Nonparametric Density Estimation
10716: Advanced Machine Learning
Pradeep Ravikumar (amending notes from Larry Wasserman)

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 typically depend on a tuning parameter h, and choosing h carefully is crucial. To emphasize the dependence on h we sometimes write $\widehat{p}_h$.

A very simple non-parametric distribution estimator is simply the empirical distribution:

$$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i},$$

but this is not very suitable as an estimate of the underlying distribution. It "overfits" to the training data by placing all probability mass on the given training points $\{X_i\}_{i=1}^n$ and zero mass even on very nearby points. It moreover does not have a density. So usually by nonparameteric density estimation, we mean something that does a bit more, in particular by "smoothing" the empirical distribution $\mathbb{P}_n$. For this reason, nonparametric density estimation is also often referred to as smoothing.

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.

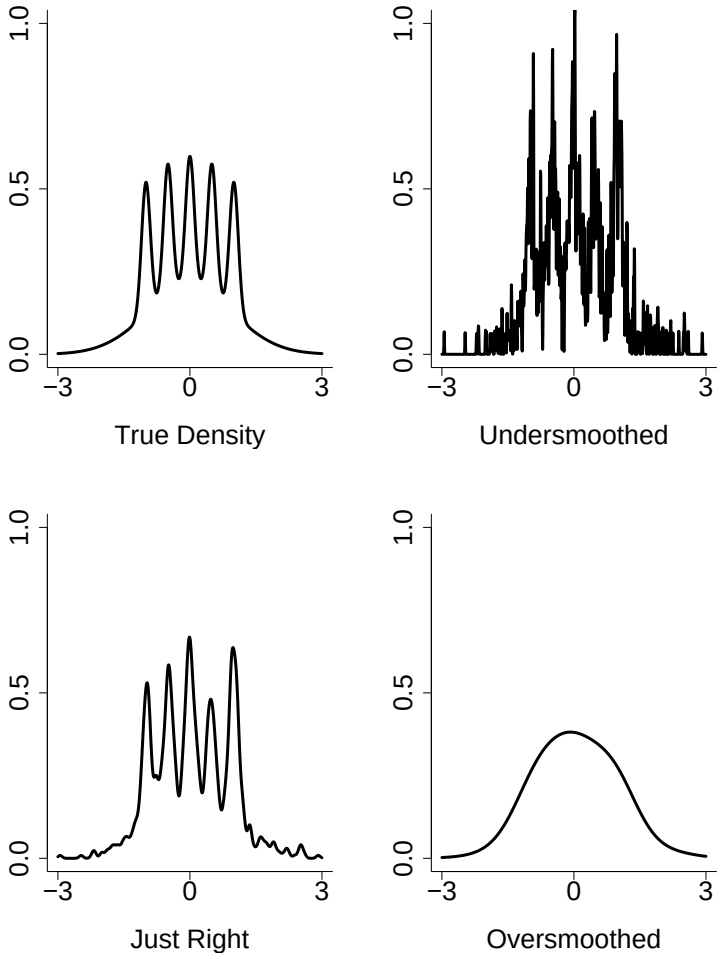

Figure 1: The Bart Simpson density from Example 1. Top left: true density. The other plots are kernel estimators based on $n = 1{,}000$ draws. Bottom left: bandwidth $h = 0.05$ chosen by leave-one-out cross-validation. Top right: bandwidth $h/10$. Bottom right: bandwidth $10h$.

2 Applications

Density estimation could be used for sampling new points (see the outpouring of creative, and perhaps even worrying, uses of such sampling in the context of images and text), and more generally, for a compact summary of data useful for downstream probabilistic reasoning. It can also be used in particular for regression, classification, and clustering. Suppose $\widehat{p}(x, y)$ is an estimate of $p(x, y)$.

Regression. We can then compute the following estimate of the regression function:

$$\begin{aligned}\widehat{m}(x) &= \int y\widehat{p}(y|x)dy \\ &= \int y\frac{\widehat{p}(y,x)}{\widehat{p}(x)}dy.\end{aligned}$$

Classification. For classification, recall the Bayes optimal classifier

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

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 classifier. Many classifiers that you are familiar with can be re-expressed this way.

Clustering. For clustering, we look for the high density regions, based on an estimate of the density. We will discuss more on this when we discuss clustering.

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

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 $D(\widehat{p}, \widehat{q})$, for some divergence D as a test statistic.

3 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 $R(p, \widehat{p}) = \mathbb{E}(L(p, \widehat{p}))$.

A key advantage of the L_2 loss is that the risk has a very mathematically convenient decomposition:

$$R(p, \widehat{p}) = \mathbb{E} \int (p(x) - \widehat{p}(x))^2 dx \tag{2}$$

$$= \int b_n^2(x) dx + \int v_n(x) dx \tag{3}$$

where $b_n(x) = \mathbb{E}(\widehat{p}(x)) - p(x)$ is the bias and $v(x) = \mathrm{Var}(\widehat{p}(x))$ is the variance.

The estimator $\widehat{p}$ typically involves "smoothing" the empirical distribution in some way. The main challenge is to determine how much smoothing to do. When the data are oversmoothed, the bias term is large and the variance is small. When the data are undersmoothed the opposite is true. This is called the *bias–variance tradeoff*. Minimizing risk corresponds to balancing bias and variance.

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

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

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

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

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, due to the density ratios.

The *minimax risk* over a class of densities $\mathcal{P}$ is

$$R_n(\mathcal{P}) = \inf_{\widehat{p}} \sup_{p \in \mathcal{P}} R(p, \widehat{p}) \tag{4}$$

and an estimator is *minimax* if its risk is equal to the minimax risk. We say that $\widehat{p}$ is *rate optimal* if

$$R(p, \widehat{p}) \asymp R_n(\mathcal{P}). \tag{5}$$

Typically the minimax rate is of the form $n^{-C/(C+d)}$ for some $C > 0$.

4 Function Spaces

A distinguishing characteristic of "non-parametric" methods is that what we are estimating is not in a finite-dimensional parametric space. Typically, it is in some infinite-dimensional function space. We briefly review some classical function spaces.

The class of Lipschitz functions $H(1, L)$ on $\mathcal{X} \subset \mathbb{R}$ is the set of functions g such that

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

A differentiable function is Lipschitz if and only if it has bounded derivatives. Conversely a Lipschitz function is differentiable almost everywhere.

Let $\mathcal{X} \subset \mathbb{R}$ and let β be an integer. The Holder space $H(\beta, L)$ is the set of functions g mapping $\mathcal{X}$ to $\mathbb{R}$ such that g is $\ell = \beta - 1$ times differentiable and satisfies

$$|g^{(\ell)}(y) - g^{(\ell)}(x)| \leq L|x - y| \quad \text{for all } x, y \in T.$$

A more intuitive perspective of this class is that its first β derivatives are all bounded.

A yet another perspective of this class is as a set of functions that are close to their Taylor series approximation upto order β. If $g \in H(\beta, L)$ and $\ell = \beta - 1$, then we can define the Taylor approximation of g at x by

$$\widetilde{g}(y) = g(y) + (y - x)g'(x) + \cdots + \frac{(y - x)^\ell}{\ell!} g^{(\ell)}(x)$$

and then $|g(y) - \widetilde{g}(y)| \leq L|y - x|^\beta$.

The definition for higher dimensions is similar. Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$.

Given a vector $s = (s_1, \ldots, s_d)$, define

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

as the s-th partial derivative. We will also use the compact notation $|s| = s_1 + \cdots + s_d$, $s! = s_1! \cdots s_d!$, $x^s = x_1^{s_1} \cdots x_d^{s_d}$.

Let β and L be positive integers. The Hölder class is then defined as:

$$H(\beta, L) = \left\{ p : \ |D^s p(x) - D^s p(y)| \leq L\|x-y\|, \quad \text{for all } s \text{ such that } |s| = \beta - 1, \text{ and all } x, y \right\}. \tag{6}$$

For example, if $d = 1$ and $\beta = 2$ (which is the most common setting) this means that

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

As before, we could also view this class as functions with bounded D^s partial derivatives, for $|s| \leq \beta$. For instance, with $\beta = 2$, the class consists of functions have bounded second derivatives.

And as before, this function class comprises functions that are close to their Taylor series approximation upto order β. Let

$$p_{x,\beta}(u) = \sum_{|s|<\beta} \frac{(u-x)^s}{s!} D^s p(x). \tag{7}$$

Then, if $p \in H(\beta, L)$, we can show that: $p(x)$ is close to its .

$$|p(u) - p_{x,\beta}(u)| \leq L\|u - x\|^\beta. \tag{8}$$

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

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

4.1 Categories of Nonparametric Density Estimators

We will discuss two broad categories of nonparametric density estimators: (a) those based on hard partitioning of the input space viz. histograms (technically not *density* estimators), and soft-partitioning of the input space viz. kernel density estimators, and (b) those based on projection onto an infinite-dimensional dimensional function space, where we will look at a particular instance called series estimators.

5 Histograms

Perhaps the simplest nonparametric distribution estimators, after the empirical distribution, are histograms. The high level idea is to discretize the data, and then simply use the MLE

of the resulting categorical distribution (which is simply the frequencies of each category in the data).

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 $N = (1/h)^d$ such bins and each has volume h^d. Denote the bins by $B_1, \ldots, B_N$. Now we can write the true density

$$p(x) = \sum_{j=1}^{N} P(X \in B_j)\, p(x|X \in B_j).$$

We can estimate $P(X \in B_j)$ via

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

as the fraction of data points in bin B_j. While we can approximate $p(x|X \in B_j)$ via the density of the uniform distribution over the bin B_j so that $p(x|X \in B_j) = 1/h^d \mathbb{I}(x \in B_j)$. Plugging these two values in, we get the histogram density estimator:

$$\widehat{p}_h(x) = \sum_{j=1}^{N} \frac{\widehat{\theta}_j}{h^d}\, I(x \in B_j). \tag{9}$$

5.1 Statistical Analysis: Histograms

Suppose that $p \in \mathcal{P}(L) := H(1, L)$ where

$$H(1,L) = \left\{ p : \ |p(x) - p(y)| \leq L\|x - y\|, \text{ for all } x, y \right\}. \tag{10}$$

Theorem 2 *The L_2 risk of the histogram estimator is bounded by*

$$\sup_{p \in H(1,L)} R(p, \widehat{p}) = \int (\mathbb{E}(\widehat{p}_h(x) - p(x))^2 \leq L^2 h^2 d + \frac{C}{nh^d}. \tag{11}$$

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,

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

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

The rate of convergence $n^{-2\beta/(2\beta+d)}$ is slow when the dimension d is large. The typical rate of convergence for parameter models is typically $d/\sqrt{n}$. To see the difference between these two rates, to get to ϵ error with non-parametric rates, we would require number of samples scaling as $n^{-2\beta/(2\beta+d)} \leq \epsilon \; n \geq (1/\epsilon)^{d/2\beta+1} = O(1/\epsilon)^d$, which scales exponentially with the dimension d. On the other hand, for parametric rates, $d/\sqrt{n} \leq \epsilon$ only requires that $n \geq (d/\epsilon)^2$, which only scales polynomially with the dimension.

This upper bound can also be shown to be tight. Specifically:

Theorem 3 *There exists a constant $C > 0$ such that*

$$\inf_{\widehat{p}} \sup_{P \in H(1,L)} \mathbb{E} \int (\widehat{p}(x) - p(x))^2 dx \geq C \left(\frac{1}{n}\right)^{\frac{2}{d+2}}. \tag{12}$$

The above result showed that the histogram estimator is close (wrt ℓ_2 loss) to the true density **in expectation**. A more powerful result would be to show that it is close with high probability. This entails analyzing

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

where $\|f\|_\infty = \sup_x |f(x)|$.

Theorem 4 *With probability at least $1 - \delta$,*

$$\|\widehat{p}_h - p\|_\infty \leq \sqrt{\frac{1}{cnh^d} \log\left(\frac{2}{\delta h^d}\right)} + L\sqrt{d}h. \tag{13}$$

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

$$\|\widehat{p}_h - p\|_\infty \leq \sqrt{c^{-1} n^{-\frac{2}{2+d}} \left[\log\left(\frac{2}{\delta}\right) + \left(\frac{2}{2+d}\right) \log n\right]} + L\sqrt{d} n^{-\frac{1}{2+d}} = O\left(\left(\frac{\log n}{n}\right)^{\frac{1}{2+d}}\right). \tag{14}$$

5.2 Adaptive histograms: Density Trees

Instead of uniformly partitioning the input domain, one can adaptively partition it. 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. 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.

6 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 $\sigma_K^2 \equiv \int x^2 K(x)dx > 0$. *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 $K(\|x\|)$. For presentational simplicity, we will overload notation for both the multivariate and univariate kernels, and if not specified, for vector x, we will use $K(x) = K(\|x\|)$.

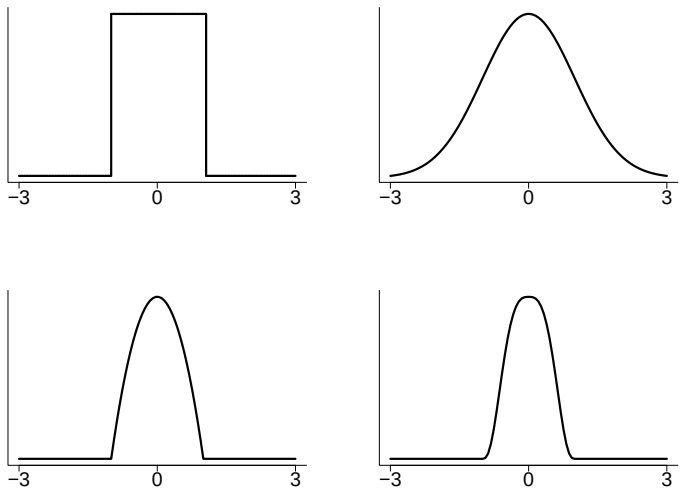

Figure 2: Examples of smoothing kernels: boxcar (top left), Gaussian (top right), Epanechnikov (bottom left), and tricube (bottom right).

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

$$\widehat{p}(x) = \frac{1}{n}\sum_{i=1}^n \frac{1}{h^d} K\left(\frac{x - X_i}{h}\right). \tag{15}$$

More generally, we define

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

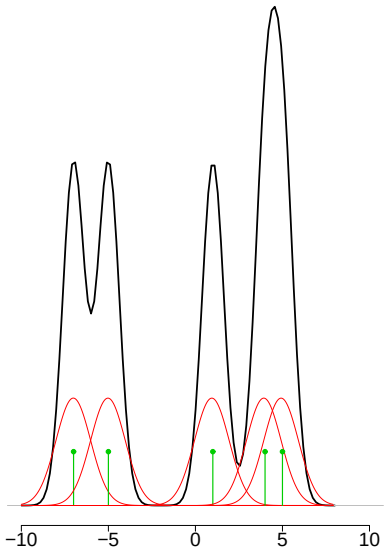

Figure 3: A kernel density estimator $\widehat{p}$. At each point x, $\widehat{p}(x)$ is the average of the kernels centered over the data points X_i. The data points are indicated by short vertical bars. The kernels are not drawn to scale.

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.

6.1 Statistical Analysis: Kernel Estimators

In this section we examine the performance 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.

Conditions on Kernel Function. In order for the kernel density estimate to be able to estimate well a smooth function in $H(\beta, L)$ for $\beta > 2$, we need a "higher order kernel".

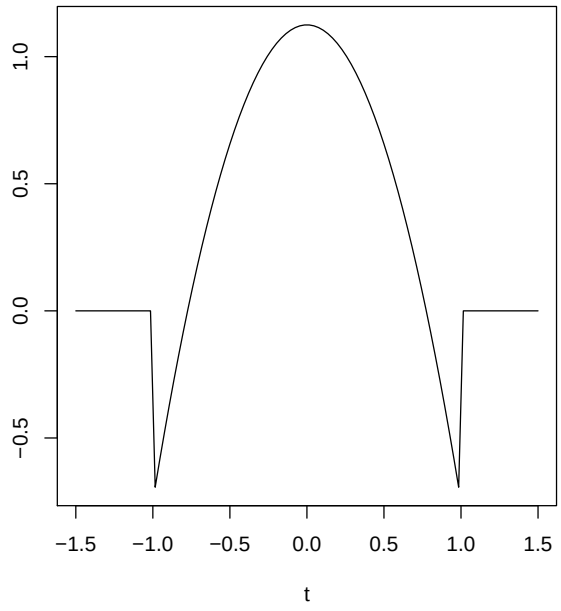

Figure 4: *A higher-order kernel function: specifically, a kernel of order 4*

Assume now that the kernel K has the form $K(x) = k(\|x\|)$ for some univariate kernel k that has support on $[-1, 1]$. A univariate kernel is said to have order β provided that: $\int k = 1$, $\int |k|^q < \infty$ for any $q \geq 1$, $\int |t|^\beta |k(t)|dt < \infty$ and $\int t^s k(t)dt = 0$ for $s < \beta$. An example of a kernel that satisfies these conditions for $\beta = 2$ is $k(x) = (3/4)(1 - x^2)$ 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; because of which such "higher order kernels" for $\beta > 2$ are not that popular. For example, a 4th-order kernel is $K(t) = \frac{3}{8}(3 - 5t^2)1\{|t| \leq 1\}$, plotted in Figure 4. Notice that it takes 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 5 *The bias of $\widehat{p}_h$ satisfies:*

$$\sup_{p \in H(\beta, L)} |p_h(x) - p(x)| \leq ch^\beta \tag{16}$$

for some c.

Next we bound the variance.

Lemma 6 *The variance of $\widehat{p}_h$ satisfies:*

$$\sup_{p \in H(\beta, L)} \text{Var}(\widehat{p}_h(x)) \leq \frac{c}{nh^d} \tag{17}$$

for some $c > 0$.

Since the mean squared error is equal to the variance plus the bias squared, together the previous two lemmas yield:

Theorem 7 *The L_2 risk is bounded above, uniformly over $H(\beta, L)$, as*

$$\sup_{p \in H(\beta,L)} \mathbb{E} \int (\widehat{p}_h(x) - p(x))^2 dx \preceq h^{2\beta} + \frac{1}{nh^d} \tag{18}$$

If $h \asymp n^{-1/(2\beta+d)}$ then

$$\sup_{p \in H(\beta,L)} \mathbb{E} \int (\widehat{p}_h(x) - p(x))^2 dx \preceq \left(\frac{1}{n}\right)^{\frac{2\beta}{2\beta+d}}. \tag{19}$$

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

6.2 Minimax Lower 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 8 *There exists C depending only on β and L such that*

$$\inf_{\widehat{p}} \sup_{p \in H(\beta,L)} \mathbb{E}_p \int (\widehat{p}(x) - p(x))^2 dx \geq C \left(\frac{1}{n}\right)^{\frac{2\beta}{2\beta+d}}. \tag{20}$$

Theorem 8 together with (19) imply that kernel estimators are rate minimax.

Concentration Analysis of Kernel Density Estimator Now we state a result which says how fast $\widehat{p}(x)$ concentrates around $p(x)$.

Theorem 9 *For all small $\epsilon > 0$,*

$$\mathbb{P}(|\widehat{p}(x) - p_h(x)| > \epsilon) \leq 2 \exp\left\{-cnh^d\epsilon^2\right\}. \tag{21}$$

Hence, for any $\delta > 0$,

$$\sup_{p \in H(\beta,L)} \mathbb{P}\left(|\widehat{p}(x) - p(x)| > \sqrt{\frac{C \log(2/\delta)}{nh^d}} + ch^\beta\right) < \delta \tag{22}$$

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

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

The first statement follows from an application of Bernstein's inequality. While the last statement follows from bias-variance calculations followed by Markov's inequality.

Concentration in L_∞. While Theorem 9 shows that, for each x, $\widehat{p}(x)$ is close to $p(x)$ with high probability; it would be nice to have 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,

$$\mathbb{P}(\|\widehat{p}_h - p_h\|_\infty > \epsilon) \leq 4\left(\frac{C}{h^{d+1}\epsilon}\right)^d \exp\left(-\frac{3n\epsilon^2 h^d}{28K(0)}\right). \tag{23}$$

A more sophisticated analysis in Giné and Guillou (2002) (which in turn replaces Bernstein's inequality in previous proof with a more sophisticated inequality due to Talagrand) yields the following:

Theorem 10 *Suppose that $p \in H(\beta, L)$. 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

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

6.3 Boundary Bias

One caveat with the kernel density estimator is what happens near the boundary of the sample space. If x is $O(h)$ close to the boundary, then the bias is $O(h)$ instead of $O(h^2)$. The main reason is that when we compute an average over nearby points; points near the boundary have more points towards directions leading away from the boundary, compared to directions towards the boundary. We will discuss more about this when we cover non-parametric regression.

There are a variety of fixes including: data reflection, transformations, boundary kernels, local likelihood. These are not as popular as simple kernel density estimation however.

6.4 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 11 *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,*

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

and

$$R = \frac{1}{4}\sigma_K^4 h_n^4 \int p''(x)^2 dx + \frac{\int K^2(x)dx}{nh} + O\left(\frac{1}{n}\right) + O(h_n^6) \tag{24}$$

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

If we differentiate (24) with respect to h and set it equal to 0, we see that the asymptotically optimal bandwidth is

$$h_* = \left(\frac{c_2}{c_1^2 A(f) n}\right)^{1/5} \tag{25}$$

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 (24), we see that if the optimal bandwidth is used then $R = O(n^{-4/5})$.

7 Picking Bandwidths of Kernel Estimators

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.

7.1 Leave One Out Cross-Validation

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

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

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

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

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

Definition 12 *The leave-one-out cross-validation estimator of risk is*

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

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

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

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

Theorem 13 (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,*

$$\frac{\int \left(p(x) - \widehat{p}_{\widehat{h}}(x)\right)^2 dx}{\inf_h \int \left(p(x) - \widehat{p}_h(x)\right)^2 dx} \xrightarrow{a.s.} 1. \tag{27}$$

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.

7.2 V-fold Cross-Validation

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 $\bar{H} = \{h_1, \ldots, h_N\}$ be a finite grid of bandwidths. For $j \in [N]$, denote

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

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

The loss of $\widehat{p}_j$ is given as: $L(p, \widehat{p}_j) = \int \widehat{p}_j^2(x) - 2\int \widehat{p}_j(x)p(x)dx$. Define the estimated risk

$$\widehat{L}_j \equiv \widehat{L}(p, \widehat{p}_j) = \int \widehat{p}_j^2(x) - \frac{2}{n}\sum_{i=1}^{n} \widehat{p}_j(Z_i). \tag{28}$$

Let $\widehat{p} = \text{argmin}_{j\in[N]}\widehat{L}(p, \widehat{p}_j)$. Schematically:

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

Theorem 14 (Wegkamp 1999) *There exists a $C > 0$ such that*

$$\mathbb{E}(\|\widehat{p} - p\|^2) \leq 2\min_{j\in[N]} \mathbb{E}(\|\widehat{p}_j - p\|^2) + \frac{C\log N}{n}.$$

A similar result can be proved for V-fold cross-validation.

7.3 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

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

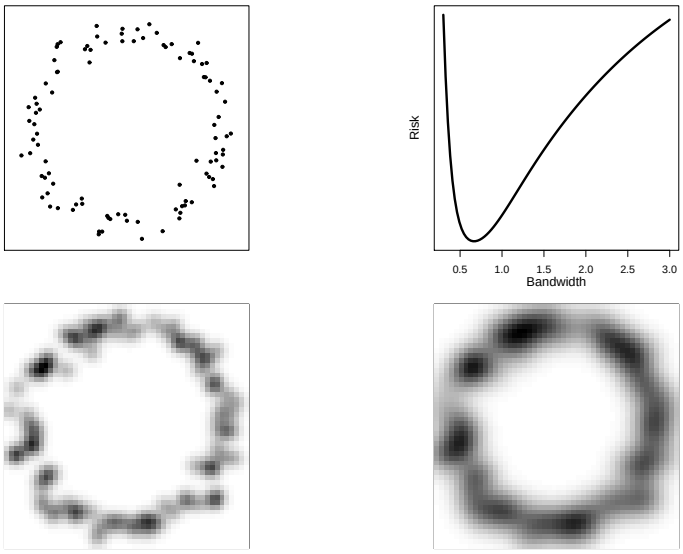

Figure 5: Synthetic two-dimensional data set. Top left: data. Top right: cross-validation function. Bottom left: kernel estimator based on the bandwidth that minimizes the cross-validation score. Bottom right: kernel estimator based on the twice the bandwidth that minimizes the cross-validation score.

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.

7.4 Picking Bandwidths to optimize L_1 instead of L_2 Risk

Here we discuss another approach to choosing h aimed at the L_1 loss. Recall that this L_1 loss between some density g and the true distribution P is given as: $\int |g(x) - p(x)|dx = 2\sup_A \left| \int_A g(x)dx - P(A) \right|$. The idea is to restrict to 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

$$\Delta(g) = \sup_{A \in \mathcal{A}} \left| \int_A g(x)dx - P(A) \right|. \tag{29}$$

Note that this yields an approximation to the L_1 risk, which optimizes over all sets, rather than just some restricted class of sets, so we have to choose these carefully. We will next discuss two approaches to specify these test classes.

7.4.1 VC Classes

Let $\mathcal{A}$ be a class of sets with VC dimension ν. As in section 7.2, 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 $\widehat{p} = \text{argmin}_{j \in [N]} \Delta_n(\widehat{p}_j)$.

Theorem 15 *For any $\delta > 0$ there exists c such that*

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

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 $d = 1$. Another caveat with the above is that $\Delta(g)$ is only an approximation of the L_1 loss, depending on the richness of the class of sets $\mathcal{A}$. Is there a small enough class of sets $\mathcal{A}$ that would be as if minimizing the L_1 loss?

7.4.2 Yatracos Classes

Devroye and Györfi (2001) use such 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 $\mathcal{A} = \left\{ A(i,j) : \ i \neq j \right\}$ where $A(i,j) = \{x : \ p_i(x) > p_j(x)\}$. Let

$$\widehat{p} = \text{argmin}_{j \in [N]} \Delta_n(p_j),$$

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 16 *The estimator $\widehat{p}$ satisfies*

$$\int |\widehat{p} - p| \leq 3 \min_j \int |p_j - p| + 4\Delta \tag{30}$$

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.

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

$$\widehat{p} = \text{argmin}_{p_h} \Delta_n(p_h).$$

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

Theorem 17 *(Devroye and Györfi, 2001.) The risk of $\widehat{p}$ satisfies*

$$\mathbb{E} \int |\widehat{p} - p| \leq c_1 \inf_h \mathbb{E} \int |\widehat{p}_h - p| + c_2 \sqrt{\frac{\log n}{n}}. \tag{31}$$

The proof involves showing that the terms on the right hand side of (30) are small. We refer the reader to Devroye and Györfi (2001) for the details.

Finding computationally efficient methods to implement this approach remains an open question.

8 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

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

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

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

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

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

9 Miscellanea

9.1 High Dimensions, Curse of Dimensionality

As discussed earlier, the non-parametric rate of convergence $n^{-C/(C+d)}$ 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.

9.2 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

$$P_h = P \star K_h$$

where $\star$ 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. We can then show:

Theorem 18 *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.

[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.

9.3 Graphical Models/Conditional Independence based methods

If we can live with some bias, we can reduce the dimensionality by imposing some (conditional) 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.

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.

9.4 Mixtures

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

9.5 Adaptive Kernels

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.

10 Summary

1. We discussed two categories of nonparametric density estimators: partition based (hard-partition based such as histograms, and soft-partition based such as kernel den-

sity estimators), and projection onto function space based (series estimators).

2. Of these, the most commonly used nonparametric density estimator is the kernel density estimator
$$\widehat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{\|x - X_i\|}{h}\right).$$

3. The kernel estimator is rate minimax over many classes of densities.

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

11 Appendix: Proofs

11.1 Proof of Theorem 2

We prove the result by bounding the bias and variance of $\widehat{p}_h$.

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

$$p_h(x) \equiv \mathbb{E}(\widehat{p}_h(x)) = \frac{\theta_j}{h^d} \tag{32}$$

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

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

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

$$\sup_{p\in\mathcal{P}(L)} R(p, \widehat{p}) = \int (\mathbb{E}(\widehat{p}_h(x) - p(x))^2 \leq L^2h^2d + \frac{C}{nh^d}. \tag{33}$$

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

11.2 Proof of Theorem 4

We now derive a concentration result for $\widehat{p}_h$ where 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

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

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

$$\mathbb{P}(|\overline{Y} - \mu| > \epsilon) \leq 2 \exp\left\{-\frac{n\epsilon^2}{2\sigma^2 + 2M\epsilon/3}\right\}. \tag{34}$$

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 $1 - \pi_n$,

$$\|\widehat{p}_h - p\|_\infty \leq \|\widehat{p}_h - p_h\|_\infty + \|p_h - p\|_\infty \leq \epsilon + L\sqrt{d}h. \tag{35}$$

Now set

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

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

$$\|\widehat{p}_h - p\|_\infty \leq \sqrt{\frac{1}{cnh^d}\log\left(\frac{2}{\delta h^d}\right)} + L\sqrt{d}h. \tag{36}$$

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

$$\|\widehat{p}_h - p\|_\infty \leq \sqrt{c^{-1}n^{-\frac{2}{2+d}}\left[\log\left(\frac{2}{\delta}\right) + \left(\frac{2}{2+d}\right)\log n\right]} + L\sqrt{d}n^{-\frac{1}{2+d}} = O\left(\left(\frac{\log n}{n}\right)^{\frac{1}{2+d}}\right). \tag{37}$$

11.3 Proof of Lemma 5 (Bias of Kernel Density Estimators)

We have

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

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

11.4 Proof of Lemma 6 (Variance of Kernel Density Estimators)

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,

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

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

11.5 Proof of Theorem 9 (Concentration of Kernel Density Estimators)

By the triangle inequality,

$$|\widehat{p}(x) - p(x)| \leq |\widehat{p}(x) - p_h(x)| + |p_h(x) - p(x)| \tag{38}$$

where $p_h(x) = \mathbb{E}(\widehat{p}(x))$. From Lemma 5, $|p_h(x) - p(x)| \leq ch^\beta$ 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 6.

Recall Bernstein's inequality: Suppose that $Y_1, \ldots, Y_n$ are iid with mean μ, $\text{Var}(Y_i) \leq \sigma^2$ and $|Y_i| \leq M$. Then

$$\mathbb{P}(|\overline{Y} - \mu| > \epsilon) \leq 2\exp\left\{-\frac{n\epsilon^2}{2\sigma^2 + 2M\epsilon/3}\right\}. \tag{39}$$

Then, 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 (38).

11.6 Proof of Theorem 11 (Asymptotics of Kernel Density Estimators)

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 $\text{Var}[\widehat{p}(x)] = n^{-1}\text{Var}[K_h(x, X)]$. Now,

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

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.

11.7 Proof of Theorem 15 (VC Approximation to L1)

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 $p_* = \mathrm{argmin}_{g\in\mathcal{P}}\Delta(g)$. Then,

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

11.8 Proof of Theorem 16 (Yatracos Approximation to L1)

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,

$$\int |\widehat{p} - p| \leq \int |p_s - p| + \int |p_s - p_i|. \tag{40}$$

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

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

The result follows from (40).