Notes On Nonparametric Density Estimation

James L. Powell
Department of Economics
University of California, Berkeley

Univariate Density Estimation via Numerical Derivatives

Consider the problem of estimating the density function $f(x)$ of a scalar, continuously-distributed i.i.d. sequence x_i at a particular point x. If the density f is in a known parametric family (e.g., Gaussian), estimation of the density reduces to estimation of the finite-dimensional parameters that characterize that particular density in the parametric family. Without a parametric assumption, though, estimation of the density f over all points in its support would involve estimation of an infinite number of parameters, known in statistics as a *nonparametric estimation* problem (though "infinite-parametric estimation" might be a more accurate title).

Since the density function $f(x)$ is the derivative of the cumulative distribution function $F(x) \equiv \Pr\{x_i \leq x\}$, and since the empirical c.d.f.

$$\hat{F}(x) \equiv \frac{1}{n}\sum_{i=1}^{n} 1\{x_i \leq x\}$$

is the natural nonparametric of the c.d.f., it seems natural to base estimation of f on the empirical c.d.f. However, while $\hat{F}$ is $\sqrt{n}$-consistent and asymptotically normal, it would be clearly nonsensical to estimate f by differentiating $\hat{F}$, since its derivative is either zero or undefined. Using the definition of the density f as the (right-) derivative of the c.d.f.,

$$f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h},$$

we might estimate the density f by a corresponding difference ratio of $\hat{F}$:

$$\begin{aligned} \hat{f}(x) &= \frac{\hat{F}(x+h) - \hat{F}(x)}{h} \\ &= \frac{1}{nh}\sum_{i=1}^{n} 1\{x < x_i \leq x+h\}, \end{aligned}$$

where the "perturbation" h, also known as a "bandwidth" or "window width") is positive but "small," depending upon the sample size (i.e., $h \equiv h_n$).

To show MSE consistency of $\hat{f}$, it suffices to choose the bandwidth sequence h_n so that the mean bias and variance of $\hat{f}$ both tend to zero as the sample size increases. Since the empirical c.d.f. $\hat{F}$ is an unbiased estimator of F – that is, $E[\hat{F}(x)] = F(x)$ – the bias of $\hat{f}$ is evidently

$$\begin{aligned} E[\hat{f}(x)] - f(x) &= \frac{F(x+h) - F(x)}{h} - f(x) \\ &\to 0 \end{aligned}$$

if

$$h = h_n \to 0 \qquad \text{as} \qquad n \to \infty.$$

The variance of $\hat{f}(x)$ is

$$\begin{aligned} V(\hat{f}(x)) &= V\left(\frac{1}{nh}\sum_{i=1}^{n} 1\{x < x_i \leq x + h\}\right) \\ &= \frac{1}{nh^2} V(1\{x < x_i \leq x + h\}) \\ &= \frac{1}{nh}\left[\frac{F(x+h) - F(x)}{h}\left(1 - (F(x+h) - F(x))\right)\right] \\ &= \frac{f(x)}{nh} + O(\frac{1}{n}), \end{aligned}$$

which will tend to zero if

$$nh = nh_n \to \infty \qquad \text{as} \qquad n \to \infty.$$

Thus, if the bandwidth sequence h_n tends to zero as n tends to infinity, but at a slower rate than $1/n$, the MSE of $\hat{f}$ will converge to zero, ensuring its (weak) consistency.

To narrow down the choice of bandwidth sequence h_n, we might want to choose it to maximize the rate of convergence of the MSE of $\hat{f}$ to zero. To do this, we need an explicit expression for its bias as a function of h. Assuming $F(x+h)$ is smooth enough to admit a second-order Taylor's series expansion around $h = 0$,

$$F(x+h) = F(x) + f(x) \cdot h + \frac{f'(x)}{2}h^2 + o(h^2),$$

the MSE of $\hat{f}$ can be expressed as

$$\begin{aligned} MSE(\hat{f}(x); h) &= \left[E(\hat{f}(x)) - f(x)\right]^2 + V(\hat{f}(x)) \\ &= \left(\frac{f'(x)}{2}h\right)^2 + o(h^2) + \frac{f(x)}{nh} + O(\frac{1}{n}). \end{aligned}$$

Because the squared bias is directly related to h, while the variance is inversely related to h, the fastest convergence of the MSE to zero occurs when the squared bias and variance converge to zero at the same

speed. (If they converge at different speeds, the slower speed dominates.) Thus the optimal bandwidth sequence h^* will satisfy

$$O\left((h^*)^2\right) = O\left(\frac{1}{nh^*}\right),$$

so

$$h^* = O\left(\left(\frac{1}{n}\right)^{1/3}\right)$$

will give the fastest rate of convergence of the MSE to zero,

$$MSE(\hat{f}(x); h^*) = O\left(\left(\frac{1}{n}\right)^{2/3}\right).$$

This rate is slower than the usual parametric rate of convergence of the MSE (if it exists) to zero, which is $O(\frac{1}{n})$ for, e.g., the sample mean.

Although the optimal *rate* of convergence of h^* to zero does not depend upon f itself, the optimal *level* would require knowledge of $f(x)$. Assuming the bandwidth sequence is of the form

$$h^* = \frac{c}{n^{1/3}},$$

we can choose c to minimize the limiting normalized MSE

$$\lim_{n\to\infty} n^{2/3} MSE(\hat{f}(x); h^*) = \left(\frac{f'(x)}{2}c\right)^2 + \frac{f(x)}{c},$$

which is minimized in c at

$$c^* = \left(\frac{(f'(x))^2}{2f(x)}\right)^{-1/3}.$$

So the optimal bandwidth sequence

$$h^* = \left(\frac{2f(x)}{(f'(x))^2 n}\right)^{1/3}$$

that minimizes the leading terms in the MSE formula is infeasible, since it requires knowledge of the density $f(x)$ itself and its first derivative. Though it can be shown (under additional conditions) that consistent estimation of the constant c^* will not affect the rate of convergence of $\hat{f}$, etc., nonparametric estimation of $f'(x)$ cannot be based upon $\hat{f}(x)$ directly (since its derivative is zero wherever it is defined), but would require a similar "numerical derivate" approach, which would entail its own bandwidth choice problem. An alternative is to "standardize" the choice of the constant term c^* for some known parametric density function; for example, if $f(x) = \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})$, where ϕ is the standard normal density, then

$$c^* = \left(\frac{x^2\phi(x)}{2}\right)^{-1/3}\sigma,$$

and estimation of the optimal constant would reduce to estimation of the standard deviation σ of the x_i distribution.

Instead of choosing the bandwidth for a specific value of x, we might want to choose it to minimize a "global" MSE criterion over all possible x values, such as the *integrated mean-squared error* criterion

$$\begin{aligned} IMSE(\hat{f}; h) &= \int_{-\infty}^{\infty} MSE(\hat{f}(x); h)dx \\ &= \int_{-\infty}^{\infty} \left(\frac{f'(x)}{2}h\right)^2 dx + \frac{1}{nh} + o(h^2) + O(\frac{1}{n}). \end{aligned}$$

The same calculations as for the pointwise optimal bandwidth h^* yield the optimal IMSE bandwidth h^+ to be

$$h^+ = \left(\frac{2}{n \int (f'(x))^2 \, dx}\right)^{1/3},$$

which still depends upon the derivative of f. (A similar expression holds for the bandwidth minimizing an *average MSE* criterion, where "dx" is replaced by "$f(x)dx$" in the formula for the IMSE.) For the Gaussian density $f(x) = \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})$, the optimal IMSE bandwidth would be

$$\begin{aligned} h^+ &= \left(\frac{2}{n \int (-x\phi(x))^2 \, dx}\right)^{-1/3} \sigma \\ &\cong \frac{2.4\sigma}{n^{1/3}}, \end{aligned}$$

so standardizing the bandwidth choice to be optimal for normal densities reduces the bandwidth choice problem to estimation of the standard deviation σ.

Multivariate Kernel Density Estimation

The numerical derivative estimator of the univariate density $f(x)$ above is a special case of a general class of nonparametric density estimators called *kernel density estimators.* Now supposing $x_i \in R^p$, we can think of "smoothing out" the empirical c.d.f. $\hat{F}(x)$ for x_i by replacing it with a convolution of $\hat{F}$ and the distribution of an independent, continuously-distributed "noise" term $h \cdot \varepsilon$, where $h = h_n$ is a small positive "bandwidth" as above and ε has a known density function $K(\varepsilon)$ (also known as the "kernel" function). For a particular realized value of x_i, the density function for $X_i \equiv x_i + h \cdot \varepsilon$ would be

$$f_{X_i}(x) = \frac{1}{h^p} K\left(\frac{x - x_i}{h}\right)$$

by the usual change-of-variables formula for multivariate densities. (Here the absolute value of the determinant of the Jacobian $d\varepsilon/dX_i'$ is h^{-p}.) Thus the kernel density estimator $\hat{f}(x)$ is the average of f_{X_i} over

the observed values of x_i in the sample,

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^p} K\left(\frac{x - x_i}{h}\right).$$

As long as $\int K(u)du = 1$, the density estimator $\hat{f}(x)$ will also integrate to one over x, as befitting a density function. The numerical derivative estimator discussed above is a special case with $p = 1$ and $K(u) = 1\{-1 \leq u < 0\}$, the density function for a $Uniform(-1, 0)$ random variable.

The MSE calculations are straightforward extensions of those for the numerical derivative estimator. The expectation of $\hat{f}$ is

$$\begin{aligned} E[\hat{f}(x)] &= E\left[\frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^p} K\left(\frac{x - x_i}{h}\right)\right] \\ &= E\left[\frac{1}{h^p} K\left(\frac{x - x_i}{h}\right)\right] \\ &= \int \frac{1}{h^p} K\left(\frac{x - z}{h}\right) f(z)dz. \end{aligned}$$

Making the change-of-variables $u = u(z) = h^{-1}(x - z)$ (with $du = h^{-p}dz$),

$$E[\hat{f}(x)] = \int K(u) f(x - hu)du,$$

which clearly tends to $f(x)$ as $h \to 0$ if f is continuous and bounded above (by dominated convergence), as long as $\int K(u)du = 1$. (The last condition implies $\int |K(u)|\, du < \infty$, a condition that will be more relevant later, when the nonnegativity restriction on $K(u)$ is relaxed). Assuming that $f(x)$ is smooth enough for $f(x - hu)$ to admit a second-order Taylor's expansion around $h = 0$, i.e.,

$$\begin{aligned} f(x - hu) &= f(x) - \frac{\partial f(x)}{\partial x'}(hu) + \frac{1}{2}(hu)' \frac{\partial^2 f(x)}{\partial x \partial x'}(hu) + o(h^2) \\ &= f(x) - h\left(\frac{\partial f(x)}{\partial x'} \cdot u\right) + \frac{h^2}{2} tr\left(\frac{\partial^2 f(x)}{\partial x \partial x'} uu'\right) + o(h^2), \end{aligned}$$

the bias of $\hat{f}$ can be expressed as

$$E[\hat{f}(x)] - f(x) = -h\left(\frac{\partial f(x)}{\partial x'} \cdot \int uK(u)du\right) + \frac{h^2}{2} tr\left(\frac{\partial^2 f(x)}{\partial x \partial x'} \cdot \int uu'K(u)du\right) + o(h^2).$$

If the "mean" of the kernel, $\int uK(u)du$, is nonzero (as with the "one-sided" numerical derivative estimator above), then the bias is $O(h)$; however, if the kernel function $K(u)$ is chosen to be symmetric about zero, or, more generally, if

$$\int uK(u)du = 0, \qquad (*)$$

then the bias is

$$E[\hat{f}(x)] - f(x) = O(h^2).$$

Similarly, the variance of $\hat{f}(x)$ can be calculated as

$$\begin{aligned}
Var(\hat{f}(x)) &= Var\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^p}K\left(\frac{x-x_i}{h}\right)\right) \\
&= \frac{1}{n}Var\left(\frac{1}{h^p}K\left(\frac{x-x_i}{h}\right)\right) \\
&= \frac{1}{n}E\left(\frac{1}{h^p}K\left(\frac{x-x_i}{h}\right)\right)^2 - \frac{1}{n}\left(E\left[\frac{1}{h^p}K\left(\frac{x-x_i}{h}\right)\right]\right)^2 \\
&= \frac{1}{n}\int\frac{1}{h^{2p}}\left[K\left(\frac{x-z}{h}\right)\right]^2 f(z)dz - \frac{1}{n}(E[\hat{f}(x)])^2 \\
&= \frac{1}{nh^p}\int[K(u)]^2 f(x-hu)du - \frac{1}{n}(E[\hat{f}(x)])^2 \\
&= \frac{f(x)}{nh^p}\int[K(u)]^2 du + o\left(\frac{1}{nh^p}\right).
\end{aligned}$$

(The second equality exploits the fact that x_i is *i.i.d.*, and the fifth makes the same change-of-variables as for the bias formula.) So the bias of $\hat{f}(x)$ is $O(h^2)$ under condition (*), and its variance is $O((nh^p)^{-1})$; the optimal bandwidth sequence h^*, which equates the rate of convergence of the squared bias and variance to zero, thus satisfies

$$O((h^*)^4) = O\left(\frac{1}{n(h^*)^p}\right)$$

so

$$h^* = O\left(\frac{1}{n}\right)^{1/(p+4)},$$

and the MSE evaluated at h^* is

$$MSE(\hat{f}(x); h^*) = O\left(\left(\frac{1}{n}\right)^{4/(p+4)}\right).$$

Note that, as p increases – i.e., the number of components in x_i, number of arguments of $f(x)$ increases – the best rate of convergence of the MSE declines, a phenomenon referenced by the catch phrase, "the curse of dimensionality."

The Asymptotic Distribution of the Kernel Density Estimator

The kernel density estimator $\hat{f}(x)$ can be rewritten as a sample average of independent, identically-distributed random variables

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} z_{in},$$

where

$$z_{in} \equiv \hat{f}(x) = \frac{1}{h^p} K\left(\frac{x - x_i}{h}\right).$$

Here the second subscript in z_{in} denotes its dependence on the sample size n through the bandwidth term $h = h_n$. Such doubly-subscripted random variables, where the range of the first subscript is bounded above by the second, are known as *triangular arrays*, and classical limit theorems must be modified to account for the changing distribution of the observations as the sample size increases.

To demonstrate asymptotic normality of $\hat{f}(x)$, a convenient central limit theorem is **Liapunov's Central Limit Theorem for Triangular Arrays.** It statest that, if the scalar random variable z_{in} is independently (but not necessarily identically) distributed with variance $Var(z_{in}) \equiv \sigma_{in}^2$ and r-th absolute central moment $E[|z_{in} - E(z_{in})|^r] \equiv \rho_{in} < \infty$ for some $r > 2$, and if

$$\frac{\left(\sum_{i=1}^n \rho_{in}\right)^{1/r}}{\left(\sum_{i=1}^n \sigma_{in}^2\right)^{1/2}} \to 0$$

as $n \to 0$ (known as the *Liapunov condition*), then

$$\overline{z}_n \equiv \frac{1}{n} \sum_{i=1}^n z_{in}$$

is asymptotically normal,

$$\frac{\overline{z}_n - E[\overline{z}_n]}{\sqrt{Var(\overline{z}_n)}} \to^d \mathcal{N}(0,1).$$

Applying this theorem to $\hat{f}(x)$, we obtain the variance of the z_{in} terms as

$$\begin{aligned}
\sigma_{in}^2 &\equiv Var(z_{in}) \\
&= Var\left(\frac{1}{h^p} K\left(\frac{x - x_i}{h}\right)\right) \\
&= \frac{f(x)}{h^p} \int [K(u)]^2 \, du + o\left(\frac{1}{h^p}\right)
\end{aligned}$$

from our earlier calculations; setting $r = 3$, we get an upper bound for the third central moment of z_{in} as

$$\begin{aligned}
\rho_{in} &\equiv E[|z_{in} - E(z_{in})|^3] \\
&\leq 8E[|z_{in}|^3] \\
&= 8E\left[\left|\frac{1}{h^p} K\left(\frac{x - x_i}{h}\right)\right|^3\right] \\
&= \frac{8f(x)}{h^{2p}} \int |K(u)|^3 \, du + o\left(\frac{1}{h^{2p}}\right),
\end{aligned}$$

where the inequality uses the expansion

$$\begin{aligned} E[|z_{in} - E(z_{in})|^3] &\leq E[(|z_{in}| + |E(z_{in})|)^3] \\ &= E[|z_{in}|^3] + 3E[|z_{in}|^2] \cdot |E(z_{in})| + 3E[|z_{in}|] \cdot |E(z_{in})|^2 + (E(z_{in}))^3 \end{aligned}$$

and the last equality makes the same change-of-variables calculations as for the variance of z_{in}. Thus, for this problem the Liapunov condition is

$$\begin{aligned} \frac{\left(\sum_{i=1}^n \rho_{in}\right)^{1/r}}{\left(\sum_{i=1}^n \sigma_{in}^2\right)^{1/2}} &\leq \frac{\left(n\frac{8f(x)}{h^{2p}} \int |K(u)|^3 \, du + o\left(\frac{n}{h^{2p}}\right)\right)^{1/3}}{\left(n\frac{f(x)}{h^p} \int [K(u)]^2 \, du + o\left(\frac{n}{h^p}\right)\right)^{1/2}} \\ &= O\left(\frac{n^{1/3}}{h^{2p/3}}\right) \cdot O\left(\frac{n^{-1/2}}{h^{-p/2}}\right) \\ &= O((nh^p)^{-1/6}) \\ &\to 0 \end{aligned}$$

if

$$nh^p \to \infty,$$

the same condition imposed to ensure that $Var(\hat{f}(x)) \to 0$ as $n \to \infty$. Under this condition, then,

$$\frac{\hat{f}(x) - E[\hat{f}(x)]}{\sqrt{Var(\hat{f}(x))}} \to^d \mathcal{N}(0,1),$$

or, subtituting the expression for $Var(\hat{f}(x)) = O((nh^p)^{-1})$ and collecting terms,

$$\sqrt{nh^p}(\hat{f}(x) - E[\hat{f}(x)]) \to^d \mathcal{N}(0, f(x) \int [K(u)]^2 \, du).$$

This is almost, but not quite, in the form of the usual expression for an asymptotic distribution of an estimator; the difference is that the expectation of the estimator $E[\hat{f}(x)]$ rather than the true value $f(x)$ is subtracted from $\hat{f}(x)$ in this expression. Writing

$$\sqrt{nh^p}(\hat{f}(x) - f(x)) = \sqrt{nh^p}(\hat{f}(x) - E[\hat{f}(x)]) + \sqrt{nh^p}(E[\hat{f}(x)] - f(x)),$$

the asymptotic normality of $\hat{f}(x)$ around $f(x)$ will require the second, bias term to converge to a constant. Inserting the expression for the bias,

$$\begin{aligned} \sqrt{nh^p}(E[\hat{f}(x)] - f(x)) &= \sqrt{nh^p}\left(\frac{h^2}{2} tr\left(\frac{\partial^2 f(x)}{\partial x \partial x'} \cdot \int uu' K(u) du\right) + o(h^2)\right) \\ &= O(\sqrt{nh^{p+4}}). \end{aligned}$$

If the bandwidth $h = h_n$ takes the form

$$h_n = c\left(\frac{1}{n}\right)^{1/(p+4)},$$

so that it converges to zero at the optimal rate, then

$$\begin{aligned}\sqrt{nh^p}(E[\hat{f}(x)] - f(x)) &\rightarrow \frac{c^{(p+4)/2}}{2} tr\left(\frac{\partial^2 f(x)}{\partial x \partial x'} \cdot \int uu' K(u) du\right) \\ &\equiv \delta(x),\end{aligned}$$

and

$$\sqrt{nh^p}(\hat{f}(x) - f(x)) \rightarrow^d \mathcal{N}(\delta(x), f(x) \int [K(u)]^2 du).$$

Here the approximating normal distribution for $\hat{f}(x)$ would be centered at $f(x) + \delta(x)$, not $f(x)$, and construction of the usual confidence regions or test statistics would be complicated by the fact that $\delta(x)$ depends upon the unknown second derivative of $f(x)$.

If the bandwidth tends to zero *faster* than the optimal rate, i.e.,

$$h^* = o\left(\frac{1}{n}\right)^{1/(p+4)},$$

then

$$\sqrt{nh^p}(E[\hat{f}(x)] - f(x)) \rightarrow 0,$$

and the bias term vanishes from the asymptotic distribution,

$$\sqrt{nh^p}(\hat{f}(x) - f(x)) \rightarrow^d \mathcal{N}(0, f(x) \int [K(u)]^2 du).$$

Often in practice this sort of "undersmoothing" – which implies the bias of $f(x)$ is negligible relative to the variance – is assumed, and confidence intervals of the form

$$f(x) \in \left[f(x) - 1.96\sqrt{\hat{f}(x) \int [K(u)]^2 du}, f(x) + 1.96\sqrt{\hat{f}(x) \int [K(u)]^2 du}\right]$$

are reported, though it is best to view the claimed 95% asymptotic coverage rate with some skepticism. Note that if the bandwidth tends to zero *slower* than the optimal rate, e.g.,

$$h^* = o\left(\frac{1}{n}\right)^{\gamma}, \qquad \gamma > \frac{1}{p+4},$$

then the bias of $\hat{f}(x)$ dominates its standard deviation, and the normalized difference $\sqrt{nh^p}(\hat{f}(x) - f(x))$ diverges.

Rescaling for Multivariate Kernels

Derivatives of the Kernel Density and Regression Estimators

Higher-Order (Bias-Reducing) Kernels