University of Illinois | Department of Economics
Spring 2017 | Econ 574 | Roger Koenker

Lecture 6
"An Introduction to Density Estimation "

A fundamental problem of nonparametric statistics is density estimation. Many of the methods we will use later in the course arise in a relatively simple form here, and consequently it is a natural place to begin. See Silverman (1986) and Devroye (1987) for rather different expositions of this topic.

Histograms – if you must

The simplest example is the univariate histogram, or bar-graph. Suppose $X_1, \dots, X_n$ is a random sample from F with density f. Partition the real line into $T = \{A_1, \dots, A_k\}$ disjoint sets, e.g., $T_n : A_i = [(i-1)h + c, ih + c]$. Let

$$\begin{aligned} A(x) &= \{A_i \in T | x \in A_i\} \\ \hat{f}_n(x) &= \frac{\hat{F}_n\{A(x)\}}{\lambda\{A(x)\}} = \frac{\#\{X_j \in A(x)\}/n}{\text{length } (A(x))} \end{aligned}$$

e.g., we might have $\hat{f}_n(x) = (nh)^{-1}\#\{X_j \in A(x)\}$.

We can define a population analogue of $\hat{f}_n(x)$ as

$$p_T(x) = \frac{P((A(x))}{\lambda(A(x))}$$

so for f continuous at x, let $h = \lambda(A(x))$ and $A(x) = [a, b]$

$$\begin{aligned} p_T(x) - f(x) &= h^{-1} \int_a^b f(y)dy - f(x) \\ &= f(\tilde{y}) - f(x) \qquad \text{for } \tilde{y} \in (a, b) \\ &\to 0 \qquad \text{as } h \to 0, \end{aligned}$$

A classical measure of performance $\hat{f}_n$ is mean squared error which can be decomposed into squared bias and variance,

$$\begin{aligned} \text{MSE } (x) &= E(\hat{f}_n(x) - f(x))^2 \\ &= (E\hat{f}_n(x) - f(x))^2 + V(\hat{f}_n(x)) \end{aligned}$$

Note

$$\begin{aligned} E\hat{f}_n(X) &= En^{-1} \sum I_A(X_i)/\lambda(A) \\ &= n^{-1} \sum P(A)/\lambda(A) = p_T(x). \end{aligned}$$

$$
\begin{aligned}
V(\sum I_A(x)/(n\lambda(A))) &= \left(\frac{1}{n\lambda}\right)^2 \sum V(I_A(x)) \\
&= (n\lambda)^{-2} nP(A)(1-P(A))
\end{aligned}
$$

for $\lambda = h$:

$$
\begin{aligned}
&= (nh^2)^{-1} P(A)(1-P(A)) \\
&= (nh)^{-1} p_T(x)(1-P(A)) \\
&\leq (nh)^{-1} p_T(x)
\end{aligned}
$$

so we have,

Thm: For $f(x)$ continuous with $h \to 0$ and $nh \to \infty$ then $\text{MSE}(x) \to 0$.

This pointwise result is promising, but we might like something stronger, e.g.,

$$
\int |f_n(x) - f(x)| dx \to 0
$$

I won't prove this, see Section 2.5 of Devroye(1987), but note that

$$
\int |\hat{f}_n - f| dx = 2 \int (f - \hat{f}_n)^+ dx
$$

and $(f - \hat{f}_n)^+ < f$ so dominated convergence gives $\int |f_n - f| dx \overset{P}{\to} 0$. The (Lebesgue) dominated convergence theorem is a standard technique for strengthening pointwise convergence to stronger forms of convergence: If $f_n \to f \;\; a.e.[\mu]$ and there exists g such that $|f_n| \leq g \;\; a.e.[\mu]$ for all n and $\int g d\mu < \infty$, then $\int f_n d\mu \to \int f d\mu$.

Bandwidth Choice and Rates of Convergence We would like to have more guidance on how to choose h. To this end consider,

$$
\text{MSE}(x) = (nh)^{-1} p_T(x) + o((nh)^{-1}) + (p_T(x) - f(x))^2
$$

where

$$
p_T(x) - f(x) = h^{-1} \int_a^b (f(y) - f(x)) dy
$$

but by the mean value theorem, $f(y) = f(x) + f'(\tilde{x}(y))(y - x)$ so,

$$
p_T(x) - f(x) = h^{-1} \int (y - x) f'(\tilde{x}(y)) dy
$$

where we have assumed $f(x)$ is absolutely continuous and $\tilde{x}(y) \in (x, y) \subset (a, b)$. Assume, further, $|f'(y)| \leq c_h(x)$ for $(x - y) < h$ so,

$$
|p_T(x) - f(x)| \leq h^{-1} \int |(y - x)||f'(\tilde{x}(y))| dy \leq \int c_h(x) dy \leq h c_h(x).
$$

Note that $\lim_{h \to 0} c_h(x) = |f'(x)|$, so

$$
\text{MSE } (x) = (nh)^{-1} f(x) + o((nh)^{-1}) + h^2 (f'(x))^2 + o(h^2)
$$

minimising with respect to h we have the first order conditions,

$$-(nh^2)^{-1}f + 2h(f')^2 = 0$$

so $h^3 = (2n)^{-1}f/(f')^2$ or $h = kn^{-1/3}$, and therefore, at the optimal bandwidth,

$$\text{MSE }(x) = n^{-2/3}[f(x)/k + k^2(f'(x))^2] + o(n^{-2/3})$$

Note that this is rather unsatisfactory by comparison with convergence rates in parametric problems where MSE $(\hat{\theta}) = O(n^{-1})$. We may regard this as "the cost of being non-parametric." We get to heaven more slowly when we don't know the correct model.

For R users, note that the R `hist` command uses $h = 1/(\log_2(n) + 1)$ which R calls Sturges rule and is sometimes also called Doane's Rule. Since the number of bars in a histogram is $m = O(h^{-1})$ we have $m = O(\log_2(n) + 1)$ bars while for optimal method we have $m = O(k^{-1}n^{1/3})$. So the number of bars increases much faster for optimal choice. For $n < 500$ it doesn't matter much but for n larger than 500 it does matter. A reasonable value for k above is 3.5. Wand (1997) has a good discussion of this. In fact, Wand (1997) serves as a good example of style and content for a 574 paper. R allows the user to specify one of these alternative rules by specifying `breaks = "Scott"` for the rule $k = 3.5\hat{\sigma}n^{-1/3}$ or `breaks = "FD"` for the rule $k = 2\tilde{\sigma}n^{-1/3}$ where $\hat{\sigma}$ is the usual standard deviation estimate, and $\tilde{\sigma}$ is the estimated interquartile range, which is generally regarded as safer, more robust, choice.

If $f'(x) = 0$, then there are obvious problems with the choice of k suggested by these MSE calculations. This would be the case, for example, if we considered $x = 0$ and f were any unimodal density symmetric about zero. One way to circumvent this problem is to explicitly admit that we don't simply want to estimate the density at a single point but that we would really like to minimize *integrated* mean squared error. We can develop an approximation for this quite easily from what we have already done.

Write

$$\int V(\hat{f}_n(x))dx = \sum_{k=-\infty}^{\infty} \int_{A_k} V(\hat{f}_n(x))dx = \frac{1}{nh} \sum_{k=-\infty}^{\infty} P(A_k)(1 - P(A_k))$$

Note that $\sum P(A_k) = \int f(x)dx = 1$ and by the mean value theorem,

$$\sum P^2(A_k) = \sum f^2(\xi_k)h^2 = h \int f^2(x)dx + o(1)$$

so the integrated variance may be approximated by

$$\int V(\hat{f}_n(x))dx = (nh)^{-1} - n^{-1} \int f^2(x)dx + o(n^{-1})$$

Now consider the integrated bias:

$$\begin{aligned} hp_T(x) &= \int (f(x) + (t-x)f'(x) + \frac{1}{2}(t-x)^2 f''(x) + \ldots)dt \\ &= hf(x) + h(\frac{h}{2} - x)f'(x) + O(h^3) \end{aligned}$$

so the bias at x is,

$$p_T(x) - f(x) = (\frac{h}{2} - x)f'(x) + O(h^2)$$

and

$$\int_{A_k} (\frac{h}{2} - x)^2 f'(x)^2 dx = f'(\xi_k)^2 \int (\frac{h}{2} - x)^2 dx = \frac{h^3}{12} f'(\xi_k)^2$$

for some $\xi_k \in A_k$, and integrated squared bias is,

$$\frac{h^3}{12} \sum_{k=-\infty}^{\infty} (f'(\xi_k))^2 = \frac{h^2}{12} \int_{-\infty}^{\infty} (f'(x))^2 dx + o(h^2)$$

Now, note that if we minimize the asymptotic integrated squared error,

$$\text{AMISE}(h) = \frac{1}{nh} + \frac{h^2}{12} \int_{-\infty}^{\infty} (f'(x))^2 dx$$

we obtain $h^* = kn^{-1/3}$, with $k = (6/\int (f')^2)^{1/3}$, So integrated squared error is also $O(n^{-2/3})$ like the pointwise MSE, but now we can consider k based on global considerations. For example, if $f \sim \mathcal{N}(0,1)$,

$$\int (f')^2 = \frac{1}{4\sqrt{\pi}}$$

so $k \approx 3.5$. If instead $f \sim \mathcal{N}(\mu, \sigma^2)$, then we have Scott's $k \approx 3.5\sigma$. This is quite reasonable unless the data are very heavy tailed in which case the estimation of σ may be problematic. (More on this later in the course.) As an alternative, Freedman and Diaconis have proposed the rule

$$h^* = 2rn^{-1/3}$$

where r is the interquartile range. This is somewhat narrower than the normal theory proposal of Scott (1992).

Kernel Density Estimation

Rosenblatt (1956) proposed the following alternative for estimating $f(x)$. Let $A = (x - h_n, x + h_n)$ and set

$$\hat{f}_n(x) = (2h_n n)^{-1} \sum I_A(X_i) = \frac{F_n(x + h_n) - F_n(x - h_n)}{2h_n}$$

Clearly $2nh_n f_n(x) \sim B(n, p_n)$ where $p_n = F^+ - F^- = F(x + h_n) - F(x - h_n)$ so,

$$E\hat{f}_n = \frac{F^+ - F^-}{2h} \to f(x) \quad \text{if} \quad h_n \to 0$$

$$V\hat{f}_n = \frac{(F^+ - F^-)(1 - F^+ + F^-)}{4nh_n^2} \to 0$$

if $h_n \to 0$ and $nh_n \to \infty$, since

$$\frac{F^+ - F^-}{2h} \to f \quad \text{and} \quad \frac{1 - F^+ - F^-}{2nh} \to 0.$$

So far this is very much like the histogram, except for the fact that the "bin" is centered at the x-value of interest. However this turns out to have a surprisingly important effect.

Asymptotic Normality of $f_n(x)$. Recall that if $Y \sim B(n,p), EY = np, VY = np(1-p)$ and by DeMoivre Laplace (and the quincunx)

$$Z_n = \frac{Y_n - np}{\sqrt{VY}} \rightsquigarrow \mathcal{N}(0,1)$$

Here, let $\Delta = F^+ - F^-$ and write,

$$\begin{aligned} Z_n &= \frac{2nh_n\hat{f} - n\Delta}{\sqrt{n\Delta(1-\Delta)}} = \frac{(2nh_n\hat{f} - n\Delta)/\sqrt{2nh_n}}{\sqrt{\Delta(1-\Delta)/2h_n}} = \frac{\sqrt{2nh_n}(\hat{f} - (\Delta/2h_n))}{\sqrt{\Delta(1-\Delta)/2h_n}} \\ &\to \frac{\sqrt{2nh_n}(\hat{f} - E\hat{f})}{\sqrt{f(x)}} \end{aligned}$$

Consider the bias, writing,

$$E\hat{f}(x) - f(x) = \frac{1}{2}\left[\left(\frac{F(x+h) - F(x)}{h} - f(x)\right) + \left(\frac{F(x) - F(x-h)}{h} - f(x)\right)\right]$$

expanding $F(x \pm h)$ around 0 and simplifying for $\eta_1, \eta_2 \in [0,1]$

$$\begin{aligned} \frac{F(x+h) - F(x)}{h} &= f(x) + \frac{f'(x)h}{2} + \frac{f''(x+\eta_1 h)h^2}{6} \\ \frac{F(x) - F(x-h)}{h} &= f(x) - \frac{f(x)h}{2} + \frac{f''(x+\eta_2 h)h^2}{6} \end{aligned}$$

so

$$E\hat{f} - f = \frac{h^2}{12}[f''(x+\eta_1 h) + f''(x+\eta_2 h)] \to \frac{h_n^2}{6} f''(x)$$

for $h_n \to 0$, provided $f''(x) < \infty$ *so this cancellation of f' effect is the big gain over histogram.*

Note that the effect of centering the kernel estimate at x, rather than using the uncentered histogram estimate, is to remove the $f'(x)$ term in the bias which is $O(h)$ and replace it with a term which is $O(h^2)$. As in the histogram case we have,

$$V\hat{f} \to \frac{f(x)}{2nh_n}$$

so

$$\text{MSE }(x) = \frac{f(x)}{2nh_n} + \frac{h_n^4}{36}(f''(x))^2 + o((nh_n)^{-1}) + o(h_n^4).$$

Thm: If $h_n = cn^{-1/5-\delta}$ for $c > 0$ and $\delta \in \left(-\frac{1}{5}, \frac{4}{5}\right)$, then

$$n^{4/5}E(\hat{f}(x) - f(x))^2 = \frac{1}{2c}f(x)n^{\delta} + \frac{c^4}{36}(f''(x))^2 n^{-4\delta} + o_p(1).$$

Proof: At one limit $\delta = -1/5 + \varepsilon$, so $h_n = cn^{-\varepsilon} \to 0$ and $nh_n = cn^{1-\varepsilon} \to \infty$, and at the other limit $\delta = 4/5 - \varepsilon$, so $h_n = cn^{-1+\varepsilon} \to 0$ and $nh_n = cn^{\varepsilon} \to \infty$.

To minimize MSE clearly $\delta = 0$ is optimal since otherwise $n^{4/5}$ MSE $\to \infty$. What about the optimal value for c? Let $\delta = 0$ and differentiating with respect to c, we have the first order conditions,

$$\frac{c^3}{9}(f''(x))^2 = \frac{f(x)}{2c^2} \quad \Rightarrow c = \left(4.5\frac{f(x)}{(f''(x))^2}\right)^{1/5}$$

For example, if $f(x) = \phi(x)$ so $f'(x) = -xf(x)$ and $f''(x) = -\phi(x) + x^2\phi(x)$, then

$$c(x) = \left(9/2 \ \ \frac{\phi(x)}{\phi''(x)^2}\right)^{1/5} = (2/9 \ \ \phi(x)(x^2-1)^2)^{-1/5}$$

so at the mean for example we have, $c(0) = 1.623$. Plotting $c(x)$ we see a rather strange scallop shape, that suggests that you would want to have wider bandwidths at ± 1, I'm rather doubtful about that, but the other implication – that bandwidth should be larger in the tails than in the center of the distribution – definitely does seem reasonable.

We have gained substantially, by centering our histogram estimate at x, now MSE $(x) = O(n^{-4/5})$ considerably better than the $O(n^{-2/3})$ for the histogram estimate. What have we lost? We now have somewhat more computation since at each x we need an estimate; this isn't too burdensome, though.

Where do we go from here? Two "features" of the Rosenblatt $\hat{f}$ seem awkward:

(i) sharp edges of the kernel
(ii) fixed bandwidth with respect to x.

In the last question of Problem Set 2, we will consider smoothing the kernel shape and later we can (perhaps) consider adaptive bandwidth choice. Silverman has a good discussion of both of these topics. Silverman is a good model for a monograph in statistics.

In problem 4 of PS 2 you are asked to review the preceding computations using a smoother form for the kernel function. Qualitatively the situation is similar. A smoother kernel has an effect on the relevant constants, but not on the rates for $h_n \to 0$ and $MSE_n \to 0$.

An interesting question, one that has not really received sufficient attention in the literature is: what determines the level of difficulty of density estimation? Here I will discuss a result of Devroye which provides an interesting partial answer to this question – in effect an analogue of the Cramér-Rao bound for parametric situations.

Def: Let $f(x)$ be a function on $[a,b]$ for any partition $T(x), x_0 = a < x_1 < \ldots < x_n = b$, let

$$V_T = \sum_{k=0}^{n-1} |f(x_{k+1}) - f(x_k)|.$$

The least upper bound of V_T over partitions T is called the *total variation* of the function f and may be denoted $V(f)$.

Some elementary facts about total variation.

F1. Monotone functions: $V(f) = |f(b) - f(a)|$

F2. Lipshitz functions: $V(f) \leq K(b-a)$

Pf: Recall that f is Lipshitz on $[a, b]$ if

$$|f(x) - f(y)| \leq K|x - y|$$

thus

$$|f(x_{k+1}) - f(x_k)| \leq K(x_{k+1} - x_k)$$

so

$$V(f) \leq K(b-a).$$

Note that if f has a derivative f' at every point $x \in [a, b]$, then by mean value theorem,

$$f(x) - f(y) = f'(z)(x - y)$$

so if $f'(z)$ is bounded, f is Lipshitz.

My favorite reference on this is Natanson (1974), but any real analysis book has a treatment.

A result that is not quite so trivial, but quite important is the following.

F3. If f is absolutely continuous, then $V(f) = \int_a^b |f'(t)|dt$

This leads to an interesting measure of roughness for functions that seems interesting from a statistical standpoint.

Def: Let f be an absolutely continuous function with derivative f' on $[a, b]$. The *roughness* of f is given by

$$R(f) = V(f')$$

Remark: Note, if f' is absolutely continuous, then

$$R(f) = V(f') = \int_a^b |f''(t)|dt$$

so linear functions have roughness 0. Of course, densities can't be linear; they can be piecewise linear, but then the kinks contribute to the roughness. Consider a triangular density, and compute its roughness.

Example: Seemingly nice functions can have $V(f) = \infty$. Take $f(x) = x\cos(\pi/2x)$ on $[0, 1]$ for $T = \{0 < \frac{1}{2n} < \frac{1}{2n-1} < \ldots < \frac{1}{3} < \frac{1}{2} < 1,$ then $V_T = 1 + \frac{1}{2} + \ldots + \frac{1}{3} + \ldots + \frac{1}{n} \approx \log n.$

Now we finally get to Devroye's result.

Thm: Let f be a density with $R(f) < \infty$ and $\int x^2 f < \infty$. If $K(\cdot)$ is a nonnegative order 2 kernel for which $\int(1+x^2)K^2 < \infty$, then

$$\inf_{h>0} E(\int |\hat{f}_n - f|) \leq (1+o(1))C(K)\gamma(f)n^{-2/5}$$

where $C(K)$ is a constant depending solely on K and

$$\gamma(f) = (R(f)(\int \sqrt{f})^4)^{1/5}$$

Remark: This reduces the difficulty factor to two salient functionals. One is the roughness of f as measured by the total variation of f', the other is the tail behavior – $\int \sqrt{f}$ can be arbitrarily large – in the Cauchy case $\int \sqrt{f} = \infty$. The best case from the point of view of R is the isosceles triangle density – only the jumps at the corners and at the mode contribute to R, but in general, we get a picture like 7.7 in Devroye. It can also be shown that

$$\inf_{h>0} E(\int |\hat{f}_n - f|) \geq (D+o(1))C(K)\gamma(f)n^{-2/5}$$

for some $D > 0$ and *all* densities f. This for some constant c, we can plot $xy^4 < c$ where $x = R(f)$ and $y = \int \sqrt{f}$ and obtain a simple way to characterize the feasible set of densities and how difficult various densities are by measuring the distance to this boundary. As Devroye says, "The lower bound for $\gamma(f)$ is really due to the fact that when one has to draw a density, one either needs to create a big tail if the density is to be smooth, or one needs a lot of oscillation if the tail is to be small."

References

Devroye, L. (1987) *A Course in Density Estimation,* Birkhauser.

Silverman, B. (1986) *Density Estimation for Statistical Data Analysis,* Chapman-Hall.

Rosenblatt, M. (1956) Remarks on some nonparametric estimates of a density function, *Annals of Math Stat*, 27, 832-837.

Wand, M. (1997) Data based choice of histogram bin width, *American Statistician*, 51, 59-64.

Scott, D.W. (1992) *Multivariate Density Estimation*, Wiley.

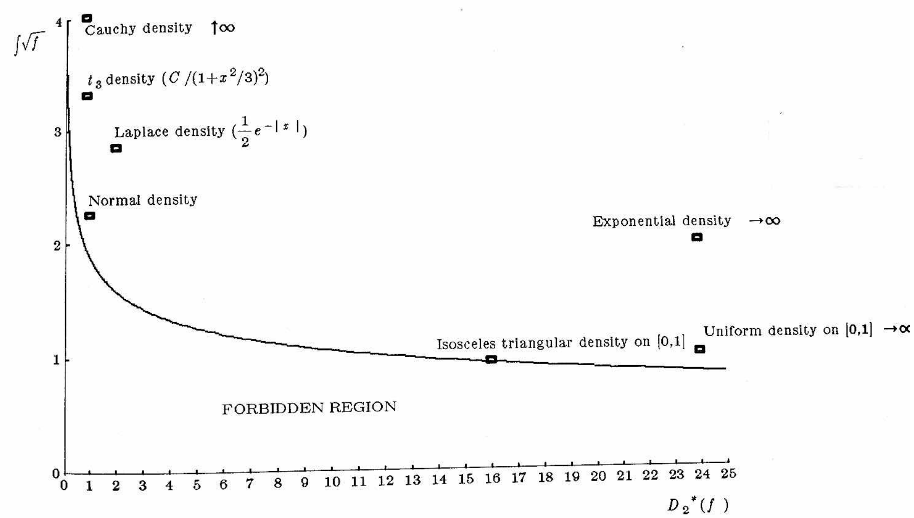

Figure 7.7.
Plane of $\int\sqrt{f}$ versus $D_2^*(f)$.