

Stat 710: Mathematical Statistics

Lecture 17

Jun Shao

Department of Statistics
University of Wisconsin
Madison, WI 53706, USA

Lecture 17: Density estimation

Why do we estimate a density?

Suppose that X_1, \dots, 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 $\hat{F}(x) = \int_{-\infty}^x \hat{f}(t) dt$, which may be better than F_n
 \hat{f} itself may be of interest

Difference quotient

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

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

where F_n is the empirical c.d.f., and $\{\lambda_n\}$ is a sequence of positive

Lecture 17: Density estimation

Why do we estimate a density?

Suppose that X_1, \dots, 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 $\hat{F}(x) = \int_{-\infty}^x \hat{f}(t) dt$, which may be better than F_n
 \hat{f} itself may be of interest

Difference quotient

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

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

where F_n is the empirical c.d.f., and $\{\lambda_n\}$ is a sequence of positive

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 λ_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* of the form

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

where w is a known Lebesgue p.d.f. on \mathcal{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 \mathcal{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

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

Kernel density estimators

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

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

where w is a known Lebesgue p.d.f. on \mathcal{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 \mathcal{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

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

Properties of kernel density estimator

If f is bounded and continuous at t , then, by the dominated convergence theorem, the bias of $\hat{f}(t)$ converges to 0 as $\lambda_n \rightarrow 0$.

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 and $w_0 = \int [w(t)]^2 dt < \infty$, the variance of $\hat{f}(t)$ is

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

Properties of kernel density estimator

Hence, if $\lambda_n \rightarrow 0$, $n\lambda_n \rightarrow \infty$, and f' is bounded and continuous at t , then

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

If $\lambda_n \rightarrow 0$, $n\lambda_n \rightarrow \infty$, and f is bounded and continuous at t and $w_0 = \int_{-\infty}^{\infty} [w(t)]^2 dt < \infty$, then

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

This can be shown as follows.

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

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

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

and

Properties of kernel density estimator

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

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 $\varepsilon > 0$,

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

which converges to 0 under the given conditions.

Properties of kernel density estimator

This proves

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

Furthermore,

$$\begin{aligned} E[\widehat{f}(t)] - f(t) &= \lambda_n^{-1} E(Y_{1n}) - f(t) \\ &= \int_{-\infty}^{\infty} w(y)[f(t - \lambda_n y) - f(t)] dy \\ &= \lambda_n \int_{-\infty}^{\infty} y w(y) f'(\xi_{t,y,n}) dy, \end{aligned}$$

where $|\xi_{t,y,n} - t| \leq \lambda_n$.

If f' is bounded and continuous at t , $\int |t| w(t) dt < \infty$, and $n\lambda_n^3 \rightarrow 0$, then

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

and

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

Other properties of density estimator

Similar to the estimation of a c.d.f., we can also study global properties of f_n or \hat{f} as an estimator of the density curve f , using a suitably defined distance between f and its density estimator.

For example, we may study the convergence of $\sup_{t \in \mathcal{R}} |\hat{f}(t) - f(t)|$ or $\int |\hat{f}(t) - f(t)|^2 dt$.

Other density estimators

There are many other density estimation methods, for example,

- the nearest neighbor method (Stone, 1977)
- the smoothing splines (Wahba, 1990)
- local polynomial
- the method of empirical likelihoods

Other properties of density estimator

Similar to the estimation of a c.d.f., we can also study global properties of f_n or \hat{f} as an estimator of the density curve f , using a suitably defined distance between f and its density estimator.

For example, we may study the convergence of $\sup_{t \in \mathcal{R}} |\hat{f}(t) - f(t)|$ or $\int |\hat{f}(t) - f(t)|^2 dt$.

Other density estimators

There are many other density estimation methods, for example,

- the nearest neighbor method (Stone, 1977)
- the smoothing splines (Wahba, 1990)
- local polynomial
- the method of empirical likelihoods

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 and kernel estimate \hat{f} , are plotted in Figure 5.1 with the curve of the true p.d.f.

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

