Chapter 9

Non-Parametric Density Function Estimation

9.1. Introduction

We have discussed several estimation techniques: method of moments, maximum likelihood, and least squares estimation. In most cases we have adopted the priveleged position of supposing that we knew *a priori* what functional form is appropriate for describing the distribution associated with the random variable. The complete description of the random variable then merely requires the estimation of some parameters. However, as we have seen in our discussion of robust M-type estimation it is often the case that we must make inferences about the noise distribution from the data themselves: we must adopt a **non-parametric** approach. In the previous lecture we introduced the idea of using the residuals from a least squares fit to the observations to guide the choice of loss function that would be appropriate for maximum likelihood estimation. The iterative process allowed us to develop an efficient estimation scheme in the presence of non-Gaussian noise. Now we turn to a more general question: given an arbitrary collection of data how might we find the pdf associated with them? What follows is a survey of methods for density estimation. Such estimates are useful in the presentation and exploration of data: they can, for example, reveal skewness in distributions, or the presence of multiple modes in the data, and may provide the basis for important physical interpretations of the observations. Exploratory data analysis is discussed in Chapter 15 of Dekking *et al.*

9.2. Comparing Data and Theory: Density Estimates and Sample Distribution Functions

Consider, for example, the data in Figure 9-1: the length of stable polarity intervals between reversals of the geomagnetic field over the past 119 Myr a subset of the data first introduced in Figure 3 of Chapter 1. You might reasonably ask why we should treat these intervals as being a kind of random variable; they are not, as in the GPS case (Figure 2 of Chapter 1), repeated measurements of the "same thing", with error added. Our justification for treating the polarity intervals as a random variable is that the time between reversals is highly variable and apparently unpredictable; a probabilistic description seems like the only way to capture this behavior. The specific description often used is to take the lengths of polarity intervals as arising from a **Poisson process**. As we saw in Chapter 3 the Poisson process describes a probabilistic behavior: in it, the probability of a reversal within any given time interval is independent of how long it has been since the last

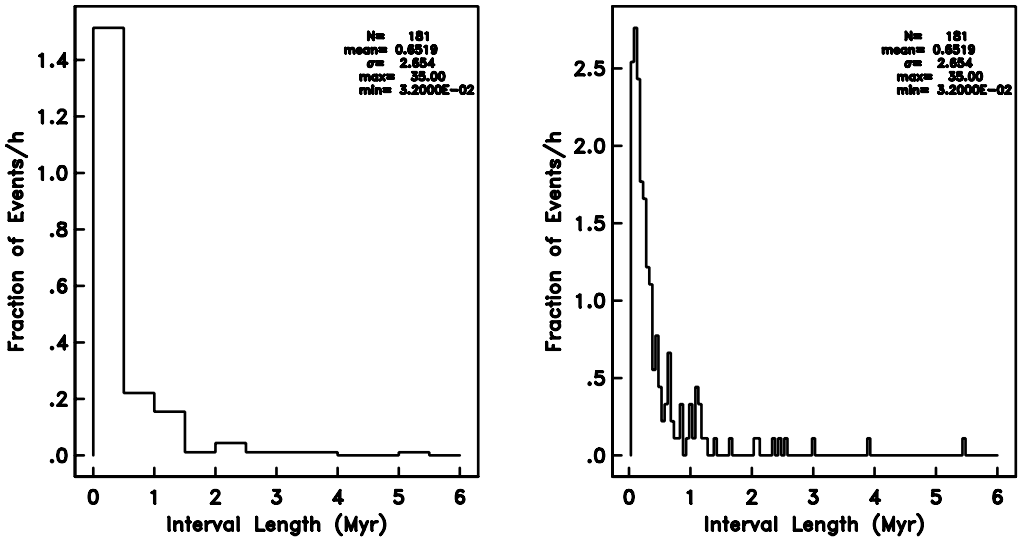

Figure 9 -1: Histograms of stable polarity interval lengths for 0-119Ma from Cande and Kent (1995) timescale. Bin sizes are 0.5 Myr (left) and 0.05 Myr (right). x-axis is truncated at 6 Myr, omitting the longest intervals.

reversal occurred. That is to say the timing of the next reversal cannot be determined *a priori*, although we have some idea of the average reversal rate. Consequently, the length of each polarity interval is a random variable, and likely to vary from one interval to the next. For the time being we ignore position in the reversal sequence and just treat the lengths of polarity intervals as a collection of numbers. The histogram shows no intervals shorter than .03 Myr, a concentration between 0 and about 1.5 Myr, and an occasional much longer interval (one is actually 35 Myr long, but the histogram is truncated at 6 Myr for plotting purposes).

The histogram seems like a natural means for actually measuring the probability of observing a particular interval length; (or range of lengths); we could compare a suitably normalized histogram with various theoretical probability density functions. To use a histogram to estimate a pdf, take an origin x_0 and a bin width h and define the bins of the histogram as the intervals $[x_0 + mh, x_0 + (m + 1)h]$ for positive and negative integers m. The histogram estimate of the pdf is then defined by

$$H_n(x) = \hat{\phi}(x) \; = \; \frac{1}{nh} \text{ (number of } x_i \text{ in the same bin as } x) \tag{1}$$

where we are using $\hat{\phi}$ to denote an estimate of ϕ. Histograms have the advantage of being simple. The main disadvantage is the discrete nature of the plot: the bin width is an intrinsic limit on resolution and the story may change depending on how we select width and boundaries of the bins. We can imagine making the answer more precise by decreasing the bin size, as in the right panel of Figure 9-1, but for a fixed number of observations this will ultimately result in many bins with no observations in them; the limiting result would be when one places a value at each observation and zeroes everywhere in between. It is not always easy to decide on an appropriate level of smoothing.

9.2:1 Choosing a Suitable Bin Size for Histograms

One way to choose the bin size for your histogram is to use the **normal reference method**. Suppose we decide to choose a bin width that minimizes the difference between $H_n(x)$ and the true pdf $\phi(x)$ using some appropriate measure. Let's use the **mean integrated square error or MISE**, defined as

$$MISE = \mathcal{E}\left[\int_{-\infty}^{\infty}[H_n(x) - \phi(x)]^2 dx\right] \tag{2}$$

For a smooth pdf ϕ and in $\lim n \to \infty$ it can be shown that the value of h needed in (1) is

$$h = C(\phi)n^{-1/3} \tag{3}$$

where

$$C(\phi) = 6^{1/3}\left[\int_{-\infty}^{\infty}[\phi'(x)]^2 dx\right]^{-1/3} \tag{4}$$

a result that clearly depends on the unknown pdf ϕ. We need to find $C(\phi)$ and the normal reference method just supposes that we can try $\phi \sim N(\mu, \sigma^2)$. This gives a simple data-based strategy for choosing the bin width h.

Exercise: Show that for the normal reference method $C(\phi) = (24\sqrt{\pi})^{1/3}\sigma$ and hence verify that $h = 24\sqrt{\pi})^{1/3}sn$.

9.3. Alternatives to the Histogram

There are many alternatives to the histogram for making density estimates*. One of the simpler ones is known as the **naive estimator**. It approximates the density function $\phi(x)$ for the random variable X, defined as

$$\phi(x) = \lim_{h\to 0}\frac{1}{2h}p[x - h < X < x + h] \tag{5}$$

by

$$\hat{\phi}(x) = \frac{1}{2nh}\sum_{i=1}^{n}\Pi\left(\frac{x - x_i}{2h}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}w\left(\frac{x - x_i}{h}\right) \tag{6}$$

where Π is the rectangle or boxcar function. That is, the estimate is constructed by placing a box of height $(2nh)^{-1}$ and width $2h$ on each observation and summing all the boxes. Unlike the histogram, this eliminates

* B. W. Silverman (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, is a very good introduction, at an intermediate level.

having to choose the box boundaries, but leaves freedom to control smoothness by the choice of h. The naive estimator suffers from being discontinuous, with jumps at $x_i \pm h$ and a derivative of zero everywhere else. This difficulty can be overcome by replacing the weight function $w(x)$ by a **kernel function** $K(x)$ with more agreeable properties. A **kernel density estimate** is given by

$$\hat{\phi}_K(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{7}$$

with $K(x)$ usually chosen as a symmetric probability density function satisfying the condition

$$\int_{-\infty}^{\infty} K(x)dx = 1; \qquad K(x) = K(-x) \tag{8}$$

Often $K(x)$ is selected so that $K(x) = 0$ for $|x| > 1$. Common examples are

Epanechnikov kernel

$$K(x) = \frac{3}{4}(1 - x^2) \qquad -1 \le x \le 1 \tag{9}$$

Triweight kernel

$$K(x) = \frac{35}{32}(1 - x^2)^3 \qquad -1 \le x \le 1 \tag{10}$$

Normal kernel

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \qquad -\infty \le x \le \infty \tag{11}$$

There are obvious advantages to kernel estimates. It is clear from the definition that provided $K(x)$ is a density estimate, $\hat{\phi}_K(x)$ will have the necessary properties for a density function. From a visual perspective the specific kernel used often seems to have only a very small impact on the resulting estimate, but it's worth remembering that $\hat{\phi}_K$ will inherit all the continuity and differentiability properties of K. The properties of the kernel can be used in considering the mathematical properties of the kernel estimator (such as bias, consistency, and efficiency). Figure 9-2(b) and (c) show two examples of kernel estimates for the inter-reversal time density estimate constructed using a Gaussian density as $K(x)$. You can see that again the amount of structure is determined by the **window width** h, in this case the standard deviation of the Gaussian. h is also known as the **smoothing parameter** or **bandwidth**. We shall encounter it again in dealing with data smoothing in a more general context.

9.3:1 Choosing the bandwidth h for $\hat{\phi}_k(x)$

Once again we can use the normal reference method in selecting the bandwidth h for kernel estimates. Suppose we decide on a kernel K and want to minimize *MISE* between $\hat{\phi}_{K,n}(x)$ and the true pdf $\phi(x)$

$$MISE = \mathcal{E}\left[\int_{-\infty}^{\infty} [\hat{\phi}_{K,n}(x) - \phi(x)]^2 dx\right] \tag{12}$$

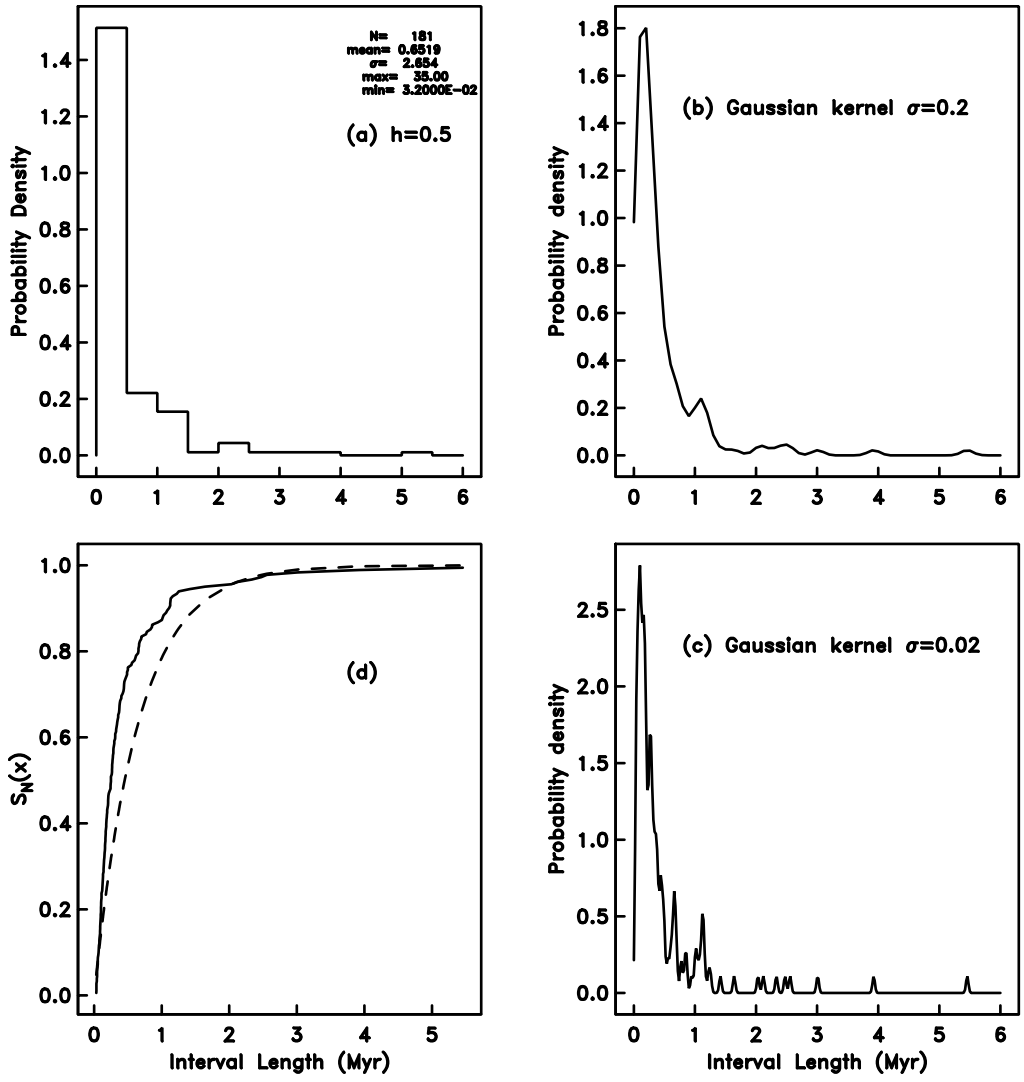

Figure 9-2: Sample distribution function and various probability density estimates for the data of Figure 9-1. Top left histogram with $h = 0.5$, top right Gaussian kernel estimate with $\sigma = 0.2$, bottom right Gaussian kernel estimate with $\sigma = .02$. Bottom left shows the sample distribution function (solid line), compared with the best-fitting exponential distribution (dashed line).

For a smooth pdf ϕ and in $\lim n \to \infty$ it can be shown that the value of h needed in (12) is

$$h = C_1(\phi)C_2(K)n^{-1/5} \tag{13}$$

where

$$C_1(\phi) = \left[\int_{-\infty}^{\infty} [\phi''(x)]^2 dx \right]^{-1/5} \tag{14}$$

and

$$C_2(K) = \frac{\left[\int_{-\infty}^{\infty}[K(x)]^2 dx\right]^{1/5}}{\left[\int_{-\infty}^{\infty}[x^2 K(x)]^2 dx\right]^{2/5}} \tag{15}$$

Again the result that depends on the unknown pdf ϕ and also on the kernel K. For a normal kernel we get $C_2(K) = (2\sqrt{\pi})^{-1/5}$ and for the normal reference method $C_1(\phi) = (8\sqrt{\pi}/3)^{1/5}\sigma$, yielding $h = (4/3)^{1/5} s n^{-1/5} = 1.06 s n^{-1/5}$.

Although kernel estimates are the most widely used density estimates they do suffer from some drawbacks, especially when applied to long-tailed distributions. As we see in Figure 9-2 the concentration of observations has an intrinsic variability with the value of $\phi(x)$: the naturally low probability of acquiring data in regions where ϕ is small results in a tendency for the density estimate to be noisy in the tails of the distribution. This effect can be mitigated by broadening the kernel (increasing h), but only at the expense of potential loss of resolution near the center of the distribution where the data are denser: the center of the distribution may appear too broad, and one runs the risk of missing details such as multiple modes that could reflect interesting physical properties inherent in the data.

Other considerations in using kernel estimates are that symmetry in the kernel might not always be desirable, for example when the data are bounded on one side, *e.g.*, always positive.

9.4. Sample Distribution Functions

Using any estimate of the probability density function as a comparison with parametric forms suffers from the difficulty that we lose information by binning or averaging in constructing the density. The information we actually have is just a sample of numbers $T_n = x_1, x_2, \ldots, x_n$ drawn from a distribution, not an actual function. If we want to compare our observations with some theoretical statistical model we can construct a kind of empirical distribution function for our sample (after sorting in ascending order)

$$S_n(x) = \frac{1}{n}[\text{number of } x_i \in T_n) < x] \tag{16}$$

so that

$$S_n(x_{(i)}) = \frac{i}{n} \tag{17}$$

This is a kind of "staircase" function with a jump at each sample value of x. Recall from Chapter 6, page10, that S_n is called the **sample distribution function**. Note that S_n has all the properties of a distribution function. As n gets larger and larger the law of large numbers (on which all of probability theory is

based) guarantees that $S_n(x) \rightarrow \Phi(x)$, the underlying distribution for the observations. The advantage of constructing $S_n(x)$ is that it provides a single valued function that we can compare with any theoretical distribution function, without having to choose bin sizes. We have already encountered S_n in the context of hypothesis testing in Chapter 5. We can think of $S_n(x_i)$, $i = 1, \ldots, n$, as a collection of observations that need to be adequately described by any statistical model we wish to adopt. In other words we should expect that $\hat{\Phi}(x_i) - S_n(x_i)$ should be small in some sense, if $\hat{\Phi}$ is a good statistical model. This was the basis of the Kolmogorov-Smirnov test - in which case the property we characterize as small is the maximum absolute deviation between the two curves. The sample distribution function for the data set considered in Figure 9-1 is given in Figure 9-2, also shown as the dashed line is the exponential distribution function, one (rather bad) candidate for a parametric distribution that might describe these observations. But clearly the fit of $S_n(x)$ to $\hat{\Phi}(x)$ is a criterion that might be used in selecting an appropriate estimate for $\hat{\phi}(x)$ (assuming that we can evaluate $\hat{\Phi}(x)$ directly by integrating $\hat{\phi}(x)$).

9.5. Adaptive Estimation: Nearest Neighbors and Variable Kernels

Data adaptive methods of density estimation have been developed in an attempt to address the problem outlined above. The resolution attainable depends on the number of observations available and this will not in general be uniform across the domain of $\phi(x)$. The general idea is to adapt the amount of smoothing in the estimate to the *local* density of data. Two methods are introduced here, known as the nearest neighbors and variable kernel techniques.

9.5:1 Nearest Neighbor Method

The basic idea of the nearest neighbor method is to control the degree of smoothing in the density estimate based on the size of a box required to contain a given number of observations (contrast this with the naive estimator which uses the number of observations falling in a box of fixed width centered at the point of interest). The size of this box is controlled using an integer k, that is considerably smaller than the sample size, a typical choice would be $k \approx n^{\frac{1}{2}}$. Suppose we have the *ordered* data sample $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. For any point x on the line we define the distance betwen x and the points on the sample by

$$d_i(x) = |x_i - x|$$

so that

$$d_1(x) \leq d_2(x) \leq d_3(x) \ldots d_n(x) \tag{18}$$

Then we define the kth **nearest neighbor density estimate** by

$$\hat{\phi}(x) = \frac{(k-1)}{2nd_k(x)} \tag{19}$$

This can be understood in terms of the number of observations one would expect in an interval $[x-r, x+r]$ with $r > 0$. We expect this to be about $2rn\phi(x)$. By definition we expect $(k-1)$ observations in $[x - d_k(x), x + d_k(x)]$, so we can estimate the density by noting that

$$k - 1 = 2d_k(x)n\hat{\phi}(x) \tag{20}$$

from which (18) can be recovered. Near the center of the distribution $d_k(x)$ will be smaller than in the tails, so we can expect the problem of undersmoothing in the tails to be reduced. Like its relative, the naive estimator, the nearest neighbor estimator is not smooth: $d_k(x)$ has a discontinuity in its derivative at every point x_i. Furthermore, although $\hat{\phi}(x)$ is positive and continuous everywhere it is not in fact a probability density. Outside $[x_{(1)}, x_{(n)}]$ we get $d_k(x) = x_{(k)} - x$ and $d_k(x) = x - x_{(n-k+1)}$ which make the tails of the $\hat{\phi}$ defined in (19) fall off like x^{-1}, that is extremely slowly: the integral of $\hat{\phi}(x)$ is infinite.

This can in principle be fixed by using a **generalized kth nearest neighbor estimate**

$$\hat{\phi}(x) = \frac{1}{nd_k(x)} \sum_{i=1}^{n} K\left(\frac{x - x_i}{d_k(x)}\right) \tag{21}.$$

In fact this is just a kernel estimate evaluated at x with window width $d_k(t)$. Overall smoothing is controlled by choice of k, with the window width at any specific point depending on the density of points surrounding it. Note, however, that the derivative of the generalized nearest neighbor estimate (GNNE) will be discontinuous at all points where $d_k(x)$ has a discontinuous derivative. In general the integrability, and behavior in the distribution tails will depend on the form of K.

9.5:2 Variable Kernel Method

The variable kernel estimate is constructed in a similar fashion to the classical kernel estimate, but the scale parameter for the bumps at each datum can vary from one data point to another. Once again we suppose that k is a positive integer, and $K(x)$ a kernel function. Define $d_{j,k}$ as the distance from the point x_j to the kth nearest point in the data set. The **variable kernel estimate** with smoothing parameter h is defined as

$$\hat{\phi}(x) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{hd_{j,k}} K\left(\frac{x - x_j}{hd_{j,k}}\right) \tag{22}$$

The role of $d_{j,k}$ is to make the kernels flatter in regions where data are sparse. For a fixed k the overall degree of smoothing depends on the parameter h. The responsiveness of the window width choice to local smoothing is determined by the choice of k. Unlike the GNNE estimate of (21) the window width does not depend on the distance from x to the data points, but depends only on the distance between data points. Also unlike the GNNE, provided that K is a pdf the kernel estimate will be too.

9.6. Maximum Penalized Likelihood Estimators

The methods of density estimation described are basically empirical techniques based on the definition of a pdf. What would happen if we applied the standard estimation techniques like MLE? Following the standard approach we could write:

$$L(\phi|x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \phi(x_i) \tag{23}$$

There is no finite maximum for $l = \log L$ over the class of density functions. The likelihood can be made arbitrarily large by taking densities that approach the sum of delta functions located at the observations. To see that this is true, consider the naive density estimate in the limit as $h \to 0$.

If one wants to use a maximum likelihood kind of approach to the problem it is necessary to place restrictions on the kinds of densities over which the likelihood is to be maximized. One possibility involves incorporating into the likelihood a term including *roughness* of the curve. Suppose we define roughness $R(\phi)$ as

$$R(\phi) = \int_{-\infty}^{\infty} (\phi'')^2$$

and a **penalized log likelihood** by

$$l_\alpha(\phi) = \sum_{i=1}^{n} \log \phi(x_i) - \alpha R(\phi) \tag{24}$$

where α is a positive smoothing parameter. We won't go into the details here, but it is possible to find the **maximum penalized likelihood density estimate** as defined by (24) over the class of functions ϕ that satisfy $\int_{-\infty}^{\infty} \phi = 1$, $\phi(x) \geq 0$ for all x and $R(\phi) < \infty$. The parameter α controls the tradeoff between smoothness and goodness of fit to the data. Small α generates a rough maximum penalized likelihood estimator. *Silverman* (1986) discusses this and some other approaches to non-parametric density estimation that we will not treat.

9.6:1 Bounded Domains

We conclude this section by noting that it is often necessary to give special consideration to the domain of definition for a density function: *e.g.*, what to do when it is always positive, or bounded on both sides. One needs to ensure that $\phi(x)$ is zero outside the domain of x. Again see *Silverman* for some basic strategies.