Introduction to Nonparametric Econometrics

Matthias Hagmann[1][2]

HEC Lausanne & FAME

April 2003

[1]HEC Lausanne, BFSH 1, CH-1015 Lausanne-Dorigny, matthias.hagmann@hec.unil.ch.

[2]I would like to thank Daniel Hoechle and Carlos Lenz for helpful comments. All remaining errors are mine.

These lecture notes provide a nontechnical introduction to the field of nonparametric econometrics. The reader is expected to know the basics in probability theory and parametric econometrics, e.g. the classical linear regression model[1]. For a more mathematical treatment of the subject, the reader is referred to Pagan and Ullah (1999), Wand and Jones (1995) and the references therein.

1 Introduction

Let $\{(Y_i, X_i), i = 1, 2, ...\}$ be a series of independent random variables. In our context, Y_i is the dependent variable and X_i the independent or explanatory variable. In this text we will just treat the univariate case, meaning that both random variables are unidimensional. Most of the theory given will also hold under some suitable assumptions in the dependent case, i.e. when (Y_i, X_i) is a stationary time series. Some aspects:

- The distribution function (*df*) of X :

$$F_X(x) = \Pr(X \leq x).$$

- The probability density function (*pdf*) of X, $f_X(x)$, is defined as

$$F_X(x) = \int_{-\infty}^{x} f_X(z)\, dz$$

or equivalently

$$f_X(x) = F_X^{(1)}(x),$$

where $F_X^{(1)}(x)$ denotes the first derivative of the *df*.

- The joint *df* of X and Y

$$F_{X,Y}(x, y) = \Pr(X \leq x, Y \leq y).$$

- The joint *pdf* of X and Y, $f_{X,Y}(x, y)$, is defined as

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(z, w)\, dzdw.$$

- The marginal *pdf* of X, $f_X(x)$ is defined as

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y)\, dy.$$

[1] A good and sufficient source for both subjects is Hill, Griffiths and Judge (2001). A more advanced introduction to mathematical statistics is given in Hogg and Craig (1995) and to econometrics in Johnston and DiNardo (1997).

- The conditional *pdf* of Y given X, $f_{Y|X}(y \mid x)$, is defined as

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

- The conditional mean of Y given X is given by

$$E(Y \mid X = x) = \int_{-\infty}^{+\infty} y f_{Y|X}(y \mid x)\, dy = g(x).$$

  Note that $g(x)$ only depends on the $x-$variable.

Whereas the joint density $f_{X,Y}(x,y)$ delivers a very general description of the joint occurrence of two (or more) random variables, the conditional density describes the probability of occurrence of the dependent variable given that the independent variable takes some specific value x. Regression theory focuses on the first moment of the conditional density, the conditional mean. The relationship between two random variables, e.g. consumption and income is described as:

$$\begin{aligned} y &= g(x) + \varepsilon \\ E(\varepsilon \mid x) &= 0. \end{aligned} \tag{1}$$

We call $g(x)$ the regression function and ε the error or disturbance term. Note that the second line is not an assumption but follows by taking conditional expectations on both sides of EQ(1):

$$\begin{aligned} E(y \mid x) &= E(g(x) \mid x) + E(\varepsilon \mid x) \\ &= g(x) + E(\varepsilon \mid x). \end{aligned} \tag{2}$$

This is true because conditioning on x means we know that value, so $g(x)$ is known as well because the g-function does only depend on x. Therefore, the expectation for the first term can be ignored. EQ(2) then immediately implies the second line in EQ(1). Intuitively: the only information we can get out of the known variable x to predict the dependent variable y is contained in the $g-$function, the disturbance term is completely uninformative.

So far we just made some statistical description of the relationship of two random variables which is in general true. But how do we estimate this relationship given a random sample observed in the real world? In general, the researcher has no idea about the form of $g(x)$. There are basically two approaches: the parametric and the nonparametric approach.

1.1 The Parametric Approach

The parametric approach assumes that the $g-$function can be described by a few parameters. The most prominent model is the linear 2-parameter model:

$$g(x) = \alpha + \beta x.$$

Other functional forms are possible like the log-linear or quadratic model:

$$\begin{aligned} g(x) &= \alpha + \beta \ln x, \\ g(x) &= \alpha + \beta x + \gamma x^2 \end{aligned}$$

where in the last example three parameters are used to describe the functional relationship between y and x. The linear form can be justified by the fact, that given Y and X are jointly normally distributed, then the conditonal expectation will take a linear form. Estimation of all these models is quite easy, the parameters can be determined by running a simple OLS-regression. As is well known, under the Gauss-Markov assumptions, OLS is the **B**est **L**inear **U**nbiased **E**stimator or briefly **BLUE**. Remember however, that in practice, seldom one of those assumptions will hold. Especially, it is hard to believe, that the researcher knows the exact functional form of $g(x)$. The model for $g(x)$ is therefore likely to be misspecified and leads to wrong inference. The best we can hope is that the functional form assumed in our model is not too far away from the true form, so that the mistakes we make are quite small.

1.2 The Nonparametric Approach

The nonparametric approach does not assume any functional form for $g(x)$, but rather goes back to the statistical definition of conditional expectation:

$$g(x) = \int_{-\infty}^{+\infty} y f_{Y|X}(y \mid x)\, dy = \frac{1}{f_X(x)} \int_{-\infty}^{+\infty} y f_{X,Y}(x, y)\, dy. \tag{3}$$

Instead of modelling g directly, we could model the joint density of (X, Y) and the marginal *pdf* of X. Choosing a parametric model for these densities would however implicitly imply the choice of a parametric model for g (e.g. choosing normality implies linearity). If we don't know the model for g, why should we know the model for the densities? The only way out of this situation is therefore not to assume any model for the densities at all and estimate them by some datadriven nonparametric way. The estimates of the joint and the marginal can then be used to get an estimate $\hat{g}(x)$ of the conditional expectation at point x. From EQ(3) :

$$\hat{g}(x) = \frac{1}{\hat{f}_X(x)} \int_{-\infty}^{+\infty} y \hat{f}_{X,Y}(x, y)\, dy.$$

These ideas make it obvious, that the estimation of *pdfs* is a central topic in nonparametric econometrics. Once a good understanding for density estimation is developed, the estimation of regression functions is "peanuts".

1.3 Semiparametric estimation

A drawback of nonparametric estimation is that the absence of a structure given by a parametric form makes estimation less precise. If the assumed parametric model is indeed true, parametric

estimation is much more precise than a corresponding nonparametric estimator. This is especially true if the explanatory variable is multivariate. This problem of nonparametric estimation is called the "curse of dimensionality" and will be explained later in the course. It intuitively means that estimation gets "exponentially more difficult" as the dimension of the regressors increases.

A way out of this dilemma is semiparametric estimation, which is as the name suggests, a combination of parametric and nonparametric estimation ideas.

1.4 Structure of these Notes

We will start in the next chapter with probability density function estimation. After briefly reviewing the main parametric workhorse "Maximum Likelihood" estimation, we will focus on the most popular nonparametric approach called kernel density estimation. The third section develops properties of the kernel density estimator whereas the fourth section deals with the choice of the smoothing parameter. The fifth section deals with bias reduction techniques for the kernel density estimator. The sixth section then introduces nonparametric regression estimation. In Section 7 we will outline some semiparametric techniques.

2 Probability Density Function Estimation

In this section we consider a univariate series of i.i.d random variables $\{X_i, i = 1, 2, ...\}$. The reader may think about income data of households in a particular year, losses of an insurance company from fire insurance contracts or, in a time series context, a series of stock returns. Before turning to nonparametric techniques, we first repeat the maximum likelihood concept.

2.1 Maximum Likelihood

In this subsection we assume that we know the parametric form of the *pdf* we want to estimate and that it can be described by some function f which is characterized by a parameter vector θ whose true value is θ_0. Maximum likelihood is the most efficient method to get an estimate of θ_0.

2.1.1 Definition of the ML Estimator

The joint *pdf* of a random sample of size N is given by

$$f_{X_1,...,X_N}(x_1, ..., x_N; \theta_0).$$

Since we are in the i.i.d case, this joint density can be factored into the product of its marginals:

$$f_{X_1,...,X_N}(x_1, ..., x_N; \theta_0) = \prod_{i=1}^{N} f_{X_i}(x_i; \theta_0). \tag{4}$$

Note that in EQ(4), the parameter θ_0 is **fixed** and characterizes the true *pdf* of the random sample. The joint density function allows to evaluate the probability of occurrence of **each** single vector $x = (x_1, ..., x_N)$. The likelihood function is something very similar like the joint density function, but the roles of x and θ_0 are exchanged. Now the sample x is observed and a **fixed** point. Instead we can **vary** the vector θ which parametrizes the *pdf*. Every different parameter value θ specifies a different joint density. We want to choose θ such that the implied joint density is maximized at the sample vector x, meaning that it is likely that the given realisation $x = (x_1, ..., x_N)$ is indeed observed from such a joint density.

The likelihood function is given by

$$L(\theta; x_1, ..., x_N) = \prod_{i=1}^{N} f_{X_i}(\theta; x_i). \tag{5}$$

The estimation of θ is obvious now: choose this parameter vector to maximize the likelihood function.

Definition 1 *Given an observed (i.i.d) sample x and some parametrized family of probability density functions $f_X(x;\theta)$, the maximum likelihood estimator $\hat{\theta}_{ML}$ is the maximizer of the likelihood function:*

$$\hat{\theta}_{ML} = \arg\max_{\theta} \prod_{i=1}^{N} f_{X_i}(\theta; x_i).$$

Remark 1 *Almost always we consider the log likelihood function which is the logarithm of the likelihood function and easier to handle:*

$$\log L(\theta; x_1, ..., x_N) = \sum_{i=1}^{N} \log f_{X_i}(\theta; x_i). \tag{6}$$

Since this transformation is monotonic, the maximizers of the likelihood and log likelihood function are identical.

2.1.2 Properties of ML Estimators

Since this is not a course about ML estimation, we just collect here the most important properties of the ML estimator:

- Consistency: $\hat{\theta}_{ML}$ converges in probability to the true parameter value θ_0. This means that for a very large sample size N, $\hat{\theta}_{ML}$ is very close to θ_0.
- Asymptotic normality: it can be shown that

  $$\sqrt{N}\left(\hat{\theta}_{ML} - \theta_0\right) \to N\left(0, I(\theta_0)^{-1}\right)$$

  where $I(\theta_0)^{-1}$ is the variance of the maximum likelihood estimator[2].
- It can be shown that any other possible method for the estimation of θ_0 yields an estimator $\hat{\theta}$ which has (for large N) a variance at least as large as the ML estimator:

  $$\text{Var}\left(\hat{\theta}\right) \geq \text{Var}\left(\hat{\theta}_{ML}\right) \text{ for large N.}$$

Obviously it is the third point which makes ML estimation very attractive since the first two are shared by other estimators as well. So in case we indeed know the true parametric density function, the best thing we can do is ML.

In practice it is however unlikely that we know the true parametric form. Figure 1 shows two examples what happens, when one misspecifies the parametric density function. Data from

[2] $I(\theta_0)$ is also called the Fisher-information matrix.

a lognormal LN$(0,1)$ and a Weibull$(1,1.5)$[3] were generated and the two *pdfs* estimated assuming that both data sets can be described by a gamma density.

Figure 1 approx. here

The figure clearly shows that the fit for the Weibull sample is, although incorrect, not too bad. The fit for the lognormal sample is however quite bad. The mode of the density is overestimtated, whereas the tails of the density are underestimated[4].

2.2 Example

Before turning to nonparametric techniques, we conclude this subsection with an example for ML estimation.

Example 1 *We have a random sample of size N and assume that a single observation is $X_i \sim N(\mu, \sigma^2)$. We have to compute EQ(6) where*

$$\begin{aligned} \log f_{X_i}(\mu, \sigma^2; x_i) &= \log\left[\left(2\pi\sigma^2\right)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}\left[x_i-\mu\right]^2\right)\right] \\ &= -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left[x_i-\mu\right]^2 \end{aligned} \tag{7}$$

The log likelihood function is then given by[5]*:*

$$\log L = -\frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left[x_i-\mu\right]^2.$$

The two FOC give us the ML estimators

$$\begin{aligned} \hat{\mu} &= \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{x}, \\ \hat{\sigma}^2 &= \frac{1}{N}\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2, \end{aligned}$$

which are just the sample mean and the sample variance[6]*.*

[3] These and the gamma density are parametric density functions which can be described by two parameters.

[4] For a real life example: suppose these data are losses from fire insurance contracts. To control their risk, insurance companies try to estimate the density for their losses. Especially the large losses are important for them. As the right hand panel in Figure 1 shows, the wrong parametric form can potentially underestimate the risk contained in the insurance portfolio.

[5] The first term in EQ(7)is not relevant since it does not contain any of the parameters of interest.

[6] Note that the estimator for the variance is slightly biased, since an unbiased estimator for σ^2 has just $(N-1)$ in the denominator. The effect is however small even for moderate sample sizes.

2.3 The Histogram

The first and simplest nonparametric description of a density is the standard histogram. To construct a histogram, one first has to determine the so called origin D_0 and the binwidth b (see Figure 2). The origin determines where one begins to construct the bins, the binwidth determines the width of a single bin. The histogram is then completely determined by the fact that the density has to integrate to one. The procedure to calculate the density at a point x is as follows:

- determine in which bin x is located. Call this $bin\,(x)$.
- count the number of observations in $bin\,(x)$ and normalize such that the density integrates to one. Formally:

$$\hat{f}\,(x) = \frac{1}{Nb} \sum_{i=1}^{N} I\,(X_i \in bin\,(x)) \tag{8}$$

  where I is the indicator function and equals one if the condition in brackets is true and zero otherwise.

Why do we have to normalize by Nb? Imagine we just have one large bin which contains all observations (Figure 3). As can be seen from EQ(8), the density takes inside the bin everywhere the value $1/b$. Since the width of the bin is b, the density integrates to one and we therefore have chosen the right normalization.

Figure 2-5 approx. here

Some remarks:

- Inside a bin, the density estimate is a constant.
- The density is not continuous when the bin changes.
- Choosing a different origin D_0, the histogram changes its appearance (Figures 2/4).
- A large binwidth produces a very smooth density (Figure 3), a small one returns a very rough and erratic picture (Figure 5).

The first three properties are clearly not a desirable feature of a density estimator. As we will see later, the last observation will turn out to be one of the most important questions in the nonparametric literature: the choice of the smoothing parameter.

2.4 The local Histogram or naive Kernel Estimator

We try in a first step to get rid of the choice of the origin. The idea is to center a local bin at each point x where one wants to estimate the density. Then one just considers points for the density calculation that are in a neighbourhood $\pm b$ of the point of interest x. Formally and similarly to EQ(8)[7]:

$$\begin{aligned}\hat{f}(x) &= \frac{1}{N \cdot (2b)} \sum_{i=1}^{N} I(x - b \leq X_i \leq x + b) \\ &= \frac{1}{N \cdot b} \sum_{i=1}^{N} \frac{1}{2} I\left(-1 \leq \frac{X_i - x}{b} \leq 1\right) \end{aligned} \tag{9}$$

which can be rewritten as

$$\hat{f}(x) = \frac{1}{Nb} \sum_{i=1}^{N} K\left(\frac{X_i - x}{b}\right) \tag{10}$$

where

$$K(u) = \begin{cases} 1/2, & |u| \leq 1 \\ 0, & |u| > 1 \end{cases}. \tag{11}$$

The estimator in EQ(9) is called the local histogram or more often the naive kernel density estimator. The K function, called kernel, displayed in Figure 6 is always positive, integrates to one and is just a uniform probability density function.

Figure 7 considers an example of this estimator, 200 random standard normal values were drawn and the naive kernel estimator applied. Although we got rid of the choice of the origin problem, we see that the density is still not continuous[8]. Figure 8 makes clear why this is the case. Estimating the density at point x_1 includes 6 observations in the local bin. Estimating a bit to the right of x_1, the point which were just on the left boundary does no longer get any weight in the new local bin and only 5 observations remain. There is therefore a discrete jump in the density as soon as one observation enters or leaves the local bin. The reason for these jumps is that the weight which is allocated to an observation is not a continuous function of its distance from the point of interest. This feature is not desirable since we would like to get a continuous density.

2.5 The Kernel Density Estimator

To get a continuous density estimate, we have to replace the uniform kernel by another kernel function which changes the weight allocated to the observation in a neighbourhood of x in a

[7] Note that width of the local bin is 2b, this explains the denominator in EQ(9) .

[8] The figure shows a rough but continuous density, this is because the program (Eviews), which created this picture, is connecting the points by lines.

continuous way. This kernel function can be any symmetric function which is always positive and integrates to one. Note that every symmetric *pdf* satisfies this requirement.

Definition 2 *A kernel estimator of $f(x)$ is defined by*

$$\hat{f}(x) = \frac{1}{Nb}\sum_{i=1}^{N} K\left(\frac{X_i - x}{b}\right)$$

where b is called the bandwidth or smoothing parameter and K is called the kernel and satisfies[9]

$$\int K(u)\,du = 1, \tag{12}$$

$$K(u) \geq 0 \quad \textit{everywhere.} \tag{13}$$

We further assume that

$$b \to 0 \textit{ as } N \to \infty \tag{14}$$

$$Nb \to \infty \textit{ as } N \to \infty \tag{15}$$

Property (12) makes sure that the estimated density is integrating to one. This follows since by a change of variable[10]

$$\begin{aligned}
\int \hat{f}(x)\,dx &= \int \frac{1}{Nb}\sum_{i=1}^{N} K\left(\frac{X_i - x}{b}\right) dx \\
&= \frac{1}{Nb}\sum_{i=1}^{N} \int K\left(\frac{X_i - x}{b}\right) dx \\
&= \frac{1}{Nb}\sum_{i=1}^{N} \int K(u)\,b\,du \\
&= \int K(u)\,du,
\end{aligned}$$

which equals only one if the used kernel function integrates to one. Property (13) ensures that $\hat{f}(x) \geq 0$, we will relax this property later. Property (14) means that as our available sample gets larger, we decrease the bandwidth and just give weight to observations in a shrinking neighbourhood around the point x, where we estimate the density. Property (15) means that the smoothing parameter is not approaching zero as fast as the sample size approaches infinity. Popular kernel functions which satisfy the above requirements are:

- Triangular kernel: $(1 - |u|)\, I(|u| \leq 1)$.

[9] The limits of integration are from now on $-\infty$ to $+\infty$ unless otherwise stated.

[10] Use $u = \frac{X_i - x}{b}$ and the fact that $K\left(\frac{X_i - x}{b}\right) = K\left(\frac{x - X_i}{b}\right)$.

- Epanechnikov kernel: $\frac{3}{4}\left(1-u^2\right) I\left(|u| \leq 1\right)$.
- Gaussian kernel: $(2\pi)^{-1/2} \exp\left(-\frac{1}{2}u^2\right)$.

Figures 9/10/11 show density estimates using those kernels and the same data as in Figure 7. An optimal bandwidth was chosen, we discuss later how this is done. Note that the rough feature present in the naive kernel estimator disappeared. We also remark, that independent of the kernel used, the density estimates look very similar. We will see that $\hat{f}(x)$ depends more on the smoothing parameter b than on K.

3 Properties of the Kernel Density Estimator

In this section we derive mean square error, bias and variance of the kernel density estimator and also show that the estimator is consistent and has an asymptotic normal distribution.

3.1 Mean Square Error (MSE)

We start with a definition.

Definition 3 *The MSE of an estimator $\hat{\theta}$ of the true value θ_0 is given by*[11]

$$\text{MSE}\left(\hat{\theta}\right) = E\left(\hat{\theta} - \theta_0\right)^2.$$

The MSE gives a good indication of how close the estimator is to the true value θ_0. We prefer estimators with small MSE. Note that at first glance, the definition of MSE seems to be the same as the variance of an estimator. Indeed, there is a relationship between these two measures which leads us to our first Lemma.

Lemma 1 $\text{MSE}\left(\hat{\theta}\right) = \text{Bias}\left(\hat{\theta}\right)^2 + \text{Var}\left(\hat{\theta}\right)$.

Recall the definition of the bias of an estimator.

Definition 4 *The bias of an estimator $\hat{\theta}$ of the true value θ_0 is defined as*

$$\text{Bias}\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta_0.$$

Example 2 *Assume we draw 1 Billion samples of 200 random values from a standard normal distribution and calculate for each sample j the density estimate $\hat{f}_j(x)$, if the average of all these evaluated estimators is not equal to $f(x)$, then our estimator has a bias. Recall that e.g. the OLS estimator is unbiased, so is the sample mean, c.f. Example 1, since*

$$\begin{aligned} E(\bar{x}) &= E\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} E(x_i) = \mu. \end{aligned}$$

[11] $\hat{\theta}$ can be for example the least squares estimator $\hat{\beta}$ in linear regression, in our context it is the kernel density estimator $\hat{f}(x)$. Also note that the notation $E\left(\hat{\theta} - \theta_0\right)^2$ means $E\left(\left(\hat{\theta} - \theta_0\right)^2\right)$ and not $\left[E\left(\hat{\theta} - \theta_0\right)\right]^2$. When we mean the latter, we will explicitely write it like this.

Let's proove now Lemma 1.

Proof. By definition:

$$\begin{aligned} MSE\left(\hat{\theta}\right) &= E\left(\hat{\theta}-\theta_0\right)^2 \\ &= E\left(\left[\hat{\theta}-E\left(\hat{\theta}\right)\right]+\left[E\left(\hat{\theta}\right)-\theta_0\right]\right)^2 \\ &= E\left[\hat{\theta}-E\left(\hat{\theta}\right)\right]^2+E\left[E\left(\hat{\theta}\right)-\theta_0\right]^2+ \\ &\quad 2\cdot E\left(\left[\hat{\theta}-E\left(\hat{\theta}\right)\right]\left[E\left(\hat{\theta}\right)-\theta_0\right]\right) \\ &= Var\left(\hat{\theta}\right)+Bias\left(\hat{\theta}\right)^2, \end{aligned}$$

which follows since as $\left[E\left(\hat{\theta}\right)-\theta_0\right]$ is constant, we can take it out of the expectation and therefore

$$\begin{aligned} E\left(\left[\hat{\theta}-E\left(\hat{\theta}\right)\right]\left[E\left(\hat{\theta}\right)-\theta_0\right]\right) &= \left[E\left(\hat{\theta}\right)-\theta_0\right]\cdot E\left[\hat{\theta}-E\left(\hat{\theta}\right)\right] \\ &= 0. \end{aligned}$$

■

3.2 Bias and Variance of the Kernel Estimator

There exists in general no unbiased nonparametric density estimator. We first derive bias and variance of the kernel density estimator and then give a lengthy intuitive discussion of these two expressions.

Starting with the bias, we have that

$$Bias\left(\hat{f}(x)\right)=E\left(\hat{f}(x)\right)-f(x).$$

The first term can be written as

$$\begin{aligned} E\left(\frac{1}{Nb}\sum_{i=1}^{N}K\left(\frac{X_i-x}{b}\right)\right) &= \frac{1}{b}E\left(K\left(\frac{X_1-x}{b}\right)\right) && (16) \\ &= \frac{1}{b}\int K\left(\frac{X_1-x}{b}\right)f(X_1)\,dX_1 && (17) \\ &= \int K(u)\,f(bu+x)\,du && (18) \end{aligned}$$

because

- (16) follows since data is i.i.d.,
- (17) follows by definition of expectation,

- (18) follows by a change of variable using $u = \frac{X_1 - x}{b}$.

The last term we can be simplified using a Taylor expansion:

$$\begin{aligned} & \int K(u) \left[f(x) + f^{(1)}(x) bu + \frac{1}{2} f^{(2)}(x) b^2 u^2 + ... \right] du \\ = \; & f(x) + \frac{b^2}{2} f^{(2)}(x) \int K(u) u^2 du + ... \end{aligned}$$

The second term in the bracket dissapeared since the kernel is symmetric, therefore $\int uK(u) du = 0$. Collecting everything together and abstracting from higher order terms[12]we get that

$$Bias\left(\hat{f}(x)\right) = \frac{b^2}{2} f^{(2)}(x) \int K(u) u^2 du.$$

Turning to the variance:

$$\begin{aligned} Var\left(\hat{f}(x)\right) &= \frac{1}{N^2 b^2} Var\left[\sum_{i=1}^{N} K\left(\frac{X_i - x}{b} \right) \right] && (19) \\ &= \frac{1}{Nb^2} Var\left(K\left(\frac{X_1 - x}{b} \right) \right) && (20) \\ &= \frac{1}{Nb^2} E\left(K\left(\frac{X_1 - x}{b} \right)^2 \right) - \frac{1}{Nb^2} \left[E\left(K\left(\frac{X_1 - x}{b} \right) \right) \right]^2 && (21) \end{aligned}$$

because

- (19) follows since $Var(aX) = a^2 Var(X)$,
- (20) follows since we have i.i.d data, so no covariance terms,
- (21) follows since $Var(X) = E(X^2) - [E(X)]^2$.

It can be shown that the second term in EQ(21) is "very small" and is therefore neglected from further calculations, the first term can be written, using similar steps as above, as

$$\begin{aligned} & \frac{1}{Nb^2} \int K\left(\frac{X_1 - x}{b} \right)^2 f(X_1) dX_1 \\ = \; & \frac{1}{Nb^2} \int K(u)^2 f(bu + x) bdu \\ = \; & \frac{1}{Nb} \int K(u)^2 \left[f(x) + f^{(1)}(x) bu + \frac{1}{2} f^{(2)}(x) b^2 u^2 + ... \right] du \\ = \; & \frac{f(x)}{Nb} \int K(u)^2 du + ..., \end{aligned}$$

again neglecting small terms. We can therefore state now our first proposition.

[12]Higher order terms are terms which are very small as soon as the sample size takes sufficiently large values. They are therefore ignored.

Proposition 1 *Under suitable conditions and abstracting from higher order terms, the bias and variance of the kernel density estimator are given by the following expressions:*

$$\text{Bias}\left(\hat{f}(x)\right) = \frac{b^2}{2} f^{(2)}(x) \cdot \int K(u) u^2 du, \tag{22}$$

$$\text{Var}\left(\hat{f}(x)\right) = \frac{f(x)}{Nb} \cdot \int K(u)^2 du. \tag{23}$$

We note that whereas the bias of the kernel density estimator is increasing with b, the variance of the estimator is getting larger as we decrease the smoothing parameter. To minimize the MSE, we would like to have both bias and variance to be small. The opposite effect of the smoothing parameter on bias and variance means that there is a trade off between these two terms. Where does this come from? Imagine we want to estimate the density at x. Then we would like to use just points very close to x to get a good description of $f(x)$, since those close points seem to represent best the local behaviour of $f(x)$. But as Figure 12 shows, a small bandwidth leads to a very variable density estimate with a lot of bumps which can hardly be a useful density estimate. We call this bumpy feature undersmoothing. On the other side, choosing the bandwidth too large is called oversmoothing. Note that although the right panel in Figure 12 looks quite well, there is way too much weight in the tails of the density for standard normal data. The right amount of smoothing lies somewhere in the middle. We will devote Section 4 to the choice of bandwidth, which is still one of the most important areas of research in the nonparametric literature.

3.3 Consistency of the Kernel Density Estimator

We show here convergence in mean square, recall the definition.

Definition 5 *An estimator $\hat{\theta}_N$ is said to converge in mean square*[13] *to the true value θ_0 if the following condition holds true:*

$$\lim_{N\to\infty} E\left(\hat{\theta}_N - \theta_0\right)^2 = 0.$$

Note that for an unbiased estimator, this just means that the variance of the estimator converges to zero as $N \to \infty$. This basically means that $\hat{\theta}$ converges to θ_0 as the sample size gets larger and larger.

Proposition 2 *Under some suitable (weak) conditions and if EQ(14) and EQ(15) hold, then the kernel density estimator is mean square consistent.*

$$\lim_{N\to\infty} E\left(\hat{f}(x) - f(x)\right)^2 = 0.$$

[13] This is also called L^2 convergence.

Proof. *From Lemma 1 we just have to show*

$$\lim_{N\to\infty}\left[\text{Bias}\left(\hat{f}(x)\right)^2+\text{Var}\left(\hat{f}(x)\right)\right]=0,$$

but from the assumptions made this follows easily:

$$\lim_{N\to\infty}\text{Bias}\left(\hat{f}(x)\right)^2=\lim_{N\to\infty}\left[\frac{b^2}{2}f^{(2)}(x)\cdot\int K(u)\,u^2du\right]^2=0$$

since $b\to 0$ *as* $N\to\infty$ *and*

$$\lim_{N\to\infty}\text{Var}\left(\hat{f}(x)\right)=\lim_{N\to\infty}\frac{f(x)}{Nb}\cdot\int K(u)^2\,du=0$$

since $Nb\to\infty$ *as* $N\to\infty$. ■

We next give a discussion of asymptotic normality of the kernel density estimator.

3.4 Asymptotic normality of the Kernel Density Estimator

Although we know the bias and variance of the kernel density estimator, we can't do valid and efficient inference, e.g. construct confidence intervals, since we don't know yet the distribution of the kernel density estimator. Recall first the central limit theorem.

Theorem 1 *If we have an i.i.d sample* $\{X_i\}_{i=1}^N$ *with* $E(X_i)=\mu$ *and* $Var(X_i)=\sigma^2$, *then*

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(X_i-\mu)\to N\left(0,\sigma^2\right). \tag{24}$$

The above means that $\sqrt{N}(\bar{x}-\mu)$ has for very large N a distribution which is normal with mean zero and variance σ^2. Note that we have two effects: by the law of large numbers, $\bar{x}$ converges to its mean. Also the variance of $\bar{x}$ converges to zero since the variance of the sample mean is σ^2/N. We need the premultiplier $\sqrt{N}$ to ensure that we don't get a point mass distribution at μ. Note that

$$Var\left(\sqrt{N}\bar{x}\right)=N\cdot Var(\bar{x})=N\cdot\frac{\sigma^2}{N}=\sigma^2,$$

so this premultiplier ensures that we get a stable variance, which does not depend on N, the number of observations we have.

We would like to get a similar result like (24) for our kernel density estimator. The discussion we give here is very intuitive but sufficient for our purposes[14]. Note that

$$\hat{f}(x)-f(x)=\left(\hat{f}(x)-E\left[\hat{f}(x)\right]\right)+\left(E\left[\hat{f}(x)\right]-f(x)\right). \tag{25}$$

[14] A mathematically correct proof can be found in Pagan and Ullah (1999).

To start, let us focus on the first bracket in EQ(25). It can be written as

$$\frac{1}{Nb}\sum_{i=1}^{N} K\left(\frac{X_i - x}{b}\right) - E\left[\hat{f}(x)\right].$$

Note that by construction, this whole term has zero expectation and the first term is a sum of i.i.d random variables. If we can stabilize the variance of this whole term as $N \to \infty$, we should by the central limit theorem get a limiting normal distribution for the kernel density estimator. Call the stabilizing factor c, then we want to have

$$\begin{aligned} & Var\left(c\left(\hat{f}(x) - E\left[\hat{f}(x)\right]\right)\right) \\ = \; & c^2\, Var\left(\hat{f}(x)\right) \\ = \; & c^2 \frac{f(x)}{Nb} \cdot \int K(u)^2\, du = const. \end{aligned}$$

It follows therefore immediately that we have to choose

$$c = \sqrt{Nb}.$$

Premultiplying EQ(25) by this term yields

$$\sqrt{Nb}\left(\hat{f}(x) - E\left[\hat{f}(x)\right]\right) + \sqrt{Nb}\left(E\left[\hat{f}(x)\right] - f(x)\right).$$

We established that the first term has a stable variance and has by the CLT a limiting normal distribution, the second term is by EQ(22) just given by

$$\begin{aligned} & \sqrt{Nb}\frac{b^2}{2} f^{(2)}(x) \cdot \int K(u)\, u^2 du \\ = \; & N^{1/2} b^{5/2} f^{(2)}(x) \cdot \int K(u)\, u^2 du. \end{aligned} \tag{26}$$

We want term (26) to go to zero as N gets very large, such that this term can be ignored. This term then has no effect on the asymptotic distribution. We therefore assume that in addition to EQ(14) and EQ(15) , also

$$Nb^5 \to 0 \tag{27}$$

holds. Summing up, we can therefore establish that

$$\sqrt{Nb}\left(\hat{f}(x) - f(x)\right) \to N\left(0, f(x) \cdot \int K(u)^2\, du\right).$$

Using this result, an approximate pointwise 95% confidence intervall for f can be written as

$$\hat{f}(x) \pm 1.96\,(Nb)^{-1/2}\left(f(x) \cdot \int K(u)^2\, du\right)^{1/2}.$$

In practice we replace $f(x)$ by its consistent estimator $\hat{f}(x)$.

Remark 2 *In parametric econometrics, the stabilizing factor or the so called rate of convergence is* $\sqrt{N}$*. In nonparametric econometrics, we just saw that the factor takes the form* $\sqrt{Nb}$*, which is slower than the parametric rate. This is because* Nb *is not converging to infinity as fast as* N *is. The intuition behind this is that in parametric estimation, the* ***whole*** *sample size is used. In nonparametrics, one is focussing on a shrinking neighbourhood of the point of interest, so just* ***part*** *of the sample is needed. This shrinking neighbourhood is described by the smoothing parameter, which explains its appearance in the nonparametric rate of convergence.*

Remark 3 *Imagine we want to estimate a* d*-dimensional density, one can show that in this case the convergence rate is no longer* $\sqrt{Nb}$ *but* $\sqrt{Nb^d}$*. So increasing the dimension means that our estimator converges even slower to its true value* $f(x)$*. This is different in parametric estimation, the rate of convergence is always the same. This situation is called "the curse of dimensionality" and is the main drawback of nonparametric techniques.*

4 Optimal Bandwidth and Kernel

We saw in the last section that the bias of the kernel density estimator can be decreased by choosing a low smoothing parameter. But at the same time, this increased the variance. In this section we derive the optimal bandwidth.

4.1 Optimal Bandwidth

Recall that we used the MSE as a goodness of fit measure for our density estimator. Using Lemma 1 and Equations (22) and (23), we can write

$$MSE\left(\hat{f}(x)\right) = \frac{C_1}{Nb} + C_2^2 b^4 \tag{28}$$

where

$$\begin{aligned} C_1 &= f(x) \cdot \int K(u)^2 du, \\ C_2 &= \frac{1}{2} f^{(2)}(x) \cdot \int K(u) u^2 du. \end{aligned}$$

We can now choose the smoothing parameter such that the MSE is minimized:

$$\frac{dMSE}{db} = -\frac{C_1}{Nb^2} + 4C_2^2 b^3 = 0$$

which implies that the optimal bandwidth at point x is given by

$$b_{opt}(x) = \left(\frac{C_1}{4C_2^2}\right)^{1/5} N^{-1/5} = \left(\frac{f(x) \cdot \int K(u)^2 du}{\left[f^{(2)}(x) \int K(u) u^2 du\right]^2}\right)^{1/5} N^{-1/5}. \tag{29}$$

Note that this optimal bandwidth varies with x, we can avoid this by minimizing a global goodness of fit measure. The most popular global measure is the integrated mean square error defined as

$$MISE = \int MSE\left(\hat{f}(x)\right) dx.$$

Following similar steps like above, one can show that the optimal global bandwidth is given by

$$b_{opt} = \left(\frac{\int K(u)^2 du}{\int f^{(2)}(x)^2 dx \cdot \left[\int K(u) u^2 du\right]^2}\right)^{1/5} N^{-1/5} \tag{30}$$

Note that the optimal local and global bandwidths depend on the sample size, the kernel and on the unknown quantities $f(x)$ and $f^{(2)}(x)$. The optimal bandwidth can therefore not be computed. Before turning to some possible solutions, let's give the intuition why $\int f^{(2)}(x)^2 dx$ appears in the expression for the optimal global bandwidth[15]. Note that when $f(x)$ is a very smooth density,

[15] Remember that the second derivative measures the curvature of a function at point x.

e.g. close to a uniform density, then this expression is close to zero and the optimal bandwidth is therefore very large. Intuitively, if the density takes similar values almost everywhere, then we can borrow information from a large neighbourhood of the point of interest. If the density is very curvy, this is not the case and we should use a small bandwidth.

Here are some solutions for the unknown values in EQ(29) and (30):

- insert pilot estimates of $f(x)$ and $f^{(2)}(x)$, using kernel estimates based on arbitrary b.
- insert pilot estimates of $f(x)$ and $f^{(2)}(x)$ based on a parametric model for f, e.g. take the normal density.
- since b should vary with the scale of X, estimate the first factor in EQ(30) by the sample standard deviation[16]. Then just choose the bandwidth proportional to $N^{-1/5}$.

The second rule is known as Silverman's rule of thumb, one can show that, combined with a gaussian kernel function, it amounts to choose

$$b = 1.06\sigma N^{-1/5},$$

where σ is the standard deviation and can be estimated from the data. This rule seems to work pretty well in practice unless the density is bimodal or strongly skewed.

4.2 Optimal Kernel

EQ(28) shows that the MSE (as well the MISE) also depends on the chosen kernel function[17]. So to make the M(I)SE as small as possible, we could choose a kernel function which minimizes the M(I)SE. It can be shown that the Epanechnikov kernel is in this sense optimal. But also, the gain of the Epanechnikov kernel versus other kernels is very small, so that researchers usually do not care about the choice of kernel. What really matters in nonparametrics is the choice of bandwidth.

4.3 Cross Validation Method of choosing Bandwidth

We outline here an automatic method which achieves for large N the optimal global bandwidth given in EQ(29) . Write

$$\hat{f}_b(x) = \frac{1}{Nb}\sum_{i=1}^{N} K\left(\frac{X_i - x}{b}\right) \tag{31}$$

[16]This means that for data ranging from -1000 to 1000, the bandwidth should be larger than for data ranging just from -1 to 1.

[17]We could substitute the optimal bandwidth given in EQ(28) back into EQ(29) to obtain the optimal MSE, which does no longer depend on the bandwidth, but on unknown density terms and the chosen kernel.

to emphasize the dependence of the kernel estimator on the smoothing parameter b. Then by analogy with parametric maximum likelihood, consider

$$L(b) = \prod_{j=1}^{N} \hat{f}_b(X_j). \tag{32}$$

Note that this function is very similar to the likelihood function given in EQ(5). So a very intuitive idea is to choose the smoothing parameter b such that this function is maximized. It turns out that this is problematic since substituting EQ(31) in EQ(32) yields

$$\begin{aligned} L(b) &= \prod_{j=1}^{N} \frac{1}{Nb} \sum_{i=1}^{N} K\left(\frac{X_i - X_j}{b}\right) \\ &= \prod_{j=1}^{N} \frac{1}{Nb} \left[K(0) + \sum_{i \neq j} K\left(\frac{X_i - X_j}{b}\right)\right]. \end{aligned} \tag{33}$$

This function is of course maximized if we choose $b = 0$, which can hardly be a good smoothing parameter. A solution to this problem is the so called leave one out estimator, defined as

$$\hat{f}_b^*(X_j) = \frac{1}{(N-1)\,b} \sum_{i \neq j}^{N} K\left(\frac{X_i - X_j}{b}\right),$$

this will prevent the appearance of the problematic $K(0)$ term in EQ(33). The cross validated bandwidth is then defined as

$$\hat{b} = \arg\max_b \prod_{j=1}^{N} \hat{f}_b^*(X_j).$$

This bandwidth selection is data-dependent, but can come close to the optimal bandwidth as described above. This procedure is relatively expensive to compute for large N, though this is increasingly possible with faster computational facilities.

5 Bias Reduction Techniques

In this section we deal with bias reduction techniques for the kernel density estimator. This topic is of particular importance in the nonparametric literature, since the bias induced by smoothing can be large.

5.1 Higher Order Kernels

Recall the development of the bias of the kernel density estimator in Section 3, EQ(18):

$$Bias\left(\hat{f}(x)\right) + f(x) = \int K(u) f(ub + x)\, du.$$

We can use a Taylor expansion to approximate the last term by

$$\begin{aligned} & \int K(u) \left[\begin{array}{c} f(x) + f^{(1)}(x)\, ub + \frac{1}{2} f^{(2)}(x) (ub)^2 + \frac{1}{6} f^{(3)}(x) (ub)^3 + \\ \frac{1}{24} f^{(4)}(x) (ub)^4 + ... \end{array} \right] du \\ = \; & f(x) + \frac{b^2}{2} f^{(2)}(x) \int u^2 K(u)\, du + \frac{b^4}{24} f^{(4)}(x) \int u^4 K(u)\, du + ..., \end{aligned}$$

which follows since the symmetry of the kernel function implies that the second and forth term in brackets are zero. Abstracting from higher order terms than those considered here, we can therefore write the bias of the kernel estimator as

$$Bias\left(\hat{f}(x)\right) = \frac{b^2}{2} f^{(2)}(x) \int u^2 K(u)\, du + \frac{b^4}{24} f^{(4)}(x) \int u^4 K(u)\, du. \tag{34}$$

Note that the second term is converging faster to zero than the first[18], this is why we ignored this term in the development of the bias in Section 3.

The so called higher order kernel technique aims at eliminating the first and larger term by choosing an appropriate kernel function. For this we need

$$\int u^2 K(u)\, du = 0.$$

It is obvious that this is only possible when $K(u)$ takes negative values, since u^2 is for sure everywhere positive. Before showing how to construct such a higher order kernel, let us remark that the above discussion can be extended to also eliminate the second term in EQ(34) , this would require

$$\int u^4 K(u)\, du = 0.$$

In general, let us give the following definition:

[18] In case this is not obvious for you: use the bandwidth selection rule in EQ(30) . EQ(34) can then be written as

$$Bias\left(\hat{f}(x)\right) = c_1 N^{-2/5} + c_2 N^{-4/5},$$

where c_i are some irrelevant constants. The second term converges faster to zero than the first.

Definition 6 *A function $K_{(r)}(u)$ is called a kernel function of order r if the following conditions hold:*

$$\int K_{(r)}(u)\,du = 1, \tag{35}$$

$$\int u^j K_{(r)}(u)\,du = 0 \quad 1 \leq j \leq r-1. \tag{36}$$

The corresponding kernel density estimator of order r is given by

$$\hat{f}_{(r)}(x) = \frac{1}{Nb}\sum_{i=1}^{N} K_{(r)}\left(\frac{X_i - x}{b}\right).$$

Again, Condition (35) ensures that the density estimator integrates to one. Condition (36) ensures that "enough" terms in the Taylor expansion of the bias cancel out. Again, abstracting from terms of higher order than those considered, we obtain the bias of the higher order kernel density estimator.

Proposition 3 $\text{Bias}\left(\hat{f}_{(r)}(x)\right) = \frac{b^r}{r!} f^{(r)}(x) \int u^r K_{(r)}(u)\,du.$

Note that the standard kernel function used so far satisfies the definition of a higher order kernel with $r = 2$. Also note that as long as the kernel function is symmetric, Condition (36) is not a restriction when j is odd.

Example 3 Let us construct here a kernel function of order $r = 3$. Note that this is what we need to set the first term in EQ(34) equal to zero. Let $\phi(u)$ denote the standard normal *pdf* and consider

$$K_{(3)}(u) = \left(c + d \cdot u^2\right)\phi(u).$$

Then the following conditions have to hold:

$$\int K_{(3)}(u)\,du = 1,$$

$$\int u^2 K_{(3)}(u)\,du = 0.$$

These two conditions can be written as[19]

$$c + d = 1,$$

$$c + 3d = 0.$$

The solutions are $c = 3/2$ and $d = -1/2$.We plot this kernel function of order three in Figure ?.

Figure 13 app. here

[19]Recall that if u is a normal random variable with zero mean and variance σ^2, then $E\left(u^4\right) = 3\,\sigma^2$.

Note that since the higher order kernel function is not everywhere positive, the density estimate can take negative values as well. This problem does only vanish if samples are very large, since the kernel estimator is consistent. In practice, although theoretically appealing, there is hardly bias improvement from choosing $r \geq 3$. Although we did not derive the exact expression for the variance of the higher order kernel here, note from EQ(23) that the variance of the kernel estimator depends on

$$\int K(u)^2 \, du,$$

which is likely to be larger for a higher order kernel than a standard kernel. Also note that using the Taylor expansion in our bias derivation, we implicitely assumed that the unknown density is r-times differentiable, which may not be true.

5.1.1 Implications for Bandwidth Selection

Abstracting from higher order terms, the MSE for the higher order kernel density estimator is given by

$$MSE\left(\hat{f}_{(r)}(x)\right) = \frac{C_1}{Nb} + C_r^2 b^r,$$

where

$$\begin{aligned} C_1 &= f(x) \cdot \int K_{(r)}(u)^2 \, du, \\ C_r &= \frac{f^{(r)}(x)}{r!} \cdot \int u^r K_{(r)}(u) \, du. \end{aligned}$$

Following identical steps as in Section 4, the optimal local bandwidth is given by

$$b_{opt(r)}(x) = \left(\frac{C_1}{2rC_r^2}\right)^{1/(2r+1)} N^{-1/(2r+1)} = const \cdot N^{-1/(2r+1)}. \tag{37}$$

Note that the optimal local bandwidth for the standard kernel estimator is obtained from EQ(37) by setting $r = 2$. When we assume that fourth order derivatives of f exist, then we can choose $r = 4$ and obtain that the optimal local bandwidth takes the form $const \cdot N^{-1/9}$. So this bandwidth converges slower to zero than that for the standard kernel density estimator. The intuition is that the since the bias is smaller, the marginal cost of increasing the bandwidth is smaller than before, but the marginal benefit (decreasing the variance) remains the same. The smoother f, the more information we can borrow from other parts of the distribution (e.g. uniform).

5.2 Multiplicative Bias Correction

An attractive method to reduce the smoothing bias is multiplicative bias reduction. Jones, Linton and Nielsen (1995) (JLN) propose to use the following nonparametric density estimator:

$$\tilde{f}(x) = \frac{1}{Nb} \sum_{i=1}^{N} \frac{\hat{f}(x)}{\hat{f}(X_i)} K\left(\frac{X_i - x}{b}\right), \tag{38}$$

where $\hat{f}(x)$ denotes the standard kernel density estimator. This procedure can be viewed as multiplicative bias correction. Note that by definition,

$$f(x) = \hat{f}(x) \cdot \alpha(x) \tag{39}$$

where

$$\alpha(x) = \frac{f(x)}{\hat{f}(x)}.$$

A natural estimator of $\alpha(x)$ is given by

$$\hat{\alpha}(x) = \frac{1}{Nb} \sum_{i=1}^{N} \frac{K\left(\frac{X_i - x}{b}\right)}{\hat{f}(X_i)}. \tag{40}$$

Combining EQ(40) and EQ(39) yields the density estimator in EQ(38) . It can be shown that this estimator has a leading term in the bias of the form $const \cdot b^4$, the $const \cdot b^2$ term is vanishing similarly as for higher order kernel methods. This happens because in EQ(38) , the denominator cancels out some of the bias which is contained in the numerator. JLN show that the appealing theoretical properties of their estimator also transfers to finite samples. Also, unlike higher order kernel density estimators, the JLN estimator is always positive.

6 Nonparametric Regression

In this section we turn to nonparametric regression techniques. After a brief discussion about multivariate density estimation, we introduce the Nadaraya-Watson estimator. We then go on to discuss local polynomial estimators.

6.1 Multivariate Kernel Density Estimation

Till now we considered in this text univariate density estimation. Recall from the introduction of this text that an estimator of the conditional expectation function is given by

$$\hat{g}(x) = \frac{1}{\hat{f}_X(x)} \int_{-\infty}^{+\infty} y \hat{f}_{X,Y}(x,y)\, dy. \tag{41}$$

We assume throughout that both Y and X are one-dimensional. We already know how to estimate the marginal density of X, but we do not know the corresponding estimator for the bivariate joint density $f_{X,Y}(x,y)$. We concentrate here on bivariate density estimators based on product kernel functions[20].

Definition 7 *The bivariate kernel density estimator at point (x,y) is defined by*

$$\hat{f}_{X,Y}(y,x) = \frac{1}{Nb^2} \sum_{i=1}^{N} K_2\left(\frac{Y_i - y}{b}, \frac{X_i - x}{b}\right). \tag{42}$$

K_2 is a product kernel defined as

$$K_2(u,v) = K(u) K(v),$$

where $K(u)$ is a univariate pdf.

[20] Multivariate kernel density estimators can be defined using any multivariate probability density function.

Note that the estimator for the marginal density of X can be obtained from EQ(42) by integrating over y.

$$\begin{aligned}
\hat{f}(x) &= \int \hat{f}_{X,Y}(y,x)\,dy \\
&= \int \frac{1}{Nb^2}\sum_{i=1}^{N} K_2\left(\frac{Y_i-y}{b},\frac{X_i-x}{b}\right)dy \\
&= \frac{1}{Nb^2}\sum_{i=1}^{N}\int K\left(\frac{Y_i-y}{b}\right)K\left(\frac{X_i-x}{b}\right)dy \\
&= \frac{1}{Nb^2}\sum_{i=1}^{N} K\left(\frac{X_i-x}{b}\right)\int K(u)\,b du \\
&= \frac{1}{Nb}\sum_{i=1}^{N} K\left(\frac{X_i-x}{b}\right),
\end{aligned}$$

which is the standard kernel density estimator. Almost all results stated in Section 3 go through also in the bivariate case by replacing Nb with Nb^2. Especially the rate of convergence, or stabilizing factor, is now $\sqrt{Nb^2}$ which is slower than the univariate convergence rate $\sqrt{Nb}$. This is the curse of dimensionality discussed in Remark 3.

It is in most cases not appropriate to choose a single bandwidth parameter for both dimensions. This problem can be solved by transforming the variables such that they have equal variance[21].

6.2 The Nadaraya-Watson Estimator

Knowing how to estimate bivariate densities, we can now derive a nonparametric estimator for the regression function $g(x)$. Using EQ(41) and substituting the nonparametric estimators for the joint and marginal density we obtain

$$\hat{g}(x) = \frac{1}{\frac{1}{Nb}\sum_{i=1}^{N} K\left(\frac{X_i-x}{b}\right)}\int y\frac{1}{Nb^2}\sum_{i=1}^{N} K\left(\frac{Y_i-y}{b}\right)K\left(\frac{X_i-x}{b}\right)dy.$$

[21] One often goes one step further and chooses the transformation such that the covariance between the x and y is eliminated as well. One then estimates the density of the transformed data and finally transforms back to obtain the density estimate of the untransformed data.

The integral term can be written as[22]

$$\begin{aligned}
&\frac{1}{Nb^2}\sum_{i=1}^{N} K\left(\frac{X_i - x}{b}\right)\int yK\left(\frac{y - Y_i}{b}\right)dy \\
= &\frac{1}{Nb^2}\sum_{i=1}^{N} K\left(\frac{X_i - x}{b}\right)\int (ub + Y_i)\,K(u)\,b\,du \\
= &\frac{1}{Nb}\sum_{i=1}^{N} K\left(\frac{X_i - x}{b}\right)\left[Y_i\int K(u)\,du + \int ubK(u)\,du\right] \\
= &\frac{1}{Nb}\sum_{i=1}^{N} Y_i K\left(\frac{X_i - x}{b}\right).
\end{aligned}$$

Collecting terms, we have therefore derived the famous Nadaraya-Watson estimator given by

$$\hat{g}(x) = \frac{\frac{1}{Nb}\sum_{i=1}^{N} Y_i K\left(\frac{X_i - x}{b}\right)}{\frac{1}{Nb}\sum_{i=1}^{N} K\left(\frac{X_i - x}{b}\right)}.$$

Note that this estimator is a weighted average of the observations $\{Y_i\}_{i=1}^{N}$, where the weight w_i for observation Y_i is given by

$$w_i = \frac{K\left(\frac{X_i - x}{b}\right)}{\sum_{i=1}^{N} K\left(\frac{X_i - x}{b}\right)}.$$

These weights sum up to one and put a large weight on an observation Y_i if the corresponding regressor X_i is close to x.

The Nadaraya-Watson estimator can also be interpreted as the solution to the following minimization problem:

$$\hat{g}(x) = \arg\min_{g}\sum_{i=1}^{N}(Y_i - g)^2 K\left(\frac{X_i - x}{b}\right).$$

Note that for each x, this is just a weighted least squares criterion, where the only explanatory variable is a (local) constant. The Nadaraya-Watson estimator can therefore be obtained by running OLS on the following regression equation:

$$\begin{aligned}
Y_i\sqrt{K\left(\frac{X_i - x}{b}\right)} &= g\sqrt{K\left(\frac{X_i - x}{b}\right)} + \sqrt{K\left(\frac{X_i - x}{b}\right)}e_i \\
Y_i^* &= gK_i^* + e_i^*
\end{aligned}$$

This way of reasoning will become especially useful when we deal with local polynomial estimators.

6.3 Properties of the Nadaraya-Watson Estimator

Bias and variance results for nonparametric regression estimators are more complex to derive than for nonparametric density estimators. We restrict ourselves therefore to the discussion of the main results.

[22] Use a change of variable: $u = \frac{y - Y_i}{b}$. The last line follows since the kernel integrates to one and is symmetric.

Proposition 4 *Under suitable assumptions and abstracting from higher order terms, bias and variance of the Nadaraya-Watson estimator are given by*

$$\begin{aligned} \text{Bias}\left(\hat{g}\left(x\right)\right) &= \frac{b^2}{2}\left[g^{(2)}\left(x\right)+2\frac{f^{(1)}\left(x\right)}{f\left(x\right)}g^{(1)}\left(x\right)\right]\int u^2K\left(u\right)du, \quad (43)\\ \text{Var}\left(\hat{g}\left(x\right)\right) &= \frac{\sigma^2}{Nbf\left(x\right)}\int K\left(u\right)^2du. \quad (44)\end{aligned}$$

Note that this expressions are quite similar as the bias and variance expressions for the kernel density estimator in EQ(22) and EQ(23) . The bias of the regression estimator contains in addition to the second derivative of the regression function also a term which depends on the slope of the regression function, the slope of the density and the density of X itself. This feature is different than in the density estimation case. We will see soon how to get rid of this term.

The variance of the Nadaraya-Watson estimator is large in areas of small $f\left(x\right)$. This means that in areas where there are few observations, it is difficult to estimate the regression function, which is very intuitive. Also the variance depends on the variance of the disturbance term. The higher the noise in the model, the more difficult is precise estimation, again this is very intuitive.

Also note from EQ(43) and EQ(44), that in terms of bandwidth choice, we have the same trade off as we had in density estimation. We do not repeat this discussion here, but turn to local polynomial estimators.

6.4 Local Polynomial Estimators

We saw in the previous section that the Nadaraya-Watson estimator can be interpreted as locally fitting a constant. We can extend this approach to local linear fitting. Note that by a Taylor expansion, we can write the standard regression model as

$$\begin{aligned} Y_i &= g\left(X_i\right)+\varepsilon_i\\ &\simeq g\left(x\right)+g^{(1)}\left(x\right)\left(X_i-x\right)+\varepsilon_i\\ &= \beta_0\left(x\right)+\beta_1\left(x\right)\left(X_i-x\right)+\varepsilon_i.\end{aligned}$$

Using this, we can form again a weighted least squares criterion function

$$Q\left(x\right)=\sum_{i=1}^{N}\left(Y_i-\beta_0\left(x\right)-\beta_1\left(x\right)\left(X_i-x\right)\right)^2K\left(\frac{X_i-x}{b}\right).$$

This can again be interpreted as running OLS on the regression equation

$$Y_i^*=\beta_0\left(x\right)K_i^*+\beta_1\left(x\right)K_i^*\left(X_i-x\right)+e_i^*,$$

where the variables are defined as above. The estimator of the regression function at point x is then given by

$$\hat{g}_{ll}(x) = \hat{\beta}_0(x)$$

It can be shown that the variance of the local linear estimator $\hat{g}_{ll}(x)$ matches the variance of the Nadaraya-Watson estimator. The bias however takes a different form:

$$Bias\left(\hat{g}_{ll}(x)\right) = \frac{b^2}{2} g^{(2)}(x) \int u^2 K(u)\, du,$$

the second term in brackets in EQ(43) has vanished. Why does this happen? Intuitively, locally fitting a line means that we can capture the local linear properties of the regression function, meaning that we can better estimate the first derivative of the regression function. This term therefore vanishes. Also note that if the true regression function is linear, the local linear estimator has no bias at all. This is because $g^{(2)}(x)$ equals zero in this case. The approach described above can be extended without problems to local polynomial fitting. However, there is a similiar feature there as in higher order kernel methods, it is questionable how well the nice theoretical properties transfer to finite samples. In practice, higher order fitting than quadratic will hardly bring any improvements.

References

[1] Hill, C.R., Griffiths W.E. and G.G. Judge (2001): "Undergraduate Econometrics", John Wiley & Sons, Inc.

[2] Hogg, R. and Craig, T. (1995): "Introduction to Mathematical Statistics", Prentice Hall.

[3] Johnston, J. and J. DiNardo (1997): "Econometric Methods", McGraw-Hill International Editions.

[4] Jones, M.C., Linton, O. and J.P. Nielsen (1995): "A simple bias reduction method for density estimation", *Biometrika,* 82, pp. 327-38.

[5] Pagan, A. and A. Ullah (1999): "Nonparametric Econometrics", Cambridge University Press.

[6] Wand, M.P and M.C. Jones (1995): "Kernel Smoothing", Chapman & Hall.