The following topics are covered today: Basic Nonparametric Regression. There are four books that you can find reference: Silverman(1986), Wand and Jones(1995), Hardle(1990), Hardle(1994)(Ch 38 in the Handbook of Econometrics vol 4).

Key Words: Density estimate, conditional expectation, kernel smoothing, k-NN(k nearest neighborhood), local polynomial, series(sieve), spline, curse of dimensionality, optimal rate of convergence, bias-variance trade-off, undersmoothing, oversmoothing, pointwise asymptotic distribution, pointwise confidence interval, uniform confidence interval, automatic selection of bandwidth, uniform rate of convergence.

Basic View: There can be many meanings to “nonparametrics”. One meaning is optimization over a set of function. For example, given the sample of observations $x_1, \ldots, x_n$, you are asked to find a distribution function under which the joint probability of $x_1, \ldots, x_n$ is maximized. You can convince yourself that this is the empirical distribution: Essentially, you are asked to find a distribution function (so that $G(x) = \frac{1}{n} \sum_{i=1}^{n} 1(x_i \leq x)$) that maximizes. Convince yourself that the solution is $G(x) = \frac{1}{n} \sum_{i=1}^{n} 1(x_i \leq x)$.

Now $\frac{1}{2} \mathbb{1}(|x| \leq 1)$ is the uniform density over $(-1,1)$, this is called the uniform kernel. Of course you can replace by another density function $K(\cdot)$, not necessary over $(-1,1)$, and plug in into where the uniform density appears, to get $\hat{f}(x) = \frac{1}{nh} \sum_{t=1}^{n} K\left(\frac{x-x_i}{h}\right)$.

Another motivation, probably easier to understand: you estimate the distribution function $F(x)$ by $\hat{F}(x) = \frac{1}{n} \sum_{t=1}^{n} 1(x_i \leq x)$ but you can’t differentiate it to get the density. So you smooth out $\hat{F}(x)$ by replacing $1(x_i \leq x)$ by $G\left(\frac{x-x_i}{h}\right)$ where $G(\cdot)$ is any smooth distribution function (so that $G(\infty) = 1, G(-\infty) = 0$), and $h \rightarrow 0$. In practice, you just take $h$ to be some small but fixed number, like 0.1. So you have $\hat{F}(x) = \frac{1}{n} \sum_{t=1}^{n} G\left(\frac{x-x_i}{h}\right)$, differentiate this to get $\hat{f}(x) = \frac{1}{nh} \sum_{t=1}^{n} K\left(\frac{x-x_i}{h}\right)$ for $K = G'(\cdot)$ the density corresponding to $G(\cdot)$.

Higher dimension: if $x \in \mathbb{R}^d$ a $d$-dimension vector, then just take $G(\cdot)$ to be a $d$-dimensional distributional and do the same thing, $\hat{F}(x) = \frac{1}{n} \sum_{t=1}^{n} G\left(\frac{x-x_i}{h}\right)$. Now when you differentiate it don’t forget that you have to do that to each of the $x_1, \ldots, x_d$, so you will pull out $d$ of $h$ to get:

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{t=1}^{n} K\left(\frac{x-x_i}{h}\right)$$
where now \(K(\cdot)\) is the multivariate density corresponding to the multivariate distribution function \(G(\cdot)\).

**Conditional expectation: motivation** You want to estimate \(E(y|x)\) or more generally \(E(g(y)|x)\) for some function \(g(\cdot)\). You can also estimate things like the conditional median of \(y\) give \(x\), say \(med(y|x)\), or any other conditional quantiles.

**Local Weighting:** The whole thing about nonparametric estimate of \(E(y|x)\) is just use observations for which \(x_i\) is close to \(x\). The obvious to do is to take a neighborhood around \(x\) and average over those \(y_i\) for which \(x_i\) is with this neighborhood. The size of the neighborhood should shrink to 0 but not too fast. More generally you take weighted average over all \(y_i\) but you give more weights to those \(y_i\) for which the \(x_i\) is close to \(x\), and you give less weight to those \(y_i\) for which the \(x_i\) is far away from \(x\). See Stone(1977).

For weights \(W_n(x, x_i)\) such that (1) \(\sum_{i=1}^{n} W_n(x, x_i) = 1\) (this is what “weight” means); (2) \(W_n(x, x_i) \to 0\) if \(x_i \neq x\); (3) \(\max_{1 \leq i \leq n} |W_n(x, x_i)| \to 0\) as \(n \to \infty\), you estimate \(E(y|x)\) by \(\sum_{i=1}^{n} W_n(x, x_i) Y_i\). Similarly, you estimate \(E(g^2(x))\) by \(\sum_{i=1}^{n} W_n(x, x_i) Y_i^2\), and \(F(y|x)\) by \(\sum_{i=1}^{n} W_n(x, x_i) 1(Y_i \leq y)\). Anythings you do parametrically, if you do that only for \(x_i\) close to \(x\), then you become “nonparametric”. You can run least square, quantile regression, and even maximum likelihood over small(shrinking) neighborhood of \(x\), and call that “nonparametric” sth sth ....

Depending on how you define the weight sequence \(W_n(x, x_i)\), the local weighting type nonparametric estimators are classified into many many different categories. The most important ones are kernel smoothing, k-nearest neighborhood, local polynomials. Asymptotically they are all equivalent to each other. Kernel smoothing is the easiest to use so this is the only thing we will deal with in details. In additions to these “local nonparametric estimate”, there are also “global” nonparametric estimates including series(sieve) estimator and splines. Series(sieve) estimation has been paid increasingly attention recently. But few people would bother with splines. The reason again is that series(sieve) is extremely easy to do, you just need to run least square! And just remember to say that the number of regressors included \(\to \infty\) as \(n \to \infty\), although in practice you just include some finite number of terms. But the focus today is kernel.

**Kernel smoothing:** If you use density weighting for the weights \(W_n(x, x_i)\), then you get the kernel estimator of \(E(y|x)\). If \(x_i\) is one-dimensional, let \(W_n(x, x_i) = \frac{1}{nh} K\left(\frac{x-x_i}{h}\right)\). By this very definition \(\sum_{i=1}^{n} W_n(x, x_i) = 1\). Then the kernel estimate is defined by:

\[
\sum_{i=1}^{n} W_n(x, x_i) Y_i = \sum_{i=1}^{n} \frac{1}{nh} K\left(\frac{x-x_i}{h}\right) Y_i = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) Y_i
\]

For multidimension \(x_i \in R^d\), use the multidimension density estimate \(W_n(x, x_i) = \frac{1}{nh^d} K\left(\frac{x-x_i}{h}\right)\), and the kernel estimator of \(E(y|x)\) is given by:

\[
\sum_{i=1}^{n} W_n(x, x_i) Y_i = \sum_{i=1}^{n} \frac{1}{nh^d} K\left(\frac{x-x_i}{h}\right) Y_i = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) Y_i
\]
Another view of kernel estimator: You can think of
\[
E(y|x) = \frac{E(y|x) f(x)}{f(x)} = \gamma(x)
\]
where \( \gamma(x) \equiv E(y|x) f(x) = \int y f(y, x) dy \). So you may estimate \( \gamma(x) \) and \( f(x) \) separately.

Of course \( \hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K \left( \frac{x-x_i}{h} \right) \). For \( \hat{\gamma}(x) \), plug in the kernel density estimate for \( x, y \):
\[
\hat{f}(x, y) = \frac{1}{nh^d} \sum_{i=1}^{n} K \left( \frac{x-x_i}{h} \right) K \left( \frac{y-y_i}{h} \right) \int y f(y, x) dy, \quad \text{and use a change of variable} \quad u = (y_i - y) / h:
\]
\[
\int y \hat{f}(y, x) dy = \frac{1}{nh^d} \sum_{i=1}^{n} K \left( \frac{x-x_i}{h} \right) \int y \frac{1}{h} \hat{K} \left( \frac{y_i - y}{h} \right) dy
\]
\[
= \frac{1}{nh^d} \sum_{i=1}^{n} K \left( \frac{x-x_i}{h} \right) \int (y_i + u) K(u) du = \frac{1}{nh^d} \sum_{i=1}^{n} K \left( \frac{x-x_i}{h} \right) y_i
\]

Yet another view for \( \gamma(x) \): you can think of \( \int y \hat{f}(y, x) dy \) as \( \int y dP \), where \( P \) is the measure over \( y \) defined by \( P(y_i \leq y, x_i = x) \). Note that \( P(y_i \leq y, x_i = x) = \frac{d}{dx} P(y_i \leq y, x_i = x) \), you can estimate \( P(y_i \leq y, x_i \leq y) \) by \( \frac{1}{n} \sum_{i=1}^{n} 1(y_i \leq y) G((x_i - x) / h) \). Differentiating this, you estimate \( \frac{d}{dx} P(y_i \leq y, x_i \leq y) \) by \( \frac{1}{nh^d} \sum_{i=1}^{n} 1(y_i \leq y) K((x_i - x) / h) \). Plug in this estimate of \( P \) into \( \int y dP \):
\[
\int y d\hat{P} = \int y d \frac{1}{nh^d} \sum_{i=1}^{n} 1(y_i \leq y) K((x_i - x) / h)
\]
\[
= \frac{1}{nh^d} \sum_{i=1}^{n} K((x_i - x) / h) \int y d(1(y_i \leq y)) = \frac{1}{nh^d} \sum_{i=1}^{n} K((x_i - x) / h) y_i
\]

Note: Now note that to study convergence rate, asymp distribution, etc, we only need to worry about \( \hat{\gamma}(x) \), and we may ignore discussing \( \hat{f}(x) \). The reason is that \( \hat{f}(x) \) is just a special case of \( \hat{\gamma}(x) \), in which we take \( y_i \equiv 1 \) identically.

Convenient forms of kernel(density) function include those on pp2303 Hardle(1994) Ch8 vol Handbook: (1) Uniform kernel \( \frac{1}{2} 1(|x| \leq 1) \); (2) Triangular kernel: \( (1 - |u|) 1(|u| \leq 1) \); and quartic, epanechnikov, gaussian, etc. Read pp2303, and try to use gauss or matlab to graph out what each of these look like. To get consistency, you want \( h \to 0 \) as \( n \to 0 \) so that you don’t have a bias, you also want \( nh^d \to \infty \) so that you have enough observations. More on this below.

Estimating Derivatives: It is also easy to estimate derivatives of \( \gamma(x) \)(including \( f(x) \)). As long as your kernel is smooth differentiable, you could just simply differentiate your \( \hat{\gamma}(x) \) and \( f(x) \). To estimate the \( k \)th derivative of \( \gamma(x) \), denoted by \( \gamma^{(k)}(x) \), use
\[
\hat{\gamma}^{(k)}(x) = \frac{1}{nh^{k+d}} \sum_{i=1}^{n} K^{(k)} \left( \frac{x-x_i}{h} \right) y_i
\]
Don’t forget that each time you differentiate wrt $x$ in $K \left( \frac{x_i - x}{h} \right)$, you will pull out one $h$, so altogether you pull out $k$ of them. Also remember that when $x$ is $d$-dimension, the derivative can be taken with respect to different combinations out of $x_1, \ldots, x_d$, as long as the degree of differentiation with respect to $x_1, \ldots, x_d$ sum to $k$. Sometimes people write $f^{(k)}$ (and similarly $K^{(k)}$) using very long notation. For example, $f^{(k)}$ is written as: for positive integers $\lambda_1, \ldots, \lambda_d$, let $\lambda_1 + \ldots + \lambda_d = k$. Then write $f^{(k)}(x)$ as $\frac{\partial^k}{\partial x_1^{\lambda_1} \ldots \partial x_d^{\lambda_d}} f(x)$.

**k-NN(nearest neighborhood) and local polynomials:** These are the other two major weighting schemes for $W_{ni}(x)$. Both of these are asymptotically equivalent to kernel weighting.

k-NN: Instead of using observations whose $x_i$ are with $h$ far away from $x$, you may also just use your $k$ closest neighbors, assuming you are standing at the point $x$. Look at the example in pp3 Hardle(1990) for the precise definition. You may weight your $k$ nearest neighbors equally, or you may give even higher weight to the ones who just live next to you and give decreasing weights to the ones increasing further away from you. You can of course use any kernel density weight $K(\cdot)$. In this case you may take the bandwidth $h$ to be the distance between $x$ and the $k$th nearest neighbor $x_i$. Giving equal weights to each of the $k$ neighbors is the same as saying using an uniform kernel density weight. If you do this then there is an easy relation between $h$ and $k$: $k \sim n (2h)^d f(x)$. For the case where $d = 1$, check to see that this relation makes sense to you by verifying the equivalence between kernel and k-NN in Table 3.2.1 pp46 Hardle(1990).

For local polynomial, read pp30-32 Hardle(1990). You must know how to run a polynomial regression using all observations, just regress $y$ on $1, x, x^2$, all cross products, cross products of all powers $\ldots$, up to degree $k$. If you run this $k$th degree polynomial regression only over a neighborhood for whose $x_i$ with $h$ distance from $x$: $1 \leq |x_i - x| \leq h$, then you get the local polynomial estimate. The degree $k$ of local polynomial corresponding to the order of the kernel, which we will talk about below.

**Series and Splines:** There is also another thing called sieve estimator, which is very similar to, and you can think that it is just same as, series estimator. The only difference between series and local polynomials is that you run the polynomials using all observations, instead of only a shrinking neighborhood $(x - h, x + h)$. However, now instead of fixing $k$, the degree of polynomial, you promise to let $k \rightarrow \infty$. But this is nothing more than a promise, you don’t actually do it. Just take a number $k$, not too small. Instead of using polynomials, which tends to be highly correlated, you may also use family of orthogonal series of functions, like trigonometric function, legendre polynomials, laguerre polynomials. If you are interested you can find more information about these function in a recent book by Judd(1998).

Splines is based on find a twice differentiable function $g(x)$ that minimizes $\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(x)^2 \, dx$, for some $\lambda > 0$. The reason to include the second term is to penalize the roughness of the estimate $g$, otherwise the minimization problem is ill-defined and you will simply take $\hat{g}(x_i) = y_i$, no good. This will give you a cubic polynomial with continuous second derivatives. Of course instead of penalizing by $\lambda \int g''(x)^2 \, dx$, you might think of all other kinds of penalizations. But this has not been much used other than simple case with one dimension $x$. If you are interested read Hardle(1990) sec 3.4 pp56.
Now back to the “local” nonparametric estimators.

**Curse of dimensionality:** For a given bandwidth(window size), the higher dimension $x$, the less data you have in a neighborhood with bandwidth $h$. Say all your $x$ is on the unit interval $(0,1)$, you only look at observations within a small bandwidth $h$, the total number of observations is $n$. If $d = 1$, then the effective number of observations you have in that $h$ neighborhood is $nh$. If $d = 2$, the effective number of observations you have is $nh^2$, so on so forth. So for general $d$, the number of observations you have in a $d$ dimension cube with length $h$ is just $nh^d$. Since $h$ is small, goes to 0, of course you get less and less observations as $d$ get larger. Another view of this is that, fixing the number of total observations, the average distance between any pair of observations gets larger and larger as $d$ increases(you find that your data becomes sparser and sparser in higher dimensions).

**Optimal rate of convergence for nonparametric estimates:** Suppose the true function $\gamma(x)$ is $p$th degree differentiable, all $p$th derivative bounded uniformly over $x$. Then the optimal bandwidth $h_{opt}$ you should use is $n^{-\frac{1}{2p+d}}$, and the best rate at which your estimate $\hat{\gamma}(x)$ can approach $\gamma(x)$ is $O_p\left(n^{-\frac{1}{2p+d}}\right)$. Now we will explain why. Reading is in sec 4 of Hardle(1990). You may consult the original Stone(1980)(1982) papers if you are interested.

The problem here is the bias and variance tradeoff. Basically, by using a bandwidth of size $h$ you are doing two things: (1) you are using observations $x_i$ that can be at most $h$ away from $x$ to approximate the true function $\gamma(x)$ at point $x$; (2) you are using all $nh^d$ observations in this small cube of length $h$. Step (1) gives you a bias in estimating $\gamma(x)$, and step (2) gives you the variance in the estimate. The smaller the $h$, the smaller the bias, but the less observations you have, thus the large the variance. The larger the $h$, the larger the bias, but then you have more observations which reduces the variance in your estimate. To get consistency, you want to get rid of both the bias and the variance. To get rid of bias, simply take $h \longrightarrow 0$, as $n \rightarrow 0$. To get rid of variance, you want to have infinite number of observations as $n \rightarrow \infty$, but the number of observations you have is $nh^d$, so you also want $nh^d \rightarrow \infty$. If you have both $h \rightarrow 0$ and $nh^d \rightarrow \infty$, then your estimate is consistent.

However, you want more than consistency. You want $\hat{\gamma}(x)$ to converge as fast as possible, to do that you need to balance the trade off between bias and variance. To do this, we need to find more precise representations of the bias and the variance.

The bias is of order $h^p$: Because you know your function $\gamma(x)$ have $p$ bounded derivatives, so you could do a taylor expansion of the approximation error $\gamma(x_i) - g(x)$ around $x$ up to degree $p - 1$. Now each term is a polynomial in $|x_i - x|$ and remember that $|x_i - x| \leq h$. Just assume that you asomehow knows the coefficients in those $p - 1$ expansion terms, so that you don’t have to worry about these terms. But the last term, the $p$th order term, an order of $|x_i - x|^p \leq h^p$ cannot be further expanded anymore, so you have to live up with it. The order of this error is $h^p$. These are different ways to get rid of the first $p - 1$ terms in the taylor expansion. Higher order kernel is one way, which we will see below. Use $p - 1$th order local polynomial is another way(which is just exactly doing a taylor expansion of order $p - 1$) but we won’t get into this.

The variation is of order $O_p\left(\frac{1}{\sqrt{nh^d}}\right)$. Just think of the simplest problem of estimating a
mean using a sample average. When you have \( n \) observations, using a central limit theorem, you know that \( \sqrt{n} (\bar{x} - \mu) \sim N(0,?/n) \). So that \( \bar{x} - \mu = O_p\left(\frac{1}{\sqrt{n}}\right) \). Now with a window size \( h \) the number of observations you have is \( nh^d \), by analogy to the simple case of \( \bar{x} \), the precision of your estimate, of course, should be \( O_p\left(\frac{1}{\sqrt{nh^d}}\right) \).

So that you total error, bias plus estimation error, is of the order of \( h^p + \left(\frac{1}{\sqrt{nh^d}}\right) \), find a \( h \) to minimize this, you can easily see that \( h_{opt} = O\left(n^{-\frac{1}{2p+d}}\right) \). A more formal way to think of the total error is MSE(mean square error) = bias^2 + variance. Now bias^2 = \( h^{2p} \), and variance = \( \frac{1}{nh^d} \). We will see these in detail for kernel regressions. Then the (pointwise) optimal rate of convergence is given by:

\[
O(h_{opt}^p) = O\left(\frac{1}{\sqrt{nh^d}}\right) = O\left(n^{-\frac{p}{2p+d}}\right)
\]

It should then be clear that it is not possible to have \( \sqrt{n} \) convergence for nonparametric estimates since \( \frac{p}{2p+d} < \frac{1}{2} \). Sometimes \( n^{1/4} \) rate of convergence is needed for getting rid of the second order terms for semiparametric estimators. You will need \( p > d/2 \) to get \( \frac{p}{2p+d} > 1/4 \). So the more regressors you have, the more smooth your regression function has to be in order to get the same rate of convergence.

**Optimal rate for derivative estimates:** What is the optimal bandwidth for estimating \( \gamma^{(k)}(x) \) and what is the best convergence rate. It turns out the optimal bandwidth is of the same order as that of estimating \( \gamma(x) \) itself. This can be seen again from the same bias and variance tradeoff.

**bias:** Now if \( \gamma(x) \) has \( p \) bounded derivatives, then \( \gamma^{(k)}(x) \) has \( p-k \) bounded derivatives. So by analogy, the bias is of the order of \( h^{p-k} \).

**variance:** Again you have \( nh^d \) observations effectively. However, each differentiation amplify the error by an order of \( \frac{1}{h} \) so after taking \( k \) derivative you amplify the error by \( \frac{1}{h^k} \), see equation (1). So the estimation error is of the order of \( \frac{1}{h^k\sqrt{nh^d}} \). The total error is of order \( h^{p-k} + \frac{1}{h^k\sqrt{nh^d}} \). Find a \( h \) to minimize this again you see that \( h_{opt} = n^{-\frac{1}{2p+d}} \). Note that \( h_{opt} \) does not depend on \( k \) at all. Then the best convergence rate is:

\[
n^{p-k} = \frac{1}{h^k\sqrt{nh^d}} = n^{-\frac{p-k}{2p+d}}
\]

The order of bias and variance will be developed in detail for kernel estimates below.

**Higher order kernels** The kernel weights should integrate to 1, as required for a density, so \( \int K(u) du = 1 \). If you use a density \( K(\cdot) \) that is symmetric around 0, then \( \int uK(u) du = 0 \). But if you insist that the kernel \( K(\cdot) \) must be a density, then \( K(u) \geq 0, \forall u \), and so \( \int u^2K(u) du > 0 \).

A higher order kernel is those \( K(\cdot) \) that has some similarity to a symmetric density but is not a density. Precisely, for these \( K(\cdot), \int K(u) u du = 0 \), but may be negative for some \( u \). This allows you to find such \( K(\cdot) \) so that \( \int K(u) u^2 du = 0 \). To see graphs of example, see
e.g figure 3.11 and 3.12 in pp60 and pp61 in Hardle(1990). Precisely, a kernel of order \( r \) is defined as those \( K(\cdot) \) for which:

\[
\int K(u)\,du = 1 \quad \int K(u)\,u^q\,du = 0, \forall q = 1, \ldots, r \quad \int |u^r K(u)|\,du < \infty
\]

for examples of these kernel can be found in table 4.5.1 pp135 Hardle(1990).

**Bias of kernel estimates:** use the change of variable \( u = \frac{x_i-x}{h} \), and bear in mind that the integration notation is over the \( d \)-dimensional \( x_i \), and finally do a taylor expansion of \( \gamma(x+uh) \),

\[
E\hat{\gamma}(x) - \gamma(x) = E\frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) Y_i - \gamma(x) = E\frac{1}{h^d} K\left(\frac{x-x_i}{h}\right) Y_i - \gamma(x)
\]

\[
= E\int \frac{1}{h^d} K\left(\frac{x-x_i}{h}\right) m(x_i) f(x_i)\,dx_i - \gamma(x) = E\int \frac{1}{h^d} K\left(\frac{x-x_i}{h}\right) \gamma(x_i)\,dx_i - \gamma(x)
\]

\[
= \int K(u) \gamma(x+uh)\,du - \gamma(x) = \sum_{j=1}^{r-1} h^j \frac{\gamma^{(j)}(x)}{j!} \int u^j k(u)\,du + h^r \frac{1}{r!} \int \gamma^{(r)}(x^*)\,u^r K(u)\,du
\]

If your kernel is of the order \( r \), then each of the integrals \( \int u^j k(u)\,du = 0 \) in the first summation. So you are left with only the last term, so that bias = \( h^r \). Note that you can do this only if your true underlying function \( \gamma(x) \) has \( p \geq r \) bounded derivatives. So if your function \( \gamma(x) \) has \( p \) bounded derivatives and your kernel is of order \( r \), then the bias = \( h^{\min(p,r)} \). Optimally, if you know your \( \gamma(x) \) is \( p \)th degree smooth, then you should use a \( p \)th degree kernel to reduce bias. But you never know what \( p \)-smoothness your function is. So why not always use a very high order kernel? That is because a high order kernel will neccessary make the variance big, as seen below.

In the basic case when you just use a symmetric density for \( K(\cdot) \), it is of order \( r = 2 \) so that the bias is of order \( h^2 \).

**Variance of kernel estimates:** This one is easier than the bias since higher order kernel does not play any role here, let \( g(x_i) = E(Y_i^2|x_i) f(x_i) \):

\[
Var\left(\frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) Y_i\right)
\]

\[
= \frac{1}{n^2 h^{2d}} \sum_{i=1}^{n} Var\left(K\left(\frac{x-x_i}{h}\right) Y_i\right) = \frac{1}{nh^d} E\left(\frac{1}{nh^d} E\left(K^2\left(\frac{x-x_i}{h}\right) Y_i^2\right) - \frac{1}{nh^d} (EK\left(\frac{x-x_i}{h}\right) Y_i)^2\right)
\]

\[
= \frac{1}{nh^d} \int \frac{1}{h^d} K^2\left(\frac{x-x_i}{h}\right) E(Y_i^2|x_i) f(x_i)\,dx_i - \frac{1}{n} \left(E\frac{1}{h^d} K\left(\frac{x-x_i}{h}\right) Y_i\right)^2
\]

\[
= \frac{1}{nh^d} \int \frac{1}{h^d} K^2\left(\frac{x-x_i}{h}\right) g(x_i)\,dx_i + O\left(\frac{1}{n}\right) = \frac{1}{nh^d} \int K^2(u)g(x+uh)\,du + O\left(\frac{1}{n}\right)
\]

\[
= \frac{1}{nh^d} \int K^2(u)g(x)\,du + \frac{1}{nh^d} \int K^2(u)g'(x^*)\,udu + O\left(\frac{1}{n}\right)
\]

\[
= \frac{1}{nh^d} \int K^2(u)g(x)\,du + O\left(\frac{1}{nh^d}\right) + O\left(\frac{1}{n}\right) = O\left(\frac{1}{nh^d}\right)
\]
Bias and Variance of derivatives of kernel estimates: You could verify the formulas for bias and variance for $\gamma^{(k)}(x)$ by carrying through the same exercise of taking expectation, change of variable in integration. An additional tools that are needed here is integration by parts, assuming $K(u)$ is bounded and all the first $k$th derivatives diminish at the boundary, you can verify that:

$$\int K^{(k)}(u) \gamma(u) \, du = (-1)^k \int K(u) \gamma^{(k)}(u) \, du.$$  

Exercise: Use integration by parts to obtain the bias formula for kernel estimates of $\gamma^{(k)}(x)$, and convince yourself that higher order kernel plays the same role here. The variance formula can be calculated as before without having to worry about integration by parts.

Asymptotic Distribution, Confidence band To show how precise your nonparametric estimate is at each point $x$, it is neccessary to derive its asymptotic distribution and obtain the confidence interval at each point $x$. Section 4 in Hardle(1990) contains detailed discussion. The derivation of the asymptotic distribution depends on what bandwidth $h$ you use, if you use $h \sim h_{opt}$, where $\sim$ means “the same order as”, then the asymptotic distribution will depend on both the bias and the variance. On the other hand, if you use $h << h_{opt}$ where $<<$ means smaller by an order of magnitude, i.e., $\frac{h}{h_{opt}} \rightarrow 0$, then there is no bias in the asymptotic distribution but the convergence rate is not the fastest. To look at this in more detail, consider the simple example for $d = 1, r = 2$, for one-dimensional $x$ and for a standard density kernel with only order 2, then $h_{opt} = n^{-\frac{2}{d+r}} = n^{-\frac{1}{2}}$. To compute the asymptotic distribution of $\sqrt{n h_{opt}} (\hat{m}(x) - m(x)) = h_{opt}^{-2} (\hat{m}(x) - m(x))$, for $\hat{m}(x) = \frac{\gamma(x)}{f(x)}$, first linear this in terms of $\hat{\gamma}(x)$ and $\hat{f}(x)$ by simply taking first order taylor expansion:

$$\hat{m}(x) - m(x) \approx \frac{1}{f(x)} (\hat{\gamma}(x) - \gamma(x)) - \frac{\gamma(x)}{f(x)^2} (\hat{f}(x) - f(x))$$

As seen above: $(E\hat{\gamma}(x) - \gamma(x)) = \frac{1}{2} h^2 \gamma''(x) \int u^2 K(u) \, du$. Note that for $\gamma(x) = m(x) f(x)$, $\gamma''(x) = m''(x) f(x) + 2f'(x) m'(x) + m(x) f''(x)$. So for the density where $m(x) \equiv 1$, $(E\hat{f}(x) - f(x)) = \frac{1}{2} h^2 f''(x)$. Therefore

$$E h_{opt}^{-2} (\hat{m}(x) - m(x)) = \frac{1}{2} \left( \frac{\gamma''(x)}{f} - \frac{m''(x)}{f} \right) = \frac{1}{2} \left( \frac{1}{f} \left( m'' f + 2 m' f' + f m'' \right) - \frac{m}{f} f''(x) \right) \int u^2 K(u) \, du$$

$$= \frac{1}{2} \frac{2 m'(x) f'(x) + m''(x) f(x)}{f(x)} \int u^2 K(u) \, du.$$  

To compute the variance, follows from the variance calculation for kernel estimation to get: $Var \left( \sqrt{n h} (\hat{\gamma}(x) - \gamma(x)) \right) \rightarrow g(x) \int K^2(u) \, du$, for $g(x) = E (y^2|x) f(x)$. For density estimate where $y \equiv 1$, $g(x) = f(x)$, so that $Var \left( \sqrt{n h} (\hat{f}(x) - f(x)) \right) \rightarrow g(x) \int K^2(u) \, du$.  

It is also neccessary to calculate the covariance between $\hat{\gamma}(x)$ and $\hat{f}(x)$, you may use the type of change-of-variable calculation to get:

$$Cov \left( \sqrt{n h} (\hat{\gamma}(x) - \gamma(x)), \sqrt{n h} (\hat{f}(x) - f(x)) \right) \rightarrow \gamma(x) \int K^2(u) \, du.$$
Then use the delta method to obtain the variance of $\text{Var}\left( \sqrt{n}h \left( \hat{m}(x) - m(x) \right) \right)$:

$$\text{Var}\left( \sqrt{n}h \left( \hat{m}(x) - m(x) \right) \right) = \text{Var}\left( \sqrt{n}h \left( \frac{1}{f} \gamma(x) - \frac{m}{f} \right) \right)$$

$$= \int K^2(u) \text{du} \left( \frac{1}{f} E \left( y^2 | x \right) f(x) - \frac{2}{f} m(x) \gamma(x) + \frac{m^2}{f^2} \right)$$

$$= \frac{1}{f(x)} \left( E \left( y^2 | x \right) - m(x)^2 \right) \int K^2(u) \text{du} = \frac{1}{f(x)} \sigma^2(x) \int K^2(u) \text{du}$$

To summarize:

$$\sqrt{n}h \left( \hat{m}(x) - m(x) \right) \xrightarrow{d} N \left( \left( \frac{m''(x) f(x) + 2m'(x) f'(x)}{2f(x)} \right) \int u^2 K(u) \text{du}, \frac{1}{f(x)} \sigma^2(x) \int K^2(u) \text{du} \right)$$

If instead of using $h_{opt} = n^{-1/5}$, you use a undersmooth bandwidth $h << n^{-1/5}$, say $h = n^{-1/6}$, then the bias term will disappear and you don’t have to worry about it anymore. In that case the asymptotic distribution simplifies to:

$$\sqrt{n}h \left( \hat{m}(x) - m(x) \right) \xrightarrow{d} N \left( 0, \frac{1}{f(x)} \sigma^2(x) \int K^2(u) \text{du} \right)$$

Therefore, if you don’t want to bother with the complicated bias term in drawing your confidence interval, then you should use a undersmooth bandwidth, like $h = n^{-1/4}$.

However, if you insist on using the optimal $h_{opt}$ and want to draw the confidence interval around your estimated $\hat{m}(x)$, then you will need to be able to estimate the bias term consistently. However, the bias term involve second derivatives $m''(x)$, or for that matter $\gamma''(x)$. And $\gamma''(x)$ can NOT be estimated consistently using $h_{opt}$. Instead, you will need to use a separate and oversmoothed bandwidth, say $g = n^{-1/6}$ to estimate $\gamma''(x)$ consistently. The reason is simply that $\hat{\gamma}''(x)$ has a bigger variance than $\hat{\gamma}(x)$, by an order of $\frac{1}{g^2}$. To reduce this larger bias you need more data and so you need a larger oversmoothed bandwidth $g$. In fact you can show that the variance of $\hat{\gamma}''(x)$ is $O \left( \frac{1}{nh^{2-1/5}} \right)$, so if you use $h_{opt} = n^{-1/5}$, the variance won’t go to 0. But if you use $g = n^{-1/6}$, then it will. This $g$ is called “pilot” bandwidth in Hardle(1990)’s terminology.

**Uniform confidence band:** The above asymptotic distribution is for deriving the Pointwise confidence band. The uniform band is the two curves for which the probability that they encompass the ENTIRE true $m(x)$ function is $\alpha%$. Constructing uniform band is difficult. You might consult Hardle(1990) sec 4.3 if you are interested. But I suggest against doing that in first reading. You need to know however, the difference between pointwise confidence band and uniform band.

**Automatic bandwidth selection:** Two ways: cross validation and penalizing functions, read Hardle(1990) pp147-155.

Cross Validation: You want a good fit of your estimate, try to minimize $\sum_{i=1}^{n} (\hat{m}(x_i) - m(x_i))^2$, but you don’t know the true $m(x_i)$, so you estimate it using $y_i$: $\min_{i=1}^{n} (\hat{m}(x_i) - y_i)^2$. But
then you are in trouble, since as $h \to 0$, $\hat{m}(x_i) = y_i$, so you have perfect fit 0, same problem that motivates the spline estimator. Another way to think about this, write

$$\sum_{i=1}^{n} (\hat{m}(x_i) - y_i)^2 = \sum_{i=1}^{n} (\hat{m}(x_i) - m(x_i) - \epsilon_i)^2 = \sum_{i=1}^{n} (\hat{m}(x_i) - m(x_i))^2 + \sum_{i=1}^{n} \epsilon_i^2 - 2\sum_{i=1}^{n} (\hat{m}(x_i) - m(x_i))\epsilon_i$$

The first term is what you want. The second term is not related to $h$ and can be ignored. The third term causes trouble, its expectation:

$$E \sum_{i=1}^{n} \frac{1}{nh} \sum_{j=1}^{n} K \left( \frac{x_i - x_j}{h} \right) \epsilon_j \epsilon_i = \frac{1}{nh} \sum_{i=1}^{n} K(0) \sigma^2 = \frac{1}{h} \sigma^2 K(0)$$

(2)

This is not zero and in fact not negligible (same order as first term). The problem is due to the correlation between $\epsilon_i$ and itself. Of course if $\epsilon_i$ is not used to estimate $m(x_i)$ then this term wouldn’t exist to cause problem, therefore leading to the leave-one-out estimate $\hat{m}_{-i}(x_i) = \frac{1}{(n-1)h} \sum_{j=1, j \neq i}^{n} K \left( \frac{x_j - x_i}{h} \right) y_i$, and define the cross-validation function that is to be minimized over $h$ as: $\min CV(h) = \sum_{i=1}^{n} (m_{-i}(x_i) - m(x_i))^2$. According to Hardle(1990) sec 5.1 this will automatically give you the optimal bandwidth, no matter how smooth your $m(x)$ is. Of course this statement needs to be constrained by the order of the kernel function you use.

Therefore the point of cross-validation is to get rid of the nonzero expectation due to correlation of $\epsilon_i$ with itself, in (2), another way to get rid of this term is just to subtract it, if you know $\sigma^2$. But you don’t know $\sigma^2$, but you can still estimate this term consistently by $K(0) \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{m}(x_i))^2$. This leads to the penalizing function methods, read pp154 Hardle(1990), to minimizer over $h$:

$$G(h) = \sum_{i=1}^{n} (\hat{m}(x_i) - y_i)^2 + 2K(0) \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{m}(x_i))^2$$

$G(h)$ has the same property as $CV(h)$.

**Bias reduction by Jacknifing:** (skip the rest of this note if you don’t find it interesting. These are Extra material) Another way to think about bias reduction is by Jacknifing. Read Hardle(1990) sec 4.6 (in particular page 143) for details. It is essentially equivalent to high order kernel. It doesn’t make any difference if you are just running a simple kernel regression. So if you are just running a simple kernel regression don’t bother with Jacknifing. However, if you have some objective function that are only convex with positive $K(\cdot)$, say, if you want to run a nonparametric quantile regression and want to reduce bias, then operationally the Jackknife method is very useful in preserving the convexity of the objective function. See for example, the recent paper Honore and Powell(1998) for illustration.

**Uniform rate of convergence:** In addition to the pointwise rate. It is sometimes useful to obtain optimal bandwidth and optimal uniform convergence rate, i.e., for $\sup_{x \in X} |\hat{\gamma}(x) - \gamma(x)|$. The necessary steps are again to consider the bias-variance tradeoff. The bias for $\sup_{x \in X} |\hat{\gamma}(x) - \gamma(x)|$ for a $r$th order kernel is the same before $h^p$, noting that the bias calculation is completely uniform in $x$. The variance calculation, on the other hand, need to
be modified. It turns out the variance is of the order $O_p \left( \left( \frac{nh^d}{\log n} \right)^{-1/2} \right)$, i.e. $\sup_{x \in X} |\hat{\gamma}(x) - E\hat{\gamma}(x)| = O_p \left( \left( \frac{nh^d}{\log n} \right)^{-1/2} \right)$. The mysterious log $n$ comes from considering the variance at an increasing number of points simultaneous and by the application of the Bernstein inequality, a form of the exponential inequality:

**Bernstein inequality:** This is one of the many exponential inequality used to bound the tail probability of sums of mean 0 random variables. To see a clean and brief discussion of it you may read pp 193 of Pollard(1984). It basically says that for $z_1, \ldots, z_n$ independent of each other, each $z_i$ bounded by $B$, $V = Var(z_1) + \ldots + Var(z_n)$, $S_n = z_1 + \ldots + z_n$, $P(|S_n| > \eta) \leq 2 \exp \left( -\frac{\eta^2}{2V + \frac{\eta}{4B\eta}} \right)$. This says that $S_n$ looks roughly like normal distribution with variance $V$ in the tails. The term $\frac{1}{\eta} M \eta$ turns out to be not important and may in fact be ignored.

**Variance for uniform rate:** To show that $\sup_{x \in X} |\hat{\gamma}(x) - E\hat{\gamma}(x)| = O_p \left( \left( \frac{nh^d}{\log n} \right)^{-1/2} \right)$, the first thing is to partition the compact set $X$ into side length $n^{-L}$ squares, for some big number $L$. If $X$ is compact, say just the unit cube, then you got about $n^{dL}$ such squares with side length $n^{-L}$. Let $x_1, j = 1, \ldots, n^{dL}$ be the center of each of the squares. As long as $L$ is big enough, the difference between $\gamma(x_1) - E\hat{\gamma}(x_1)$ and $\gamma(x_2) - E\hat{\gamma}(x_2)$ for $x_1$ and $x_2$ in the same square is small enough and may be ignored, because now the difference is of order $n^{-L} h^{-d} \ll \left( \frac{nh^d}{\log n} \right)^{-1/2}$. Then roughly speaking, $\sup_{x \in X} |\hat{\gamma}(x) - E\hat{\gamma}(x)| \approx \sup_{x, j=1, \ldots, n^{dL}} |\hat{\gamma}(x_j) - E\hat{\gamma}(x_j)|$.

Now for each $x_j$, $\hat{\gamma}(x_j) - E\hat{\gamma}(x_j)$ is sums of mean 0 random variables, the sum of variance $V$, as have been calculated above, is about $\frac{1}{nh^d} D$, for some constant $D$. Assuming both $Y_i$ and $K(\cdot)$ are bounded, then $B = \frac{1}{nh^d} E$ for some $E > 0$. remember also that saying that

$$\sup_{x \in X} |\hat{\gamma}(x) - E\hat{\gamma}(x)| = O_p \left( \left( \frac{nh^d}{\log n} \right)^{-1/2} \right)$$

means

$$\lim_{n \to \infty} P \left( \sup_{x \in X} |\hat{\gamma}(x) - E\hat{\gamma}(x)| > \left( \frac{nh^d}{\log n} \right)^{-1/2} M \right) \xrightarrow{M \to \infty} 0.$$ 

Now time to use the Bernstein inequality:

$$P \left( \sup_{x \in X} |\hat{\gamma}(x) - E\hat{\gamma}(x)| > \left( \frac{nh^d}{\log n} \right)^{-1/2} M \right) < P \left( \sup_{x, j = 1, \ldots, n^{dL}} |\hat{\gamma}(x_j) - E\hat{\gamma}(x_j)| > \left( \frac{nh^d}{\log n} \right)^{-1/2} M \right) \leq n^{dL} \exp \left( -\frac{1}{2} \frac{M^2 \log n}{nh^d} + \frac{1}{2} \frac{E}{nh^d} M \sqrt{\log n} \right) \sim n^{dL} \exp \left( -\frac{M}{2D} \log n \right) = 2 \exp \left( \frac{dL - M}{2D} \log n \right) \xrightarrow{M \to \infty, n \to \infty} 0.$$

Then minimize the sum of bias + error $= h^p + O_p \left( \left( \frac{nh^d}{\log n} \right)^{-1/2} \right)$ you will find the optimal bandwidth and the best convergence rate.