Lecture 10: Density estimation and nonparametric regression

Density estimation

Suppose that \( X_1, \ldots, X_n \) are i.i.d. random variables from \( F \) and that \( F \) is unknown but has a Lebesgue p.d.f. \( f \).

Estimation of \( F \) can be done by estimating \( f \).

Note that estimators of \( F \) derived in §5.1.1 and §5.1.2 do not have Lebesgue p.d.f.’s.

Having a density estimator \( \hat{f} \), \( F \) can be estimated by \( \hat{F}(x) = \int_{-\infty}^{x} f(t) dt \), which may be better than \( F_n \).

\( \hat{f} \) itself may be of interest

Difference quotient

Since \( f(t) = F'(t) \), a simple estimator of \( f(t) \) is the difference quotient

\[
f_n(t) = \frac{F_n(t + \lambda_n) - F_n(t - \lambda_n)}{2\lambda_n}, \quad t \in \mathbb{R},
\]

where \( F_n \) is the empirical c.d.f. and \( \{\lambda_n\} \) is a sequence of positive constants.
Properties of difference quotient

Since $2n\lambda_n f_n(t)$ has the binomial distribution $Bi(F(t + \lambda_n) - F(t - \lambda_n), n)$,

$$E[f_n(t)] \rightarrow f(t) \quad \text{if } \lambda_n \rightarrow 0 \text{ as } n \rightarrow \infty$$

and

$$\text{Var}(f_n(t)) \rightarrow 0 \quad \text{if } \lambda_n \rightarrow 0 \text{ and } n\lambda_n \rightarrow \infty.$$ 

Thus, we should choose $\lambda_n$ converging to 0 slower than $n^{-1}$.

If we assume that $\lambda_n \rightarrow 0$, $n\lambda_n \rightarrow \infty$, and $f$ is continuously differentiable at $t$, then it can be shown (exercise) that

$$\text{mse}_{f_n(t)}(F) = \frac{f(t)}{2n\lambda_n} + o\left(\frac{1}{n\lambda_n}\right) + O(\lambda_n^2)$$

and, under the additional condition that $n\lambda_n^3 \rightarrow 0$,

$$\sqrt{n\lambda_n}[f_n(t) - f(t)] \rightarrow_d N(0, \frac{1}{2} f(t)).$$
Kernel density estimators

A useful class of estimators is the class of *kernel density estimators*:

$$\hat{f}(t) = \frac{1}{n\lambda_n} \sum_{i=1}^{n} w \left( \frac{t-X_i}{\lambda_n} \right),$$

where $w$ is a known Lebesgue p.d.f. on $\mathbb{R}$ and is called the kernel.

If we choose $w(t) = \frac{1}{2} I_{[-1,1]}(t)$, then $\hat{f}(t)$ is essentially the same as the so-called histogram.

Properties of kernel density estimator

$\hat{f}$ is a Lebesgue density on $\mathbb{R}$, since

$$\int_{-\infty}^{\infty} \hat{f}(t) dt = \frac{1}{n\lambda_n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} w \left( \frac{t-X}{\lambda_n} \right) dt = \int_{-\infty}^{\infty} w(y) dy = 1.$$ 

The bias of $\hat{f}(t)$ as an estimator of $f(t)$ is

$$E[\hat{f}(t)] - f(t) = \frac{1}{\lambda_n} \int w \left( \frac{t-Z}{\lambda_n} \right) f(z) dz - f(t)$$

$$= \int w(y) [f(t-\lambda_n y) - f(t)] dy$$
If $f$ is bounded and continuous at $t$, then, by the dominated convergence theorem, the bias of $\hat{f}(t)$ converges to 0 as $\lambda_n \to 0$.

If $f'$ is bounded and continuous at $t$ and $\int |t| w(t) dt < \infty$, then the bias of $\hat{f}(t)$ is $O(\lambda_n)$.

If $f''$ is bounded and continuous at $t$, $\int t w(t) dt = 0$, and $0 < \int t^2 w(t) dt < \infty$ (2nd order kernel), then the bias of $\hat{f}(t)$ is $O(\lambda_n^2)$.

If $f$ is bounded and continuous at $t$ and $w_0 = \int [w(t)]^2 dt < \infty$, then

$$\text{Var}(\hat{f}(t)) = \frac{1}{n\lambda_n^2} \text{Var} \left( w \left( \frac{t-X_1}{\lambda_n} \right) \right)$$

$$= \frac{1}{n\lambda_n^2} \int \left[ w \left( \frac{t-z}{\lambda_n} \right) \right]^2 f(z) dz$$

$$- \frac{1}{n} \left[ \frac{1}{\lambda_n} \int w \left( \frac{t-z}{\lambda_n} \right) f(z) dz \right]^2$$

$$= \frac{1}{n\lambda_n} \int [w(y)]^2 f(t - \lambda_n y) dy + O \left( \frac{1}{n} \right)$$

$$= \frac{w_0 f(t)}{n\lambda_n} + o \left( \frac{1}{n\lambda_n} \right)$$
Hence, if $w_0 < \infty$, $f'$ is bounded and continuous at $t$, then

$$\text{mse}_{\hat{f}(t)}(F) = \frac{w_0 f(t)}{n\lambda_n} + O(\lambda_n^2)$$

and the best rate $n^{-2/3}$ is achieved when $\lambda_n$ has order $n^{-1/3}$.

If $w_0 < \infty$, $f''$ is bounded and continuous at $t$ and $\int tw(t)dt = 0$, then

$$\text{mse}_{\hat{f}(t)}(F) = \frac{w_0 f(t)}{n\lambda_n} + O(\lambda_n^4)$$

and the best rate $n^{-4/5}$ is achieved when $\lambda_n$ has order $n^{-1/5}$.

If $\lambda_n \to 0$, $n\lambda_n \to \infty$, $f$ is bounded and continuous at $t$ and $w_0 < \infty$, then

$$\sqrt{n\lambda_n}\{\hat{f}(t) - E[\hat{f}(t)]\} \to_d N(0, w_0 f(t)).$$

This can be shown as follows.

Let $Y_{in} = w\left(\frac{t - X_i}{\lambda_n}\right)$.

Then $Y_{1n}, \ldots, Y_{nn}$ are independent and identically distributed with

$$E(Y_{1n}) = \int_{-\infty}^{\infty} w\left(\frac{t - x}{\lambda_n}\right) f(x)dx.$$
\[= \lambda_n \int_{-\infty}^{\infty} w(y) f(t - \lambda_n y) \, dy\]
\[= O(\lambda_n)\]

\[\text{Var}(Y_{1n}) = \int_{-\infty}^{\infty} \left[ w \left( \frac{t-x}{\lambda_n} \right) \right]^2 f(x) \, dx\]
\[- \left[ \int_{-\infty}^{\infty} w \left( \frac{t-x}{\lambda_n} \right) f(x) \, dx \right]^2\]
\[= \lambda_n \int_{-\infty}^{\infty} [w(y)]^2 f(t - \lambda_n y) \, dy + O(\lambda_n^2)\]
\[= \lambda_n w_0 f(t) + o(\lambda_n),\]

since \(f\) is bounded and continuous at \(t\) and \(w_0 = \int_{-\infty}^{\infty} [w(t)]^2 \, dt < \infty\).

Then
\[\text{Var}(\hat{f}(t)) = \frac{1}{n^2 \lambda_n^2} \sum_{i=1}^{n} \text{Var}(Y_{in}) = \frac{w_0 f(t)}{n \lambda_n} + o \left( \frac{1}{n \lambda_n} \right).\]

Note that \(\hat{f}(t) - E\hat{f}(t) = \sum_{i=1}^{n} [Y_{in} - E(Y_{in})]/(n \lambda_n).\)
To apply Lindeberg’s central limit theorem to $\hat{f}(t)$, we find, for $\varepsilon > 0$,

$$\frac{E(Y_{1n}^2 I_{\{\lfloor Y_{1n} - E(Y_{1n}) \rfloor > \varepsilon \sqrt{n\lambda_n}\}}} {\lambda_n}$$

$$= \int_{|w(y) - E(Y_{1n})| > \varepsilon \sqrt{n\lambda_n}} [w(y)]^2 f(t - \lambda_n y) dy,$$

Since $E(Y_{1n}) = O(\lambda_n)$, if $\lambda_n \to 0$ and $n\lambda_n \to \infty$, the set

$\{ |w(y) - E(Y_{1n})| > \varepsilon \sqrt{n\lambda_n} \}$

shrinks to empty as $n \to \infty$. This proves that Lindeberg’s condition is satisfied and thus

$$\sqrt{n\lambda_n} \{ \hat{f}(t) - E[\hat{f}(t)] \} \to_d N(0, w_0 f(t)).$$

Furthermore, if

$$E[\hat{f}(t)] - f(t) = O(\lambda_n)$$

then

$$\sqrt{n\lambda_n} \{ E[\hat{f}(t)] - f(t) \} = O\left( \sqrt{n\lambda_n \lambda_n} \right) \to 0$$

if $n\lambda_n^3 \to 0$, which implies that

$$\sqrt{n\lambda_n} \{ \hat{f}(t) - f(t) \} \to_d N(0, w_0 f(t)).$$
If
\[ E[\hat{f}(t)] - f(t) = O(\lambda_n^2) \]
then
\[ \sqrt{n\lambda_n}\{E[\hat{f}(t)] - f(t)\} = O\left(\sqrt{n\lambda_n}\lambda_n^2\right) \rightarrow 0 \]
if \( n\lambda_n^5 \rightarrow 0 \), which implies that
\[ \sqrt{n\lambda_n}\{f(t) - f(t)\} \rightarrow_d N(0, w_0f(t)) \).

In any case, the best choice of \( \lambda_n \) for the mse does not satisfy \( n\lambda_n^3 \rightarrow 0 \) or \( n\lambda_n^5 \rightarrow 0 \).

Example 5.4

An i.i.d. sample of size \( n = 200 \) was generated from \( N(0, 1) \).
Density curve estimates, difference quotient \( f_n \) (short dashed curve) and kernel estimate \( \hat{f} \) (long dashed curve), are plotted in Figure 5.1 with the curve of the true p.d.f. (solid curve)

For the kernel estimate, \( w(t) = \frac{1}{2} e^{-|t|} \) is used and \( \lambda_n = 0.4 \).

From Figure 5.1, it seems that the kernel estimate is much better than the difference quotient.
Figure 5.1. Density estimates in Example 5.4

- True p.d.f.
- Estimator (5.26)
- Estimator (5.29)
Nonparametric regression

In many applications we want to estimate the regression function

$$\mu(t) = E(Y_i|t) = E(Y_i|X_i = t)$$

based on a random sample \((Y_1, X_1), \ldots, (Y_n, X_n)\) from a population with a pdf \(f(x, y)\).

In nonparametric regression, we do not specify any form of \(\mu(t)\) except that it is a smooth function of \(t\).

A nonparametric estimator of \(\mu(t)\) based on a kernel \(w(t)\) is

$$\hat{\mu}(t) = \frac{\sum_{i=1}^{n} Y_i w\left(\frac{t - X_i}{\lambda_n}\right)}{\sum_{i=1}^{n} w\left(\frac{t - X_i}{\lambda_n}\right)}, \quad t \in \mathbb{R}$$

From the previous discussion on the kernel estimation of the pdf of \(X_i\), \(f(t)\), the denominator divided by \(n\lambda_n\) converges in probability to \(f(t)\) if \(\lambda_n \to 0\) and \(n\lambda_n \to \infty\).

Hence, for the consistency of \(\hat{\mu}(t)\) as an estimator of \(\mu(t)\), it suffices to show that, for any \(t \in \mathbb{R}\),
\[ h_n(t) = \frac{1}{n \lambda_n} \sum_{i=1}^{n} Y_i w \left( \frac{t - X_i}{\lambda_n} \right) \rightarrow_p \int yf(t, y) dy \]

Consider first the expectation:

\[ E[h_n(t)] = \frac{1}{\lambda_n} E \left[ Y_i w \left( \frac{t - X_i}{\lambda_n} \right) \right] = \frac{1}{\lambda_n} \int \int yw \left( \frac{t - x}{\lambda_n} \right) f(x, y) dx dy \]

\[ = \int \int yw(z) f(t - \lambda_n z, y) dz dy \]

Suppose that \( f(x, y) \) is continuous and \( f(x, y) \leq c(y)g(y) \), where \( g(y) \) is the pdf of \( Y_i \) and \( c(y) \) is a function of \( y \) satisfies

\[ E[|Y_i|c(Y_i)] = \int |y|c(y)g(y) dy < \infty \]

Then, if \( \lambda_n \rightarrow 0 \) as \( n \rightarrow \infty \), by the dominated convergence theorem,

\[ \lim_{n \rightarrow \infty} E[h_n(t)] = \lim_{n \rightarrow \infty} \int \int yw(z) f(t - \lambda_n z, y) dz dy \]

\[ = \int \int yw(z) f(t, y) dz dy \]
\[
= \int w(z)\,dz \int yf(t, y)\,dy \\
= \int yf(t, y)\,dy
\]

Thus, it remains to show that the variance of \( h_n(t) \) converges to 0 under some conditions.

\[
\text{Var}(h_n(t)) = \frac{1}{n\lambda_n^2} \text{Var} \left( Y_i w \left( \frac{t-X_i}{\lambda_n} \right) \right) \\
\leq \frac{1}{n\lambda_n^2} \mathbb{E} \left[ Y_i w \left( \frac{t-X_i}{\lambda_n} \right) \right]^2 \\
= \frac{1}{n\lambda_n^2} \int \int y^2 \left[ w \left( \frac{t-x}{\lambda_n} \right) \right]^2 f(x, y)\,dx\,dy \\
= \frac{1}{n\lambda_n} \int \int y^2 [w(z)]^2 f(t - \lambda_n z, y)\,dz\,dy
\]

Suppose that \( f(x, y) \) is continuous and \( f(x, y) \leq c(y)g(y) \), where \( g(y) \) is the pdf of \( Y_i \) and \( c(y) \) is a function of \( y \) satisfies

\[
\mathbb{E}[Y_i^2 c(Y_i)] = \int y^2 c(y)g(y)\,dy < \infty
\]
Also, assume \( w_0 = \int [w(z)]^2 dz < \infty \) and \( E(Y_i^2) < \infty \).

Then
\[
\lim_{n \to \infty} \int \int y^2 [w(z)]^2 f(t - \lambda_n z, y)dzdy = \int \int y^2 [w(z)]^2 f(t, y)dzdy \\
= \int [w(z)]^2 dz \int y^2 f(t, y)dy < \infty
\]

Hence,
\[
\text{Var}(h_n(t)) = O \left( \frac{1}{n\lambda_n} \right)
\]
which converges to 0 if \( n\lambda_n \to \infty \).

Under some more conditions, similar to the estimation of \( f(t) \), for any \( t \in \mathbb{R} \), we can show that for some function \( \sigma^2(t) \),
\[
\sqrt{n\lambda_n} [\hat{\mu}(t) - \mu(t)] \text{ converges in distribution to } N(0, \sigma^2(t))
\]
Note that \( \hat{\mu}(t) \) is a ratio estimator \( h_n(t)/\hat{f}(t) \).
Kernel estimators of \( \mu(t) = E(Y_i | X_i = t) \) have convergence rates slower than \( n^{-1/2} \).

However, the convergence rate is \( n^{-1/2} \) if we average kernel estimators.

For example, we can estimate \( \mu = E(Y_i) = E[E(Y_i | X_i)] = E[\mu(X_i)] \) by

\[
\hat{\mu} = \frac{\sum_{j=1}^{n} \sum_{i=1}^{n} Y_i w \left( \frac{X_j - X_i}{\lambda_n} \right)}{\sum_{j=1}^{n} \sum_{i=1}^{n} w \left( \frac{X_j - X_i}{\lambda_n} \right)}
\]

a ratio of V-statistics (but the kernel of V-statistics depending on \( \lambda_n \)).

Under some conditions, it can be shown that

\[
\sqrt{n}(\hat{\mu} - \mu) \text{ converges in distribution to } N(0, \sigma^2)
\]

for some \( \sigma^2 \).

Conditions on \( \lambda_n \): for some constant \( C > 0 \),

\[
\lambda_n = Cn^{-s}, \quad \frac{1}{2} < s < 1 \quad \text{or} \quad \frac{1}{4} < s < 1 \quad \text{if } \int tw(t)dt = 0
\]

This is not the best choice \( (s = 1/3 \text{ or } 1/5) \) for estimating \( \mu(t) \) with a fixed \( t \).
$k$-nearest neighbor ($k$-NN) estimators

The kernel estimator

$$\hat{\mu}(t) = \frac{\sum_{i=1}^{n} Y_i w\left(\frac{t - X_i}{\lambda_n}\right)}{\sum_{i=1}^{n} w\left(\frac{t - X_i}{\lambda_n}\right)}, \quad t \in \mathbb{R}$$

is a weighted average of $Y_i$’s in a fixed neighborhood around $t$, determined in shape by the kernel $w$ and the bandwidth $\lambda_n$. The $k$-NN estimator is a weighted average in a varying neighborhood defined through those $X_i$’s which are among the $k$-nearest neighbors of $t$ in Euclidean distance:

$$\tilde{\mu}(t) = \sum_{i=1}^{n} Y_i W_{ki}(t)$$

where

$$W_{ki} = \begin{cases} 
1/k & i \in X_i \text{ is one of the } k \text{ nearest observations to } t \\
0 & \text{otherwise}
\end{cases}$$

Example

$$(X_i, Y_i)'s = (1,5), (7,12), (3,1), (2,0),(5,4)$$
\( n = 5, \, k = 3, \, t = 4. \)
The 3 nearest neighbors to \( t = 4 \) are 3 \((i = 3)\), 2 \((i = 4)\), 5 \((i = 5)\)
\( W_{k1}(4) = 0, \, W_{k2}(4) = 0, \, W_{k3}(4) = 1/3, \, W_{k4}(4) = 1/3, \, W_{k5}(4) = 1/3 \)
Thus, \( \hat{\mu} = (1 + 0 + 4)/3 = 5/3. \)

**Asymptotic theory**
- To reduce noise we need let \( k \) tend to infinity as a function of \( n \).
- To keep the approximation error (bias) low we need the neighborhood around \( t \) shrinks asymptotically to 0.
- \( k/n \approx \lambda_n \), the bandwidth in kernel estimation; i.e., we need \( k \to \infty \) and \( k/n \to 0 \).

**Theorem**
If \((X_1, Y_1), \ldots, (X_n, Y_n)\) are i.i.d. with \( E(Y_1^2) < \infty \), \( X_1 \) sim Lebesgue p.d.f. \( f \), and \( \mu(t) = E(Y_1|X_1 = t) \), then, for some \( \sigma^2(t) \),

\[
E(\hat{\mu}(t) - \mu(t)) = \frac{\left(\mu''f + 2\mu'f'\right)(t)}{24f(t)^3} \left(\frac{k}{n}\right) + o\left(\frac{k}{n}\right)
\]

\[
\text{Var}(\hat{\mu}(t)) = \frac{\sigma^2(t)}{k} + o\left(\frac{1}{k}\right)
\]