Lecture Notes on Nonparametrics

Bruce E. Hansen
University of Wisconsin

Spring 2009

1 Introduction

Parametric means finite-dimensional. **Non-parametric** means infinite-dimensional.

The differences are profound.

Typically, parametric estimates converge at a $n^{-1/2}$ rate. Non-parametric estimates typically converge at a rate slower than $n^{-1/2}$.

Typically, in parametric models there is no distinction between the true model and the fitted model. In contrast, non-parametric methods typically distinguish between the true and fitted models.

Non-parametric methods make the complexity of the fitted model depend upon the sample. The more information is in the sample (i.e., the larger the sample size), the greater the degree of complexity of the fitted model. Taking this seriously requires a distinct distribution theory.

Non-parametric theory acknowledges that fitted models are approximations, and therefore are inherently misspecified. Misspecification implies estimation bias. Typically, increasing the complexitiy of a fitted model decreases this bias but increases the estimation variance. Nonparametric methods acknowledge this trade-off and attempt to set model complexity to minimize an overall measure of fit, typically mean-squared error (MSE).

There are many nonparametric statistical objects of potential interest, including density functions (univariate and multivariate), density derivatives, conditional density functions, conditional distribution functions, regression functions, median functions, quantile functions, and variance functions. Sometimes these nonparametric objects are of direct interest. Sometimes they are of interest only as an input to a second-stage estimation problem. If this second-stage problem is described by a finite dimensional parameter we call the estimation problem **semiparametric**.

Nonparametric methods typically involve some sort of approximation or **smoothing** method. Some of the main methods are called **kernels**, **series**, and **splines**.

Nonparametric methods are typically indexed by a **bandwidth** or **tuning parameter** which controls the degree of complexity. The choice of bandwidth is often critical to implementation. Data-dependent rules for determination of the bandwidth are therefore essential for nonparametric methods. Nonparametric methods which require a bandwidth, but do not have an explicit data-dependent rule for selecting the bandwidth, are incomplete. Unfortunately this is quite common, due to the difficulty in developing rigorous rules for bandwidth selection. Often in these cases the bandwidth is selected based on a related statistical problem. This is a feasible yet worrisome compromise.

Many nonparametric problems are generalizations of univariate density estimation. We will start with this simple setting, and explore its theory in considerable detail.

2 Kernel Density Estimation

2.1 Discrete Estimator

Let X be a random variable with continuous distribution $F(x)$ and density $f(x) = \frac{d}{dx}F(x)$. The goal is to estimate $f(x)$ from a random sample $\{X_1, ..., X_n\}$.

The distribution function $F(x)$ is naturally estimated by the EDF $\hat{F}(x) = n^{-1}\sum_{i=1}^{n} 1(X_i \leq x)$. It might seem natural to estimate the density $f(x)$ as the derivative of $\hat{F}(x)$, $\frac{d}{dx}\hat{F}(x)$, but this estimator would be a set of mass points, not a density, and as such is not a useful estimate of $f(x)$.

Instead, consider a discrete derivative. For some small $h > 0$, let

$$\hat{f}(x) = \frac{\hat{F}(x+h) - \hat{F}(x-h)}{2h}$$

We can write this as

$$\begin{aligned} \frac{1}{2nh}\sum_{i=1}^{n} 1(x+h < X_i \leq x+h) &= \frac{1}{2nh}\sum_{i=1}^{n} 1\left(\frac{|X_i - x|}{h} \leq 1\right) \\ &= \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) \end{aligned}$$

where

$$k(u) = \begin{cases} \frac{1}{2}, & |u| \leq 1 \\ 0 & |u| > 1 \end{cases}$$

is the uniform density function on $[-1, 1]$.

The estimator $\hat{f}(x)$ counts the percentage of observations which are clsoe to the point x. If many observations are near x, then $\hat{f}(x)$ is large. Conversely, if only a few X_i are near x, then $\hat{f}(x)$ is small. The **bandwidth** h controls the degree of smoothing.

$\hat{f}(x)$ is a special case of what is called a kernel estimator. The general case is

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right)$$

where $k(u)$ is a **kernel function**.

2.2 Kernel Functions

A **kernel function** $k(u) : \mathbb{R} \to \mathbb{R}$ is any function which satisfies $\int_{-\infty}^{\infty} k(u)du = 1$.

A **non-negative** kernel satisfies $k(u) \geq 0$ for all u. In this case, $k(u)$ is a probability density function.

The **moments** of a kernel are $\kappa_j(k) = \int_{-\infty}^{\infty} u^j k(u)du$.

A **symmetric** kernel function satisfies $k(u) = k(-u)$ for all u. In this case, all odd moments are zero. Most nonparametric estimation uses symmetric kernels, and we focus on this case.

The **order** of a kernel, ν, is defined as the order of the first non-zero moment. For example, if $\kappa_1(k) = 0$ and $\kappa_2(k) > 0$ then k is a second-order kernel and $\nu = 2$. If $\kappa_1(k) = \kappa_2(k) = \kappa_3(k) = 0$ but $\kappa_4(k) > 0$ then k is a fourth-order kernel and $\nu = 4$. The order of a symmetric kernel is always even.

Symmetric non-negative kernels are second-order kernels.

A kernel is **higher-order kernel** if $\nu > 2$. These kernels will have negative parts and are not probability densities. They are also refered to as **bias-reducing kernels**.

Common second-order kernels are listed in the following table

Table 1: Common Second-Order Kernels

Kernel	**Equation**	$R(k)$	$\kappa_2(k)$	$eff(k)$
Uniform	$k_0(u) = \frac{1}{2}1\left(\lvert u\rvert \leq 1\right)$	$1/2$	$1/3$	1.0758
Epanechnikov	$k_1(u) = \frac{3}{4}\left(1-u^2\right)1\left(\lvert u\rvert \leq 1\right)$	$3/5$	$1/5$	1.0000
Biweight	$k_2(u) = \frac{15}{16}\left(1-u^2\right)^2 1\left(\lvert u\rvert \leq 1\right)$	$5/7$	$1/7$	1.0061
Triweight	$k_3(u) = \frac{35}{32}\left(1-u^2\right)^3 1\left(\lvert u\rvert \leq 1\right)$	$350/429$	$1/9$	1.0135
Gaussian	$k_\phi(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right)$	$1/2\sqrt{\pi}$	1	1.0513

In addition to the kernel formula we have listed its roughness $R(k)$, second moment $\kappa_2(k)$, and its efficiency $eff(k)$, the last which will be defined later. The **roughness** of a function is

$$R(g) = \int_{-\infty}^{\infty} g(u)^2 du.$$

The most commonly used kernels are the Epanechnikov and the Gaussian.

The kernels in the Table are special cases of the polynomial family

$$k_s(u) = \frac{(2s+1)!!}{2^{s+1}s!}\left(1-u^2\right)^s 1\left(\lvert u\rvert \leq 1\right)$$

where the double factorial means $(2s+1)!! = (2s+1)(2s-1)\cdots 5\cdot 3\cdot 1$. The Gaussian kernel is obtained by taking the limit as $s \to \infty$ after rescaling. The kernels with higher s are smoother, yielding estimates $\hat{f}(x)$ which are smoother and possessing more derivatives. Estimates using the Gaussian kernel have derivatives of all orders.

For the purpose of nonparametric estimation the scale of the kernel is not uniquely defined. That is, for any kernel $k(u)$ we could have defined the alternative kernel $k^*(u) = b^{-1}k(u/b)$ for some constant $b > 0$. These two kernels are equivalent in the sense of producing the same density estimator, so long as the bandwidth is rescaled. That is, if $\hat{f}(x)$ is calculated with kernel k and bandwidth h, it is numerically identically to a calculation with kernel k^* and bandwidth $h^* = h/b$. Some authors use different definitions for the same kernels. This can cause confusion unless you are attentive.

Higher-order kernels are obtained by multiplying a second-order kernel by an $(\nu/2-1)$'th order polynomial in u^2. Explicit formulae for the general polynomial family can be found in B. Hansen (Econometric Theory, 2005), and for the Gaussian family in Wand and Schucany (Canadian Journal of Statistics, 1990). 4th and 6th order kernels of interest are given in Tables 2 and 3.

Table 2: Fourth-Order Kernels

Kernel	**Equation**	$R(k)$	$\kappa_4(k)$	$eff(k)$
Epanechnikov	$k_{4,1}(u) = \frac{15}{8}\left(1 - \frac{7}{3}u^2\right) k_1(u)$	$5/4$	$-1/21$	1.0000
Biweight	$k_{4,2}(u) = \frac{7}{4}\left(1 - 3u^2\right) k_2(u)$	$805/572$	$-1/33$	1.0056
Triweight	$k_{4,3}(u) = \frac{27}{16}\left(1 - \frac{11}{3}u^2\right) k_3(u)$	$3780/2431$	$-3/143$	1.0134
Gaussian	$k_{4,\phi}(u) = \frac{1}{2}\left(3 - u^2\right) k_\phi(u)$	$27/32\sqrt{\pi}$	-3	1.0729

Table 3: Sixth-Order Kernels

Kernel	**Equation**	$R(k)$	$\kappa_6(k)$	$eff(k)$
Epanechnikov	$k_{6,1}(u) = \frac{175}{64}\left(1 - 6u^2 + \frac{33}{5}u^4\right) k_1(u)$	$1575/832$	$5/429$	1.0000
Biweight	$k_{6,2}(u) = \frac{315}{128}\left(1 - \frac{22}{3}u^2 + \frac{143}{15}u^4\right) k_2(u)$	$29295/14144$	$1/143$	1.0048
Triweight	$k_{6,2}(u) = \frac{297}{128}\left(1 - \frac{26}{3}u^2 + 13u^4\right) k_3(u)$	$301455/134368$	$1/221$	1.0122
Gaussian	$k_{6,\phi}(u) = \frac{1}{8}\left(15 - 10u^2 + u^4\right) k_\phi(u)$	$2265/2048\sqrt{\pi}$	15	1.0871

2.3 Density Estimator

We now discuss some of the numerical properties of the kernel estimator

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right)$$

viewed as a function of x.

First, if $k(u)$ is non-negative then it is easy to see that $\hat{f}(x) \geq 0$. However, this is not guarenteed if k is a higher-order kernel. That is, in this case it is possible that $\hat{f}(x) < 0$ for some values of x. When this happens it is prudent to zero-out the negative bits and then rescale:

$$\tilde{f}(x) = \frac{\hat{f}(x)1\left(\hat{f}(x) \geq 0\right)}{\int_{-\infty}^{\infty} \hat{f}(x)1\left(\hat{f}(x) \geq 0\right) dx}.$$

$\tilde{f}(x)$ is non-negative yet has the same asymptotic properties as $\hat{f}(x)$. Since the integral in the denominator is not analytically available this needs to be calculated numerically.

Second, $\hat{f}(x)$ integrates to one. To see this, first note that by the change-of-variables $u = (X_i - x)/h$ which has Jacobian h,

$$\int_{-\infty}^{\infty} \frac{1}{h} k\left(\frac{X_i - x}{h}\right) dx = \int_{-\infty}^{\infty} k\left(u\right) du = 1.$$

The change-of variables $u = (X_i - x)/h$ will be used frequently, so it is useful to be familiar with this transformation. Thus

$$\int_{-\infty}^{\infty} \hat{f}(x)dx = \int_{-\infty}^{\infty} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h}k\left(\frac{X_i - x}{h}\right) dx = \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} \frac{1}{h}k\left(\frac{X_i - x}{h}\right) dx = \frac{1}{n}\sum_{i=1}^{n} 1 = 1$$

as claimed. Thus $\hat{f}(x)$ is a valid density function when k is non-negative.

Third, we can also calculate the numerical moments of the density $\hat{f}(x)$. Again using the change-of-variables $u = (X_i - x)/h$, the mean of the estimated density is

$$\begin{aligned}
\int_{-\infty}^{\infty} x\hat{f}(x)dx &= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} x\frac{1}{h}k\left(\frac{X_i - x}{h}\right) dx \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} (X_i + uh)\, k(u)\, du \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i \int_{-\infty}^{\infty} k(u)\, du + \frac{1}{n}\sum_{i=1}^{n} h \int_{-\infty}^{\infty} uk(u)\, du \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i
\end{aligned}$$

the sample mean of the X_i.

The second moment of the estimated density is

$$\begin{aligned}
\int_{-\infty}^{\infty} x^2\hat{f}(x)dx &= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} x^2\frac{1}{h}k\left(\frac{X_i - x}{h}\right) dx \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} (X_i + uh)^2\, k(u)\, du \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + \frac{2}{n}\sum_{i=1}^{n} X_i h \int_{-\infty}^{\infty} k(u)du + \frac{1}{n}\sum_{i=1}^{n} h^2 \int_{-\infty}^{\infty} u^2 k(u)\, du \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + h^2\kappa_2(k).
\end{aligned}$$

It follows that the variance of the density $\hat{f}(x)$ is

$$\begin{aligned}
\int_{-\infty}^{\infty} x^2\hat{f}(x)dx - \left(\int_{-\infty}^{\infty} x\hat{f}(x)dx\right)^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + h^2\kappa_2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \\
&= \hat{\sigma}^2 + h^2\kappa_2(k)
\end{aligned}$$

where $\hat{\sigma}^2$ is the sample variance. Thus the density estimate inflates the sample variance by the factor $h^2\kappa_2(k)$.

These are the numerical mean and variance of the estimated density $\hat{f}(x)$, not its sampling

mean and variance.

2.4 Estimation Bias

It is useful to observe that expectations of kernel transformations can be written as integrals which take the form of a convolution of the kernel and the density function:

$$\mathrm{E}\frac{1}{h}k\left(\frac{X_i - x}{h}\right) = \int_{-\infty}^{\infty} \frac{1}{h}k\left(\frac{z - x}{h}\right) f(z)dz$$

Using the change-of variables $u = (z - x)/h$, this equals

$$\int_{-\infty}^{\infty} k\left(u\right) f(x + hu)du.$$

By the linearity of the estimator we see

$$\mathrm{E}\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\frac{1}{h}k\left(\frac{X_i - x}{h}\right) = \int_{-\infty}^{\infty} k\left(u\right) f(x + hu)du$$

The last expression shows that the expected value is an average of $f(z)$ locally about x.

This integral (typically) is not analytically solvable, so we approximate it using a Taylor expansion of $f(x + hu)$ in the argument hu, which is valid as $h \to 0$. For a ν'th-order kernel we take the expansion out to the ν'th term

$$\begin{aligned} f\left(x + hu\right) &= f(x) + f^{(1)}(x)hu + \frac{1}{2}f^{(2)}(x)h^2u^2 + \frac{1}{3!}f^{(3)}(x)h^3u^3 + \cdots \\ &\quad + \frac{1}{\nu!}f^{(\nu)}(x)h^\nu u^\nu + o\left(h^\nu\right). \end{aligned}$$

The remainder is of smaller order than h^ν as $h \to \infty$, which is written as $o(h^\nu)$. (This expansion assumes $f^{(\nu+1)}(x)$ exists.)

Integrating term by term and using $\int_{-\infty}^{\infty} k\left(u\right) du = 1$ and the definition $\int_{-\infty}^{\infty} k\left(u\right) u^j du = \kappa_j(k)$,

$$\begin{aligned} \int_{-\infty}^{\infty} k\left(u\right) f\left(x + hu\right) du &= f(x) + f^{(1)}(x)h\kappa_1(k) + \frac{1}{2}f^{(2)}(x)h^2\kappa_2(k) + \frac{1}{3!}f^{(3)}(x)h^3\kappa_3(k) + \cdots \\ &\quad + \frac{1}{\nu!}f^{(\nu)}(x)h^\nu\kappa_\nu(k) + o\left(h^\nu\right) \\ &= f(x) + \frac{1}{\nu!}f^{(\nu)}(x)h^\nu\kappa_\nu(k) + o\left(h^\nu\right) \end{aligned}$$

where the second equality uses the assumption that k is a ν'th order kernel (so $\kappa_j(k) = 0$ for $j < \nu$).

This means that

$$\begin{aligned} \mathrm{E}\hat{f}(x) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\frac{1}{h}k\left(\frac{X_i - x}{h}\right) \\ &= f(x) + \frac{1}{\nu!}f^{(\nu)}(x)h^{\nu}\kappa_{\nu}(k) + o\left(h^{\nu}\right). \end{aligned}$$

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

$$Bias(\hat{f}(x)) = \mathrm{E}\hat{f}(x) - f(x) = \frac{1}{\nu!}f^{(\nu)}(x)h^{\nu}\kappa_{\nu}(k) + o\left(h^{\nu}\right).$$

For second-order kernels, this simplifies to

$$Bias(\hat{f}(x)) = \frac{1}{2}f^{(2)}(x)h^{2}\kappa_{2}(k) + O\left(h^{4}\right).$$

For second-order kernels, the bias is increasing in the square of the bandwidth. Smaller bandwidths imply reduced bias. The bias is also proportional to the second derivative of the density $f^{(2)}(x)$. Intuitively, the estimator $\hat{f}(x)$ smooths data local to $X_i = x$, so is estimating a smoothed version of $f(x)$. The bias results from this smoothing, and is larger the greater the curvature in $f(x)$.

When higher-order kernels are used (and the density has enough derivatives), the bias is proportional to h^{ν}, which is of lower order than h^2. Thus the bias of estimates using higher-order kernels is of lower order than estimates from second-order kernels, and this is why they are called bias-reducing kernels. This is the advantage of higher-order kernels.

2.5 Estimation Variance

Since the kernel estimator is a linear estimator, and $k\left(\frac{X_i - x}{h}\right)$ is iid,

$$\begin{aligned} \mathrm{var}\left(\hat{f}(x)\right) &= \frac{1}{nh^2}\mathrm{var}\left(k\left(\frac{X_i - x}{h}\right)\right) \\ &= \frac{1}{nh^2}\mathrm{E}k\left(\frac{X_i - x}{h}\right)^2 - \frac{1}{n}\left(\frac{1}{h}\mathrm{E}k\left(\frac{X_i - x}{h}\right)\right)^2 \end{aligned}$$

From our analysis of bias we know that $\frac{1}{h}\mathrm{E}k\left(\frac{X_i - x}{h}\right) = f(x) + o(1)$ so the second term is $O\left(\frac{1}{n}\right)$. For the first term, write the expectation as an integral, make a change-of-variables and a first-order

Taylor expansion

$$\begin{aligned}\frac{1}{h}\mathrm{E}k\left(\frac{X_i-x}{h}\right)^2 &= \frac{1}{h}\int_{-\infty}^{\infty}k\left(\frac{z-x}{h}\right)^2 f(z)dz \\ &= \int_{-\infty}^{\infty}k\left(u\right)^2 f\left(x+hu\right)du \\ &= \int_{-\infty}^{\infty}k\left(u\right)^2 \left(f\left(x\right)+O\left(h\right)\right)du \\ &= f\left(x\right)R(k)+O\left(h\right)\end{aligned}$$

where $R(k)=\int_{-\infty}^{\infty}k\left(u\right)^2 du$ is the roughness of the kernel. Together, we see

$$\mathrm{var}\left(\hat{f}(x)\right)=\frac{f\left(x\right)R(k)}{nh}+O\left(\frac{1}{n}\right)$$

The remainder $O\left(\frac{1}{n}\right)$ is of smaller order than the $O\left(\frac{1}{nh}\right)$ leading term, since $h^{-1}\to\infty$.

2.6 Mean-Squared Error

A common and convenient measure of estimation precision is the mean-squared error

$$\begin{aligned}MSE(\hat{f}(x)) &= \mathrm{E}\left(\hat{f}(x)-f(x)\right)^2 \\ &= Bias(\hat{f}(x))^2+\mathrm{var}\left(\hat{f}(x)\right) \\ &\simeq \left(\frac{1}{\nu!}f^{(\nu)}(x)h^{\nu}\kappa_{\nu}(k)\right)^2+\frac{f\left(x\right)R(k)}{nh} \\ &= \frac{\kappa_{\nu}^2(k)}{\left(\nu!\right)^2}f^{(\nu)}(x)^2h^{2\nu}+\frac{f\left(x\right)R(k)}{nh} \\ &= AMSE(\hat{f}(x))\end{aligned}$$

Since this approximation is based on asymptotic expansions this is called the asymptotic mean-squared-error (AMSE). Note that it is a function of the sample size n, the bandwidth h, the kernel function (through κ_{ν} and $R(k)$), and varies with x as $f^{(\nu)}(x)$ and $f(x)$ vary.

Notice as well that the first term (the squared bias) is increasing in h and the second term (the variance) is decreasing in nh. For $MSE(\hat{f}(x))$ to decline as $n\to\infty$ both of these terms must get small. Thus as $n\to\infty$ we must have $h\to 0$ and $nh\to\infty$. That is, the bandwidth must decrease, but not at a rate faster than sample size. This is sufficient to establish the pointwise consistency of the estimator. That is, for all x, $\hat{f}(x)\to_p f(x)$ as $n\to\infty$. We call this pointwise convergence as it is valid for each x individually. We discuss uniform convergence later.

A global measure of precision is the asymptotic mean integrated squared error (AMISE)

$$\begin{aligned} AMISE &= \int_{-\infty}^{\infty} AMSE(\hat{f}(x))dx \\ &= \frac{\kappa_\nu^2(k)}{(\nu!)^2} R\left(f^{(\nu)}\right) h^{2\nu} + \frac{R(k)}{nh}. \end{aligned}$$

where $R(f^{(\nu)}) = \int_{-\infty}^{\infty} \left(f^{(\nu)}(x)\right)^2 dx$ is the roughness of $f^{(\nu)}$.

2.7 Asymptotically Optimal Bandwidth

The AMISE formula expresses the MSE as a function of h. The value of h which minimizes this expression is called the asymptotically optimal bandwidth. The solution is found by taking the derivative of the AMISE with respect to h and setting it equal to zero:

$$\begin{aligned} \frac{d}{dh} AMISE &= \frac{d}{dh}\left(\frac{\kappa_\nu^2(k)}{(\nu!)^2} R\left(f^{(\nu)}\right) h^{2\nu} + \frac{R(k)}{nh}\right) \\ &= 2\nu h^{2\nu-1} \frac{\kappa_\nu^2(k)}{(\nu!)^2} R\left(f^{(\nu)}\right) - \frac{R(k)}{nh^2} \\ &= 0 \end{aligned}$$

with solution

$$\begin{aligned} h_0 &= C_\nu(k,f)\, n^{-1/(2\nu+1)} \\ C_\nu(k,f) &= R\left(f^{(\nu)}\right)^{-1/(2\nu+1)} A_\nu(k) \\ A_\nu(k) &= \left(\frac{(\nu!)^2 R(k)}{2\nu\kappa_\nu^2(k)}\right)^{1/(2\nu+1)} \end{aligned}$$

The optimal bandwidth is propotional to $n^{-1/(2\nu+1)}$. We say that the optimal bandwidth is of order $O\left(n^{-1/(2\nu+1)}\right)$. For second-order kernels the optimal rate is $O\left(n^{-1/5}\right)$. For higher-order kernels the rate is slower, suggesting that bandwidths are generally larger than for second-order kernels. The intuition is that since higher-order kernels have smaller bias, they can afford a larger bandwidth.

The constant of proportionality $C_\nu(k,f)$ depends on the kernel through the function $A_\nu(k)$ (which can be calculated from Table 1), and the density through $R(f^{(\nu)})$ (which is unknown).

If the bandwidth is set to h_0, then with some simplification the AMISE equals

$$AMISE_0(k) = (1+2\nu)\left(\frac{R\left(f^{(\nu)}\right)\kappa_\nu^2(k) R(k)^{2\nu}}{(\nu!)^2 (2\nu)^{2\nu}}\right)^{1/(2\nu+1)} n^{-2\nu/(2\nu+1)}.$$

For second-order kernels, this equals

$$AMISE_0(k) = \frac{5}{4}\left(\kappa_2^2(k)R(k)^4 R\left(f^{(2)}\right)\right)^{1/5} n^{-4/5}.$$

As ν gets large, the convergence rate approaches the parametric rate n^{-1}. Thus, at least asymptotically, the slow convergence of nonparametric estimation can be mitigated through the use of higher-order kernels.

This seems a bit magical. What's the catch? For one, the improvement in convergence rate requires that the density is sufficiently smooth that derivatives exist up to the $(\nu+1)$'th order. As the density becomes increasingly smooth, it is easier to approximate by a low-dimensional curve, and gets closer to a parametric-type problem. This is exploiting the smoothness of f, which is inherently unknown. The other catch is that there is a some evidence that the benefits of higher-order kernels only develop when the sample size is fairly large. My sense is that in small samples, a second-order kernel would be the best choice, in moderate samples a 4th order kernel, and in larger samples a 6th order kernel could be used.

2.8 Asymptotically Optimal Kernel

Given that we have picked the kernel order, which kernel should we use? Examining the expression $AMISE_0$ we can see that for fixed ν the choice of kernel affects the asymptotic precision through the quantity $\kappa_\nu(k)\,R(k)^\nu$. All else equal, $AMISE$ will be minimized by selecting the kernel which minimizes this quantity. As we discussed earlier, only the shape of the kernel is important, not its scale, so we can set $\kappa_\nu = 1$. Then the problem reduces to minimization of $R(k) = \int_{-\infty}^{\infty} k(u)^2 du$ subject to the constraints $\int_{-\infty}^{\infty} k(u)du = 1$ and $\int_{-\infty}^{\infty} u^\nu k(u)du = 1$. This is a problem in the calculus of variations. It turns out that the solution is a scaled of $k_{\nu,1}$ (see Muller (Annals of Statistics, 1984)). As the scale is irrelevant, this means that for estimation of the density function, the higher-order Epanechikov kernel $k_{\nu,1}$ with optimal bandwidth yields the lowest possible AMISE. For this reason, the Epanechikov kernel is often called the "optimal kernel".

To compare kernels, its relative efficiency is defined as

$$\begin{aligned} eff(k) &= \left(\frac{AMISE_0(k)}{AMISE_0(k_{\nu,1})}\right)^{(1+2\nu)/2\nu} \\ &= \frac{\left(\kappa_\nu^2(k)\right)^{1/2\nu} R(k)}{\left(\kappa_\nu^2(k_{\nu,1})\right)^{1/2\nu} R(k_{\nu,1})} \end{aligned}$$

The ratios of the AMISE is raised to the power $(1+2\nu)/2\nu$ as for large n, the AMISE will be the same whether we use n observations with kernel $k_{\nu,1}$ or $n\,eff(k)$ observations with kernel k. Thus the penalty $eff(k)$ is expressed as a percentage of observations.

The efficiencies of the various kernels are given in Tables 1-3. Examining the second-order kernels, we see that relative to the Epanechnikov kernel, the uniform kernel pays a penalty of about 7%, the Gaussian kernel a penalty of about 5%, the Triweight kernel about 1.4%, and the Biweight

kernel less than 1%. Examining the 4th and 6th-order kernels, we see that the relative efficiency of the Gaussian kernel deteriorates, while that of the Biweight and Triweight slightly improves.

The differences are not big. Still, the calculation suggests that the Epanechnikov and Biweight kernel classes are good choices for density estimation.

2.9 Rule-of-Thumb Bandwidth

The optimal bandwidth depends on the unknown quantity $R\left(f^{(\nu)}\right)$. Silverman proposed that we try the bandwidth computed by replacing $R\left(f^{(\nu)}\right)$ in the optimal formula by $R\left(g_{\hat{\sigma}}^{(\nu)}\right)$ where g_σ is a reference density – a plausible candidate for f, and $\hat{\sigma}^2$ is the sample standard deviation. The standard choice is to set $g_\sigma = \phi_{\hat{\sigma}}$, the $N(0, \hat{\sigma}^2)$ density. The idea is that if the true density is normal, then the computed bandwidth will be optimal. If the true density is reasonably close to the normal, then the bandwidth will be close to optimal. While not a perfect solution, it is a good place to start looking.

For any density g, if we set $g_\sigma(x) = \sigma^{-1} g(x/\sigma)$, then $g_\sigma^{(\nu)}(x) = \sigma^{-1-\nu} g^{(\nu)}(x/\sigma)$. Thus

$$\begin{aligned} R\left(g_\sigma^{(\nu)}\right)^{-1/(2\nu+1)} &= \left(\int g_\sigma^{(\nu)}(x)^2 dx\right)^{-1/(2\nu+1)} \\ &= \left(\sigma^{-2-2\nu}\int g^{(\nu)}(x/\sigma)^2 dx\right)^{-1/(2\nu+1)} \\ &= \left(\sigma^{-1-2\nu}\int g^{(\nu)}(x)^2 dx\right)^{-1/(2\nu+1)} \\ &= \sigma R\left(g^{(\nu)}\right)^{-1/(2\nu+1)}. \end{aligned}$$

Furthermore,

$$\left(R\left(\phi^{(\nu)}\right)\right)^{-1/(2\nu+1)} = 2\left(\frac{\pi^{1/2}\nu!}{(2\nu)!}\right)^{1/(2\nu+1)}.$$

Thus

$$R\left(\phi_{\hat{\sigma}}^{(\nu)}\right)^{-1/(2\nu+1)} = 2\hat{\sigma}\left(\frac{\pi^{1/2}\nu!}{(2\nu)!}\right)^{1/(2\nu+1)}.$$

The rule-of-thumb bandwidth is then $h = \hat{\sigma} C_\nu(k)\, n^{-1/(2\nu+1)}$ where

$$\begin{aligned} C_\nu(k) &= R\left(\phi^{(\nu)}\right)^{-1/(2\nu+1)} A_\nu(k) \\ &= 2\left(\frac{\pi^{1/2}(\nu!)^3 R(k)}{2\nu(2\nu)!\kappa_\nu^2(k)}\right)^{1/(2\nu+1)} \end{aligned}$$

We collect these constants in Table 4.

Table 4: Rule of Thumb Constants

Kernel	$\nu = 2$	$\nu = 4$	$\nu = 6$
Epanechnikov	2.34	3.03	3.53
Biweight	2.78	3.39	3.84
Triweight	3.15	3.72	4.13
Gaussian	1.06	1.08	1.08

Silverman Rule-of-Thumb: $h = \hat{\sigma} C_\nu(k) n^{-1/(2\nu+1)}$ where $\hat{\sigma}$ is the sample standard deviation, ν is the order of the kernel, and $C_\nu(k)$ is the constant from Table 4.

If a Gaussian kernel is used, this is often simplified to $h = \hat{\sigma} n^{-1/(2\nu+1)}$. In particular, for the standard second-order normal kernel, $h = \hat{\sigma} n^{-1/5}$.

2.10 Density Derivatives

Consider the problem of estimating the r'th derivative of the density:

$$f^{(r)}(x) = \frac{d^r}{dx^r} f(x).$$

A natural estimator is found by taking derivatives of the kernel density estimator. This takes the form

$$\hat{f}^{(r)}(x) = \frac{d^r}{dx^r} \hat{f}(x) = \frac{1}{nh^{1+r}} \sum_{i=1}^{n} k^{(r)}\left(\frac{X_i - x}{h}\right)$$

where

$$k^{(r)}(x) = \frac{d^r}{dx^r} k(x).$$

This estimator only makes sense if $k^{(r)}(x)$ exists and is non-zero. Since the Gaussian kernel has derivatives of all orders this is a common choice for derivative estimation.

The asymptotic analysis of this estimator is similar to that of the density, but with a couple of extra wrinkles and noticably different results. First, to calculate the bias we observe that

$$\mathrm{E}\frac{1}{h^{1+r}} k^{(r)}\left(\frac{X_i - x}{h}\right) = \int_{-\infty}^{\infty} \frac{1}{h^{1+r}} k^{(r)}\left(\frac{z - x}{h}\right) f(z) dz$$

To simplify this expression we use integration by parts. As the integral of $h^{-1} k^{(r)}\left(\frac{z-x}{h}\right)$ is $-k^{(r-1)}\left(\frac{z-x}{h}\right)$, we find that the above expression equals

$$\int_{-\infty}^{\infty} \frac{1}{h^r} k^{(r-1)}\left(\frac{z - x}{h}\right) f^{(1)}(z) dz.$$

Repeating this a total of r times, we obtain

$$\int_{-\infty}^{\infty} \frac{1}{h} k\left(\frac{z - x}{h}\right) f^{(r)}(z) dz.$$

Next, apply the change of variables to obtain

$$\int_{-\infty}^{\infty} k(u) f^{(r)}(x+hu)dz.$$

Now expand $f^{(r)}(x+hu)$ in a ν'th-order Taylor expansion about x, and integrate the terms to find that the above equals

$$f^{(r)}(x) + \frac{1}{\nu!} f^{(r+\nu)}(x) h^{\nu} \kappa_{\nu}(k) + o(h^{\nu})$$

where ν is the order of the kernel. Hence the asymptotic bias is

$$\begin{aligned} Bias(\hat{f}^{(r)}(x)) &= \mathrm{E}\hat{f}^{(r)}(x) - f^{(r)}(x) \\ &= \frac{1}{\nu!} f^{(r+\nu)}(x) h^{\nu} \kappa_{\nu}(k) + o(h^{\nu}). \end{aligned}$$

This of course presumes that f is differentiable of order at least $r+\nu+1$.

For the variance, we find

$$\begin{aligned} \mathrm{var}\left(\hat{f}^{(r)}(x)\right) &= \frac{1}{nh^{2+2r}} \mathrm{var}\left(k^{(r)}\left(\frac{X_i - x}{h}\right)\right) \\ &= \frac{1}{nh^{2+2r}} \mathrm{E}k^{(r)}\left(\frac{X_i - x}{h}\right)^2 - \frac{1}{n}\left(\frac{1}{nh^{1+r}} \mathrm{E}k^{(r)}\left(\frac{X_i - x}{h}\right)\right)^2 \\ &= \frac{1}{nh^{2+2r}} \int_{-\infty}^{\infty} k^{(r)}\left(\frac{z-x}{h}\right)^2 f(z)dz - \frac{1}{n} f^{(r)}(x)^2 + O\left(\frac{1}{n}\right) \\ &= \frac{1}{nh^{1+2r}} \int_{-\infty}^{\infty} k^{(r)}(u)^2 f(x+hu)\, du + O\left(\frac{1}{n}\right) \\ &= \frac{f(x)}{nh^{1+2r}} \int_{-\infty}^{\infty} k^{(r)}(u)^2 du + O\left(\frac{1}{n}\right) \\ &= \frac{f(x) R(k^{(r)})}{nh^{1+2r}} + O\left(\frac{1}{n}\right). \end{aligned}$$

The AMSE and AMISE are

$$AMSE(\hat{f}^{(r)}(x)) = \frac{f^{(r+\nu)}(x)^2 h^{2\nu} \kappa_{\nu}^2(k)}{(\nu!)^2} + \frac{f(x) R(k^{(r)})}{nh^{1+2r}}$$

and

$$AMISE(\hat{f}^{(r)}(x)) = \frac{R\left(f^{(r+\nu)}\right) h^{2\nu} \kappa_{\nu}^2(k)}{(\nu!)^2} + \frac{R(k^{(r)})}{nh^{1+2r}}.$$

Note that the order of the bias is the same as for estimation of the density. But the variance is now of order $O\left(\frac{1}{nh^{1+2r}}\right)$ which is much larger than the $O\left(\frac{1}{nh}\right)$ found earlier.

The asymptotically optimal bandwidth is

$$\begin{aligned} h_r &= C_{r,\nu}(k,f)\, n^{-1/(1+2r+2\nu)} \\ C_{r,\nu}(k,f) &= R\left(f^{(r+\nu)}\right)^{-1/(1+2r+2\nu)} A_{r,\nu}(k) \\ A_{r,\nu}(k) &= \left(\frac{(1+2r)\,(\nu!)^2\, R(k^{(r)})}{2\nu\kappa_\nu^2(k)}\right)^{1/(1+2r+2\nu)} \end{aligned}$$

Thus the optimal bandwidth converges at a slower rate than for density estimation. Given this bandwidth, the rate of convergence for the AMISE is $O\left(n^{-2\nu/(2r+2\nu+1)}\right)$, which is slower than the $O\left(n^{-2\nu/(2\nu+1)}\right)^{-4/5}$) rate when $r=0$.

We see that we need a different bandwidth for estimation of derivatives than for estimation of the density. This is a common situation which arises in nonparametric analysis. The optimal amount of smoothing depends upon the object being estimated, and the goal of the analysis.

The AMISE with the optimal bandwidth is

$$AMISE(\hat{f}^{(r)}(x)) = (1+2r+2\nu)\left(\frac{\kappa_\nu^2(k)}{(\nu!)^2(1+2r)}\right)^{(2r+1)/(1+2r+2\nu)}\left(\frac{R\left(k^{(r)}\right)}{2\nu}\right)^{2\nu/(1+2r+2\nu)} n^{-2\nu/(1+2r+2\nu)}.$$

We can also ask the question of which kernel function is optimal, and this is addressed by Muller (1984). The problem amounts to minimizing $R\left(k^{(r)}\right)$ subject to a moment condition, and the solution is to set k equal to $k_{\nu,r+1}$, the polynomial kernel of ν'th order and exponent $r+1$. Thus to a first derivative it is optimal to use a member of the Biweight class and for a second derivative a member of the Triweight class.

The relative efficiency of a kernel k is then

$$\begin{aligned} eff(k) &= \left(\frac{AMISE_0(k)}{AMISE_0(k_{\nu,r+1})}\right)^{(1+2\nu+2r)/2\nu} \\ &= \left(\frac{\kappa_\nu^2(k)}{\kappa_\nu^2(k_{\nu,r+1})}\right)^{(1+2r)/2\nu} \frac{R\left(k^{(r)}\right)}{R\left(k_{\nu,r+1}^{(r)}\right)}. \end{aligned}$$

The relative efficiencies of the various kernels are presented in Table 5. (The Epanechnikov kernel is not considered as it is inappropriate for derivative estimation, and similarly the Biweight kernel for $r=2$). In contrast to the case $r=0$, we see that the Gaussian kernel is highly inefficient, with the efficiency loss increasing with r and ν. These calculations suggest that when estimating density derivatives it is important to use the appropriate kernel.

Table 5: Relative Efficiency $eff(k)$

		Biweight	Triweight	Gaussian
$r=1$	$\nu=2$	1.0000	1.0185	1.2191
	$\nu=4$	1.0000	1.0159	1.2753
	$\nu=6$	1.0000	1.0136	1.3156
$r=2$	$\nu=2$		1.0000	1.4689
	$\nu=4$		1.0000	1.5592
	$\nu=6$		1.0000	1.6275

The Silverman Rule-of-Thumb may also be applied to density derivative estimation. Again using the reference density $g_\sigma = \phi_\sigma$, we find the rule-of-thumb bandwidth is $h = C_{r,\nu}(k)\,\hat{\sigma} n^{-1/(2r+2\nu+1)}$ where

$$C_{r,\nu}(k) = 2\left(\frac{\pi^{1/2}(1+2r)(\nu!)^2(r+\nu)!R\left(k^{(r)}\right)}{2\nu\kappa_\nu^2(k)(2r+2\nu)!}\right)^{1/(2r+2\nu+1)}.$$

The constants $C_{r,v}$ are collected in Table 6. For all kernels, the constants $C_{r,\nu}$ are similar but slightly decreasing as r increases.

Table 6: Rule of Thumb Constants

		Biweight	Triweight	Gaussian
$r=1$	$\nu=2$	2.49	2.83	0.97
	$\nu=4$	3.18	3.49	1.03
	$\nu=6$	3.44	3.96	1.04
$r=2$	$\nu=2$		2.70	0.94
	$\nu=4$		3.35	1.00
	$\nu=6$		3.84	1.02

2.11 Multivariate Density Estimation

Now suppose that X_i is a q-vector and we want to estimate its density $f(x) = f(x_1, ..., x_q)$. A multivariate kernel estimator takes the form

$$\hat{f}(x) = \frac{1}{n|H|}\sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right)$$

where $K(u)$ is a multivariate kernel function depending on a bandwidth vector $H = (h_1, ..., h_q)'$ and $|H| = h_1 h_2 \cdots h_q$. A multivariate kernel satisfies That is,

$$\int K(u)(du) = \int K(u) du_1 \cdots du_q = 1$$

Typically, $K(u)$ takes the product form:

$$K(u) = k(u_1)\,k(u_2)\cdots k(u_q).$$

As in the univariate case, $\hat{f}(x)$ has the property that it integrates to one, and is non-negative if $K(u) \geq 0$. When $K(u)$ is a product kernel then the marginal densities of $\hat{f}(x)$ equal univariate kernel density estimators with kernel functions k and bandwidths h_j.

With some work, you can show that when $K(u)$ takes the product form, the bias of the estimator is

$$Bias(\hat{f}(x)) = \frac{\kappa_\nu(k)}{\nu!} \sum_{j=1}^{q} \frac{\partial^\nu}{\partial x_j^\nu} f(x) h_j^\nu + o\left(h_1^\nu + \cdots + h_q^\nu\right)$$

and the variance is

$$\begin{aligned} \operatorname{var}\left(\hat{f}(x)\right) &= \frac{f(x)\, R(K)}{n\,|H|} + O\left(\frac{1}{n}\right) \\ &= \frac{f(x)\, R(k)^q}{n h_1 h_2 \cdots h_q} + O\left(\frac{1}{n}\right). \end{aligned}$$

Hence the AMISE is

$$AMISE\left(\hat{f}(x)\right) = \frac{\kappa_\nu^2(k)}{(\nu!)^2} \int \left(\sum_{j=1}^{q} \frac{\partial^\nu}{\partial x_j^\nu} f(x) h_j^\nu \right)^2 (dx) + \frac{R(k)^q}{n h_1 h_2 \cdots h_q}$$

There is no closed-form solution for the bandwidth vector which minimizes this expression. However, even without doing do, we can make a couple of observations.

First, the AMISE depends on the kernel function only through $R(k)$ and $\kappa_\nu^2(k)$, so it is clear that for any given ν, the optimal kernel minimizes $R(k)$, which is the same as in the univariate case.

Second, the optimal bandwidths will all be of order $n^{-1/(2\nu+q)}$ and the optimal AMISE of order $n^{-2\nu/(2\nu+q)}$. This rates are slower than the univariate ($q = 1$) case. The fact that dimension has an adverse effect on convergence rates is called the **curse of dimensionality**. Many theoretical papers circumvent this problem through the following trick. Suppose you need the AMISE of the estimator to converge at a rate $O\left(n^{-1/2}\right)$ or faster. This requires $2\nu/(2\nu + q) > 1/2$, or $q < 2\nu$. For second-order kernels ($\nu = 2$) this restricts the dimension to be 3 or less. What some authors will do is slip in an assumption of the form: "Assume $f(x)$ is differentiable of order $\nu + 1$ where $\nu > q/2$," and then claim that their results hold for all q. The trouble is that what the author is doing is imposing greater smoothness as the dimension increases. This doesn't really avoid the curse of dimensionality, rather it hides it behind what appears to be a technical assumption. The bottom line is that nonparametric objects are much harder to estimate in higher dimensions, and that is why it is called a "curse".

To derive a rule-of-thumb, suppose that $h_1 = h_2 = \cdots = h_q = h$. Then

$$AMISE\left(\hat{f}(x)\right) = \frac{\kappa_\nu^2(k) R\left(\nabla^\nu f\right)}{(\nu!)^2} h^{2\nu} + \frac{R(k)^q}{n h^q}$$

where

$$\nabla^\nu f(x) = \sum_{j=1}^{q} \frac{\partial^\nu}{\partial x_j^\nu} f(x).$$

We find that the optimal bandwidth is

$$h_0 = \left(\frac{(\nu!)^2 \, q R(k)^q}{2\nu\kappa_\nu^2(k) R\left(\nabla^\nu f\right)} \right)^{1/(2\nu+q)} n^{-1/(2\nu+q)}$$

For a rule-of-thumb bandwidth, we replace f by the multivariate normal density ϕ. We can calculate that

$$R\left(\nabla^\nu \phi\right) = \frac{q}{\pi^{q/2} 2^{q+\nu}} \left((2\nu-1)!! + (q-1)\left((\nu-1)!!\right)^2 \right).$$

Making this substitution, we obtain $h_0 = C_\nu\left(k,q\right) n^{-1/(2\nu+q)}$where

$$C_\nu\left(k,q\right) = \left(\frac{\pi^{q/2} 2^{q+\nu-1} \left(\nu!\right)^2 R(k)^q}{\nu\kappa_\nu^2(k) \left((2\nu-1)!! + (q-1)\left((\nu-1)!!\right)^2 \right)} \right)^{1/(2\nu+q)}.$$

Now this assumed that all variables had unit variance. Rescaling the bandwidths by the standard deviation of each variable, we obtain the rule-of-thumb bandwidth for the j'th variable:

$$h_j = \hat{\sigma}_j C_\nu\left(k,q\right) n^{-1/(2\nu+q)}.$$

Numerical values for the constants $C_\nu(k,q)$ are given in Table 7 for $q = 2, 3, 4$.

Table 7: Rule of Thumb Constants

$\nu = 2$	$q = 2$	$q = 3$	$q = 4$
Epanechnikov	2.20	2.12	2.07
Biweight	2.61	2.52	2.46
Triweight	2.96	2.86	2.80
Gaussian	1.00	0.97	0.95
$\nu = 4$			
Epanechnikov	3.12	3.20	3.27
Biweight	3.50	3.59	3.67
Triweight	3.84	3.94	4.03
Gaussian	1.12	1.16	1.19
$\nu = 6$			
Epanechnikov	3.69	3.83	3.96
Biweight	4.02	4.18	4.32
Triweight	4.33	4.50	4.66
Gaussian	1.13	1.18	1.23

2.12 Least-Squares Cross-Validation

Rule-of-thumb bandwidths are a useful starting point, but they are inflexible and can be far from optimal.

Plug-in methods take the formula for the optimal bandwidth, and replace the unknowns by estimates, e.g. $R\left(\hat{f}^{(\nu)}\right)$. But these initial estimates themselves depend on bandwidths. And each situation needs to be individually studied. Plug-in methods have been thoroughly studied for univariate density estimation, but are less well developed for multivariate density estimation and other contexts.

A flexible and generally applicable data-dependent method is cross-validation. This method attempts to make a direct estimate of the squared error, and pick the bandwidth which minimizes this estimate. In many senses the idea is quite close to model selection based on a information criteria, such as Mallows or AIC.

Given a bandwidth h and density estimate $\hat{f}(x)$ of $f(x)$, define the mean integrated squared error (MISE)

$$MISE(h) = \int \left(\hat{f}(x) - f(x)\right)^2 (dx) = \int \hat{f}(x)^2 (dx) - 2\int \hat{f}(x) f(x) (dx) + \int f(x)^2 (dx)$$

Optimally, we want $\hat{f}(x)$ to be as close to $f(x)$ as possible, and thus for $MISE(h)$ to be as small as possible.

As $MISE(h)$ is unknown, cross-validation replaces it with an estimate.

The goal is to find an estimate of $MISE(h)$, and find the h which minimizes this estimate.

As the third term in the above expression does not depend on the bandwidth h, it can be ignored.

The first term can be directly calculated.

For the univariate case

$$\begin{aligned}
\int \hat{f}(x)^2 dx &= \int \left(\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right)\right)^2 dx \\
&= \frac{1}{n^2 h^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \int k\left(\frac{X_i - x}{h}\right) k\left(\frac{X_j - x}{h}\right) dx
\end{aligned}$$

The convolution of k with itself is $\bar{k}(x) = \int k(u)\, k(x-u)\, du = \int k(u)\, k(u-x)\, du$ (by symmetry of k). Then making the change of variables $u = \dfrac{X_i - x}{h}$,

$$\begin{aligned}
\frac{1}{h}\int k\left(\frac{X_i - x}{h}\right) k\left(\frac{X_i - x}{h}\right) dx &= \int k(u)\, k\left(u - \frac{X_i - X_j}{h}\right) du \\
&= \bar{k}\left(\frac{X_i - X_j}{h}\right).
\end{aligned}$$

Hence

$$\int \hat{f}(x)^2 dx = \frac{1}{n^2 h} \sum_{i=1}^{n} \sum_{j=1}^{n} \bar{k}\left(\frac{X_i - X_j}{h}\right).$$

Discussion of $\bar{k}(x)$ can be found in the following section.

In the multivariate case,

$$\int \hat{f}(x)^2 dx = \frac{1}{n^2 |H|} \sum_{i=1}^{n} \sum_{j=1}^{n} \bar{K}\left(H^{-1}(X_i - X_j)\right)$$

where $\bar{K}(u) = \bar{k}(u_1) \cdots \bar{k}(u_q)$

The second term in the expression for $MISE(h)$ depends on $f(x)$ so is unknown and must be estimated. An integral with respect to $f(x)$ is an expectation with respect to the random variable X_i. While we don't know the true expectation, we have the sample, so can estimate this expectation by taking the sample average. In general, a reasonable estimate of the integral $\int g(x)f(x)dx$ is $\frac{1}{n}\sum_{i=1}^{n} g(X_i)$, suggesting the estimate $\frac{1}{n}\sum_{i=1}^{n} \hat{f}(X_i)$. In this case, however, the function $\hat{f}(x)$ is itself a function of the data. In particular, it is a function of the observation X_i. A way to clean this up is to replace $\hat{f}(X_i)$ with the "leave-one-out" estimate $\hat{f}_{-i}(X_i)$, where

$$\hat{f}_{-i}(x) = \frac{1}{(n-1)|H|} \sum_{j \neq i} K\left(H^{-1}(X_j - x)\right)$$

is the density estimate computed without observation X_i, and thus

$$\hat{f}_{-i}(X_i) = \frac{1}{(n-1)|H|} \sum_{j \neq i} K\left(H^{-1}(X_j - X_i)\right).$$

That is, $\hat{f}_{-i}(X_i)$ is the density estimate at $x = X_i$, computed with the observations except X_i. We end up suggesting to estimate $\int \hat{f}(x)f(x)dx$ with

$$\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i) = \frac{1}{n(n-1)|H|} \sum_{i=1}^{n} \sum_{j \neq i} K\left(H^{-1}(X_j - X_i)\right)$$

. It turns out that this is an unbiased estimate, in the sense that

$$\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i)\right) = \mathrm{E}\left(\int \hat{f}(x)f(x)dx\right)$$

To see this, the LHS is

$$\begin{aligned}
\mathrm{E}\hat{f}_{-n}(X_n) &= \mathrm{E}\left(\mathrm{E}\left(\hat{f}_{-n}(X_n) \mid X_1, ..., X_{n-1}\right)\right) \\
&= \mathrm{E}\left(\int \hat{f}_{-n}(x) f(x)\,(dx)\right) \\
&= \int \mathrm{E}\left(\hat{f}(x)\right) f(x)\,(dx) \\
&= \mathrm{E}\left(\int \hat{f}(x) f(x)\,(dx)\right)
\end{aligned}$$

the second-to-last equality exchanging integration, and since $\mathrm{E}\left(\hat{f}(x)\right)$ depends only in the bandwidth, not the sample size.

Together, the least-squares cross-validation criterion is

$$CV(h_1, ..., h_q) = \frac{1}{n^2 |H|} \sum_{i=1}^{n} \sum_{j=1}^{n} \bar{K}\left(H^{-1}(X_i - X_j)\right) - \frac{2}{n(n-1)|H|} \sum_{i=1}^{n} \sum_{j \neq i} K\left(H^{-1}(X_j - X_i)\right).$$

Another way to write this is

$$\begin{aligned}
CV(h_1, ..., h_q) &= \frac{\bar{K}(0)}{n|H|} + \frac{1}{n^2 |H|} \sum_{i=1}^{n} \sum_{j \neq i} \bar{K}\left(H^{-1}(X_i - X_j)\right) - \frac{2}{n(n-1)|H|} \sum_{i=1}^{n} \sum_{j \neq i} K\left(H^{-1}(X_j - X_i)\right) \\
&\simeq \frac{R(k)^q}{n|H|} + \frac{1}{n^2 |H|} \sum_{i=1}^{n} \sum_{j \neq i} \left(\bar{K}\left(H^{-1}(X_i - X_j)\right) - 2K\left(H^{-1}(X_j - X_i)\right)\right)
\end{aligned}$$

using $\bar{K}(0) = \bar{k}(0)^q$ and $\bar{k}(0) = \int k(u)^2$, and the approximation is by replacing $n-1$ by n.

The cross-validation bandwidth vector are the value $\hat{h}_1, ..., \hat{h}_q$ which minimizes $CV(h_1, ..., h_q)$. The cross-validation function is a complicated function of the bandwidhts, so this needs to be done numerically.

In the univariate case, h is one-dimensional this is typically done by plotting (a grid search). Pick a lower and upper value $[h_1, h_2]$, define a grid on this set, and compute $CV(h)$ for each h in the grid. A plot of $CV(h)$ against h is a useful diagnostic tool.

The $CV(h)$ function can be misleading for small values of h. This arises when there is data rounding. Some authors define the cross-validation bandwidth as the largest local minimer of $CV(h)$ (rather than the global minimizer). This can also be avoided by picking a sensible initial range $[h_1, h_2]$. The rule-of-thumb bandwidth can be useful here. If h_0 is the rule-of-thumb bandwidth, then use $h_1 = h_0/3$ and $h_2 = 3h_0$ or similar.

We we discussed above, $CV(h_1, ..., h_q) + \int f(x)^2\,(dx)$ is an unbiased estimate of $MISE(h)$. This by itself does not mean that $\hat{h}$ is a good estimate of h_0, the minimizer of $MISE(h)$, but it

turns out that this is indeed the case. That is,

$$\frac{\hat{h} - h_0}{h_0} \to_p 0$$

Thus, $\hat{h}$ is asymptotically close to h_0, but the rate of convergence is very slow.

The CV method is quite flexible, as it can be applied for any kernel function.

If the goal, however, is estimation of density derivatives, then the CV bandwidth $\hat{h}$ is not appropriate. A practical solution is the following. Recall that the asymptotically optimal bandwidth for estimation of the density takes the form $h_0 = C_\nu(k, f)\, n^{-1/(2\nu+1)}$ and that for the r'th derivative is $h_r = C_{r,\nu}(k, f)\, n^{-1/(1+2r+2\nu)}$. Thus if the CV bandwidth $\hat{h}$ is an estimate of h_0, we can estimate $C_\nu(k, f)$ by $\hat{C}_\nu = \hat{h} n^{1/(2\nu+1)}$. We also saw (at least for the normal reference family) that $C_{r,\nu}(k, f)$ was relatively constant across r. Thus we can replace $C_{r,\nu}(k, f)$ with $\hat{C}_\nu$ to find

$$\begin{aligned}
\hat{h}_r &= \hat{C}_\nu n^{-1/(1+2r+2\nu)} \\
&= \hat{h} n^{1/(2\nu+1) - 1/(1+2r+2\nu)} \\
&= \hat{h} n^{(1+2r+2\nu)/(2\nu+1)(1+2r+2\nu) - (2\nu+1)/(1+2r+2\nu)(2\nu+1)} \\
&= \hat{h} n^{2r/((2\nu+1)(1+2r+2\nu))}
\end{aligned}$$

Alternatively, some authors use the rescaling

$$\hat{h}_r = \hat{h}^{(1+2\nu)/(1+2r+2\nu)}$$

2.13 Convolution Kernels

If $k(x) = \phi(x)$ then $\bar{k}(x) = \exp(-x^2/4)/\sqrt{4\pi}$.

When $k(x)$ is a higher-order Gaussian kernel, Wand and Schucany (Canadian Journal of Statistics, 1990, p. 201) give an expression for $\bar{k}(x)$.

For the polynomial class, because the kernel $k(u)$ has support on $[-1, 1]$, it follows that $\bar{k}(x)$ has support on $[-2, 2]$ and for $x \geq 0$ equals $\bar{k}(x) = \int_{x-1}^{1} k(u)k(x-u)du$. This integral can be easily solved using algebraic software (Maple, Mathematica), but the expression can be rather cumbersome.

For the 2nd order Epanechnikov, Biweight and Triweight kernels, for $0 \leq x \leq 2$,

$$\bar{k}_1(x) = \frac{3}{160}(2-x)^3\left(x^2 + 6x + 4\right)$$

$$\bar{k}_2(x) = \frac{5}{3584}(2-x)^5\left(x^4 + 10x^3 + 36x^2 + 40x + 16\right)$$

$$\bar{k}_3(x) = \frac{35}{1757\,184}(2-x)^7\left(5x^6 + 70x^5 + 404x^4 + 1176x^3 + 1616x^2 + 1120x + 320\right)$$

These functions are symmetric, so the values for $x < 0$ are found by $\bar{k}(x) = \bar{k}(-x)$.

For the 4th, and 6th order Epanechnikov kernels, for $0 \leq x \leq 2$,

$$\bar{k}_{4,1}(x) = \frac{5}{2048}(2-x)^3\left(7x^6+42x^5+48x^4-160x^3-144x^2+96x+64\right)$$

$$\bar{k}_{6,1}(x) = \frac{105}{3407\,872}(2-x)^3\left(495x^{10}+2970x^9+2052x^8-19\,368x^7-32\,624x^6\right.$$
$$\left.+53\,088x^5+68\,352x^4-48\,640x^3-46\,720x^2+11\,520x+7680\right)$$

2.14 Asymptotic Normality

The kernel estimator is the sample average

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|H|}K\left(H^{-1}(X_i - x)\right).$$

We can therefore apply the central limit theorem.

But the convergence rate is not $\sqrt{n}$. We know that

$$\text{var}\left(\hat{f}(x)\right) = \frac{f(x)R(k)^q}{nh_1h_2\cdots h_q} + O\left(\frac{1}{n}\right).$$

so the convergence rate is $\sqrt{nh_1h_2\cdots h_q}$. When we apply the CLT we scale by this, rather than the conventional $\sqrt{n}$.

As the estimator is biased, we also center at its expectation, rather than the true value

Thus

$$\begin{aligned}\sqrt{nh_1h_2\cdots h_q}\left(\hat{f}(x) - E\hat{f}(x)\right) &= \frac{\sqrt{nh_1h_2\cdots h_q}}{n}\sum_{i=1}^{n}\frac{1}{|H|}K\left(H^{-1}(X_i-x)\right) - E\left(\frac{1}{|H|}K\left(H^{-1}(X_i-x)\right)\right)\\ &= \frac{\sqrt{h_1h_2\cdots h_q}}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{1}{|H|}K\left(H^{-1}(X_i-x)\right) - E\left(\frac{1}{|H|}K\left(H^{-1}(X_i-x)\right)\right)\right)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_{ni}\end{aligned}$$

where

$$Z_{ni} = \sqrt{h_1h_2\cdots h_q}\left(\frac{1}{|H|}K\left(H^{-1}(X_i-x)\right) - E\left(\frac{1}{|H|}K\left(H^{-1}(X_i-x)\right)\right)\right)$$

We see that

$$\text{var}(Z_{ni}) \simeq f(x)R(k)^q$$

Hence by the CLT,

$$\sqrt{nh_1h_2\cdots h_q}\left(\hat{f}(x) - E\hat{f}(x)\right) \to_d N\left(0, f(x)R(k)^q\right).$$

We also know that

$$E(\hat{f}(x)) = f(x) + \frac{\kappa_\nu(k)}{\nu!}\sum_{j=1}^{q}\frac{\partial^\nu}{\partial x_j^\nu}f(x)h_j^\nu + o\left(h_1^\nu + \cdots + h_q^\nu\right)$$

So another way of writing this is

$$\sqrt{nh_1h_2\cdots h_q}\left(\hat{f}(x) - f(x) - \frac{\kappa_\nu(k)}{\nu!}\sum_{j=1}^{q}\frac{\partial^\nu}{\partial x_j^\nu}f(x)h_j^\nu\right) \to_d N\left(0, f\left(x\right)R(k)^q\right).$$

In the univariate case this is

$$\sqrt{nh}\left(\hat{f}(x) - f(x) - \frac{\kappa_\nu(k)}{\nu!}f^{(2)}(x)h^\nu\right) \to_d N\left(0, f\left(x\right)R(k)\right)$$

This expression is most useful when the bandwidth is selected to be of optimal order, that is $h = Cn^{-1/(2\nu+1)}$, for then $\sqrt{nh}h^\nu = C^{\nu+1/2}$ and we have the equivalent statement

$$\sqrt{nh}\left(\hat{f}(x) - f(x)\right) \to_d N\left(C^{\nu+1/2}\frac{\kappa_\nu(k)}{\nu!}f^{(2)}(x), f\left(x\right)R(k)\right)$$

This says that the density estimator is asymptotically normal, with a non-zero asymptotic bias and variance.

Some authors play a dirty trick, by using the assumption that h is of smaller order than the optimal rate, e.g. $h = o\left(n^{-1/(2\nu+1)}\right)$. For then then obtain the result

$$\sqrt{nh}\left(\hat{f}(x) - f(x)\right) \to_d N\left(0, f\left(x\right)R(k)\right)$$

This appears much nicer. The estimator is asymptotically normal, with mean zero! There are several costs. One, if the bandwidth is really seleted to be sub-optimal, the estimator is simply less precise. A sub-optimal bandwidth results in a slower convergence rate. This is not a good thing. The reduction in bias is obtained at in increase in variance. Another cost is that the asymptotic distribution is misleading. It suggests that the estimator is unbiased, which is not honest. Finally, it is unclear how to pick this sub-optimal bandwidth. I call this assumption a dirty trick, because it is slipped in by authors to make their results cleaner and derivations easier. This type of assumption should be avoided.

2.15 Pointwise Confidence Intervals

The asymptotic distribution may be used to construct pointwise confidence intervals for $f(x)$. In the univariate case conventional confidence intervals take the form

$$\hat{f}(x) \pm 2\left(\hat{f}\left(x\right)R(k)/\left(nh\right)\right)^{1/2}.$$

These are not necessarily the best choice, since the variance equals the mean. This set has the unfortunate property that it can contain negative values, for example.

Instead, consider constructing the confidence interval by inverting a test statistic. To test $H_0 : f(x) = f_0$, a t-ratio is

$$t(f_0) = \frac{\hat{f}(x) - f_0}{\sqrt{nhf_0R(k)}}.$$

We reject H_0 if $|t(f_0)| > 2$. By the no-rejection rule, an asymptotic 95% confidence interval for f is the set of f_0 which do reject, i.e. the set of f such that $|t(f)| \leq 2$. This is

$$C(x) = \left\{ f : \left| \frac{\hat{f}(x) - f}{\sqrt{nhfR(k)}} \right| \leq 2 \right\}$$

This set must be found numerically.