Lecture 10: Density estimation and nonparametric regression

Density estimation

Suppose that $X_1, ..., X_n$ 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 $\widehat{f}$, F can be estimated by $\widehat{F}(x) = \int_{-\infty}^{x} f(t)dt$, which may be better than F_n
$\widehat{f}$ itself may be of interest

Difference quotient

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

$$f_n(t) = \frac{F_n(t+\lambda_n) - F_n(t-\lambda_n)}{2\lambda_n}, \qquad t \in \mathscr{R},$$

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 $Bi(F(t+\lambda_n)-F(t-\lambda_n),n)$,

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

and

$$\mathrm{Var}\big(f_n(t)\big) \to 0 \qquad \text{if } \lambda_n \to 0 \text{ and } n\lambda_n \to \infty.$$

Thus, we should choose λ_n converging to 0 slower than n^{-1}.
If we assume that $\lambda_n \to 0$, $n\lambda_n \to \infty$, and f is continuously differentiable at t, then it can be shown (exercise) that

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

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

$$\sqrt{n\lambda_n}[f_n(t)-f(t)] \to_d N\big(0, \tfrac{1}{2}f(t)\big).$$

Kernel density estimators

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

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

where w is a known Lebesgue p.d.f. on $\mathscr{R}$ and is called the kernel.
If we choose $w(t) = \frac{1}{2} I_{[-1,1]}(t)$, then $\widehat{f}(t)$ is essentially the same as the so-called histogram.

Properties of kernel density estimator

$\widehat{f}$ is a Lebesgue density on $\mathscr{R}$, since

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

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

$$\begin{aligned} E[\widehat{f}(t)] - f(t) &= \frac{1}{\lambda_n} \int w\left(\frac{t-z}{\lambda_n}\right) f(z)dz - f(t) \\ &= \int w(y)[f(t-\lambda_n y) - f(t)]dy \end{aligned}$$

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

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

$$
\begin{aligned}
\mathrm{Var}(\widehat{f}(t)) &= \frac{1}{n\lambda_n^2} \mathrm{Var}\left(w\left(\frac{t-X_1}{\lambda_n}\right)\right) \\
&= \frac{1}{n\lambda_n^2} \int \left[w\left(\frac{t-z}{\lambda_n}\right)\right]^2 f(z)dz \\
&\quad - \frac{1}{n}\left[\frac{1}{\lambda_n}\int w\left(\frac{t-z}{\lambda_n}\right) f(z)dz\right]^2 \\
&= \frac{1}{n\lambda_n} \int [w(y)]^2 f(t-\lambda_n y)dy + O\left(\frac{1}{n}\right) \\
&= \frac{w_0 f(t)}{n\lambda_n} + o\left(\frac{1}{n\lambda_n}\right)
\end{aligned}
$$

Hence, if $w_0 < \infty$, f' is bounded and continuous at t, then

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

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

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

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

and the best rate $n^{-4/5}$ is achieved when λ_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

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

This can be shown as follows.
Let $Y_{in} = w\left(\frac{t - X_i}{\lambda_n}\right)$.
Then $Y_{1n}, ..., Y_{nn}$ are independent and identically distributed with

$$E(Y_{1n}) = \int_{-\infty}^{\infty} w\left(\frac{t-x}{\lambda_n}\right) f(x) dx$$

$$
\begin{aligned}
&= \lambda_n \int_{-\infty}^{\infty} w(y) f(t - \lambda_n y) dy \\
&= O(\lambda_n)
\end{aligned}
$$

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

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

$$
\mathrm{Var}\big(\widehat{f}(t)\big) = \frac{1}{n^2 \lambda_n^2} \sum_{i=1}^{n} \mathrm{Var}(Y_{in}) = \frac{w_0 f(t)}{n \lambda_n} + o\left(\frac{1}{n \lambda_n} \right).
$$

Note that $\widehat{f}(t) - E\widehat{f}(t) = \sum_{i=1}^{n} [Y_{in} - E(Y_{in})]/(n\lambda_n)$.

To apply Lindeberg's central limit theorem to $\widehat{f}(t)$, we find, for $\varepsilon > 0$,

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

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

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

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

Furthermore, if

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

then

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

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

$$\sqrt{n\lambda_n}\{\widehat{f}(t) - f(t)]\} \to_d N\big(0, w_0 f(t)\big).$$

If

$$E[\widehat{f}(t)] - f(t) = O(\lambda_n^2)$$

then

$$\sqrt{n\lambda_n}\{E[\widehat{f}(t)] - f(t)\} = O\left(\sqrt{n\lambda_n}\lambda_n^2\right) \to 0$$

if $n\lambda_n^5 \to 0$, which implies that

$$\sqrt{n\lambda_n}\{\widehat{f}(t) - f(t)]\} \to_d N\big(0, w_0 f(t)\big).$$

In any case, the best choice of λ_n for the mse does not satisfy $n\lambda_n^3 \to 0$ or $n\lambda_n^5 \to 0$.

Example 5.4

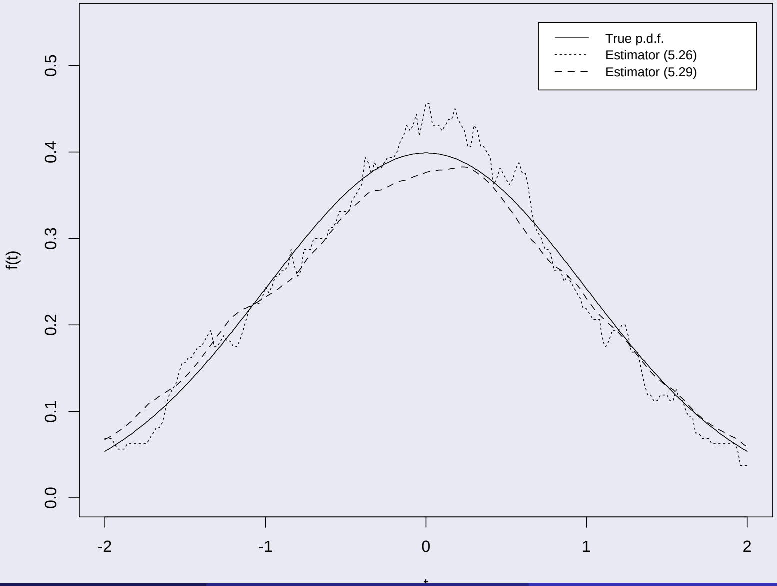
An i.i.d. sample of size $n = 200$ was generated from $N(0,1)$.

Density curve estimates, difference quotient f_n (short dashed curve) and kernel estimate $\widehat{f}$ (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



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 $(Y_1, X_1), ..., (Y_n, X_n)$ 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

$$\widehat{\mu}(t) = \sum_{i=1}^{n} Y_i w\left(\frac{t - X_i}{\lambda_n}\right) \Big/ \sum_{i=1}^{n} w\left(\frac{t - X_i}{\lambda_n}\right), \qquad t \in \mathscr{R}$$

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 $\widehat{\mu}(t)$ as an estimator of $\mu(t)$, it suffices to show that, for any $t \in \mathscr{R}$,

$$h_n(t) = \frac{1}{n\lambda_n} \sum_{i=1}^{n} Y_i w\left(\frac{t - X_i}{\lambda_n}\right) \to_p \int yf(t,y)dy$$

Consider first the expectation:

$$\begin{aligned} E[h_n(t)] &= \frac{1}{\lambda_n} E\left[Y_i w\left(\frac{t - X_i}{\lambda_n}\right)\right] \\ &= \frac{1}{\lambda_n} \int\int yw\left(\frac{t - x}{\lambda_n}\right) f(x,y)dxdy \\ &= \int\int yw(z) f(t - \lambda_n z, y)dzdy \end{aligned}$$

Suppose that $f(x,y)$ is continuous and $f(x,y) \le c(y)g(y)$, 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 $\lambda_n \to 0$ as $n \to \infty$, by the dominated convergence theorem,

$$\begin{aligned} \lim_{n\to\infty} E[h_n(t)] &= \lim_{n\to\infty} \int\int yw(z) f(t - \lambda_n z, y)dzdy \\ &= \int\int yw(z) f(t,y)dzdy \end{aligned}$$

$$= \int w(z)\,dz \int yf(t,y)dy$$
$$= \int yf(t,y)dy$$

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

$$\begin{aligned}
\mathrm{Var}(h_n(t)) &= \frac{1}{n\lambda_n^2}\mathrm{Var}\left(Y_i w\left(\tfrac{t-X_i}{\lambda_n}\right)\right) \\
&\leq \frac{1}{n\lambda_n^2} E\left[Y_i w\left(\tfrac{t-X_i}{\lambda_n}\right)\right]^2 \\
&= \frac{1}{n\lambda_n^2}\int\int y^2\left[w\left(\tfrac{t-x}{\lambda_n}\right)\right]^2 f(x,y)dxdy \\
&= \frac{1}{n\lambda_n}\int\int y^2[w(z)]^2 f(t-\lambda_n z, y)dzdy
\end{aligned}$$

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 $E[Y_i^2 c(Y_i)] = \int y^2 c(y)g(y)dy < \infty$

Also, assume $w_0 = \int [w(z)]^2 dz < \infty$ and $E(Y_i^2) < \infty$.
Then

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

Hence,

$$\mathrm{Var}(h_n(t)) = O\left(\frac{1}{n\lambda_n}\right)$$

which converges to 0 if $n\lambda_n \to \infty$.

Under some more conditions, similar to the estimation of $f(t)$, for any $t \in \mathscr{R}$, we can show that for some function $\sigma^2(t)$,

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

Note that $\widehat{\mu}(t)$ is a ratio estimator $h_n(t)/\widehat{f}(t)$.

Averaging kernel estimators

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

$$\widehat{\mu} = \sum_{j=1}^{n}\sum_{i=1}^{n} Y_i w\left(\frac{X_j - X_i}{\lambda_n}\right) \Big/ \sum_{j=1}^{n}\sum_{i=1}^{n} w\left(\frac{X_j - X_i}{\lambda_n}\right)$$

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

Under some conditions, it can be shown that

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

for some σ^2.

Conditions on λ_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 ($s = 1/3$ or $1/5$) for estimating $\mu(t)$ with a fixed t.

k-nearest neighbor (k-NN) estimators

The kernel estimator

$$\widehat{\mu}(t) = \sum_{i=1}^{n} Y_i w\left(\frac{t-X_i}{\lambda_n}\right) \Big/ \sum_{i=1}^{n} w\left(\frac{t-X_i}{\lambda_n}\right), \qquad t \in \mathscr{R}$$

is a weighted average of Y_i's in a fixed neighorhood around t, determined in shape by the kernel w and the bandwidth λ_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:

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

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, $\widetilde{\mu}=(1+0+4)/3=5/3$.

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 $(X_1, Y_1), ..., (X_n, Y_n)$ 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 $\sigma^2(t)$,

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

$$\mathrm{Var}(\widetilde{\mu}(t)) = \frac{\sigma^2(t)}{k} + o\left(\frac{1}{k}\right)$$