Eco 523 - Class 1

Nonparametric Econometrics

Carolina Caetano*

Contents

1 Introduction

In principle, parametric and nonparametric estimators are not comparable: they are made for different purposes and as such should be evaluated by different standards. When somebody says that they don't like nonparametric estimators, they show a lack of understanding of the issue.

Most nonparametric techniques used are forms of weighted regressions and can also be considered inside the parametric framework. This means that you can use a nonparametric estimator even when you assume a parametric structure of the function. However, in those cases, some parametric estimators like the OLS, GMM, etc. possess certain optimality properties, and a nonparametric estimator would likely perform worse, in the sense that it would have a higher mean squared error (MSE) than a well specified parametric estimator. Hence, if you are willing to bet on a parametric structure, you should use the best

*Special thanks to Liquan Huang and Ryan Tierney for help with plots and editing.

parametric estimator you can find for that particular structure. Your risk is this: if you are wrong, then your estimates will be biased, and no amount of data can alleviate this.

On the other hand, you may choose not to assume a given parametric structure of the function. In such cases, you can still use the typical parametric techniques. Parametric estimators that assume the wrong structure will have MSEs that don't even converge to zero as the sample increases. However, we can still use a parametric estimator, but have a different understanding of what we are doing. We can promise that as the sample increases, we will relax the functional form we are assuming. For example, we can use OLS, and as the sample increases, we keep increasing the number of regressors, using interactions, squares, cubes, etc. If this is the promise, then at least the MSE of such estimator will converge to zero as the sample increases (sure, there are some assumptions required for this, but much milder than an entire parametric structure). However, there are other techniques that were specifically designed to unearth completely unknown structures, and will do a better job at it than this approach.

There are two general classes of nonparametric estimators: fully nonparametric estimators and the semi-nonparametric estimators. Fully nonparametric estimators are local. Loosely speaking, this means that the estimators are functions only of a neighborhood of data points and not of the entire data set. Kernel estimators and local polynomials are examples of these. Semi-nonparametric estimators try to estimate the entire model at once, like parametric estimators. Series estimators are examples of these.

2 Kernel based estimation

2.1 Density estimation

2.1.1 Estimating the CDF of a random variable X

The CDF is defined as

$$F(x) = P(X \leqslant x)$$

If we have a random sample $X_1, X_2, \ldots, X_n$, we can substitute the theoretical proportion of X's below x by the empirical proportion:

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i \leqslant x)$$

This is a nonparametrically efficient estimator, and by the CLT we can show that

$$\sqrt{n}(\hat{F}(x) - F(x)) \xrightarrow{d} N\left(0, F(x)(1 - F(x))\right)$$

The reason why we can estimate the value of $F(x)$ so easily is because $F(x)$ can be

derived in the context of a random variable with a geometric distribution. We have two possible events. Either $X_i \leqslant x$ or $X_i > x$, and these events happen with probabilities $F(x)$ and $1 - F(x)$.

2.1.2 Estimating a density

The problem of estimating a density is that the corresponding event is $X = x$, which has zero probability. We can't use a simple count estimator. We will go instead by the definition of the derivative:

$$f(x) = \frac{d}{dx}F(x) = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h}$$

A good idea to estimate $f(x)$ is to substitute

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

and promise that as $n \to \infty$, $h \to 0$. Hence

$$\begin{aligned}\hat{f}(x) &= \frac{1}{2h}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(X_i \leqslant x+h) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(X_i < x-h)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2h}\mathbf{1}(x-h \leqslant X_i \leqslant x+h)\end{aligned}$$

Define the following function:

$$k(v) = \frac{1}{2}\mathbf{1}(|v| \leqslant 1)$$

then

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

The CDF estimator calculated the proportion of observations below x, whereas this density estimator calculates the proportion of observations in a small interval. It gives weight $1/2h$ to those observations between $x - h$ and $x + h$, and zero to all the others.

However, there is no reason why we should require that all observations between $x - h$ and $x + h$ have the same weight. We could weight observations differently as a function of their distance to x; it is just a matter of choosing a different weighting function k. We call those weighting functions **kernels**. The weighting function should integrate to one, so that the weight of the entire population is equal to one. Hence, density functions make

good candidates for kernels. For instance:

$$k(v) = \frac{1}{2}\mathbf{1}(|v| \leqslant 1) \quad \text{is known as the uniform kernel}$$
$$k(v) = (1 - |v|)\mathbf{1}(|v| \leqslant 1) \quad \text{is known as the triangular kernel}$$
$$k(v) = \frac{3}{4}(1 - v^2)\mathbf{1}(|v| \leqslant 1) \quad \text{is known as the Epanechnikov kernel}$$
$$k(v) = \frac{1}{\sqrt{2\pi}} e^{\frac{-v^2}{2}} \quad \text{is known as the Gaussian kernel}$$

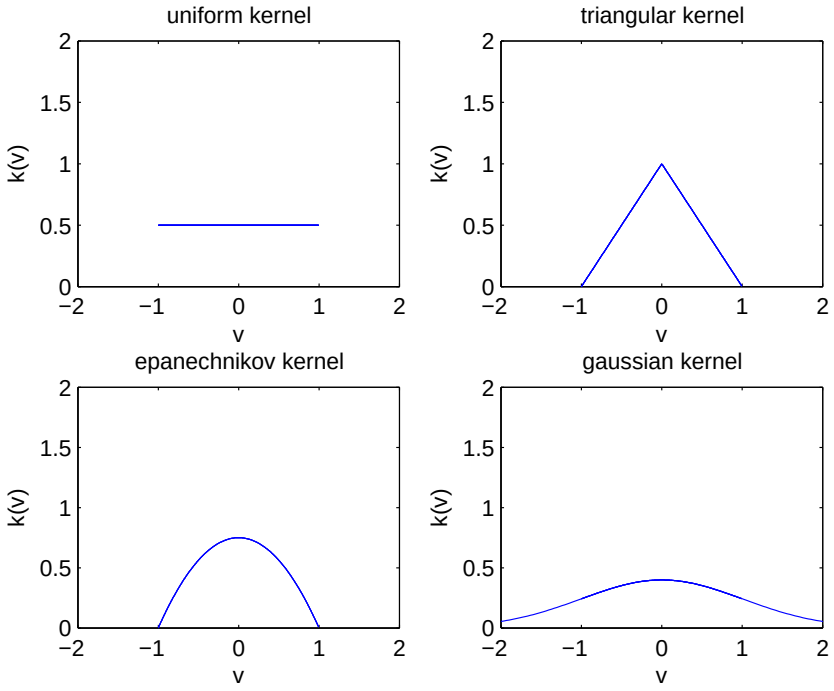

Figure 1: Various Kernels

The Gaussian kernel has infinite support, and that is often not a desirable property. If we are estimating what is essentially a derivative, why would we use information so far away? The triangular kernel is not differentiable at 0, and that is also something often undesirable. Recall that consistency and asymptotic distribution proofs often use the mean value theorem (or Taylor's theorem) which require differentiability. All the kernels above are always non-negative, but there are kernels that assume negative values and sometimes have better properties (they are known as higher-order kernels). In practice, however, these kernels are seldom used. The kernel with the best properties among the ones above is the

Epanechnikov. The fundamental properties a kernel must possess are

$$(i) \quad \int k(v)dv = 1$$
$$(ii) \quad k(v) = k(-v) \quad \left(\Longrightarrow \int vk(v)dv = 0 \right)$$

These properties are used to guarantee consistency and asymptotic normality of $\hat{f}(x)$, but we will not enter into details of density estimation. Observe, however, that all the kernels above satisfy the fundamental conditions. In fact, many different choices of kernels can serve as weight functions and estimate the density $f(x)$ adequately. The kernel does affect the efficiency of estimation, but is of no great importance in that regard. Of much greater concern is the choice of h, but we will deal with this on Class 2.

2.1.3 Multivariate density estimation

If we would like to estimate the density of a vector in a multivariate probability space, we can do it in exactly the same way. Let $x = (x_1, \ldots, x_q)$. We would like to estimate $f(x)$. Hence, we could use a function to weigh the distance between the observations X and the point x. Which distance? Any distance, for example the euclidean norm. However, in practice it is common (and often desirable) to weigh the observations by how distant each of the coordinates X_{is} are from the corresponding x_s, and then multiply the weights. These are known as product kernels. Let $v = (v_1, \ldots, v_q)$,

$$K(v) = k(v_1) \cdot k(v_2) \cdot \ \ldots \ \cdot k(v_q)$$

hence, the density estimator is

$$\hat{f}(x) = \frac{1}{nh_1 \ldots h_q} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) = \frac{1}{nh_1 \ldots h_q} \sum_{i=1}^{n} k\left(\frac{X_{i1} - x_1}{h_1}\right) \cdot k\left(\frac{X_{i2} - x_2}{h_2}\right) \cdots k\left(\frac{X_{iq} - x_q}{h_q}\right)$$

By partial integration, it is easy to see that if k satisfies the fundamental properties, then so does the product kernel K.

2.2 Kernel Regression

This is a course on non-parametric regression, and we finally arrive at our first strategy for this. Suppose that the model is

$$Y = g(X) + u$$

where $g(X) = E(Y|X)$. In order to estimate $E(Y|X)$, consider:

$$\begin{aligned} E(Y|X=x) &= \int y f_{Y|X}(y|X=x)dy \\ &= \int y \frac{f_{Y,X}(y,x)}{f_X(x)}dy \\ &= \frac{\int y f_{Y,X}(y,x)dy}{f_X(x)} \end{aligned}$$

We saw before that a good estimator of $f_X(x)$ is

$$\hat{f}_X(x) = \frac{1}{nh_1 \dots h_q} \sum_{i=1}^{n} K\left(\frac{X_i - x}{\mathbf{h}}\right)$$

In order to estimate $\int y f_{Y,X}(y,x)dy$, we begin by substituting $f_{Y,X}(y,x)$ with an estimator:

$$\begin{aligned} \int y \hat{f}_{Y,X}(y,x)dy &= \int y \frac{1}{nh_0 h_1 \dots h_q} \sum_{i=1}^{n} k\left(\frac{Y_i - y}{h_0}\right) K\left(\frac{X_i - x}{\mathbf{h}}\right) dy \\ &= \frac{1}{nh_0 h_1 \dots h_q} \sum_{i=1}^{n} K\left(\frac{X_i - x}{\mathbf{h}}\right) \int y k\left(\frac{Y_i - y}{h_0}\right) dy \end{aligned}$$

By the change of variables: $u = \frac{Y_i - y}{h_0}$,

$$\begin{aligned} \int y \hat{f}_{Y,X}(y,x)dy &= \frac{1}{nh_0 h_1 \dots h_q} \sum_{i=1}^{n} K\left(\frac{X_i - x}{\mathbf{h}}\right) \int (Y_i - h_0 u) h_0 k(v)dv \\ &= \frac{1}{nh_1 \dots h_q} \sum_{i=1}^{n} K\left(\frac{X_i - x}{\mathbf{h}}\right) \left[Y_i \int k(v)dv - h_0 \int u k(v)dv\right] \\ &= \frac{1}{nh_1 \dots h_q} \sum_{i=1}^{n} K\left(\frac{X_i - x}{\mathbf{h}}\right) Y_i \end{aligned}$$

Therefore, a natural estimator of $E(Y|X=x) = g(x)$ is

$$\hat{g}(x) = \frac{\frac{1}{nh_1 \dots h_q} \sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{\mathbf{h}}\right)}{\frac{1}{nh_1 \dots h_q} \sum_{i=1}^{n} K\left(\frac{X_i - x}{\mathbf{h}}\right)} = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{\mathbf{h}}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{\mathbf{h}}\right)}$$

Sometimes you will see the following notation:

$$k_h(X_i - x) = \frac{K\left(\frac{X_i - x}{\mathbf{h}}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{\mathbf{h}}\right)}$$

and therefore

$$\hat{g}(x) = \sum_{i=1}^{n} Y_i \, k_h(X_i - x)$$

This estimator is known as the Nadaraya-Watson kernel estimator, proposed by both of them in different papers in 1964. Notice that this estimator is a point estimator, that is, it estimates $E(Y|X = x)$, and it is different for each different value of x.

Intuition: the estimator above is simply a weighted average of the Y_i, where more weight is given to the observations such that X_i is close to x. The relative weights depend on the shape of the kernel k, but much more importantly, on the magnitude of h. Take a look at the uniform kernel:

$$\begin{aligned} \text{Uniform Kernel:} \quad & k(v) = \frac{1}{2}\mathbf{1}(|v| \leqslant 1) \\ \Longrightarrow \; & k\left(\frac{X_{is} - x_s}{h_s}\right) = \frac{1}{2}\mathbf{1}\left(\left|\frac{X_{is} - x_s}{h_s}\right| \leqslant 1\right) = \frac{1}{2}\mathbf{1}\left(x_s - h_s \leqslant X_{is} \leqslant x_s + h_s\right) \end{aligned}$$

This kernel gives equal weight to all observations that are at a distance of h_s or less from x_s.

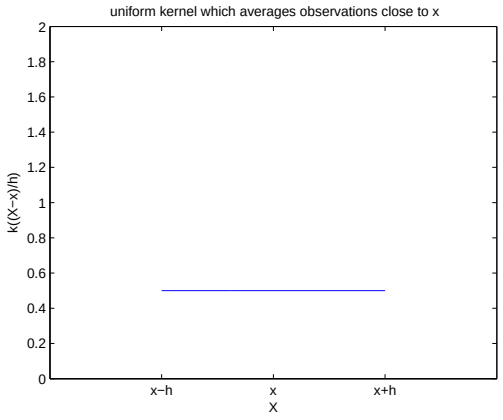

Figure 2: Uniform kernel averaging observations at a distance of h or less from x

This means that $\hat{g}(x)$, using the uniform kernel, is a simple average of the Y_i's such that X_{is} is distant from x_s by at most h_s for all $s = 1, \ldots, q$. That is a simple average inside of a (multidimensional) rectangle around x.

The smaller the h_s's the better, because the we are averaging only observations that are truly close to x. This gives an excellent intuition of how h influences the bias of $\hat{g}(x)$. The larger the h_s's, the larger the bias, because we are averaging using observations too

far away.

However, if h_s is too small, we may have too few observations to perform the average. Hence, although we will be averaging over observations that are truly close to x, they may be so few that the variability of our estimate is too high. This gives an excellent intuition of how h influences the variance of $\hat{g}(x)$. The larger the h_s's, smaller the variance of the estimate, because we are averaging using more observations.

Hence, in order to get a good estimate, we have to balance h so that it is small, but not too small. Also, as the sample size increases, we can decrease h, but slowly enough so that we are still averaging over a fair amount of observations. This is exactly what will be required for the asymptotic results of the Nadaraya-Watson estimator, as we will see later.

> **The role of *h*:** The larger h, the larger the bias (because observations far away are used), and the smaller the variance (because more observations are used).

Next, we examine the role of the kernel. Consider a different kernel, for example:

$$\text{Epanechnikov kernel:} \quad k(v) = \frac{3}{4}(1-v^2)\mathbf{1}(|v| \leqslant 1)$$

$$\implies k\left(\frac{X_{is}-x_s}{h_s}\right) = \frac{3}{4}\left(1-\left(\frac{X_{is}-x_s}{h_s}\right)^2\right)\mathbf{1}\left(x_s - h_s \leqslant X_{is} \leqslant x_s + h_s\right)$$

This kernel also averages only observations that are at a distance of h_s or less from x_s. However, it gives more weight to the observations that are closer to x_s as shown in Figure 3. For the same bandwidth, this kernel gives higher weight to observations that are closer to x. This will therefore decrease the bias. This is also achieved by other kernels, such as the Gaussian and the triangular. However, a kernel that gives too much weight to the

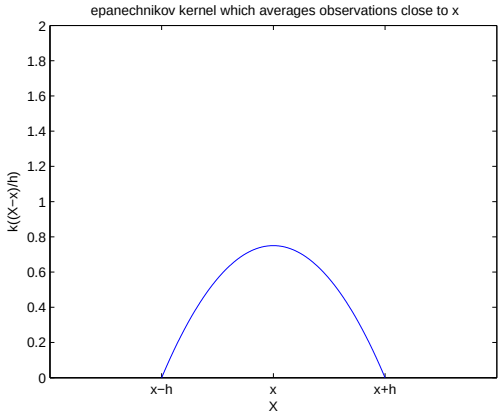

Figure 3: Epanechnikov kernel averaging observations at a distance of h or less from x

observations that are too close to x is, at the same time, ignoring the information from farther away points, which affects the variance in the same way that a smaller h does.

> **The role of the kernel:** The larger the kurtosis (the "peakedness") of the kernel, the smaller the bias (because observations far away are weighted less), and the larger the variance (because fewer observations are weighted highly).

The Epanechnikov kernel does a nice job of balancing the over-weight in the closest observations, while still weighting farther observations enough so that it doesn't increase the variance too much. It is also a nice differentiable kernel with finite support. This kernel is often excellent, not always the best, but kernel choice is rarely an important issue in practice.

The kernel regression method is very local, taking its cues directly from the data. In essence it is a true nonparametric method, assuming very little about the data structure even in smaller samples. In the next class we will introduce the local polynomial method, which is as local as the kernel method, but has even better properties to reduce bias without increasing the variance too much.

2.2.1 How to program the Nadaraya-Watson estimator

First, notice that the Epanechnikov kernel in Stata may not be correct. Be aware of this, and check the formula (and then let me know, as it has been a while since I last checked). The estimator can be rewritten as

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) 1 \cdot Y_i}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) 1 \cdot 1}$$

This is the same as a weighted OLS regression of the Y_i onto a constant, using $K\left(\frac{X_i - x}{h}\right)$ as weights. You can program it in Stata exactly like this. Alternatively, the Nadaraya-Watson estimator can be programed as an IV regression of the Y_i onto a constant variable equal to 1 using the $K\left(\frac{X_i - x}{h}\right)$ as instruments.

To program it in a matrix-based program, define:

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \iota = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, K = \begin{bmatrix} K\left(\frac{X_1 - x}{h}\right) & & 0 \\ & \ddots & \\ 0 & & K\left(\frac{X_n - x}{h}\right) \end{bmatrix}$$

Then,

$$\hat{g}(x) = (\iota^T K \iota)^{-1} \iota^T K Y.$$

Programming the $n \times n$ matrix K may overburden Matlab if the sample size is large. Alternatively, we may proceed with the IV approach, which although less intuitive, avoids this issue. Define

$$K = \begin{bmatrix} K\left(\frac{X_1 - x}{h}\right) \\ \vdots \\ K\left(\frac{X_n - x}{h}\right) \end{bmatrix},$$

the estimator can be programmed as

$$\hat{g}(x) = (K^T \iota)^{-1} K^T Y.$$

From the point of view of computational burden, this estimator requires one regression per value of x. If we are interested in nonparametric estimation at a point, then this method is excellent. However, very often nonparametric estimators are used as plugins inside of sums, for example, and we have to estimate $f(X_i)$ for every value of X_i in the sample. This can add up to a very large number of regressions. This is not a problem *per se*, since computers can run thousands of simple regressions in a reasonable amount of time, but it may become a very serious problem for standard error estimation, as we will see on the next class.

2.2.2 Asymptotic behavior of the Nadaraya-Watson kernel estimator

Theorem: Suppose that

1. Let $\mathcal{G}$ denote the class of Borel measurable functions that have a finite second moment. Assume that $g(\cdot) = E(Y|X = \cdot)$ belongs to $\mathcal{G}$.
2. $E(Y^2) < \infty$, and define $\sigma^2(x) := Var(u|X = x)$.
3. X is a random vector with a density function $f_X(x) = f(x)$.
4. Let x be interior to the support of the distribution of X, and let $f(x) > 0$ on the support of the distribution of X.
5. $g(x)$ and $f(x)$ are three times continuously differentiable.
6. The kernel $k(\cdot)$ is bounded, symmetric, and integrates to 1.
7. The sample (Y_i, X_i), $i = 1, \ldots, n$ is i.i.d.
8. As $n \to \infty$, $h_s \to 0$ (for all $s = 1, \ldots, q$), $nh_1 \ldots h_q \to \infty$ and $(nh_1 \ldots h_q) \sum_{s=1}^{q} h_s^6 \to 0$

Then,

$$\sqrt{nh_1 \dots h_q}\left(\hat{g}(x) - g(x) - \sum_{s=1}^{q} h_s^2 B_s(x)\right) \xrightarrow{d} N\left(0, \frac{\sigma^2(x)}{f(x)}\left(\int k(v)^2 dv\right)^q\right)$$

where

$$B_s(x) = \frac{\int v^2 k(v) dv}{2f(x)}\left[2\frac{\partial f(x)}{\partial x_s}\frac{\partial g(x)}{\partial x_s} + f(x)\frac{\partial^2 g(x)}{\partial x_s^2}\right]$$

***

In order to understand this result, we must first analyze the assumptions:

1. This assumption states that $E(Y|X = x)$ exists and is well defined. It is an abstract assumption of no practical concern.

2. This assumption is also abstract and of no practical concern.

3. This assumption states that X is a continuously distributed random variable. This is a real restriction on reality. Even if the data is discrete, there must be a reasonable belief that the data generating process is continuous.

4. In practice, the requirement that $f(x) > 0$ translates into "there is enough data in both sides of x." The interior point requirement is more serious. In theory any interior point would qualify. In practice interior points are determined by the bandwidth, i.e. a point is interior if it's at least h units distant from the boundary. Suppose x is too close to the left boundary. In this case, the kernel will be cut at the left boundary, and therefore will weigh proportionally more observations to the right of x. If the function is decreasing, then it has a tendency to average proportionally more observations that are lower than $g(x)$, and therefore the estimator will be positively biased (see figure 3). See the problem in an empirical example in figure 4. The bias can be quite large, so kernel regression is not recommended for estimation near the boundary.

5. The real constraint in reality here is continuity. There must be a reasonable belief that $g(x)$ and $f(x)$ are continuous. In theory this should rule out estimation at concentration points, because $f(x)$ would not be continuous there. In practice, researchers tend to ignore this. The continuity of $g(x)$ is usually justified by the theory, and it is a real constraint in practice.

6. The kernel assumption is never a problem, because the kernel is chosen by the researcher. In theory, this constraint rules out the Dirac delta distribution (though it is

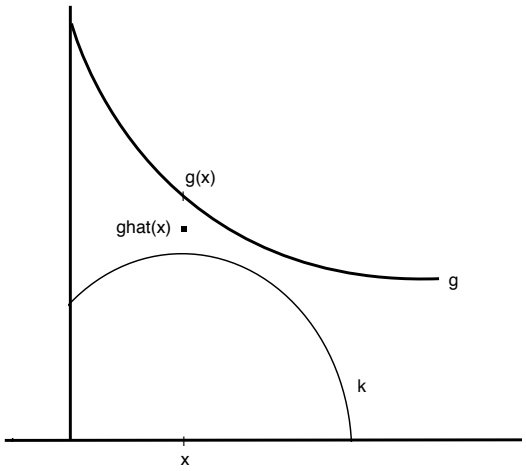

Figure 3: Kernel regression when x is less than h distance from the boundary.

the asymptotic limit of all kernels), but how would someone use a Dirac delta kernel in practice anyway? Choose a well known kernel, such as the Epanechnikov.

7. Random sampling is a real constraint, but we are all used to this.

8. Requirements on the bandwidth behavior as $n \to \infty$ are only theoretical. We make a promise of decreasing h at established rates when the sample size increases. In practice we have a given sample size, and we have to chose a bandwidth with no real guidance from this theorem. It is not trivial to choose the bandwidth, and we will consider this problem in the next class.

Summarizing, in practice the researcher has to have a random sample, be interested in the estimation of the $E(Y|X = x)$ at an interior point in the support, X has to be arguably continuously distributed (even if the data is not), and $g(x)$ must be arguably continuous. Finally, the researcher has to choose the bandwidth with care.

Now to the understanding of the theorem. The statement implies that

$$Bias(\hat{g}(x)) = \sum_{s=1}^{q} h_s^2 \int v^2 k(v) dv \left[\frac{1}{f(x)} \frac{\partial f(x)}{\partial x_s} \frac{\partial g(x)}{\partial x_s} + \frac{1}{2} \frac{\partial^2 g(x)}{\partial x_s^2} \right]$$

$$Var(\hat{g}(x)) = \frac{1}{n h_1 \dots h_q} \frac{\sigma^2(x)}{f(x)} \left(\int k(v)^2 dv \right)^q$$

We begin by looking at what affects the bias:

1. The first term that influences the bias are the bandwidths h_s. The smaller the bandwidths, the smaller the bias. This is the same as what we discussed before:

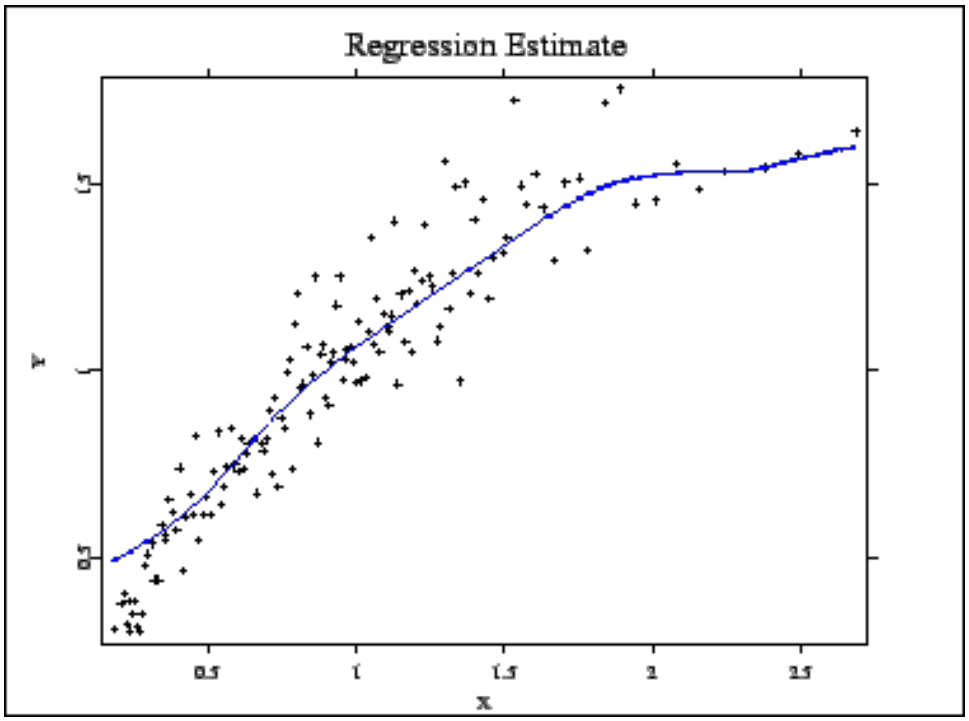

Figure 4: Nadaraya-Watson estimated regression curve. Observe the bias at the lower boundary, caused by the high derivative of the data coupled with the imbalance of data (since there is no data below the boundary.

when the bandwidth is small, the estimator is averaging only over observations that are extremely close to x, and therefore their $g(X_i)$ is closer to $g(x)$, decreasing the bias.

2. The term $\int v^2 k(v)dv$ is the same as the variance of a variable with zero mean and density k. The more concentrated around zero is the kernel, the smaller the variance of such a variable. This is exactly what we discussed before, and why the kernels that are triangle-like are better than the more uniform kernels with respect to the bias.

3. The interaction of the derivatives of f and g affects the bias. Remember that we are averaging the Y_i of the observations around x. A high derivative of $f(x)$ means that there is considerably more data to one side of x. If the derivative of g is large, then the average will have more observations that are either much above or below the true value of $g(x)$, and therefore increase the bias. The larger $f(x)$, the more likely we are to get more data that is very close to x, which would exacerbate the discrepancies.

4. The concavity (or convexity) of g affect the bias. If g is extremely concave, the observations around x will all be below the true $g(x)$, increasing the bias.

Now we consider what affects the variance

1. First, observe that the speed of convergence $\sqrt{nh_1 \dots h_q}$ affects the variance, because it is divided by $nh_1 \dots h_q$. In a parametric context, the variance gets divided by n instead, and therefore the variance tends to be smaller (think of $(h_1 \dots h_q)^{-1}$ as the cost of not knowing the functional form). The more explanatory variables the model has, the larger the variance. In order to offset the effect of the bandwidth, the requirements on the sample size get exponentially high with the inclusion of more covariates. This is the curse of dimensionality well known in nonparametric estimation. In fact, Silverman (1986) has a famous table where he shows the sample sizes required to estimate a multivariate density with a 10% mean squared error:

q	5	6	7	8	9	10
n	768	2790	10,700	43,700	187,000	842,000

2. The variance of u affects the variance in the same way that the variance of the unobservable term always affects the variance of estimators.

3. The larger $f(x)$, the smaller the variance. This is natural, the larger $f(x)$, the more observations close to x are available to be used in the estimation. More observations in the actual average means smaller variance.

4. The term $\int k(v)^2 dv$ is larger when $k(v)$ assumes consistently high values above 1. Because k integrates to one, this is more severe the more k concentrates the weight close to x. This is natural, because such kernels average over very few observations, which increases variance. The problem is exacerbated the higher the number of covariates.