Kernel Density Estimation

Kaden Bieger and Raven McKnight

1 Introduction

nonparametric statistics is a rapidly developing field which represents a large departure from the content covered in a traditional statistics course. Broadly speaking, nonparametric 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. nonparametric methods are not based on such parameters.

Kernel density estimation is a common technique within the subfield of nonparametric statistics used to estimate probability density functions. In practice, we rarely know much at all about the true distribution of our sampled data. nonparametric methods such as kernel density estimation allow us to bypass assumptions of normality and so on that may be unreasonable to make.

Kernel density estimation was developed seperately by Parzen (1962) and Rosenblatt (1956), giving the method the name "Parsen-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 for the function of kernel density estimation. Section 3 outlines major results regarding kernel density estimation, including the derivation of the expected value, variance, bias, and MSE. We end section 3 with a look towards the open question of bandwidth selection. Section four offers brief concluding remarks and summarizes practical applications.

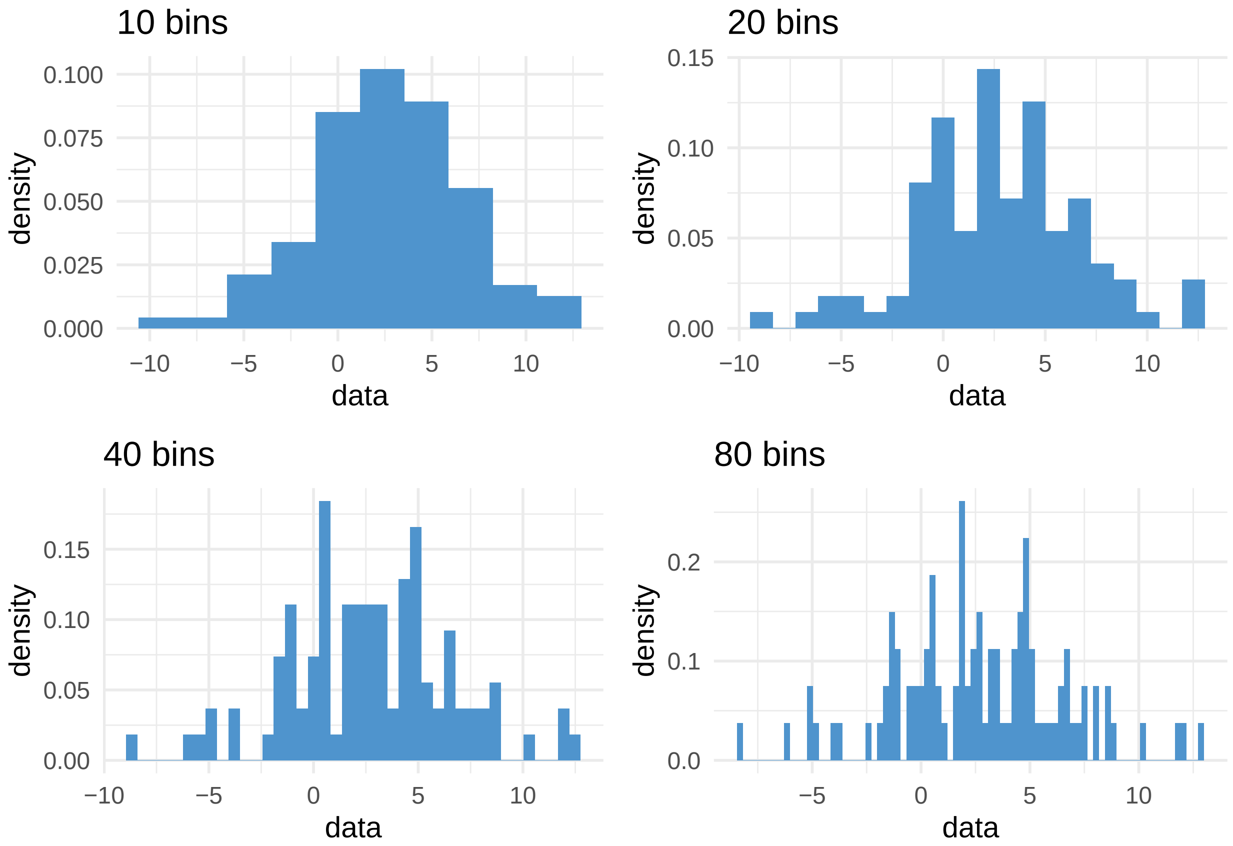
2 Intuition

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" as $b_0, b_1, \ldots, b_k$ such that bin i is defined $(b_{i-1}, b_i]$. Then we can estimate the density in bin i by the proportion 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

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

This is a more formal way of writing what we intuitvely know about histograms. While they are an excellent tool for learning about our data, they have several inherent flaws. First and foremost, histograms are incredibly 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 binwidth.



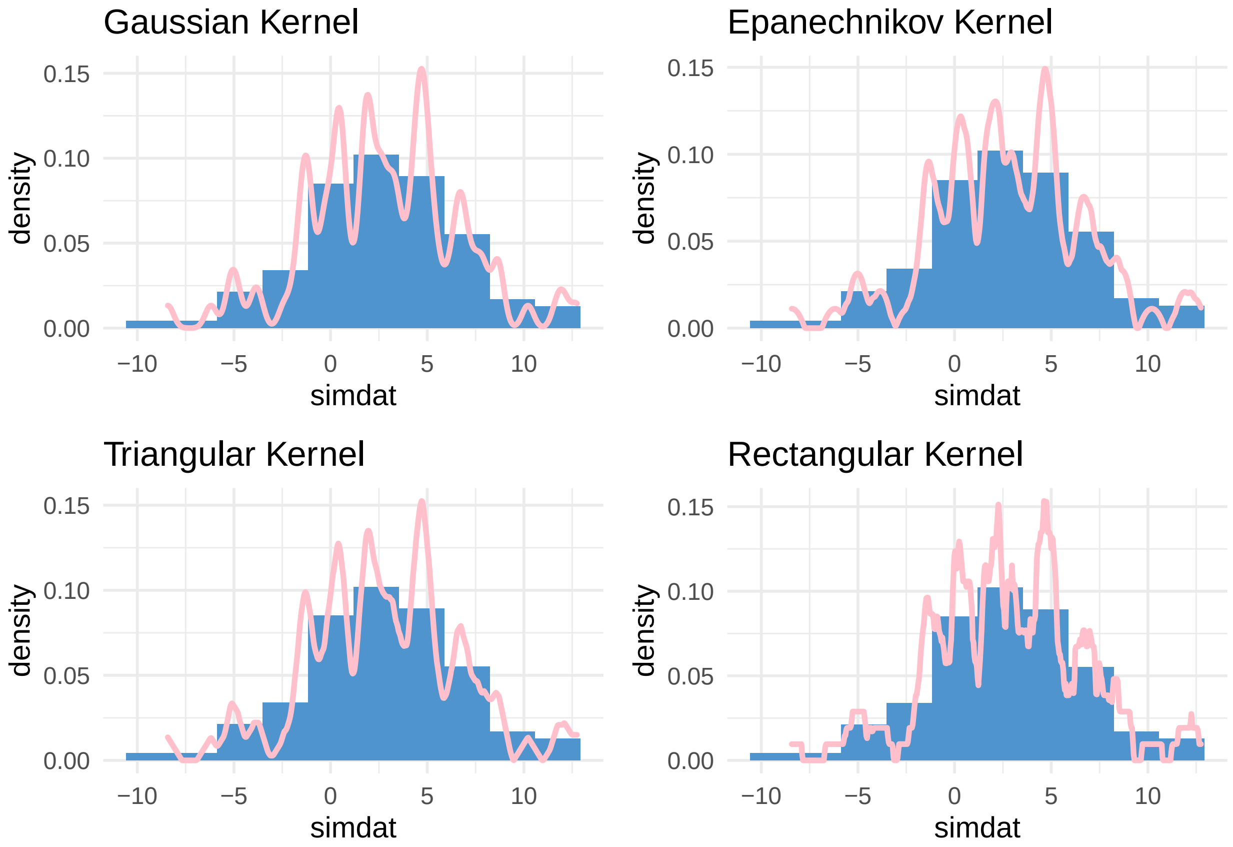
In addition to being sensitive to bins, histograms are flawed because they are excessively *local* estimates. In other words, an observation of 9.999999 will not be considered in the bin (10, 11]. In Section 3, we formally describe kernel density estimation as a more rigorous method of density estimation.

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. Kernel functions and the underlying densities they estimate f have three main properties:

1. K is symmetric about 0 and integrates over its support to 1.
2. $\int xK(x)dx < \infty$ and $\int |x|(K(x))^2dx < \infty$.
3. The probability density function $f : \mathbb{R} \to \mathbb{R}$ must be *Lipschitz Continuous.* The full definition of Lipschitz continuity is beyond the scope of this paper. Intuitively, however, Lipschitz continuity limits the aboslute 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 $\exists M \in \mathbb{R}, |f(x) - f(y)| \leq M|x - y|, \forall x, y \in \mathbb{R}$.

Several familiar probability density functions meet these requirements, including the normal (Gaussian) and uniform (rectangular or box) PDF or kernel. There are many possible kernel functions, but some of the most common are the Gaussian, Epanechnikov, Triangular, and Rectangular. Figure 2 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 simplicty and familiarity’s sake.



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

$$\begin{aligned} \hat{f}_n(s) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}I_{[s-h/2,s+h/2]}(x) = \frac{1}{nh}\sum_{i=1}^{n}I_{[-h/2,h/2]}(x-s) \\ &= \frac{1}{nh}\sum_{i=1}^{n}I_{[-1/2,1/2]}\left(\frac{x-s}{h}\right) \end{aligned} \tag{2}$$

In this notation, $\left(\frac{x-s}{h}\right)$ is essentially defining a "distance" away from our point of interest s. This term represents the weighting of each point x such that points closer to s 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 s and ammend bounding issues such than an observation of 9.999999 can be appropriately considered in the interval (10, 11].

To acheive this weighting, or smoothing, we apply our kernel function K such that the estimate is written

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

for each s. As we implement kernel density estimation, each x_i will eventually be s. Thus, the above equation could be rewritten with s subsituted for x_i.

The amount of smoothing is determined by the magnitude of h, or our *bandwidth*. In Figure 2, the bandwidth is low enough that we can observe (for example) the normal curve around each observation. We discuss bandwidth further in Section 3.2.

3.1 Expected value, bias, variance, and MSE

- a note: I think we should narrate these proofs more, but I haven't found a nice way to do so yet

Naturally, we would like to consider the bias and mean squared error of estimators produced by kernel density estimation. 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 $x_1, x_2, \dots, x_n \overset{iid}{\sim} f$, we define the expected value of our estimator as follows:

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

This first line follows simply from the definition stated above. To solve further, we use u substitution and a 2nd order Taylor expansion. First, we let $u = \frac{x-s}{h}$ and substitute.

$$\mathbf{E}[\hat{f}_n(s)] = \frac{1}{h}\int K\left(u\right) f(hs+u)dx$$

next, we apply the 2nd order Taylor expansion for $f(hs+u)$ about $h=0$. Ommitting several steps of algebra, our Taylor expansion can be written

$$\begin{aligned} f(hu+s) &= f(s) + \frac{f'(s)}{1!}(u)(h-0) + \frac{f''(s)}{2!}(u^2)(h-0)^2 + o(h^2) \\ &= f(s) + huf'(s) + \frac{h^2u^2}{2}f''(s) + o(h^2) \end{aligned}$$

where $o(h^2)$ is some function which approaches zero as h approaches infinity. We plug the Taylor expansion into our expected value above and simplify via algebra.

$$\mathbf{E}[\hat{f}_n(s)] = \int K(u)\left[f(s) + huf'(s) + \frac{h^2u^2}{2}f''(s) + o(h^2)\right]du$$

$$\begin{aligned} &= f(s)\int K(u)du + hf'(s)\int uK(u)du + \frac{h^2}{2}f''(s)\int u^2K(u)du + o(h^2) \\ &= f(s) + \frac{h^2}{2}f''(s)\int u^2K(u)du + o(h^2) \end{aligned}$$

We can plug this into the definition of bias such that

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

where $t = \int u^2K(u)du$.

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 derivatio of bias, including u substitution and Taylor expansions.

First, we plug our estimator into the formula for variance. The following is possible because K is symmetric about zero.

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

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

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

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 defintion of mean squared error. Then we can solve for MSE as follows:

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

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 sauared 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.

3.2 Bandwidth

The previous derivations have prepared us to discuss the matter of bandwidth selection. As we see above, our bias and MSE is 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.

- most active field
- introduce optimization problem
- some more visualizations about how estimates change based on bandwidth
- I don't think this is going to take up the other 5 pages so likely will end up "narrating" the proofs a bit more?

4 conclusion

- probably with an eye towards practical applications

5 References

more to come

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 Nonparametric Estimates of a Density Function." The Annals of Mathematical Statistics, 832–37.

Wasserman, Larry. 2006. All of Nonparametric Statistics. Springer Science & Business Media.