Density Estimation
36-708

1 Introduction

Let $X_1, \ldots, X_n$ be a sample from a distribution P with density p. The goal of nonparametric density estimation is to estimate p with as few assumptions about p as possible. We denote the estimator by $\widehat{p}$. The estimator will depend on a smoothing parameter h and choosing h carefully is crucial. To emphasize the dependence on h we sometimes write $\widehat{p}_h$.

Density estimation used for: regression, classification, clustering and unsupervised prediction. For example, if $\widehat{p}(x, y)$ is an estimate of $p(x, y)$ then we get the following estimate of the regression function:

$$\widehat{m}(x) = \int y\widehat{p}(y|x)dy$$

where $\widehat{p}(y|x) = \widehat{p}(y, x)/\widehat{p}(x)$. For classification, recall that the Bayes rule is

$$h(x) = I(p_1(x)\pi_1 > p_0(x)\pi_0)$$

where $\pi_1 = \mathbb{P}(Y = 1)$, $\pi_0 = \mathbb{P}(Y = 0)$, $p_1(x) = p(x|y = 1)$ and $p_0(x) = p(x|y = 0)$. Inserting sample estimates of π_1 and π_0, and density estimates for p_1 and p_0 yields an estimate of the Bayes rule. For clustering, we look for the high density regions, based on an estimate of the density. Many classifiers that you are familiar with can be re-expressed this way. Unsupervised prediction is discussed in Section 9. In this case we want to predict X_{n+1} from $X_1, \ldots, X_n$.

Example 1 (Bart Simpson) *The top left plot in Figure 1 shows the density*

$$p(x) = \frac{1}{2}\phi(x;\, 0, 1) + \frac{1}{10}\sum_{j=0}^{4}\phi(x;\, (j/2) - 1, 1/10) \tag{1}$$

where $\phi(x; \mu, \sigma)$ denotes a Normal density with mean μ and standard deviation σ. Marron and Wand (1992) call this density "the claw" although we will call it the Bart Simpson density. Based on 1,000 draws from p, we computed a kernel density estimator, described later. The estimator depends on a tuning parameter called the bandwidth. The top right plot is based on a small bandwidth h which leads to undersmoothing. The bottom right plot is based on a large bandwidth h which leads to oversmoothing. The bottom left plot is based on a bandwidth h which was chosen to minimize estimated risk. This leads to a much more reasonable density estimate.

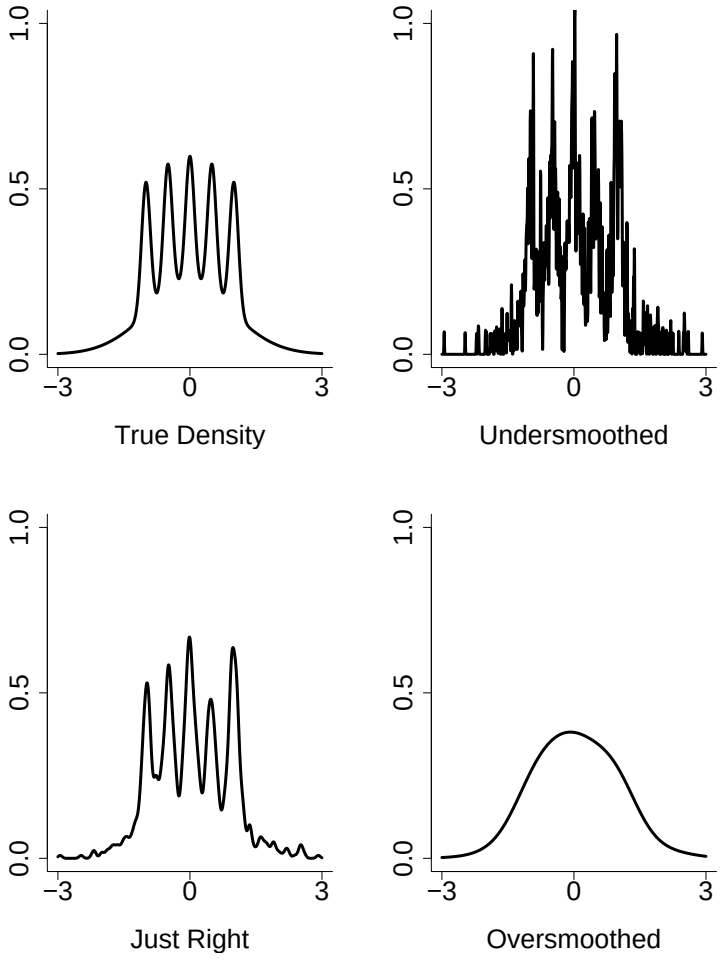

Figure 1: The Bart Simpson density from Example 1. Top left: true density. The other plots are kernel estimators based on $n = 1{,}000$ draws. Bottom left: bandwidth $h = 0.05$ chosen by leave-one-out cross-validation. Top right: bandwidth $h/10$. Bottom right: bandwidth $10h$.

2 Loss Functions

The most commonly used loss function is the L_2 loss

$$\int (\widehat{p}(x) - p(x))^2 dx = \int \widehat{p}^2(x)dx - 2\int \widehat{p}(x)p(x) + \int p^2(x)dx.$$

The risk is $R(p, \widehat{p}) = \mathbb{E}(L(p, \widehat{p}))$.

Devroye and Györfi (1985) make a strong case for using the L_1 norm

$$\|\widehat{p} - p\|_1 \equiv \int |\widehat{p}(x) - p(x)|dx$$

as the loss instead of L_2. The L_1 loss has the following nice interpretation. If P and Q are distributions define the total variation metric

$$d_{TV}(P, Q) = \sup_A |P(A) - Q(A)|$$

where the supremum is over all measurable sets. Now if P and Q have densities p and q then

$$d_{TV}(P, Q) = \frac{1}{2}\int |p - q| = \frac{1}{2}\|p - q\|_1. \quad \mathbf{H}$$

Thus, if $\int |p-q| < \delta$ then we know that $|P(A)-Q(A)| < \delta/2$ for all A. Also, the L_1 norm is transformation invariant. Suppose that T is a one-to-one smooth function. Let $Y = T(X)$. Let p and q be densities for X and let $\widetilde{p}$ and $\widetilde{q}$ be the corresponding densities for Y. Then

$$\int |p(x) - q(x)|dx = \int |\widetilde{p}(y) - \widetilde{q}(y)|dy. \quad \mathbf{H}$$

Hence the distance is unaffected by transformations. The L_1 loss is, in some sense, a much better loss function than L_2 for density estimation. But it is much more difficult to deal with. For now, we will focus on L_2 loss. But we may discuss L_1 loss later.

Another loss function is the Kullback-Leibler loss $\int p(x) \log p(x)/q(x)dx$. This is not a good loss function to use for nonparametric density estimation. The reason is that the Kullback-Leibler loss is completely dominated by the tails of the densities. **H**

3 Histograms

Perhaps the simplest density estimators are histograms. For convenience, assume that the data $X_1, \ldots, X_n$ are contained in the unit cube $\mathcal{X} = [0, 1]^d$ (although this assumption is not crucial). Divide $\mathcal{X}$ into bins, or sub-cubes, of size h. **We discuss methods for choosing**

h **later.** There are $N \approx (1/h)^d$ such bins and each has volume h^d. Denote the bins by $B_1, \ldots, B_N$. The histogram density estimator is

$$\widehat{p}_h(x) = \sum_{j=1}^{N} \frac{\widehat{\theta}_j}{h^d} I(x \in B_j) \tag{2}$$

where

$$\widehat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} I(X_i \in B_j)$$

is the fraction of data points in bin B_j. Now we bound the bias and variance of $\widehat{p}_h$. We will assume that $p \in \mathcal{P}(L)$ where

$$\mathcal{P}(L) = \left\{ p : \ |p(x) - p(y)| \leq L\|x - y\|, \text{ for all } x, y \right\}. \tag{3}$$

First we bound the bias. Let $\theta_j = P(X \in B_j) = \int_{B_j} p(u)du$. For any $x \in B_j$,

$$p_h(x) \equiv \mathbb{E}(\widehat{p}_h(x)) = \frac{\theta_j}{h^d} \tag{4}$$

and hence

$$p(x) - p_h(x) = p(x) - \frac{\int_{B_j} p(u)du}{h^d} = \frac{1}{h^d} \int (p(x) - p(u))du.$$

Thus,

$$|p(x) - p_h(x)| \leq \frac{1}{h^d} \int |p(x) - p(u)|du \leq \frac{1}{h^d} Lh\sqrt{d} \int du = Lh\sqrt{d}$$

where we used the fact that if $x, u \in B_j$ then $\|x - u\| \leq \sqrt{d}h$.

Now we bound the variance. Since p is Lipschitz on a compact set, it is bounded. Hence, $\theta_j = \int_{B_j} p(u)du \leq C \int_{B_j} du = Ch^d$ for some C. Thus, the variance is

$$\mathsf{Var}(\widehat{p}_h(x)) = \frac{1}{h^{2d}} \mathsf{Var}(\widehat{\theta}_j) = \frac{\theta_j(1 - \theta_j)}{nh^{2d}} \leq \frac{\theta_j}{nh^{2d}} \leq \frac{C}{nh^d}.$$

We conclude that the L_2 risk is bounded by

$$\sup_{p \in \mathcal{P}(L)} R(p, \widehat{p}) = \int (\mathbb{E}(\widehat{p}_h(x) - p(x))^2 \leq L^2 h^2 d + \frac{C}{nh^d}. \tag{5}$$

The upper bound is minimized by choosing $h = \left(\frac{C}{L^2 nd}\right)^{\frac{1}{d+2}}$. (Later, we shall see a more practical way to choose h.) With this choice,

$$\sup_{P \in \mathcal{P}(L)} R(p, \widehat{p}) \leq C_0 \left(\frac{1}{n}\right)^{\frac{2}{d+2}}$$

where $C_0 = L^2 d(C/(L^2 d))^{2/(d+2)}$.

Later, we will prove the following theorem which shows that this upper bound is tight. Specifically:

Theorem 2 *There exists a constant $C > 0$ such that*

$$\inf_{\widehat{p}} \sup_{P \in \mathcal{P}(L)} \mathbb{E} \int (\widehat{p}(x) - p(x))^2 dx \geq C \left(\frac{1}{n}\right)^{\frac{2}{d+2}}. \tag{6}$$

3.1 Concentration Analysis For Histograms

Let us now derive a concentration result for $\widehat{p}_h$. We will bound

$$\sup_{P \in \mathcal{P}} P^n(\|\widehat{p}_h - p\|_\infty > \epsilon)$$

where $\|f\|_\infty = \sup_x |f(x)|$. Assume that $\epsilon \leq 1$. First, note that

$$\mathbb{P}(\|\widehat{p}_h - p_h\|_\infty > \epsilon) = \mathbb{P}\left(\max_j \left|\frac{\widehat{\theta}_j}{h^d} - \frac{\theta_j}{h^d}\right| > \epsilon\right) = \mathbb{P}(\max_j |\widehat{\theta}_j - \theta_j| > h^d \epsilon) \leq \sum_j \mathbb{P}(|\widehat{\theta}_j - \theta_j| > h^d \epsilon).$$

Using Bernstein's inequality and the fact that $\theta_j(1-\theta_j) \leq \theta_j \leq Ch^d$,

$$\begin{aligned}
\mathbb{P}(|\widehat{\theta}_j - \theta_j| > h^d \epsilon) &\leq 2\exp\left(-\frac{1}{2}\frac{n\epsilon^2 h^{2d}}{\theta_j(1-\theta_j) + \epsilon h^d/3}\right) \\
&\leq 2\exp\left(-\frac{1}{2}\frac{n\epsilon^2 h^{2d}}{Ch^d + \epsilon h^d/3}\right) \\
&\leq 2\exp\left(-cn\epsilon^2 h^d\right)
\end{aligned}$$

where $c = 1/(2(C + 1/3))$. By the union bound and the fact that $N \leq (1/h)^d$,

$$\mathbb{P}(|\widehat{\theta}_j - \theta_j| > h^d \epsilon) \leq 2h^{-d} \exp\left(-cn\epsilon^2 h^d\right) \equiv \pi_n.$$

Earlier we saw that $\sup_x |p(x) - p_h(x)| \leq L\sqrt{d}h$. Hence, with probability at least $1 - \pi_n$,

$$\|\widehat{p}_h - p\|_\infty \leq \|\widehat{p}_h - p_h\|_\infty + \|p_h - p\|_\infty \leq \epsilon + L\sqrt{d}h. \tag{7}$$

Now set

$$\epsilon = \sqrt{\frac{1}{cnh^d} \log\left(\frac{2}{\delta h^d}\right)}.$$

Then, with probability at least $1-\delta$,

$$\|\widehat{p}_h - p\|_\infty \leq \sqrt{\frac{1}{cnh^d}\log\left(\frac{2}{\delta h^d}\right)} + L\sqrt{d}h. \tag{8}$$

Choosing $h = (c_2/n)^{1/(2+d)}$ we conclude that, with probability at least $1-\delta$,

$$\|\widehat{p}_h - p\|_\infty \leq \sqrt{c^{-1}n^{-\frac{2}{2+d}}\left[\log\left(\frac{2}{\delta}\right) + \left(\frac{2}{2+d}\right)\log n\right]} + L\sqrt{d}n^{-\frac{1}{2+d}} = O\left(\left(\frac{\log n}{n}\right)^{\frac{1}{2+d}}\right). \tag{9}$$

4 Kernel Density Estimation

A one-dimensional smoothing kernel is any smooth function K such that $\int K(x)\,dx = 1$, $\int xK(x)dx = 0$ and $\sigma_K^2 \equiv \int x^2K(x)dx > 0$. *Smoothing kernels* should not be confused with *Mercer kernels* which we discuss later. Some commonly used kernels are the following:

Boxcar:	$K(x) = \frac{1}{2}I(x)$	**Gaussian:**	$K(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$
Epanechnikov:	$K(x) = \frac{3}{4}(1-x^2)I(x)$	**Tricube:**	$K(x) = \frac{70}{81}(1-\lvert x\rvert^3)^3 I(x)$

where $I(x) = 1$ if $|x| \leq 1$ and $I(x) = 0$ otherwise. These kernels are plotted in Figure 2. Two commonly used multivariate kernels are $\prod_{j=1}^d K(x_j)$ and $K(\|x\|)$.

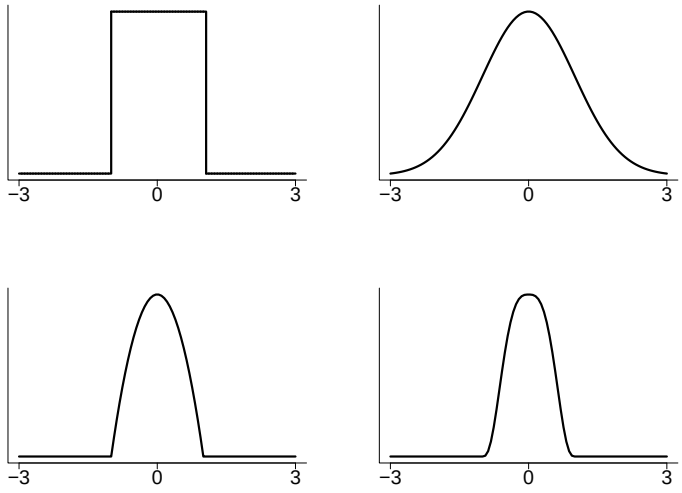

Figure 2: Examples of smoothing kernels: boxcar (top left), Gaussian (top right), Epanechnikov (bottom left), and tricube (bottom right).

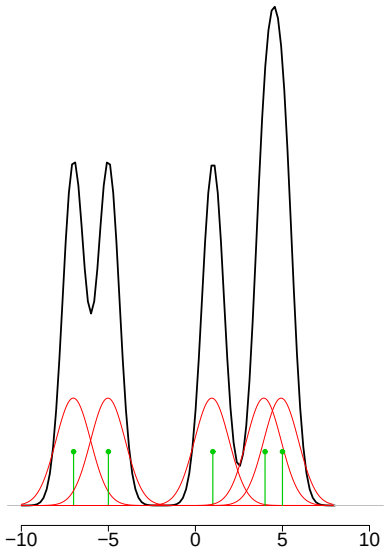

Figure 3: A kernel density estimator $\widehat{p}$. At each point x, $\widehat{p}(x)$ is the average of the kernels centered over the data points X_i. The data points are indicated by short vertical bars. The kernels are not drawn to scale.

Suppose that $X \in \mathbb{R}^d$. Given a kernel K and a positive number h, called the bandwidth, the kernel density estimator is defined to be

$$\widehat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{\|x - X_i\|}{h}\right). \tag{10}$$

More generally, we define

$$\widehat{p}_H(x) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - X_i)$$

where H is a positive definite bandwidth matrix and $K_H(x) = |H|^{-1/2}K(H^{-1/2}x)$. For simplicity, we will take $H = h^2 I$ and we get back the previous formula.

Sometimes we write the estimator as $\widehat{p}_h$ to emphasize the dependence on h. In the multivariate case the coordinates of X_i should be standardized so that each has the same variance, since the norm $\|x - X_i\|$ treats all coordinates as if they are on the same scale.

The kernel estimator places a smoothed out lump of mass of size $1/n$ over each data point X_i; see Figure 3. The choice of kernel K is not crucial, but the choice of bandwidth h is important. Small bandwidths give very rough estimates while larger bandwidths give smoother estimates.

4.1 Risk Analysis

In this section we examine the accuracy of kernel density estimation. We will first need a few definitions.

Assume that $X_i \in \mathcal{X} \subset \mathbb{R}^d$ where $\mathcal{X}$ is compact. Let β and L be positive numbers. Given a vector $s = (s_1, \ldots, s_d)$, define $|s| = s_1 + \cdots + s_d$, $s! = s_1! \cdots s_d!$, $x^s = x_1^{s_1} \cdots x_d^{s_d}$ and

$$D^s = \frac{\partial^{s_1 + \cdots + s_d}}{\partial x_1^{s_1} \cdots \partial x_d^{s_d}}.$$

Let β be a positive integer. Define the Hölder class

$$\Sigma(\beta, L) = \left\{ g : \ |D^s g(x) - D^s g(y)| \leq L\|x-y\|, \quad \text{for all } s \text{ such that } |s| = \beta - 1, \text{ and all } x, y \right\}. \tag{11}$$

For example, if $d = 1$ and $\beta = 2$ this means that

$$|g'(x) - g'(y)| \leq L\,|x - y|, \quad \text{for all } x, y.$$

The most common case is $\beta = 2$; roughly speaking, this means that the functions have bounded second derivatives.

If $g \in \Sigma(\beta, L)$ then $g(x)$ is close to its Taylor series approximation:

$$|g(u) - g_{x,\beta}(u)| \leq L\|u - x\|^\beta \tag{12}$$

where **H**

$$g_{x,\beta}(u) = \sum_{|s| \leq \beta} \frac{(u-x)^s}{s!} D^s g(x). \tag{13}$$

In the common case of $\beta = 2$, this means that

$$\left| p(u) - [p(x) + (x-u)^T \nabla p(x)] \right| \leq L||x-u||^2.$$

Assume now that the kernel K has the form $K(x) = G(x_1) \cdots G(x_d)$ where G has support on $[-1, 1]$, $\int G = 1$, $\int |G|^p < \infty$ for any $p \geq 1$, $\int |t|^\beta |K(t)| dt < \infty$ and $\int t^s K(t) dt = 0$ for $s \leq \beta$.

An example of a kernel that satisfies these conditions for $\beta = 2$ is $G(x) = (3/4)(1 - x^2)$ for $|x| \leq 1$. Constructing a kernel that satisfies $\int t^s K(t) dt = 0$ for $\beta > 2$ requires using kernels that can take negative values.

Let $p_h(x) = \mathbb{E}[\widehat{p}_h(x)]$. The next lemma provides a bound on the bias $p_h(x) - p(x)$.

Lemma 3 *The bias of $\widehat{p}_h$ satisfies:*

$$\sup_{p\in\Sigma(\beta,L)} |p_h(x) - p(x)| \leq ch^\beta \tag{14}$$

for some c.

Proof. We have

$$\begin{aligned}
|p_h(x) - p(x)| &= \int \frac{1}{h^d} K(\|u-x\|/h)p(u)du - p(x) \\
&= \left|\int K(\|v\|)(p(x+hv) - p(x))dv\right| \\
&\leq \left|\int K(\|v\|)(p(x+hv) - p_{x,\beta}(x+hv))dv\right| + \left|\int K(\|v\|)(p_{x,\beta}(x+hv) - p(x))dv\right|.
\end{aligned}$$

The first term is bounded by $Lh^\beta \int K(s)|s|^\beta$ since $p \in \Sigma(\beta, L)$. The second term is 0 from the properties on K since $p_{x,\beta}(x+hv) - p(x)$ is a polynomial of degree β (with no constant term). □

Next we bound the variance.

Lemma 4 *The variance of $\widehat{p}_h$ satisfies:*

$$\sup_{p\in\Sigma(\beta,L)} \text{Var}(\widehat{p}_h(x)) \leq \frac{c}{nh^d} \tag{15}$$

for some $c > 0$.

Proof. We can write $\widehat{p}(x) = n^{-1}\sum_{i=1}^n Z_i$ where $Z_i = \frac{1}{h^d} K\left(\frac{\|x-X_i\|}{h}\right)$. Then,

$$\begin{aligned}
\text{Var}(Z_i) &\leq \mathbb{E}(Z_i^2) = \frac{1}{h^{2d}} \int K^2\left(\frac{\|x-u\|}{h}\right) p(u)du = \frac{h^d}{h^{2d}} \int K^2\left(\|v\|\right) p(x+hv)dv \\
&\leq \frac{\sup_x p(x)}{h^d} \int K^2(\|v\|)dv \leq \frac{c}{h^d}
\end{aligned}$$

for some c since the densities in $\Sigma(\beta, L)$ are uniformly bounded. The result follows. □

Since the mean squared error is equal to the variance plus the bias squared we have:

Theorem 5 *The L_2 risk is bounded above, uniformly over $\Sigma(\beta, L)$, by $h^{4\beta} + \frac{1}{nh^d}$ (up to constants). If $h \asymp n^{-1/(2\beta+d)}$ then*

$$\sup_{p\in\Sigma(\beta,L)} \mathbb{E}\int (\widehat{p}_h(x) - p(x))^2 dx \preceq \left(\frac{1}{n}\right)^{\frac{2\beta}{2\beta+d}}. \tag{16}$$

When $\beta = 2$ and $h \asymp n^{-1/(4+d)}$ we get the rate $n^{-4/(4+d)}$.

4.2 Minimax Bound

According to the next theorem, there does not exist an estimator that converges faster than $O(n^{-2\beta/(2\beta+d)})$. We state the result for integrated L_2 loss although similar results hold for other loss functions and other function spaces. We will prove this later in the course.

Theorem 6 *There exists C depending only on β and L such that*

$$\inf_{\widehat{p}} \sup_{p \in \Sigma(\beta,L)} \mathbb{E}_p \int (\widehat{p}(x) - p(x))^2 dx \geq C \left(\frac{1}{n}\right)^{\frac{2\beta}{2\beta+d}}. \tag{17}$$

Theorem 6 together with (16) imply that kernel estimators are rate minimax.

4.3 Concentration Analysis of Kernel Density Estimator

Now we state a result which says how fast $\widehat{p}(x)$ concentrates around $p(x)$. First, recall Bernstein's inequality: Suppose that $Y_1, \ldots, Y_n$ are iid with mean μ, $\mathrm{Var}(Y_i) \leq \sigma^2$ and $|Y_i| \leq M$. Then

$$\mathbb{P}(|\overline{Y} - \mu| > \epsilon) \leq 2 \exp\left\{ -\frac{n\epsilon^2}{2\sigma^2 + 2M\epsilon/3} \right\}. \tag{18}$$

Theorem 7 *For all small $\epsilon > 0$,*

$$\mathbb{P}(|\widehat{p}(x) - p_h(x)| > \epsilon) \leq 2 \exp\left\{ -cnh^d \epsilon^2 \right\}. \tag{19}$$

Hence, for any $\delta > 0$,

$$\sup_{p \in \Sigma(\beta,L)} \mathbb{P}\left(|\widehat{p}(x) - p(x)| > \sqrt{\frac{C \log(2/\delta)}{nh^d}} + ch^\beta \right) < \delta \tag{20}$$

for some constants C and c. If $h \asymp n^{-1/(2\beta+d)}$ then

$$\sup_{p \in \Sigma(\beta,L)} \mathbb{P}\left(|\widehat{p}(x) - p(x)|^2 > \frac{c}{n^{2\beta/(2\beta+d)}} \right) < \delta.$$

Note that the last statement follows from the bias-variance calculation followed by Markov's inequality. The first statement does not.

Proof. By the triangle inequality,

$$|\widehat{p}(x) - p(x)| \leq |\widehat{p}(x) - p_h(x)| + |p_h(x) - p(x)| \tag{21}$$

where $p_h(x) = \mathbb{E}(\widehat{p}(x))$. From Lemma 3, $|p_h(x) - p(x)| \leq ch^\beta$ for some c. Now $\widehat{p}(x) = n^{-1}\sum_{i=1}^n Z_i$ where

$$Z_i = \frac{1}{h^d} K\left(\frac{\|x - X_i\|}{h}\right).$$

Note that $|Z_i| \leq c_1/h^d$ where $c_1 = K(0)$. Also, $\text{Var}(Z_i) \leq c_2/h^d$ from Lemma 4. Hence, by Bernstein's inequality,

$$\mathbb{P}(|\widehat{p}(x) - p_h(x)| > \epsilon) \;\; \leq \;\; 2\exp\left\{-\frac{n\epsilon^2}{2c_2h^{-d} + 2c_1h^{-d}\epsilon/3}\right\} \leq 2\exp\left\{-\frac{nh^d\epsilon^2}{4c_2}\right\}$$

whenever $\epsilon \leq 3c_2/c_1$. If we choose $\epsilon = \sqrt{C\log(2/\delta)/(nh^d)}$ where $C = 4c_2$ then

$$\mathbb{P}\left(|\widehat{p}(x) - p_h(x)| > \sqrt{\frac{C}{nh^d}}\right) \leq \delta.$$

The result follows from (21). □

4.4 Concentration in L_∞

Theorem 7 shows that, for each x, $\widehat{p}(x)$ is close to $p(x)$ with high probability. We would like a version of this result that holds uniformly over all x. That is, we want a concentration result for

$$\|\widehat{p} - p\|_\infty = \sup_x |\widehat{p}(x) - p(x)|.$$

We can write

$$\|\widehat{p}_h - p\|_\infty \leq \|\widehat{p}_h - p_h\|_\infty + \|p_h - p\|_\infty \leq \|\widehat{p}_h - p_h\|_\infty + ch^\beta.$$

We can bound the first term using something called *bracketing* together with Bernstein's theorem to prove that,

$$\mathbb{P}(\|\widehat{p}_h - p_h\|_\infty > \epsilon) \leq 4\left(\frac{C}{h^{d+1}\epsilon}\right)^d \exp\left(-\frac{3n\epsilon^2h^d}{28K(0)}\right). \tag{22}$$

An alternative approach is to replace Bernstein's inequality with a more sophisticated inequality due to Talagrand. We follow the analysis in Giné and Guillou (2002). Let

$$\mathcal{F} = \left\{K\left(\frac{x - \cdot}{h}\right), x \in \mathbb{R}^d, h > 0\right\}.$$

We assume there exists positive numbers A and v such that

$$\sup_P N(\mathcal{F}_h, L_2(P), \epsilon\|F\|_{L_2(P)}) \leq \left(\frac{A}{\epsilon}\right)^v, \tag{23}$$

where $N(T, d, \epsilon)$ denotes the ϵ-covering number of the metric space (T, d), F is the envelope function of $\mathcal{F}$ and the supremum is taken over the set of all probability measures on $\mathbb{R}^d$. The quantities A and v are called the VC characteristics of $\mathcal{F}_h$.

Theorem 8 (Giné and Guillou 2002) *Assume that the kernel satisfies the above property.*

1. *Let $h > 0$ be fixed. Then, there exist constants $c_1 > 0$ and $c_2 > 0$ such that, for all small $\epsilon > 0$ and all large n,*

$$\mathbb{P}\left\{\sup_{x\in\mathbb{R}^d} |\widehat{p}_h(x) - p_h(x)| > \epsilon\right\} \leq c_1 \exp\left\{-c_2 n h^d \epsilon^2\right\}. \tag{24}$$

2. *Let $h_n \to 0$ as $n \to \infty$ in such a way that $\frac{nh_n^d}{|\log h_n^d|} \to \infty$. Let*

$$\epsilon_n \geq \sqrt{\frac{|\log h_n|}{nh_n^d}}. \tag{25}$$

Then, for all n large enough, (24) holds with h and ϵ replaced by h_n and ϵ_n, respectively.

The above theorem imposes minimal assumptions on the kernel K and, more importantly, on the probability distribution P, whose density is not required to be bounded or smooth, and, in fact, may not even exist. Combining the above theorem with Lemma 3 we have the following result.

Theorem 9 *Suppose that $p \in \Sigma(\beta, L)$. Fix any $\delta > 0$. Then*

$$\mathbb{P}\left(\sup_x |\widehat{p}(x) - p(x)| > \sqrt{\frac{C\log n}{nh^d}} + ch^\beta\right) < \delta$$

for some constants C and c where C depends on δ. Choosing $h \asymp \log n/n^{-1/(2\beta+d)}$ we have

$$\mathbb{P}\left(\sup_x |\widehat{p}(x) - p(x)|^2 > \frac{C\log n}{n^{2\beta/(2\beta+d)}}\right) < \delta.$$

4.5 Boundary Bias

We have ignored what happens near the boundary of the sample space. If x is $O(h)$ close to the boundary, the bias is $O(h)$ instead of $O(h^2)$. There are a variety of fixes including: data reflection, transformations, boundary kernels, local likelihood.

4.6 Confidence Bands and the CLT

Consider first a single point x. Let $s_n(x) = \sqrt{\text{Var}(\widehat{p}_h(x))}$. The CLT implies that

$$Z_n(x) \equiv \frac{\widehat{p}_h(x) - p_h(x)}{s_n(x)} \rightsquigarrow N(0, \tau^2(x)) \qquad \mathbf{H}$$

for some $\tau(x)$. This is true even if $h = h_n$ is decreasing. Specifically, suppose that $h_n \to 0$ and $nh_n \to \infty$. Note that $Z_n(x) = \sum_{i=1}^n L_{ni}$, say. According to Lyapounov's CLT, $\sum_{i=1}^n L_{ni} \rightsquigarrow N(0,1)$ as long as

$$\lim_{n\to\infty} \sum_{i=1}^n \mathbb{E}[L_{n,i}|^{2+\delta} = 0$$

for some $\delta > 0$. But this is does not yield a confidence interval for $p(x)$. To see why, let us write

$$\frac{\widehat{p}_h(x) - p(x)}{s_n(x)} = \frac{\widehat{p}_h(x) - p_h(x)}{s_n(x)} + \frac{p_h(x) - p(x)}{s_n(x)} = Z_n(x) + \frac{\text{bias}}{\sqrt{\text{var}}(x)}.$$

Assuming that the optimize the risk by balancing the bias and the variance, the second term is some constant c. So

$$\frac{\widehat{p}_h(x) - p(x)}{s_n(x)} \rightsquigarrow N(c, \tau^2(x)).$$

This means that the usual confidence interval $\widehat{p}_h(x) \pm z_{\alpha/2} s(x)$ will not cover $p(x)$ with probability tending to $1-\alpha$. One fix for this is to undersmooth the estimator. (We sacrifice risk for coverage.) An easier approach is just to interpret $\widehat{p}_h(x) \pm z_{\alpha/2} s(x)$ as a confidence interval for the smoothed density $p_h(x)$ instead of $p(x)$.

But this only gives an interval at one point. To get a confidence band we use the bootstrap. Let P_n be the empirical distribution of $X_1, \ldots, X_n$. The idea is to estimate the distribution

$$F_n(t) = \mathbb{P}\Big(\sqrt{nh^d}||\widehat{p}_h(x) - p_h(x)||_\infty \le t\Big)$$

with the bootstrap estimator

$$\widehat{F}_n(t) = \mathbb{P}\Big(\sqrt{nh^d}||\widehat{p}^*_h(x) - \widehat{p}_h(x)||_\infty \le t \;\Big|\; X_1, \ldots, X_n\Big)$$

where $\widehat{p}^*_h$ is constructed from the bootstrap sample $X_1^*, \ldots, X_n^* \sim P_n$. Later in the course, we will show that

$$\sup_t |F_n(t) - \widehat{F}_n(t)| \xrightarrow{P} 0.$$

Here is the algorithm.

1. Let P_n be the empirical distribution that puts mass $1/n$ at each data point X_i.

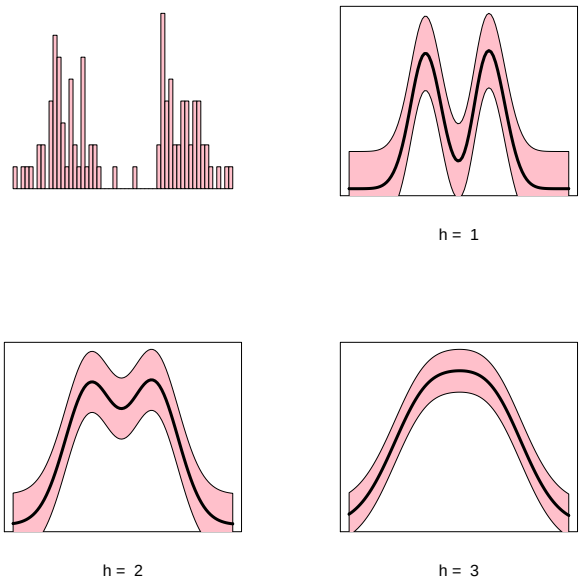

Figure 4: 95 percent bootstrap confidence bands using various bandwidths.

2. Draw $X_1^*, \ldots, X_n^* \sim P_n$. This is called a bootstrap sample.
3. Compute the density estimator $\widehat{p}_h^*$ based on the bootstrap sample.
4. Compute $R = \sup_x \sqrt{nh^d}||\widehat{p}_h^* - \widehat{p}_h||_\infty$.
5. Repeat steps 2-4 B times. This gives $R_1, \ldots, R_B$.
6. Let z_α be the upper α quantile of the R_j's. Thus

$$\frac{1}{B}\sum_{j=1}^{B} I(R_j > z_\alpha) \approx \alpha.$$

7. Let

$$\ell_n(x) = \widehat{p}_h(x) - \frac{z_\alpha}{\sqrt{nh^d}}, \quad u_n(x) = \widehat{p}_h(x) + \frac{z_\alpha}{\sqrt{nh^d}}.$$

Theorem 10 *Under appropriate (very weak) conditions, we have*

$$\liminf_{n\to\infty} \mathbb{P}\big(\ell_n(x) \le p_h(x) \le u(x) \quad \text{for all } x\big) \ge 1-\alpha.$$

See Figure 4.

If you want a confidence band for p you need to reduce the bias (undersmooth). A simple way to do this is with *twicing*. Suppose that $\beta = 2$ and that we use the kernel estimator $\widehat{p}_h$. Note that,

$$\mathbb{E}[\widehat{p}_h(x)] = p(x) + C(x)h^2 + o(h^2)$$
$$\mathbb{E}[\widehat{p}_{2h}(x)] = p(x) + C(x)4h^2 + o(h^2)$$

for some $C(x)$. That is, the leading term of the bias is $b(x) = C(x)h^2$. So if we define

$$\widehat{b}(x) = \frac{\widehat{p}_{2h}(x) - \widehat{p}_h(x)}{3}$$

then

$$\mathbb{E}[\widehat{b}(x)] = b(x).$$

We define the bias-reduced estimator

$$\widetilde{p}_h(x) = \widehat{p}_h(x) - \widehat{b}(x) = \frac{4}{3}\left(\widehat{p}_h(x) - \frac{1}{4}\widehat{p}_{2h}\right).$$

A confidence set centered at $\widetilde{p}_h$ will be asymptotically valid but will not be an optimal estimator. This is a fundamental conflict between estimation and inference.

5 Cross-Validation

In practice we need a data-based method for choosing the bandwidth h. To do this, we will need to estimate the risk of the estimator and minimize the estimated risk over h. Here, we describe two cross-validation methods.

5.1 Leave One Out

A common method for estimating risk is leave-one-out cross-validation. Recall that the loss function is

$$\int (\widehat{p}(x) - p(x))^2 dx = \int \widehat{p}^2(x)dx - 2\int \widehat{p}(x)p(x)dx + \int p^2(x)dx.$$

The last term does not involve $\widehat{p}$ so we can drop it. Thus, we now define the loss to be

$$L(h) = \int \widehat{p}^2(x)\, dx - 2\int \widehat{p}(x)p(x)dx.$$

The risk is $R(h) = \mathbb{E}(L(h))$.

Definition 11 *The leave-one-out cross-validation estimator of risk is*

$$\widehat{R}(h) = \int \left(\widehat{p}_{(-i)}(x)\right)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{p}_{(-i)}(X_i) \tag{26}$$

where $\widehat{p}_{(-i)}$ is the density estimator obtained after removing the i^{th} observation.

It is easy to check that $\mathbb{E}[\widehat{R}(h)] = R(h)$. **H**

When the kernel is Gaussian, the cross-validation score can be written, after some tedious algebra, as follows. Let $\phi(z;\sigma)$ denote a Normal density with mean 0 and variance σ^2. Then,

$$\widehat{R}(h) = \frac{\phi^d(0;\sqrt{2}h)}{(n-1)} + \frac{n-2}{n(n-1)^2}\sum_{i\neq j}\prod_{\ell=1}^{d}\phi(X_{i\ell}-X_{j\ell};\sqrt{2}h) \tag{27}$$

$$-\frac{2}{n(n-1)}\sum_{i\neq j}\prod_{\ell=1}^{d}\phi(X_{i\ell}-X_{j\ell};h). \tag{28}$$

The estimator $\widehat{p}$ and the cross-validation score can be computed quickly using the fast Fourier transform; see pages 61–66 of Silverman (1986).

For histograms, it is easy to work out the leave-one-out cross-validation in close form:

$$\widehat{R}(h) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)h}\sum_j \widehat{\theta}_j^2. \quad \textbf{H}$$

A further justification for cross-validation is given by the following theorem due to Stone (1984).

Theorem 12 (Stone's theorem) *Suppose that p is bounded. Let $\widehat{p}_h$ denote the kernel estimator with bandwidth h and let $\widehat{h}$ denote the bandwidth chosen by cross-validation. Then,*

$$\frac{\int \left(p(x)-\widehat{p}_{\widehat{h}}(x)\right)^2 dx}{\inf_h \int \left(p(x)-\widehat{p}_h(x)\right)^2 dx} \xrightarrow{a.s.} 1. \tag{29}$$

The bandwidth for the density estimator in the bottom left panel of Figure 1 is based on cross-validation. In this case it worked well but of course there are lots of examples where there are problems. Do not assume that, if the estimator $\widehat{p}$ is wiggly, then cross-validation has let you down. The eye is not a good judge of risk.

There are cases when cross-validation can seriously break down. In particular, if there are ties in the data then cross-validation chooses a bandwidth of 0.

5.2 Data Splitting

An alternative to leave-one-out is V-fold cross-validation. A common choice is $V = 10$. Fir simplicity, let us consider here just splitting the data in two halves. This version of cross-validation comes with stronger theoretical guarantees. Let $\widehat{p}_h$ denote the kernel estimator

based on bandwidth h. For simplicity, assume the sample size is even and denote the sample size by $2n$. Randomly split the data $X = (X_1, \ldots, X_{2n})$ into two sets of size n. Denote these by $Y = (Y_1, \ldots, Y_n)$ and $Z = (Z_1, \ldots, Z_n)$.[1] Let $\mathcal{H} = \{h_1, \ldots, h_N\}$ be a finite grid of bandwidths. Let

$$\widehat{p}_j(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_j^d} K\left(\frac{\|x - Y_i\|}{h}\right).$$

Thus we have a set $\mathcal{P} = \{\widehat{p}_1, \ldots, \widehat{p}\}$ of density estimators.

We would like to minimize $L(p, \widehat{p}_j) = \int \widehat{p}_j^2(x) - 2\int \widehat{p}_j(x)p(x)dx$. Define the estimated risk

$$\widehat{L}_j \equiv \widehat{L}(p, \widehat{p}_j) = \int \widehat{p}_j^2(x) - \frac{2}{n}\sum_{i=1}^{n} \widehat{p}_j(Z_i). \tag{30}$$

Let $\widehat{p} = \text{argmin}_{g\in\mathcal{P}} \widehat{L}(p, g)$. Schematically:

$$X = (X_1, \ldots, X_{2n}) \stackrel{\text{split}}{\Longrightarrow} \begin{array}{l} Y \to \{\widehat{p}_1, \ldots, \widehat{p}_N\} = \mathcal{P} \\ \\ Z \to \{\widehat{L}_1, \ldots, \widehat{L}_N\} \end{array}$$

Theorem 13 (Wegkamp 1999) *There exists a $C > 0$ such that*

$$\mathbb{E}(\|\widehat{p} - p\|^2) \leq 2\min_{g\in\mathcal{P}} \mathbb{E}(\|g - p\|^2) + \frac{C\log N}{n}.$$

This theorem can be proved using concentration of measure techniques that we discuss later in class. A similar result can be proved for V-fold cross-validation.

5.3 Asymptotic Expansions

In this section we consider some asymptotic expansions that describe the behavior of the kernel estimator. We focus on the case $d = 1$.

Theorem 14 *Let $R_x = \mathbb{E}(p(x) - \widehat{p}(x))^2$ and let $R = \int R_x\, dx$. Assume that p'' is absolutely continuous and that $\int p'''(x)^2 dx < \infty$. Then,*

$$R_x = \frac{1}{4}\sigma_K^4 h_n^4 p''(x)^2 + \frac{p(x)\int K^2(x)dx}{nh_n} + O\left(\frac{1}{n}\right) + O(h_n^6)$$

[1] It is not necessary to split the data into two sets of equal size. We use the equal split version for simplicity.

and

$$R = \frac{1}{4}\sigma_K^4 h_n^4 \int p''(x)^2 dx + \frac{\int K^2(x)dx}{nh} + O\left(\frac{1}{n}\right) + O(h_n^6) \tag{31}$$

where $\sigma_K^2 = \int x^2 K(x)dx$.

Proof. Write $K_h(x, X) = h^{-1}K\left((x - X)/h\right)$ and $\widehat{p}(x) = n^{-1}\sum_i K_h(x, X_i)$. Thus, $\mathbb{E}[\widehat{p}(x)] = \mathbb{E}[K_h(x, X)]$ and $\mathrm{Var}[\widehat{p}(x)] = n^{-1}\mathrm{Var}[K_h(x, X)]$. Now,

$$\begin{aligned}
\mathbb{E}[K_h(x, X)] &= \int \frac{1}{h} K\left(\frac{x - t}{h}\right) p(t)\, dt \\
&= \int K(u) p(x - hu)\, du \\
&= \int K(u)\left[p(x) - hup'(x) + \frac{h^2u^2}{2}p''(x) + \cdots\right] du \\
&= p(x) + \frac{1}{2}h^2 p''(x) \int u^2 K(u)\, du \cdots
\end{aligned}$$

since $\int K(x)\, dx = 1$ and $\int x\, K(x)\, dx = 0$. The bias is

$$\mathbb{E}[K_{h_n}(x, X)] - p(x) = \frac{1}{2}\sigma_K^2 h_n^2 p''(x) + O(h_n^4).$$

By a similar calculation,

$$\mathrm{Var}[\widehat{p}(x)] = \frac{p(x)\int K^2(x)\, dx}{n\, h_n} + O\left(\frac{1}{n}\right).$$

The first result then follows since the risk is the squared bias plus variance. The second result follows from integrating the first result. □

If we differentiate (31) with respect to h and set it equal to 0, we see that the asymptotically optimal bandwidth is

$$h_* = \left(\frac{c_2}{c_1^2 A(f) n}\right)^{1/5} \tag{32}$$

where $c_1 = \int x^2 K(x)dx$, $c_2 = \int K(x)^2 dx$ and $A(f) = \int f''(x)^2 dx$. This is informative because it tells us that the best bandwidth decreases at rate $n^{-1/5}$. Plugging h_* into (31), we see that if the optimal bandwidth is used then $R = O(n^{-4/5})$.

6 High Dimensions

The rate of convergence $n^{-2\beta/(2\beta+d)}$ is slow when the dimension d is large. In this case it is hopeless to try to estimate the true density p precisely in the L_2 norm (or any similar norm).

We need to change our notion of what it means to estimate p in a high-dimensional problem. Instead of estimating p precisely we have to settle for finding an adequate approximation of p. Any estimator that finds the regions where p puts large amounts of mass should be considered an adequate approximation. Let us consider a few ways to implement this type of thinking.

Biased Density Estimation. Let $p_h(x) = \mathbb{E}(\widehat{p}_h(x))$. Then

$$p_h(x) = \int \frac{1}{h^d} K\left(\frac{\|x-u\|}{h}\right) p(u)du$$

so that the mean of $\widehat{p}_h$ can be thought of as a smoothed version of p. Let $P_h(A) = \int_A p_h(u)du$ be the probability distribution corresponding to p_h. Then

$$P_h = P \oplus K_h$$

where $\oplus$ denotes convolution[2] and K_h is the distribution with density $h^{-d}K(\|u\|/h)$. In other words, if $X \sim P_h$ then $X = Y + Z$ where $Y \sim P$ and $Z \sim K_h$. This is just another way to say that P_h is a blurred or smoothed version of P. *p_h need not be close in L_2 to p but still could preserve most of the important shape information about p.* Consider then choosing a fixed $h > 0$ and estimating p_h instead of p. This corresponds to ignoring the bias in the density estimator. From Theorem 8 we conclude:

Theorem 15 *Let $h > 0$ be fixed. Then $\mathbb{P}(\|\widehat{p}_h - p_h\|_\infty > \epsilon) \leq Ce^{-nc\epsilon^2}$. Hence,*

$$\|\widehat{p}_h - p_h\|_\infty = O_P\left(\sqrt{\frac{\log n}{n}}\right).$$

The rate of convergence is fast and is independent of dimension. How to choose h is not clear.

Independence Based Methods. If we can live with some bias, we can reduce the dimensionality by imposing some independence assumptions. The simplest example is to treat the components $(X_1, \ldots, X_d)$ as if they are independent. In that case

$$p(x_1, \ldots, x_d) = \prod_{j=1}^{d} p_j(x_j)$$

and the problem is reduced to a set of one-dimensional density estimation problems.

[2]If $X \sim P$ and $Y \sim Q$ are independent, then the distribution of $X + Y$ is denoted by $P \star Q$ and is called the convolution of P and Q.

An extension is to use a forest. We represent the distribution with an undirected graph. A graph with no cycles is a forest. Let E be the edges of the graph. Any density consistent with the forest can be written as

$$p(x) = \prod_{j=1}^{d} p_j(x_j) \prod_{(j,k)\in E} \frac{p_{j,k}(x_j, x_k)}{p_j(x_j)p_k(x_k)}.$$

To estimate the density therefore only require that we estimate one and two-dimensional marginals. But how do we find the edge set E? Some methods are discussed in Liu et al (2011) under the name "Forest Density Estimation." A simple approach is to connect pairs greedily using some measure of correlation.

Density Trees. Ram and Gray (2011) suggest a recursive partitioning scheme similar to decision trees. They split each coordinate dyadically, in a greedy fashion. The density estimator is taken to be piecewise constant. They use an L_2 risk estimator to decide when to split. This seems promising. The ideas seems to have been re-discovered in Yand and Wong (arXiv:1404.1425) and Liu and Wong (arXiv:1401.2597). Density trees seem very promising. It would be nice if there was an R package to do this and if there were more theoretical results.

7 Example

Figure 5 shows a synthetic two-dimensional data set, the cross-validation function and two kernel density estimators. The data are 100 points generated as follows. We select a point randomly on the unit circle then add Normal noise with standard deviation 0.1 The first estimator (lower left) uses the bandwidth that minimizes the leave-one-out cross-validation score. The second uses twice that bandwidth. The cross-validation curve is very sharply peaked with a clear minimum. The resulting density estimate is somewhat lumpy. This is because cross-validation is aiming to minimize L_2 error which does not guarantee that the estimate is smooth. Also, the dataset is small so this effect is more noticeable. The estimator with the larger bandwidth is noticeably smoother. However, the lumpiness of the estimator is not necessarily a bad thing.

8 Derivatives

Kernel estimators can also be used to estimate the derivatives of a density.[3] Let $D^{\otimes r}p$ denote the r^{th} derivative p. We are using Kronecker notation. Let $D^{\otimes 0}p = p$, $D^{\otimes 1}f$ is the gradient

[3]In this section we follow Chacon and Duong (2013), *Electronic Journal of Statistics*, 7, 499-532.

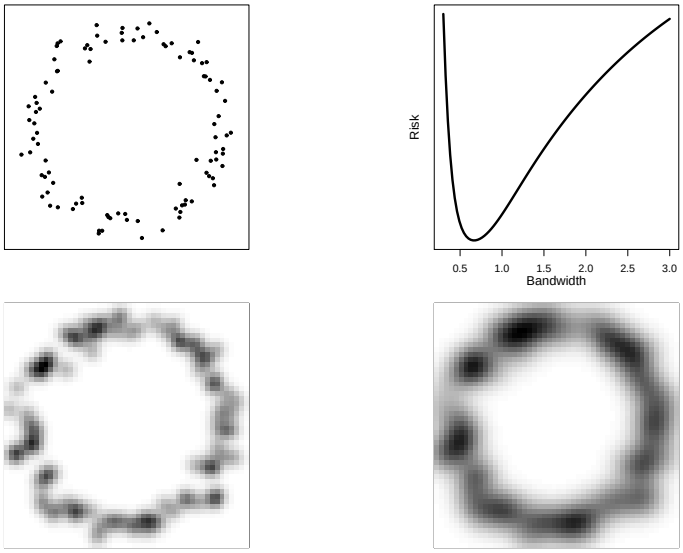

Figure 5: Synthetic two-dimensional data set. Top left: data. Top right: cross-validation function. Bottom left: kernel estimator based on the bandwidth that minimizes the cross-validation score. Bottom right: kernel estimator based on the twice the bandwidth that minimizes the cross-validation score.

of p, and $D^{\otimes 2}p = \text{vec}\mathcal{H}$ where $\mathcal{H}$ is the Hessian. We also write this as $p^{(r)}$ when convenient.

Let H be a bandwidth matrix and let

$$\widehat{p}_H(x) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - X_i)$$

where $K_H(x) = |H|^{-1/2}K(H^{-1/2}x)$. We define

$$\widehat{p^{(r)}}(x) = D^{\otimes r}\widehat{p}_H(x) = \frac{1}{n}\sum_{i=1}^{n} D^{\otimes r}K_H(x - X_i).$$

For computation, it is useful to note that

$$D^{\otimes r}K_H(x) = |H|^{-1/2}(H^{-1/2})^{\otimes r}\ D^{\otimes r}K_H(H^{-1/2}x).$$

The asymptotic mean squared error is derived in Chacon, Duong and Wand (2011) and is given by

$$\frac{1}{n}|H^{-1/2}|\text{tr}((H^{-1})^{\otimes r}R(D^{\otimes r}(K))) + \frac{m_2^2(K)}{4}\text{tr}((I_{d^r}\otimes \text{vec}^T H)R(D^{\otimes(\text{r}+2)}p)(I_{d^r}\otimes \text{vec}(H)))$$

where $R(g) = \int g(x)g^T(x)dx$, $m_2(K) = \int xx^T K(x)dx$. The optimal H has entries of order $n^{-2/(d+2r+4)}$ which yield an asymptotic mean squared error of order $n^{-4/(d+2r+4)}$. In dimension $d = 1$, the risk looks like this as a function of r:

r	risk
0	$n^{-4/5}$
1	$n^{-4/7}$
2	$n^{-4/9}$

We see that estimating derivatives is harder than estimating the density itself.

Chacon and Duong (2013) derive an estimate of the risk:

$$\text{CV}_r(H) = (-1)^r|H|^{-1/2}\text{vec}^T(H^{-1})^{\otimes r}G_n$$

where

$$G_n = \frac{1}{n^2}\sum_{i,j} D^{\otimes 2r}\overline{K}(H^{-1/2}(X_i - X_j)) - \frac{2}{n(n-1)}\sum_{i\neq j} D^{\otimes 2r}K(H^{-1/2}(X_i - X_j))$$

and $\overline{K} = K \star K$. We can now minimize CV over H. It would be nice if someone wrote an R package to do this. I think the ks package does much of this.

One application of this that we consider later in the course is mode-based clustering. Here, we use density estimation to find the modes of the density. We associate clusters with these modes. We can also test for a mode by testing if $D^2p(x) < 0$ at the estimated modes.

9 Unsupervised Prediction and Anomaly Detection

We can use density estimation to do unsupervised prediction and anomaly detection. The basic idea is due to Vovk, and was developed in a statistical framework in Lei, Robins and Wasserman (2014).

Suppose we observe $Y_1, \ldots, Y_n \sim P$. We want to predict Y_{n+1}. We will construct a level α test for the null hypothesis $H_0 : Y_{n+1} = y$. We do this for every value of y. Then we invert the test, that is, we set C_n to be the set of y's that are not rejected. It folows that

$$\mathbb{P}(Y_{n+1} \in C_n) \geq 1 - \alpha.$$

The prediction set C_n is finite sample and distribution-free.

Fix a value y. Let $A = (Y_1, \ldots, Y_n, y)$ be the *augmented dataset*. That is, we set $Y_{n+1} = y$. Let $\widehat{p}_A$ be a density estimate based on A. Consider the vector

$$\widehat{p}_A(Y_1), \ldots, \widehat{p}_A(Y_{n+1}).$$

Under H_0, the rank of these values is uniformly distributed. That is, for each i,

$$\mathbb{P}(\widehat{p}_A(Y_i) \leq \widehat{p}_A(y)) = \frac{1}{n+1}.$$

A p-value for the test is

$$\pi(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} I(\widehat{p}_A(Y_i) \leq \widehat{p}_A(y)).$$

The prediction set is

$$C_n = \Big\{ y : \ \pi(y) \geq \alpha \Big\}.$$

Computing C_n is tedious. Fortunately, Jing, Robins and Wasserman (2014) show that there is a simpler set that still has the correct coverage (but is slightly larger). The set is constructed as follows. Let $Z_i = \widehat{p}(Y_i)$. Order these observations

$$Z_{(1)} \leq \cdots \leq Z_{(n)}.$$

Let $k = \lfloor (n+1)\alpha \rfloor$ and let

$$t = Z_{(k)} - \frac{K(0)}{nh^d}.$$

Define

$$C_n^+ = \Big\{ y : \ \widehat{p}(y) \geq t \Big\}.$$

Lemma 16 *We have that $C_n \subset C_n^+$ and hence*

$$\mathbb{P}(Y_{n+1} \in C_n) \geq 1 - \alpha.$$

Finally, we note that any Y_i with a small p-value can be regarded as an outlier (anomaly).

The above method is exact. We can also use a simpler, asympotic approach. With $Z_{(k)}$ defined above, set $\widehat{C} = \{y : \ \widehat{p}(y) \geq t\}$ where now $t = Z_{(k)}$. From Cadre, Pelletier and Pudlo (2013) we have that

$$\sqrt{nh^d}\ \mu(\widehat{C} \Delta C) \xrightarrow{P} c$$

for some constant c where C is the true $1-\alpha$ level set. Hence, $P(Y_{n+1} \in \widehat{C}) = 1 - \alpha + o_P(1)$.

10 Manifolds and Singularities

Sometimes a distribution is concentrated near a lower-dimensional set. This causes problems for density estimation. In fact the density, as we usually think of it, may not be defined.

As a simple example, suppose P is supported on the unit circle in $\mathbb{R}^2$. The distribution P is *singular* with respect to Lebesgue measure μ. This means that there are sets A with $P(A) > 0$ even though $\mu(A) = 0$. Effectively, this means that the density is infinite. To see this, consider a point x on the circle. Let $B(x, \epsilon)$ be a ball of radius ϵ centered at x. Then

$$p(x) = \lim_{\epsilon \to 0} \frac{\mathbb{P}(B(x,\epsilon))}{\mu(B(x,\epsilon))} \to \infty. \qquad \textbf{H}$$

Note also that the L_2 loss does not make any sense. If you tried to use cross-validation, you would find that the estimated risk is minimized at $h = 0$. **H**

A simple solution is to focus on estimating the smoothed density $p_h(x)$ which is well-defined for every $h > 0$. More sophisticated ideas are based on topological data analysis which we discuss later in the course.

11 Series Methods

We have emphasized kernel density estimation. There are many other density estimation methods. Let us briefly mention a method based on basis functions. For simplicity, suppose that $X_i \in [0, 1]$ and let $\phi_1, \phi_2, \ldots$ be an orthonormal basis for

$$\mathcal{F} = \{f : [0,1] \to \mathbb{R}, \ \int_0^1 f^2(x)dx < \infty\}.$$

Thus

$$\int \phi_j^2(x)dx = 1, \qquad \int \phi_j(x)\phi_k(x)dx = 0.$$

An example is the cosine basis:

$$\phi_0(x) = 1, \quad \phi_j(x) = \sqrt{2}\,\cos(2\pi jx), \ j = 1, 2, \ldots,$$

If $p \in \mathcal{F}$ then

$$p(x) = \sum_{j=1}^{\infty} \beta_j \phi_j(x)$$

where $\beta_j = \int_0^1 p(x)\phi_j(x)dx$. An estimate of p is $\widehat{p}(x) = \sum_{j=1}^{k} \widehat{\beta}_j \phi_j(x)$ where

$$\widehat{\beta}_j = \frac{1}{n}\sum_{i=1}^{n} \phi_j(X_i).$$

The number of terms k is the smoothing parameter and can be chosen using cross-validation.

It can be shown that

$$R = \mathbb{E}[\int (\widehat{p}(x) - p(x))^2 dx] = \sum_{j=1}^{n} \mathrm{Var}(\widehat{\beta}_j) + \sum_{j=k+1}^{\infty} \beta_j^2. \qquad \mathbf{H}$$

The first term is of order $O(k/n)$. To bound the second term (the bias) one usually assumes that p is a *Sobolev space of order* q which means that $p \in \mathcal{P}$ with

$$\mathcal{P} = \left\{ p \in \mathcal{F} : \ p = \sum_j \beta_j \phi_j \ : \ \sum_{j=1}^{\infty} \beta_j^2 j^{2q} < \infty \right\}.$$

In that case it can be shown that

$$R \approx \frac{k}{n} + \left(\frac{1}{k}\right)^{2q}. \qquad \mathbf{H}$$

The optimal k is $k \approx n^{1/(2q+1)}$ with risk

$$R = O\left(\frac{1}{n}\right)^{\frac{2q}{2q+1}}.$$

11.1 L_1 Methods

Here we discuss another approach to choosing h aimed at the L_1 loss. The idea is to select a class of sets $\mathcal{A}$— which we call test sets— and choose h to make $\int_A \widehat{p}_h(x)dx$ close to $P(A)$ for all $A \in \mathcal{A}$. That is, we would like to minimize

$$\Delta(g) = \sup_{A \in \mathcal{A}} \Big| \int_A g(x)dx - P(A) \Big|. \tag{33}$$

VC Classes. Let $\mathcal{A}$ be a class of sets with VC dimension ν. As in section 5.2, split the data X into Y and Z with $\mathcal{P} = \{\widehat{p}_1, \ldots, \widehat{p}_N\}$ constructed from Y. For $g \in \mathcal{P}$ define

$$\Delta_n(g) = \sup_{A \in \mathcal{A}} \left| \int_A g(x)dx - P_n(A) \right|$$

where $P_n(A) = n^{-1} \sum_{i=1}^n I(Z_i \in A)$. Let $\widehat{p} = \text{argmin}_{g \in \mathcal{P}} \Delta_n(g)$.

Theorem 17 *For any $\delta > 0$ there exists c such that*

$$\mathbb{P}\left(\Delta(\widehat{p}) > \min_j \Delta(\widehat{p}_j) + 2c\sqrt{\frac{\nu}{n}} \right) < \delta.$$

Proof. We know that

$$\mathbb{P}\left(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > c\sqrt{\frac{\nu}{n}} \right) < \delta.$$

Hence, except on an event of probability at most δ, we have that

$$\begin{aligned} \Delta_n(g) &= \sup_{A \in \mathcal{A}} \left| \int_A g(x)dx - P_n(A) \right| \leq \sup_{A \in \mathcal{A}} \left| \int_A g(x)dx - P(A) \right| + \sup_{A \in \mathcal{A}} \left| P_n(A) - P(A) \right| \\ &\leq \Delta(g) + c\sqrt{\frac{\nu}{n}}. \end{aligned}$$

By a similar argument, $\Delta(g) \leq \Delta_n(g) + c\sqrt{\frac{\nu}{n}}$. Hence, $|\Delta(g) - \Delta_n(g)| \leq c\sqrt{\frac{\nu}{n}}$ for all g. Let $p_* = \text{argmin}_{g \in \mathcal{P}} \Delta(g)$. Then,

$$\Delta(p) \leq \Delta(\widehat{p}) \leq \Delta_n(\widehat{p}) + c\sqrt{\frac{\nu}{n}} \leq \Delta_n(p_*) + c\sqrt{\frac{\nu}{n}} \leq \Delta(p_*) + 2c\sqrt{\frac{\nu}{n}}.$$

□

The difficulty in implementing this idea is computing and minimizing $\Delta_n(g)$. Hjort and Walker (2001) presented a similar method which can be practically implemented when $d = 1$.

Yatracos Classes. Devroye and Györfi (2001) use a class of sets called a Yatracos class which leads to estimators with some remarkable properties. Let $\mathcal{P} = \{p_1, \ldots, p_N\}$ be a set of densities and define the Yatracos class of sets $\mathcal{A} = \Big\{ A(i,j) : \ i \neq j \Big\}$ where $A(i,j) = \{x : p_i(x) > p_j(x)\}$. Let

$$\widehat{p} = \text{argmin}_{g \in \mathcal{G}} \Delta(g)$$

where

$$\Delta_n(g) = \sup_{A \in \mathcal{A}} \left| \int_A g(u)du - P_n(A) \right|$$

and $P_n(A) = n^{-1} \sum_{i=1}^n I(Z_i \in A)$ is the empirical measure based on a sample $Z_1, \ldots, Z_n \sim p$.

Theorem 18 *The estimator $\widehat{p}$ satisfies*

$$\int |\widehat{p} - p| \leq 3 \min_j \int |p_j - p| + 4\Delta \tag{34}$$

where $\Delta = \sup_{A \in \mathcal{A}} \left| \int_A p - P_n(A) \right|$.

The term $\min_j \int |p_j - p|$ is like a bias while term Δ is like the variance.

Proof. Let i be such that $\widehat{p} = p_i$ and let s be such that $\int |p_s - p| = \min_j \int |p_j - p|$. Let $B = \{p_i > p_s\}$ and $C = \{p_s > p_i\}$. Now,

$$\int |\widehat{p} - p| \leq \int |p_s - p| + \int |p_s - p_i|. \tag{35}$$

Let $\mathcal{B}$ denote all measurable sets. Then,

$$\begin{aligned}
\int |p_s - p_i| &= 2 \max_{A \in \{B,C\}} \left| \int_A p_i - \int_A p_s \right| \leq 2 \sup_{A \in \mathcal{A}} \left| \int_A p_i - \int_A p_s \right| \\
&\leq 2 \sup_{A \in \mathcal{A}} \left| \int_A p_i - P_n(A) \right| + 2 \sup_{A \in \mathcal{A}} \left| \int_A p_s - P_n(A) \right| \\
&\leq 4 \sup_{A \in \mathcal{A}} \left| \int_A p_s - P_n(A) \right| \\
&\leq 4 \sup_{A \in \mathcal{A}} \left| \int_A p_s - \int_A p \right| + 4 \sup_{A \in \mathcal{A}} \left| \int_A p - P_n(A) \right| \\
&= 4 \sup_{A \in \mathcal{A}} \left| \int_A p_s - \int_A p \right| + 4\Delta \leq 4 \sup_{A \in \mathcal{B}} \left| \int_A p_s - \int_A p \right| + 4\Delta \\
&= 2 \int |p_s - p| + 4\Delta.
\end{aligned}$$

The result follows from (35). □

Now we apply this to kernel estimators. Again we split the data X into two halves $Y = (Y_1, \ldots, Y_n)$ and $Z = (Z_1, \ldots, Z_n)$. For each h let

$$\widehat{p}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{\|x - Y_i\|}{h} \right).$$

Let

$$\mathcal{A} = \Big\{ A(h, \nu) : \ h, \nu > 0, \ h \neq \nu \Big\}$$

where $A(h, \nu) = \{x : \ \widehat{p}_h(x) > \widehat{p}_\nu(x)\}$. Define

$$\Delta_n(g) = \sup_{A \in \mathcal{A}} \left| \int_A g(u) du - P_n(A) \right|$$

where $P_n(A) = n^{-1} \sum_{i=1}^n I(Z_i \in A)$ is the empirical measure based on Z. Let

$$\widehat{p} = \text{argmin}_{g \in \mathcal{G}} \Delta(g).$$

Under some regularity conditions on the kernel, we have the following result.

Theorem 19 *(Devroye and Györfi, 2001.) The risk of $\widehat{p}$ satisfies*

$$\mathbb{E} \int |\widehat{p} - p| \leq c_1 \inf_h \mathbb{E} \int |\widehat{p}_h - p| + c_2 \sqrt{\frac{\log n}{n}}. \tag{36}$$

The proof involves showing that the terms on the right hand side of (34) are small. We refer the reader to Devroye and Györfi (2001) for the details.

Recall that $d_{TV}(P, Q) = \sup_A |P(A) - Q(A)| = (1/2) \int |p(x) - q(x)| dx$ where the supremum is over all measurable sets. The above theorem says that the estimator does well in the total variation metric, even though the method only used the Yatracos class of sets. Finding computationally efficient methods to implement this approach remains an open question.

12 Mixtures

Another approach to density estimation is to use mixtures. We will discuss mixture modelling when we discuss clustering.

13 Two-Sample Hypothesis Testing

Density estimation can be used for two sample testing. Given $X_1, \ldots, X_n \sim p$ and $Y_1, \ldots, Y_m \sim q$ we can test $H_0 : p = q$ using $\int (\widehat{p} - \widehat{q})^2$ as a test statistic. More interestingly, we can test locally $H_0 : p(x) = q(x)$ at each x. See Duong (2013) and Kim, Lee and Lei (2018). Note that under H_0, the bias cancels from $\widehat{p}(x) - \widehat{q}(x)$. Also, some sort of multiple testing correction is required.

14 Functional Data and Quasi-Densities

In some problems, X is not just high dimensional, it is infinite dimensional. For example suppose that each X_i is a curve. An immediate problem is that the concept of a density is no longer well defined. On a Euclidean space, the density p for a probability measure is

the function that satisfies $P(A) = \int_A p(u)d\mu(u)$ for all measurable A where μ is Lebesgue measure. Formally, we say that p is the Radon-Nikodym derivative of P with respect to the dominating measure μ. Geometrically, we can think of p as

$$p(x) = \lim_{\epsilon \to 0} \frac{\mathbb{P}(\|X - x\| \leq \epsilon)}{V(\epsilon)}$$

where $V(\epsilon) = \epsilon^d \pi^{d/2}/\Gamma(d/2+1)$ is the volume of a sphere of radius ϵ. Under appropriate conditions, these two notions of density agree. (This is the Lebesgue density theorem.)

When the outcome space $\mathcal{X}$ is a set of curves, there is no dominating measure and hence there is no density. Instead, we define the density geometrically by

$$q_\epsilon(x) = \mathbb{P}(\xi(x, X) \leq \epsilon)$$

for a small ϵ where ξ is some metric on $\mathcal{X}$. However we cannot divide by $V(\epsilon)$ and let ϵ tend to 0 since the dimension d is infinite.

One way around this is to use a fixed ϵ and work with the unnormalized density q_ϵ. For the purpose of finding high-density regions this may be adequate. An estimate of q_ϵ is

$$\widehat{q}_\epsilon(x) = \frac{1}{n}\sum_{i=1}^{n} I(\xi(x, X_i) \leq \epsilon).$$

An alternative is to expand X_i into a basis: $X(t) \approx \sum_{j=1}^{k} \beta_j \psi_j(t)$. A density can be defined in terms of the β_j's.

Example 20 *Figure 6 shows the tracks (or paths) of 40 North Atlantic tropical cyclones (TC). The full dataset, consisting of 608 from 1950 to 2006 is shown in Figure 7. Buchman, Lee and Schafer (2009) provide a thorough analysis of the data. We refer the reader to their paper for the full details.*[4]

Each data point— that is, each TC track— is a curve in $\mathbb{R}^2$. Various questions are of interest: Where are the tracks most common? Is the density of tracks changing over time? Is the track related to other variables such as windspeed and pressure?

Each curve X_i can be regarded as mapping $X_i : [0, T_i] \to \mathbb{R}^2$ where $X_i(t) = (X_{i1}(t), X_{i2}(t))$ is the position of the TC at time t and T_i is the lifelength of the TC. Let

$$\Gamma_i = \Big\{(X_{i1}(t), X_{i2}(t)) : \ 0 \leq t \leq T_i\Big\}$$

be the graph of X_i. In other words, Γ_i is the track, regarded as a subset of points in $\mathbb{R}^2$. We will use the Hausdorff metric to measure the distance between curves. The Hausdorff

[4]Thanks to Susan Buchman for providing the data.

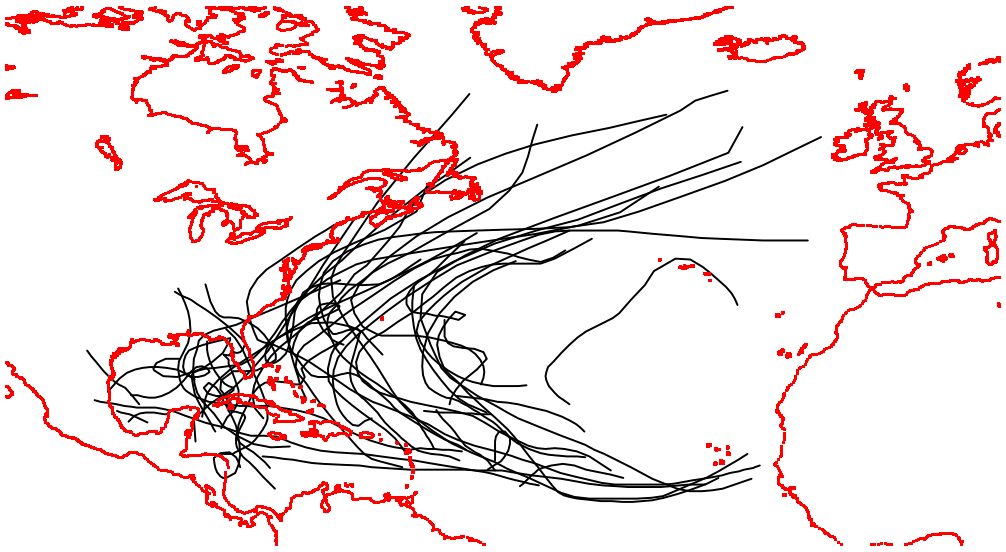

Figure 6: Paths of 40 tropical cyclones in the North Atlantic.

distance between two sets A is B is

$$d_H(A,B) = \inf\{\epsilon : \ A \subset B^\epsilon \ \text{ and } \ B \subset A^\epsilon\} \tag{37}$$

$$= \max\left\{\sup_{x\in A}\inf_{y\in B}\|x-y\|, \ \sup_{x\in B}\inf_{y\in A}\|x-y\|\right\} \tag{38}$$

where $A^\epsilon = \bigcup_{x\in A} B(x,\epsilon)$ is called the enlargement of A and $B(x,\epsilon) = \{y : \ \|y-x\| \le \epsilon\}$. We use the unnormalized kernel estimator

$$\widehat{q}_\epsilon(\gamma) = \frac{1}{n}\sum_{i=1}^{n} I(d_H(\gamma,\Gamma_i) \le \epsilon).$$

Figure 8 shows the 10 TC's with highest local density and the the 10 TC's with lowest local density using $\epsilon = 16.38$. This choice of ϵ corresponds to the 10th percentile of the values $\{d_H(X_i,X_j) : \ i \neq j\}$. The high density tracks correspond to TC's in the gulf of Mexico with short paths. The low density tracks correspond to TC's in the Atlantic with long paths.

15 Miscellanea

Another method for selecting h which is sometimes used when p is thought to be very smooth is the plug-in method. The idea is to take the formula for the mean squared error (equation 31), insert a guess of p'' and then solve for the optimal bandwidth h. For example, if $d = 1$

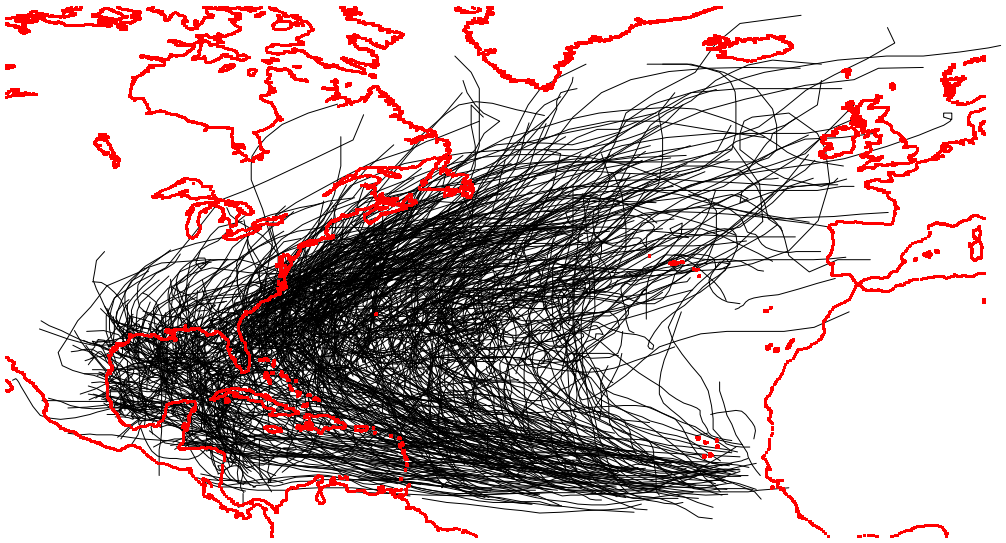

Figure 7: Paths of 608 tropical cyclones in the North Atlantic.

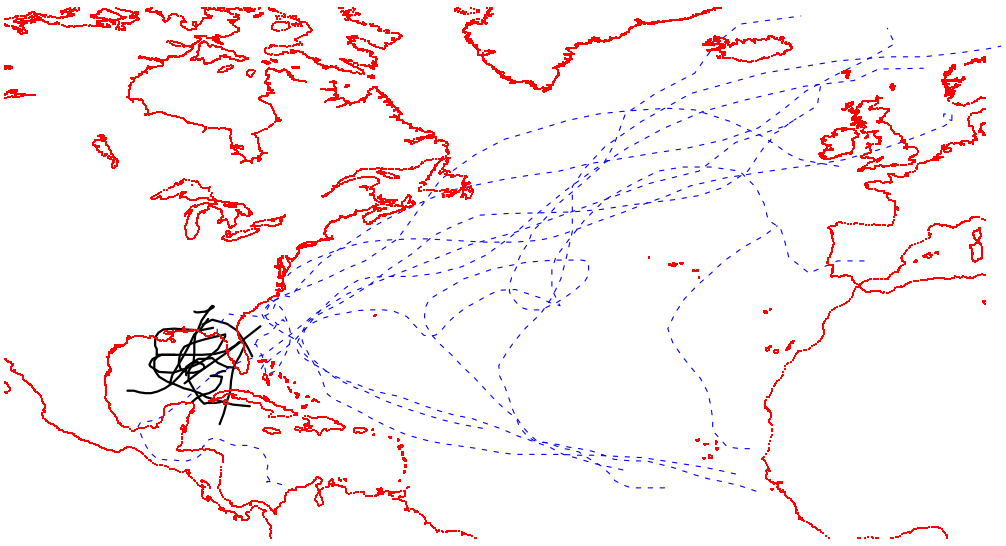

Figure 8: 10 highest density paths (black) and 10 lowest density paths (blue).

and under the idealized assumption that p is a univariate Normal this yields $h_* = 1.06\sigma n^{-1/5}$. Usually, σ is estimated by $\min\{s, Q/1.34\}$ where s is the sample standard deviation and Q is the interquartile range.[5] This choice of h_* works well if the true density is very smooth and is called the Normal reference rule.

Since we don't want to necessarily assume that p is very smooth, it is usually better to estimate h using cross-validation. See Loader (1999) for an interesting comparison between cross-validation and plugin methods.

A generalization of the kernel method is to use adaptive kernels where one uses a different bandwidth $h(x)$ for each point x. One can also use a different bandwidth $h(x_i)$ for each data point. This makes the estimator more flexible and allows it to adapt to regions of varying smoothness. But now we have the very difficult task of choosing many bandwidths instead of just one.

Density estimation is sometimes used to find unusual observations or outliers. These are observations for which $\widehat{p}(X_i)$ is very small.

16 Summary

1. A commonly used nonparametric density estimator is the kernel estimator

$$\widehat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{\|x - X_i\|}{h}\right).$$

2. The kernel estimator is rate minimax over certain classes of densities.

3. Cross-validation methods can be used for choosing the bandwidth h.

[5] Recall that the interquartile range is the 75th percentile minus the 25th percentile. The reason for dividing by 1.34 is that $Q/1.34$ is a consistent estimate of σ if the data are from a $N(\mu, \sigma^2)$.