11.2.1, page 427).

**Lemma 3 (Lyapounov CLT)** Suppose that for each \( n \), \( w_{n1}, \ldots, w_{nn} \) are independent. Assume that \( E w_{ni} = 0 \) and \( \sigma_{ni}^2 = E w_{ni}^2 < \infty \), and define \( s_n^2 = \sum_{i=1}^{n} \sigma_{ni}^2 \). Suppose further that for some \( \delta > 0 \) the following condition holds:

\[ \lim_{n \to \infty} \frac{1}{s_n^{2+\delta}} E |w_{ni}|^{2+\delta} = 0. \] (8)

Then,

\[ \sum_{i=1}^{n} w_{ni}/s_n \to_d N(0,1). \]

The condition (8) is called Lyapounov’s condition. When the data are not just independent but iid, the Lyapounov’s condition can be simplified as follows (Davidson, 1994, Theorem 23.12 on page 373).

**Lemma 4** The Lyapounov’s condition is satisfied when \( w_{n1}, \ldots, w_{nn} \) are iid, \( \sigma_n^2 = E w_{ni}^2 > 0 \) uniformly in \( n \), and \( \lim_{n \to \infty} E |w_{ni}|^{2+\delta}/n^{\delta/2} = 0 \) for some \( \delta > 0 \).

**Proof.** Since the data are iid,

\[ s_n^2 = n\sigma_n^2, \text{ and} \]

\[ s_n^{2+\delta} = (n^{1/2}\sigma_n)^{2+\delta} \]

\[ = n^{1+\delta/2}\sigma_n^{2+\delta}. \]

We have

\[ \sum_{i=1}^{n} \frac{1}{s_n^{2+\delta}} E |w_{ni}|^{2+\delta} = \sigma_n^{-2-\delta} n^{-1-\delta/2} \sum_{i=1}^{n} E |w_{ni}|^{2+\delta} \]

\[ = \sigma_n^{-2-\delta} n^{-\delta/2} E |w_{ni}|^{2+\delta}. \]

Therefore, the Lyapounov’s condition is satisfied if \( n^{-\delta/2} E |w_{ni}|^{2+\delta} \to 0 \), since \( \sigma_n \) is uniformly bounded away from zero by the assumption. \( \blacksquare \)

Assuming that \( \lim_{n \to \infty} \sigma_n^2 \) exists, in the iid case, the result of Lyapounov CLT can be stated as follows.

**Corollary 2** Suppose that for each \( n \), \( w_{n1}, \ldots, w_{nn} \) are iid, \( E w_{ni} = 0 \), \( \lim_{n \to \infty} E w_{ni}^2 > 0 \) and finite, and \( \lim_{n \to \infty} E |w_{ni}|^{2+\delta}/n^{\delta/2} = 0 \) for some \( \delta > 0 \). Then,

\[ \frac{1}{n^{1/2}} \sum_{i=1}^{n} w_{ni} \to_d N \left( 0, \lim_{n \to \infty} E w_{ni}^2 \right). \]
Next, we prove asymptotic normality of the kernel density estimator.

**Theorem 2** Suppose that \( nh_n \to \infty \) and \((nh)^{1/2} h^2 \to 0\). Assume further that \( f(x) > 0 \). Then, under Assumptions 1 and 2,

\[
(nh_n)^{1/2} \left( \hat{f}_n(x) - f(x) \right) \to_d N \left( 0, f(x) \int K^2(u) \, du \right). \tag{9}
\]

Furthermore, for \( x_1 \neq x_2 \), \((nh_n)^{1/2} \left( \hat{f}_n(x_1) - f(x_1) \right)\) and \((nh_n)^{1/2} \left( \hat{f}_n(x_2) - f(x_2) \right)\) are asymptotically independent.

**Proof.**

\[
(nh_n)^{1/2} \left( \hat{f}_n(x) - f(x) \right) = \frac{1}{n^{1/2}} \sum_{i=1}^{n} \left( \frac{1}{h_n} K \left( \frac{X_i - x}{h_n} \right) - \frac{1}{h_n} EK \left( \frac{X_i - x}{h_n} \right) \right)
+ (nh_n)^{1/2} \left( \frac{1}{h_n} EK \left( \frac{X_i - x}{h_n} \right) - f(x) \right). \tag{10}
\]

By Theorem 1(a),

\[
\frac{1}{h_n} EK \left( \frac{X_i - x}{h_n} \right) - f(x) = O(h^2).
\]

Define

\[
w_{ni} = \frac{1}{h_n^{1/2}} K \left( \frac{X_i - x}{h_n} \right) - \frac{1}{h_n^{1/2}} EK \left( \frac{X_i - x}{h_n} \right).
\]

Then,

\[
(nh_n)^{1/2} \left( \hat{f}_n(x) - f(x) \right) = \frac{1}{n^{1/2}} \sum_{i=1}^{n} w_{ni} + O_p \left( (nh)^{1/2} h^2 \right)
= \frac{1}{n^{1/2}} \sum_{i=1}^{n} w_{ni} + o_p(1),
\]

where the equality in the second line is by the assumption that \((nh)^{1/2} h^2 \to 0\). It is now left to verify the conditions of Corollary 2.

By the definition of \(w_{ni}\),

\[
Ew_{ni} = 0.
\]

Next,

\[
Ew_{ni}^2 = \frac{1}{h_n} EK^2 \left( \frac{X_i - x}{h_n} \right) - \frac{1}{h_n} \left( EK \left( \frac{X_i - x}{h_n} \right) \right)^2. \tag{11}
\]

As in the proof of Lemma 1 and by the dominated convergence theorem,

\[
EK \left( \frac{X_i - x}{h_n} \right) = h_n \int K(u) f(x + uh_n) \, du
= O(h_n), \tag{12}
\]

so that the second summand in (11) is \(O(h_n)\) and asymptotically negligible. For the first term in (11), we can use the change of variable argument again:

\[
\frac{1}{h_n} EK^2 \left( \frac{X_i - x}{h_n} \right) = \frac{1}{h_n} \int K^2 \left( \frac{u - x}{h_n} \right) \, du
= \int K^2(u) f(x + uh_n) \, du,
\]

\[
\to f(x) \int K^2(u) \, du, \tag{13}
\]
where the last result is by the dominated convergence theorem. The results in (11)-(13) together imply that
\[
\lim_{n \to \infty} Ew^2_{ni} = f(x) \int K^2(u) \, du.
\]

Lastly, we show that \( E |w_{ni}|^{2+\delta} / n^{\delta/2} \to 0 \). We will use the \( c_r \) inequality (Davidson, 1994, Theorem 9.28 on page 140) in order to deal with \( E |w_{ni}|^{2+\delta} \): for \( r > 0 \),
\[
E \left| \sum_{i=1}^{m} X_i \right|^r \leq c_r \sum_{i=1}^{m} E |X_i|^r,
\]
where \( c_r = 1 \) when \( r \leq 1 \), and \( c_r = m^{r-1} \) when \( r \geq 1 \). Now, by the \( c_r \) inequality,
\[
E |w_{ni}|^{2+\delta} \leq 2^{1+\delta} \left( \frac{1}{h_n^{1+\delta/2}} E \left| K \left( \frac{X_i - x}{h_n} \right) \right|^{2+\delta} + \frac{1}{h_n^{1+\delta/2}} \left| EK \left( \frac{X_i - x}{h_n} \right) \right|^{2+\delta} \right).
\]
By (12),
\[
\frac{1}{h_n^{1+\delta/2}} \left| EK \left( \frac{X_i - x}{h_n} \right) \right|^{2+\delta} = O \left( h_n^{1+\delta/2} \right).
\]
Further,
\[
\frac{1}{h_n^{1+\delta/2}} E \left| K \left( \frac{X_i - x}{h_n} \right) \right|^{2+\delta} = \frac{1}{h_n^{1+\delta/2}} \int \left| K \left( \frac{u - x}{h_n} \right) \right|^{2+\delta} f(u) \, du
= \frac{1}{h_n^{\delta/2}} \int |K(u)|^{2+\delta} f(x + uh_n) \, du
= O \left( \frac{1}{h_n^{\delta/2}} \right),
\]
where the equality in the last line is again by the dominated convergence theorem. Hence,
\[
\frac{E |w_{ni}|^{2+\delta}}{n^{\delta/2}} = O \left( \left( \frac{1}{nh_n} \right)^{\delta/2} \right).
\]
This completes the proof of (9).

In order to show asymptotic independence of \( \hat{f}_n(x_1) \) and \( \hat{f}_n(x_2) \), consider their asymptotic covariance:
\[
\frac{1}{h_n} EK \left( \frac{X_i - x_1}{h_n} \right) K \left( \frac{X_i - x_2}{h_n} \right) = \frac{1}{h_n} \int K \left( \frac{u - x_1}{h_n} \right) K \left( \frac{u - x_2}{h_n} \right) f(u) \, du
= \int K(u) K \left( \frac{u + x_1 - x_2}{h_n} \right) f(x + uh_n) \, du.
\]
Since the kernel function is compactly supported and \( \lim_{n \to \infty} (x_1 - x_2) / h_n = \infty \),
\[
\lim_{n \to \infty} K \left( \frac{u + x_1 - x_2}{h_n} \right) = 0,
\]
and by the dominated convergence theorem,
\[
\lim_{n \to \infty} \int K(u) K \left( \frac{u + x_1 - x_2}{h_n} \right) f(x + uh_n) \, du = 0.
\]
Asymptotic independence then follows by the Cramer-Wold device. ■
From (10), one can see that the assumption \( (nh_n)^{1/2} h_n^2 \to 0 \) is used to make the bias asymptotically negligible. Consequently, there is under-smoothing relatively to the MSE-optimal bandwidth, and the bias goes to zero at a faster rate than the variance. Suppose that the bandwidth is chosen according to \( h_n = cn^{-\alpha} \). Then,

\[
(nh_n)^{1/2} h_n^2 \sim n^{1/2} n^{-\alpha/2} n^{-2\alpha} = n^{(1-5\alpha)/2},
\]

and for \( (nh_n)^{1/2} h_n^2 \to 0 \) to hold, we need that

\[
1 - 5\alpha < 0 \text{ or } \alpha > 1/5.
\]

Thus, for asymptotic normality, the bandwidth is \( o(n^{-1/5}) \), while the MSE-optimal bandwidth is \( h_n = cn^{-1/5} \).

A more general statement of the asymptotic normality result that also includes the bias result, i.e. without imposing under-smoothing is

\[
(nh_n)^{1/2} \left( \hat{f}_n(x) - f(x) - 0.5h_n f^{(2)}(x) \int u^2 K(u) du \right) \to_d N \left( 0, f(x) \int K^2(u) du \right).
\]

The result in (14) holds provided that \( nh_n \to \infty \) and does not require that \( (nh)^{1/2} h^2 \to 0 \). In particular, if one chooses \( h_n = ah_n^{-1/5} \), then

\[
(nh_n)^{1/2} \left( \hat{f}_n(x) - f(x) \right) \to_d N \left( \frac{1}{2} a^{5/2} f^{(2)}(x) \int u^2 K(u) du, f(x) \int K^2(u) du \right),
\]

and the kernel density estimator is asymptotically biased.

**Multivariate kernel density estimation and the curse of dimensionality**

Suppose now that \( \{X_i : i = 1, \ldots, n\} \) is a collection of iid random \( d \)-vectors drawn from a distribution with a joint PDF \( f(x_1, \ldots, x_d) \). The univariate kernel density estimator can be extended to the multivariate case as follows:

\[
\hat{f}_n(x_1, \ldots, x_d) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{d} \frac{1}{h_n} K \left( \frac{X_{ij} - x_j}{h_n} \right) = \frac{1}{nh_n^d} \sum_{i=1}^{n} \prod_{j=1}^{d} K \left( \frac{X_{ij} - x_j}{h_n} \right),
\]

(note \( h_n^d \) in the denominator instead of \( h_n \)). One can see that the multivariate kernel density estimator is an extension of univariate kernel smoothing to \( d \) dimensions or \( d \) variables.

In the multivariate case, one can establish results similar to those of the univariate case. To simplify the notation, let

\[
x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^d,
\]

and write

\[
f(x_1, \ldots, x_d) = f(x).
\]

Also for \( u = (u_1, \ldots, u_d)' \in \mathbb{R}_d \), let

\[
K_d(u) = \prod_{j=1}^{d} K(u_j),
\]

\[9\]
so that \( \hat{f}_n(x_1, \ldots, x_d) = \hat{f}_n(x) = (nh_n^d)^{-1} \sum_{i=1}^n K_d((X_i - x)/h_n) \). Note that

\[
\int K_d(u) du = \int K(u_1) du_1 \times \ldots \times K(u_d) du_d = \left( \int K(u_1) du_1 \right)^d = 1,
\]

where the second equality follows by Assumption 2(a).

Similar results to those shown for the univariate estimator can be established in the multivariate case.

**Assumption 3 (a)** Suppose \( \{X_i : i = 1, \ldots, n\} \) is a collection of iid random vectors drawn from the distribution with a joint PDF \( f \).

(b) In the neighborhood \( N_x \) of \( x \), \( f \) is bounded and twice continuously differentiable with bounded partial derivatives.

**Theorem 3** Suppose that \( h_n \to 0 \) and \( nh_n \to \infty \) as \( n \to \infty \). Then, under Assumptions 2 and 3,

(a) \( E\hat{f}_n(x) - f(x) = f(x) + \frac{1}{2} h_n^2 \left( \int u_1^2 K(u_1) du_1 \right) \sum_{j=1}^d \frac{\partial^2 f(x)}{\partial x_j^2} + o(h_n^2) \).

(b) \( \text{Var} \left( \hat{f}_n(x) \right) = \frac{1}{nh_n^2} f(x) \int K^2(u_1) du_1 + O \left( 1/(nh_n^{d-2}) + 1/n \right) \).

**Proof.** For part (a),

\[
E\hat{f}_n(x) = \frac{1}{h_n^d} \int K_d \left( \frac{u - x}{h_n} \right) f(u) du = \int K_d(v) f(x + h_nv) dv,
\]

where we used the change of variable \( v = (u - x)/h_n, u = x + h_nv, du_j = h_n dv_j \) for \( j = 1, \ldots, d \), and \( du = du_1 \times \ldots \times du_d, dv = dv_1 \times \ldots \times dv_d \). Next,

\[
f(x + h_nv) = f(x) + h_nv \frac{\partial f(x)}{\partial x} + \frac{1}{2} h_n^2 v \frac{\partial^2 f(x(v))}{\partial x \partial x'} v,
\]

where \( x_n(v) \) denotes the mean-value satisfying \( \|x_n(v) - x\| \leq h_n \|v\| \), i.e. it lies between \( x \) and \( x + h_nv \). Since the kernel function is symmetric around zero,

\[
\int v K_d(v) dv = 0.
\]

By Assumption 3(b) and the same arguments as in the proof of Theorem 1(a),

\[
\frac{\partial^2 f(x_n(v))}{\partial x \partial x'} - \frac{\partial^2 f(x)}{\partial x \partial x'} = o(1).
\]

Hence,

\[
E\hat{f}_n(x) = f(x) + \frac{1}{2} h_n^2 \int v \frac{\partial^2 f(x)}{\partial x \partial x'} v K_d(v) dv + o(h_n^2)
\]

\[
= f(x) + \frac{1}{2} h_n^2 \sum_{i=1}^d \sum_{j=1}^d v_i v_j \frac{\partial^2 f(x)}{\partial x_i \partial x_j} K_d(v) dv + o(h_n^2)
\]

\[
= f(x) + \frac{1}{2} h_n^2 \sum_{j=1}^d \int v_j^2 K(v_j) dv_j + \frac{1}{2} h_n^2 \sum_{i=1}^d \sum_{j \neq i} \frac{\partial^2 f(x)}{\partial x_i \partial x_j} \int v_i K(v_i) dv_i \int v_j K(v_j) dv_j + o(h_n^2)
\]

\[
= f(x) + \frac{1}{2} h_n^2 \left( \int v_1^2 K(v_1) dv_1 \right) \sum_{j=1}^d \frac{\partial^2 f(x)}{\partial x_j^2} + o(h_n^2),
\]
where the equality in the last line holds due to the symmetry of the kernel function $K(u)$ around zero.

For part (b),

$$
\text{Var}(\hat{f}_n(x)) = \frac{1}{n} \text{Var} \left( \frac{1}{h_n^d} K_d \left( \frac{X_i - x}{h_n} \right) \right)
= \frac{1}{n} \left[ \frac{1}{h_n^{2d}} EK_d^2 \left( \frac{X_i - x}{h_n} \right) - \left( \frac{1}{h_n^d} EK_d \left( \frac{X_i - x}{h_n} \right) \right)^2 \right]
= \frac{1}{n} \frac{1}{h_n^{2d}} EK_d^2 \left( \frac{X_i - x}{h_n} \right) - (f(x) + O(h_n^2))^2
= \frac{1}{n} \frac{1}{h_n^{2d}} EK_d^2 \left( \frac{X_i - x}{h_n} \right) - f^2(x) + O(h_n^2)
= \frac{1}{n} \frac{1}{h_n^{2d}} EK_d^2 \left( \frac{X_i - x}{h_n} \right) + O\left( \frac{1}{n} \right)
= \frac{1}{n} \frac{1}{h_n^{2d}} \int K_d^2(u) f(x + h_n u) du + O\left( \frac{1}{n} \right)
= \frac{1}{n} \frac{1}{h_n^d} \int K_d^2(u) f(x + h_n u) du + O\left( \frac{1}{n} \right)
= \frac{1}{n} \frac{1}{h_n^d} \int K_d^2(u) \left( f(x) + h_n u' \frac{\partial f(x)}{\partial x} + \frac{1}{2} h_n^2 u' \frac{\partial^2 f(x_n(v))}{\partial x \partial x'} u \right) du + O\left( \frac{1}{n} \right)
= \frac{1}{n} \frac{1}{h_n^d} f(x) \left( \int K^2(u) du \right)^d + O\left( \frac{h_n^2}{nh_n^d} + \frac{1}{n} \right),
$$

where the last line follows since $\int u K^2(u) du = 0$ due to the symmetry of the kernel function around zero, and because $\int K^2(u) du = (\int K^2(u) du)^d$.

The bias and variance calculations imply that

$$
\hat{f}_n(x) = f(x) + O_p \left( \frac{1}{\sqrt{nh_n^d}} + h_n^2 \right),
$$

and therefore the rate of convergence slows down with the number of variables $d$. In the nonparametric literature, this is referred to as the curse of dimensionality.

One can derive the MSE-point-optimal bandwidth considering the leading terms in the bias and variance expressions implied by Theorem 3:

$$
\text{MSE}_n(x) = \frac{1}{4} h_n^4 \left( \int u_1^2 K(u_1) du_1 \right)^2 \left( \sum_{j=1}^d \frac{\partial^2 f(x)}{\partial x_j^2} \right)^2 + \frac{f(x) \left( \int K^2(u_1) du_1 \right)^d}{nh_n^d}.
$$

Minimizing the MSE with respect to $h_n$, we obtain:

$$
h_n^3 \left( \int u_1^2 K(u_1) du_1 \right)^2 \left( \sum_{j=1}^d \frac{\partial^2 f(x)}{\partial x_j^2} \right)^2 = \frac{d f(x) \left( \int K^2(u_1) du_1 \right)^d}{nh_n^{d+1}}.
$$

Therefore, the MSE optimal bandwidth is given by

$$
h_n = \left( \frac{d f(x) \left( \int K^2(u_1) du_1 \right)^d}{\left( \int u_1^2 K(u_1) du_1 \right)^2 \left( \sum_{j=1}^d \frac{\partial^2 f(x)}{\partial x_j^2} \right)^2} \right)^{1/(d+4)} n^{-1/(d+4)}.\]
One can see that the rate of the optimal bandwidth, \( n^{-1/(d+4)} \), increases with the number of variables \( d \), i.e. one should use larger values for the bandwidth when there are more variables.

One can extend the univariate CLT to the multivariate case as follows:

\[
(nh_n^d)^{1/2} \left( \hat{f}_n(x) - f(x) - \frac{1}{2} h_n^2 \left( \int u_1^2 K(u_1) du_1 \right) \sum_{j=1}^d \frac{\partial^2 f(x)}{\partial x_j^2} h_j n \right) \rightarrow_d N \left( 0, f(x) \left( \int K^2(u_1) du_1 \right)^d \right).
\]

To eliminate the asymptotic bias, one has to choose an under-smoothing bandwidth so that

\[
(nh_n^d)^{1/2} h_n^2 \rightarrow 0.
\]

One can also incorporate different bandwidth values for different variables:

\[
\hat{f}_n(x_1, \ldots, x_d) = \frac{1}{n h_{11}h_{22} \ldots h_{dd}} \sum_{i=1}^n \prod_{j=1}^d K \left( \frac{X_{ij} - x_j}{h_{jj}} \right).
\]

The bias and variance results in this case take the following form:

\[
E \hat{f}_n(x) = f(x) + \frac{1}{2} \left( \int u_1^2 K(u_1) du_1 \right) \sum_{j=1}^d \frac{\partial^2 f(x)}{\partial x_j^2} h_j^2 n + o \left( h_{11}^2 + \ldots + h_{dd}^2 \right).
\]

\[
Var \left( \hat{f}_n(x) \right) = \frac{f(x) \left( \int K^2(u_1) du_1 \right)^d}{n h_{11} \times \ldots \times h_{dd}} + O \left( \frac{h_{11}^2 + \ldots + h_{dd}^2}{n h_{11} \times \ldots \times h_{dd}} + \frac{1}{n} \right).
\]

Analogously, the CLT statement can be modified as

\[
(nh_{11} \times \ldots \times h_{dd})^{1/2} \left( \hat{f}_n(x) - f(x) - \frac{1}{2} \left( \int u_1^2 K(u_1) du_1 \right) \sum_{j=1}^d \frac{\partial^2 f(x)}{\partial x_j^2} h_j^2 n \right) \rightarrow_d N \left( 0, f(x) \left( \int K^2(u_1) du_1 \right)^d \right).
\]
Bibliography


