Kernel Density Estimation

Kaden Bieger and Raven McKnight

1 Introduction

Non-parametric statistics is a rapidly developing field which represents a large departure from the content covered in a traditional statistics course. Broadly speaking, non-parametric methods allow us to relax assumptions about our data. In a course such as mathematical statistics, we generally assume our data comes from a normal distribution, or at the very least from a distribution with mean μ and variance σ^2. Non-parametric methods do not rely on such parameters.

Kernel density estimation is a common technique within the sub-field of non-parametric statistics used to estimate probability density functions. In practice, we rarely know much at all about the true distribution of our sampled data. non-parametric methods such as kernel density estimation allow us to bypass assumptions of normality which are often unreasonable to make.

Kernel density estimation was developed separately by Parzen (1962) and Rosenblatt (1956), giving the method the name "Parzen-Rosenblatt window method" in related fields, such as econometrics. The method has practical applications in many disciplines, some of which we discuss in our conclusion.

In Section 2 of this paper, we build intuition about the purpose of kernel density estimation. Section 3 outlines major results regarding kernel density estimation, including the derivation of the expected value, variance, bias, and mean squared error. We end section 3 with a look towards the open question of bandwidth selection. Section 4 offers brief concluding remarks and introduces practical applications.

2 Intuition

Non-parametric density estimation may sound alien to audiences familiar with parametric statistical methods. However, density estimation is a common step in many exploratory data analyses. The **histogram** is perhaps the most well-known method for estimating probability density functions.

Given n observations, our goal is determine a likely probability distribution for our data. In the case of a histogram, we split our n observations into k "bins" or intervals of equal width. We denote the boundaries of these bins $b_0, b_1, \ldots, b_k$ such that bin i is defined $(b_{i-1}, b_i]$. Then we can estimate the density in bin i by the number of observations n_i that fall into the interval $(b_{i-1}, b_i]$. Therefore, the histogram estimate for data generating function $f(x)$ is given at point x is

$$\hat{f}(x) = \frac{1}{n} \sum_{i=0}^{k} \frac{1}{b_i - b_{i-1}} 1_{(b_{i-1}, b_i]}(x) \tag{1}$$

where $b_i - b_{i-1}$ is the width of one bin. This is a more formal way of writing what we intuitively know about histograms. While they are an excellent tool for learning about our data, they have several inherent flaws. First and foremost, histograms are dependent on the number of bins we select. Consider the example below (Figure 1). Even given a relatively normally distributed sample, the distribution of our data becomes less clear as we add additional bins. In practice, there is no accepted standard for selecting bins or bin width.

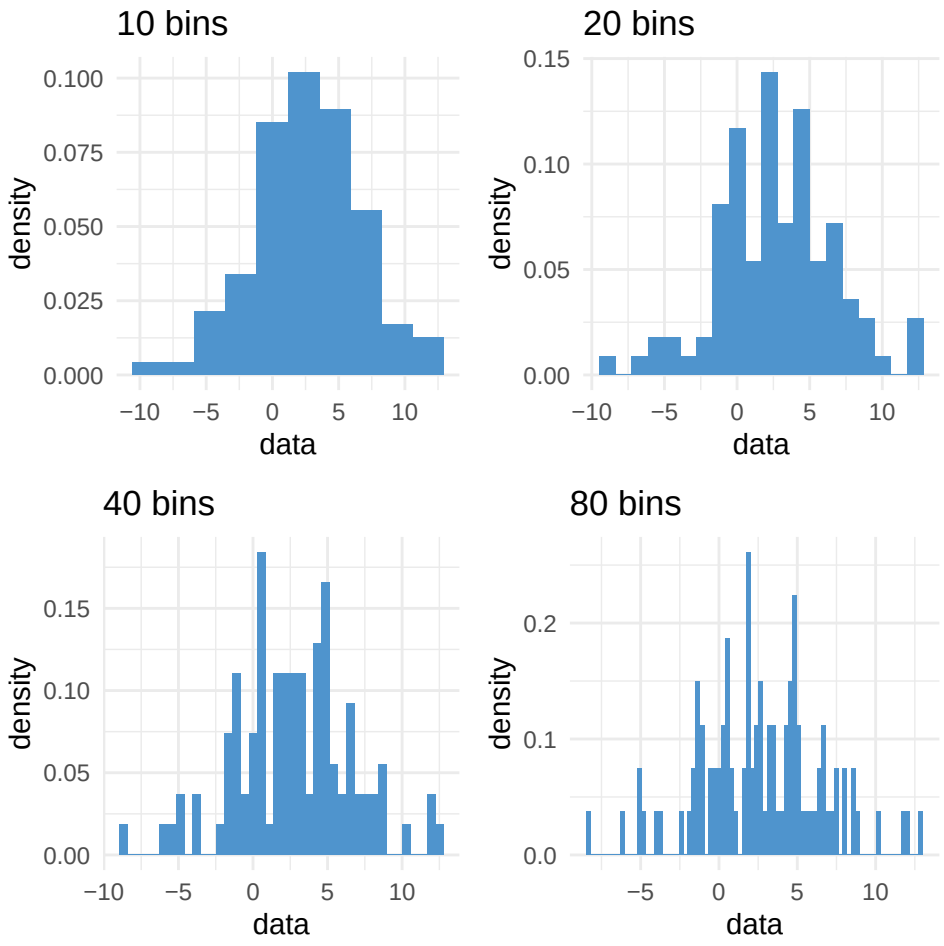

Figure 1: Example of histogram with various number of bins

In addition to being sensitive to bins, histograms are flawed because they are excessively *local* estimates. In other words, an observation of 9.999999 will not be considered in the bin (10, 11]. In Section 3, we formally describe kernel density estimation as a more rigorous method of density estimation.

3 Kernel Density Estimation

Kernel density estimation produces smooth, non-parametric density estimates using a weighting function called a **kernel**. We denote a kernel function as K. The procedure of kernel density estimation is fairly straight forward. First, a kernel is centered at each data point (Figure 2, left). In this example, the kernel is a standard normal curve. Second, each kernel is summed together (Figure 2, right).

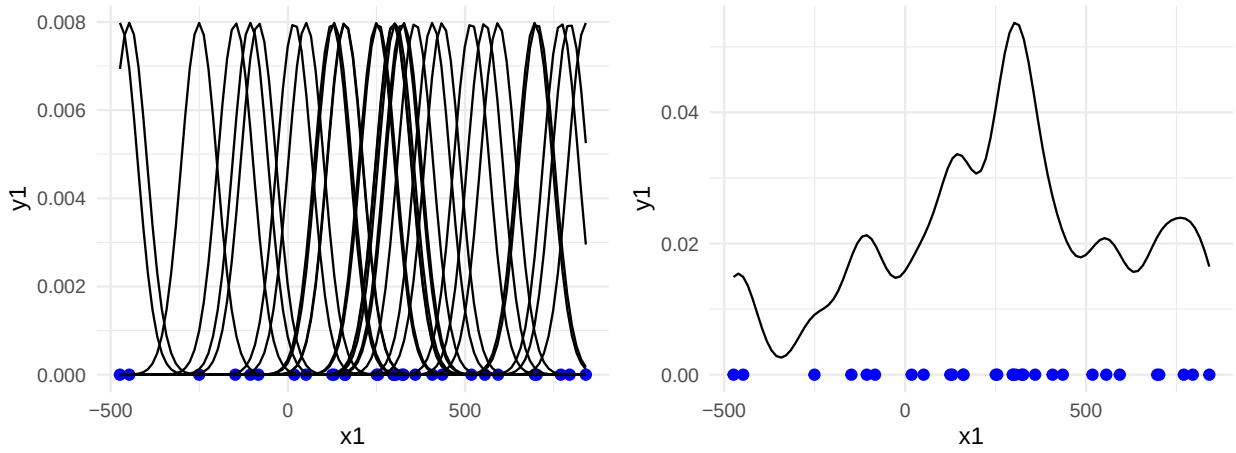

Figure 2: Steps 1 and 2 of KDE procedure

Finally, we divide the summed kernels by n to yield a density estimate which integrates to one (Figure 3). Formally, this process can be expressed as

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} K(x - x_i)$$

where $x - x_i$ centers the kernel function on each x_i. For example, using a standard normal kernel, we have

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}}\exp(\frac{1}{2}(x - x_i)) = \frac{1}{\sqrt{2\pi}n}\sum_{i=1}^{n} \exp(\frac{1}{2}(x - x_i))$$

There are many choices of kernel function other than normal (Gaussian). Kernel functions and the underlying densities they estimate f must have three main properties:

1. K is symmetric about 0 and integrates over its support to 1.
2. $\int xK(x)dx < \infty$ and $\int |x|(K(x))^2 dx < \infty$.
3. The probability density function $f : \mathbb{R} \to \mathbb{R}$ must be *Lipschitz Continuous*. Intuitively, Lipschitz continuity limits the absolute value of the slope of a function between any two points. In other words, a Lipschitz continuous function cannot change "too fast." In notation, this can be written $\exists M \in \mathbb{R}, |f(x) - f(y)| \leq M|x - y|, \forall x, y \in \mathbb{R}$. In the context of kernel density estimation, this essentially means that we cannot estimate functions with excessive curvature.

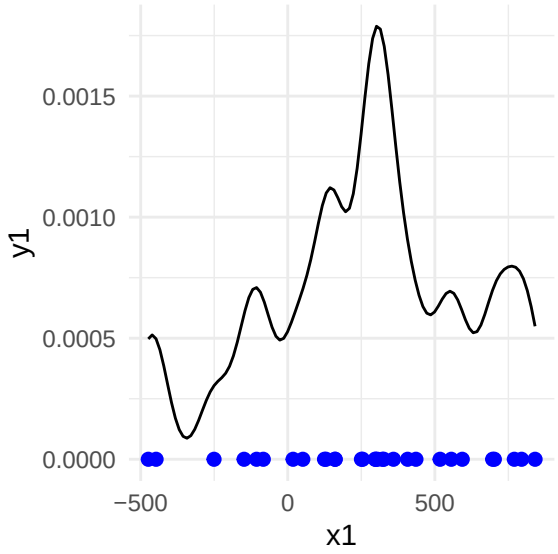

Figure 3: Step 3 of KDE procedure

Several other familiar probability density functions meet these requirements, including the uniform (rectangular or box) PDF or kernel. Some of the most common kernel functions are the Gaussian, Epanechnikov, Triangular, and Rectangular. Figure 4 illustrates kernel density estimates for the same observations using these four common kernel types. It can be shown that the Epanechnikov kernel is Mean Squared Error (MSE) optimal, although the Gaussian kernel is often used for simplicity and familiarity's sake. The Gaussian kernel is the default in many software applications, including R's `density` function.

3.1 Kernel Denisity Estimation and Bandwidth

The *bandwidth* of a kernel density estimate is denoted h, and determines the amount of smoothing, similar to the bin width of a histogram. Bandwidth h is completes the formula above to give our formal definition of a kernel density estimate as follows:

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

The entire expression is divided by h to ensure it integrates to one.

Notice that the larger h, the smaller the value inside the kernel function, making the effect of the distance between x and x_i smaller. Note also that in kernel density estimation the closer x is to an observation x_i (or multiple x_is), the higher the curve. Thus, larger h makes each kernel wider, resulting in a smoother, flatter curve.

We can see this in the following graph (Figure 5), where purple represents $h = 0.5$, blue represents $h = 1$, green represents $h = 2$, yellow represents $h = 4$, and red represents $h = 6$:

We discuss optimization of h in Section 3.3.

Given a kernel function, kernel density estimation works similarly to the histogram described in Section 2. Given an observation x, we center the kernel function at x and count the observations within a selected bandwidth h centered at point x. The interval surrounding x is written $[x-h/2, x+h/2]$. With observations $x = (x_1, x_2, \ldots, x_n)$, the kernel density estimate at point x is defined as follows.

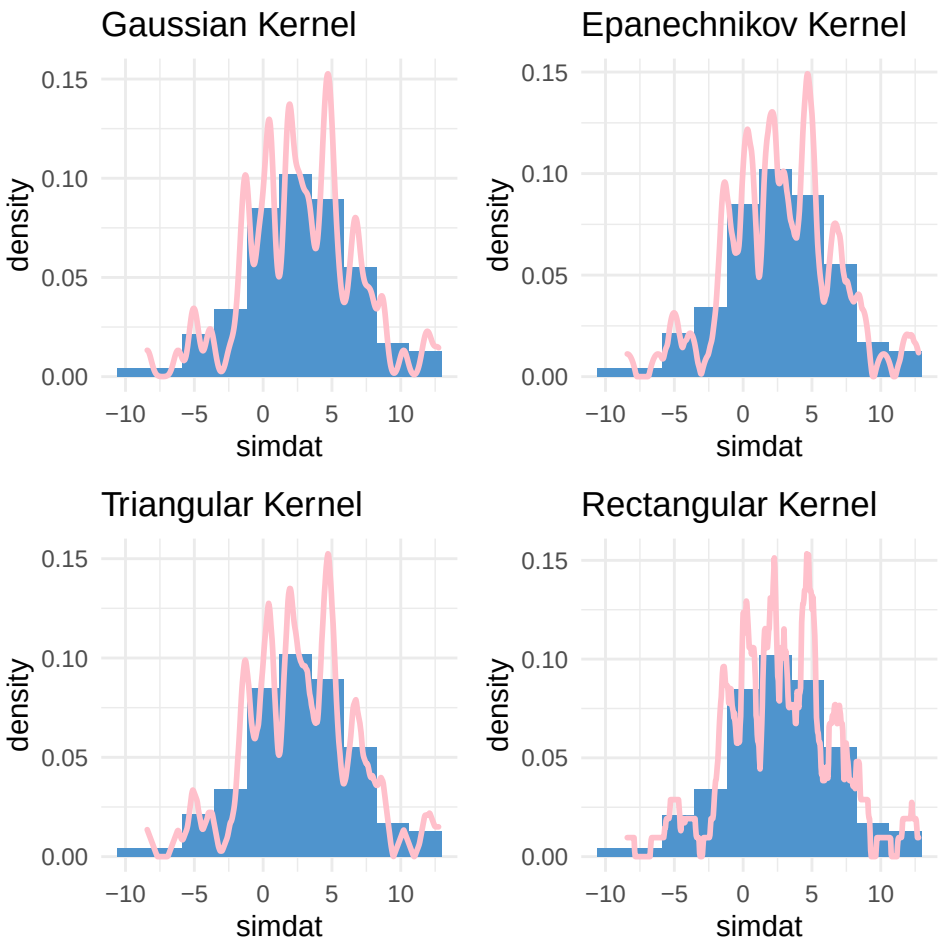

Figure 4: Example of various kernel functions

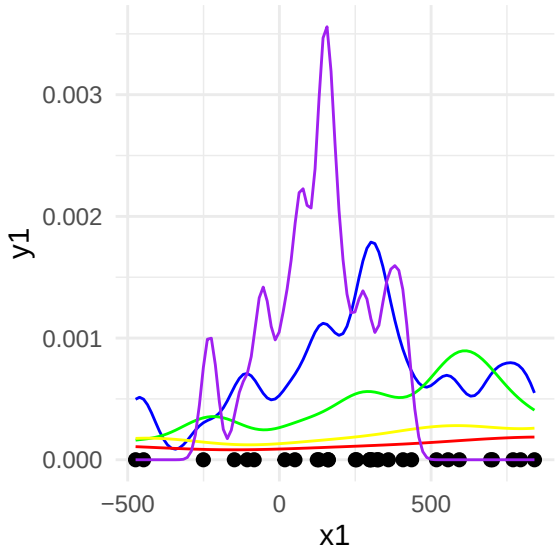

Figure 5: Kernels with various bandwidths

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K_{[x-h/2,x+h/2]}(x) = \frac{1}{nh}\sum_{i=1}^{n}K_{[-h/2,h/2]}(x-x_i)$$
$$= \frac{1}{nh}\sum_{i=1}^{n}K_{[-1/2,1/2]}\left(\frac{x-x_i}{h}\right) \tag{2}$$

In this notation, $\left(\frac{x-x_i}{h}\right)$ is essentially defining a "distance" away from our point of interest x. This term represents the weighting of each point x_i such that points closer to x are weighted more heavily than those further away. In this way, kernel density estimation is similar to a moving average. Instead of fixed bins, our bins "move" with x and amend bounding issues such than an observation of 9.999999 can be appropriately considered in the interval (10, 11]. For simplicity's sake, we often write kernel density estimators as follows, omitting the subscript on K.

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right) \tag{3}$$

3.2 Expected value, bias, variance, and MSE

Naturally, we would like to consider the bias and mean squared error of kernel density estimators. In this section, we outline proofs deriving the expected value, bias, and mean squared error for kernel density estimates. We follow a simplified version of proofs presented in Wasserman (2006). The proof relies on the three properties of kernels described above.

Given observations $x_1, x_2, \ldots, x_n \overset{iid}{\sim} f$, we define the expected value of our estimator as follows:

$$\mathbf{E}[\hat{f}(x)] = \mathbf{E}\left[\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right)\right] = \frac{1}{h}\mathbf{E}\left[K\left(\frac{x-X}{h}\right)\right] = \frac{1}{h}\int K\left(\frac{x-X}{h}\right)f(x)dx$$

This first line follows simply from the definitions of kernel density estimators and expected value. Here, we can use a change of variables and let $w = X$ such that

$$\mathbf{E}[\hat{f}(x)] = \frac{1}{h}\int K\left(\frac{x-w}{h}\right)f(w)dw$$

Next, we let $u = \frac{x-w}{h}$ and substitute. note we can rearrange $u = \frac{x-w}{h}$ to get $w = x - hu$.

$$\mathbf{E}[\hat{f}(x)] = \frac{1}{h}\int K\left(u\right)f(x-hu)du$$

Then, we apply the 2nd order Taylor expansion for $f(x-hu)$ about $h = 0$. Recall that a Taylor expansion is an infinite sum of terms approximating a function at a point. A Taylor series at point a can be written

$$\hat{f}(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3\ldots$$

and so on. Omitting several steps of algebra, our Taylor expansion can be written

$$f(x-hu) = f(x) + \frac{f'(x)}{1!}(u)(h-0) + \frac{f''(x)}{2!}(u^2)(h-0)^2 + o(h^2)$$
$$= f(x) - huf'(x) + \frac{h^2u^2}{2}f''(x) + o(h^2)$$

where $o(h^2)$ is some function which approaches zero *more quickly* than h^2 as h^2 approaches 0. Technically, this little-o notation indications that $o(h^2)$ becomes negligible compared to h^2 as $h \to 0$. We plug the Taylor expansion into our expected value above and simplify via algebra. Because $K(u)$ integrates to 1 over its support and has mean 0, this expression simplifies nicely.

$$\begin{aligned}
\mathbf{E}[\hat{f}(x)] &= \int K(u)\left[f(x) - huf'(x) + \frac{h^2u^2}{2}f''(x) + o(h^2)\right]du \\
&= f(x)\int K(u)du + hf'(x)\int uK(u)du + \frac{h^2}{2}f''(x)\int u^2K(u)du + o(h^2) \\
&= f(x) + \frac{h^2}{2}f''(x)\int u^2K(u)du + o(h^2)
\end{aligned}$$

We can plug this into the definition of bias such that

$$\begin{aligned}
\mathbf{Bias}(\hat{f}(x)) &= E[\hat{f}(x)] - f(x) \\
&= \frac{h^2}{2}f''(x)\int u^2K(u)du + o(h^2) \\
&= \frac{t \cdot h^2}{2}f''(x) + o(h^2)
\end{aligned}$$

where $t = \int u^2K(u)du$. Using this t notation is not necessary but it does simplify notation when writing proofs for mean squared error below. Given that our bandwidth h is in the numerator of this bias, we can see that we have less bias with lower h. This is relatively intuitive: the lower our bandwidth, the more our estimates rely on the data itself.

In addition to bias, we want to measure the mean squared error of our estimates. To do so, we must first derive the variance of $\hat{f}_n(s)$. We will walk through a derivation for the upper bound of the variance of our estimate. This derivation uses similar tools to the derivation of bias, including change of variables, u substitution and Taylor expansions.

First, we plug the definition of a kernel density estimator into the definition of variance.

$$\begin{aligned}
\mathbf{Var}(\hat{f}(x)) &= \mathbf{Var}\left(\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right)\right) \\
&= \frac{1}{nh^2}\left(\mathbf{E}\left[K^2\left(\frac{x-x_i}{h}\right)\right] - \mathbf{E}\left[K\left(\frac{x-x_i}{h}\right)\right]^2\right) \\
&\leq \frac{1}{nh^2}\mathbf{E}\left[K^2\left(\frac{x-X}{h}\right)\right] \\
&= \frac{1}{nh^2}\int K^2\left(\frac{x-w}{h}\right)f(w)dw
\end{aligned}$$

As above, we substitute $u = \frac{x-t}{h}$ and plug in a 1st order Taylor expansion of $f(x-hu)$.

$$
\begin{aligned}
\mathbf{Var}(\hat{f}(x)) &\leq \frac{1}{nh^2}\int K^2(u)f(x-hu)hdu \\
&= \frac{1}{nh}\int K^2(u)f(x-hu)du \\
&= \frac{1}{nh}\int K^2(u)[f(x)+huf'(x)+o(h)]du \\
&= \frac{1}{nh}\left(f(x)\int K^2(u)du + hf'(x)\int uK^2(u)du + o(h)\right) \\
\mathbf{Var}(\hat{f}(x)) &\leq \frac{f(x)}{nh}\int K^2(u)du + o\left(\frac{1}{nh}\right) \\
&= \frac{z}{nh}f(x) + o\left(\frac{1}{nh}\right)
\end{aligned}
$$

where $z = \int K^2(u)du$. Based on this definition, we can see that variance decreases as either sample size n of bandwidth h increase.

With our derivations of bias and variance complete, we can find the mean squared error of our estimator. This step follows more simply after the previous two proofs. We can simply plug our results from the proofs above into the definition of mean squared error. Then we can solve for MSE as follows:

$$
\begin{aligned}
\mathbf{MSE}(\hat{f}(x)) &= Bias^2(\hat{f}(x)) + Var(\hat{f}(x)) \\
&= \left(\frac{th^2}{2}f''(x) + o(h^2)\right)^2 + \frac{z}{nh}f(x) + o\left(\frac{1}{nh}\right) \\
&= \frac{t^2h^4}{4}\left[f''(x)\right]^2 + \frac{z}{nh}f(x) + \frac{th^2}{2}f''(x)o(h^2) + o(h^4) + o\left(\frac{1}{nh}\right)
\end{aligned}
$$

where $t = \int u^2K(u)du$ and $z = \int K^2(u)du$. Recall that the function $o(h)$ goes to zero as h goes to infinity. Therefore, the asymptotic mean squared error (AMSE) is simply the first two terms of the statement above. For simplicity, we often optimize the asymptotic mean squared error rather than the standard MSE. We discuss AMSE and optimal bandwidth below.

3.3 Bandwidth Selection

The previous derivations have prepared us to discuss the matter of bandwidth selection. As we see above, our estimate, bias, and MSE are largely dependent upon the size of our bandwidth h. Naturally, then, selection of an appropriate bandwidth is a major open question in the field of kernel density estimation.

Bandwidth selection is the most active field of kernel density estimation, and it remains an open question how to choose the best value of h.Answering this question is beyond the scope of this paper, however we introduce some tools often used for optimizing h. The first is Asymptotic Mean Squared Error, or AMSE. First, we can define AMSE

$$
\mathbf{AMSE}\hat{f}(x) = \frac{t^2h^4}{4}\left[f''(x)\right]^2 + \frac{z}{nh}f(x)
$$

which follows simply from the definition of MSE above. Next, we can optimize AMSE at any point x in terms of h. This optimization is a matter of taking a partial derivative of AMSE and setting equal to zero:

$$\frac{d}{dh}\Big[AMSE\Big] = 0$$
$$\frac{d}{dh}\left[\frac{t^2h^4}{4}\big[f''(x)\big]^2 + \frac{z}{nh}f(x)\right] = 0$$
$$\frac{4h^3t^2}{4}\big[f''(x)\big]^2 - \frac{z}{nh^2}f(x) = 0$$
$$h^3t^2\big[f''(x)\big]^2 = \frac{z}{nh^2}f(x)$$
$$h^5 = \frac{zf(x)}{nt^2\big[f''(x)\big]^2}$$
$$h = \left[\frac{zf(x)}{nt^2\big[f''(x)\big]^2}\right]^{1/5}$$

We must also confirm that the root is a local minimum as follows:

$$\frac{d}{dh}\left[\frac{d}{dh}AMSE\right]$$
$$= \frac{d}{dh}\left[h^3t^2\big[f''(x)\big]^2 - \frac{z}{nh^2}f(x)\right]$$
$$= 3h^2t^2\big[f(x)\big]^2 + \frac{2z}{nh^3}f(x)$$

The first term must be greater than zero because everything in it is squared except for 3, which is greater than zero. The second term must also be greater than zero because h, n, $f(x)$ (a pdf) and $z = \int K^2(u)du$ all must be greater than zero. Thus, $h = \left[\frac{zf(s)}{nt^2\left[f''(s)\right]^2}\right]^{1/5}$ minimizes AMSE.

Notice that this optimization depends on x, which doesn't allow us to simply insert this h into the equation we've been working with, $f(x) = \frac{1}{Nh}\sum_{i=1}^{N} K(\frac{x-x_i}{h})$.

This procedure is fairly simple for defining h at any given point. However, it is slightly more complicated to optimize h over an entire sample. To do so, we often consider Asymptotic Mean *Integrated* Squared error, or AMISE. As the name suggests, we define AMISE as the integral of MSE:

$$\textbf{AMISE}(\hat{f}(x)) = \int \left(\frac{t^2h^4}{4}\left[f''(x)\right]^2 + \frac{z}{nh}f(x)\right)dx$$

It is possible to optimize AMISE in terms of h to get the optimal bandwidth across an entire sample, although that is beyond the scope of this paper. Optimizing AMISE in terms of h shows that the optimal h is dependent on $\int \left[f''(x)\right]^2 dx$, or the curvature of the underlying probability density function. Most of the ongoing researching in kernel density estimation is focused on estimated AMISE, the underlying curvature, and the optimal h.

4 Conclusion

Kernel density estimation has many practical applications in applied research. For example, using the solution to the heat equation as a kernel function, and locations of heat x_i, the kernel density estimate represents the amount of heat generated. More generally, kernel density estimation is useful in any applied setting where the data generating function is unknown (this is almost always!). Multidimensional kernel

density estimation can also be used with two or more variables of by choosing a three or more dimensional kernel and centering it on each data point.

Some limitations of this technique lie in the selection of h. Additionally, kernel density estimators struggle to make reasonable estimates when data is multimodal or has natural bounds (ie age or height which must necessarily be positive). Some practitioners also argue that kernel density estimation obscures the sample too much, and this can be true: it is critical to remember that this procedure yields *estimates*, not truth.

Despite potential challenges, kernel density estimation is becoming an increasingly popular way to avoid the pitfalls of assuming a distribution. The question of optimizing bandwidth is an active field of research which will likely see advances in coming years.

5 References

Benedetti, Jacqueline. 1977. "On the Non-parametric Estimation of Regression Functions." Journal of the Royal Statistical Society 39 (2): 248-253.

Gasser, T; Muller, H-G; & Mammitzsch, V. 1985. "Kernels for Non-parametric Curve Estimation" Journal of the Royal Statistical Society 47 (2): 238-252

Parzen, Emanuel. 1962. "On Estimation of a Probability Density Function and Mode." The Annals of Mathematical Statistics 33 (3): 1065–76.

Rosenblatt, Murray. 1956. "Remarks on Some Non-parametric Estimates of a Density Function." The Annals of Mathematical Statistics, 832–37.

Wasserman, Larry. 2006. All of Non-parametric Statistics. Springer Science & Business Media.