LECTURE 7: NONPARAMETRIC SMOOTHING IN REGRESSION: KERNEL REGRESSION REVISITED

Text reference: Shalizi Chapter 4 and Sections 1.5 and 9.3

Overview (Nonparametric Smoothing): This lecture is intended to give some motivation for nonparametric smoothing in regression, as well as point out some of the associated challenges. It's difficult, in general, to give a precise definition of "nonparametric" inference but the basic idea is to use data to infer an unknown quantity *while making as few assumptions as possible*. Usually, this means using statistical models that are *infinite-dimensional*, versus "parametric" models that can be parametrized by a finite number of parameters.

In the context of regression, if we assume that the (unknown) regression function $r \in \mathcal{F}$ where $\mathcal{F}$ is finite-dimensional — such as, the set of straight lines, $\mathcal{F}_{lin} = \{\beta_0 + \beta_1 x : \beta_0, \beta_1 \in \mathbb{R}\}$ — then we have a *parametric regression model*. If we assume that $r \in \mathcal{F}$ where $\mathcal{F}$ is not finite-dimensional — such as, the set of all functions that are not "too wiggly" as defined by the Sobolev space $\mathcal{F}_{sob} = \{r : \int (r''(x))^2 dx < \infty\}$ — then we have a *nonparametric regression model*.

As previously discussed, nonparametric regression models are more flexible than parametric models with a smaller bias, but there's a price we pay

in terms of variance, even if we choose the smoothing/tuning parameters in the model in an optimal way. In this lecture, we are going to revisit one nonparametric regression estimator, namely the kernel smoother, and discuss the bias-variance tradeoff for a kernel smoother.

Review: Kernel Regression

Recall our basic setup: We are given n pairs of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, where

$$Y_i = r(X_i) + \epsilon_i, \quad i = 1, \ldots n,$$

and

$$r(x) = \mathbb{E}(Y|X = x) = \int y f(y|x) dy.$$

Our goal is to estimate the unknown regression function r with some function $\widehat{r}$. Assume for now that each $X_i \in \mathbb{R}$ (i.e., the predictors are 1-dimensional).

We have discussed considering $\widehat{r}$ in the class of so-called **linear smoothers**, i.e. regression estimators $\widehat{r}$ which has the form $\widehat{r}(x) = \sum_i \ell_i(x) Y_i$ for some choice of weights $\ell_i(x)$. Indeed, linear regression, k-nearest-neighbors regression and splines are special cases of linear smoothers.

Here we will revisit another important linear smoother, namely **kernel smoothing** a.k.a **kernel regression** or **Nadaraya-Watson regression**, that takes a weighted average of the Y_i's, giving higher weight to those points near x. The starting point is to define a "kernel" function $K : \mathbb{R} \to \mathbb{R}$. For

our purposes, the word **kernel** refers to any (usually smooth) function K such that $K(x) \geq 0$ and

$$\int K(x)\,dx = 1, \quad \int xK(x)dx = 0 \quad \text{and} \quad \sigma_K^2 \equiv \int x^2 K(x)dx > 0.$$

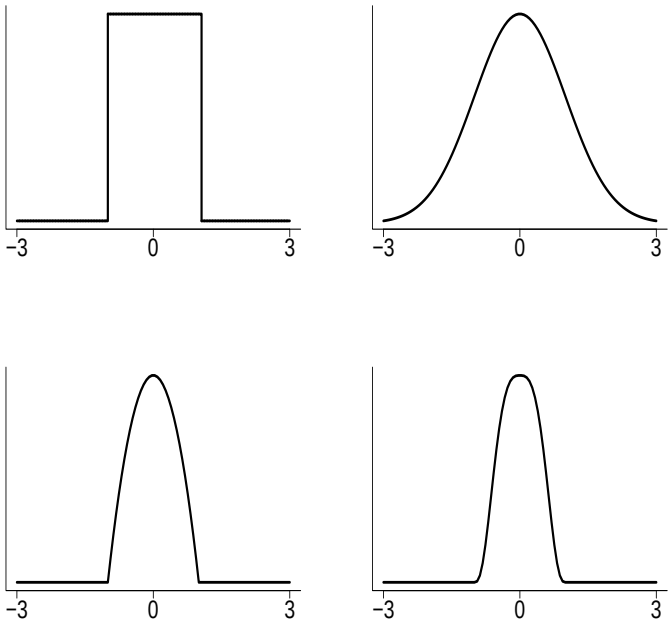

Figure 1: Examples of kernels: boxcar (top left), Gaussian (top right), Epanechnikov (bottom left), and tricube (bottom right).

Some commonly used kernels are the following:

$$\begin{aligned}
\textbf{the boxcar kernel}: \quad & K(x) = \frac{1}{2} I(x), \\
\textbf{the Gaussian kernel}: \quad & K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \\
\textbf{the Epanechnikov kernel}: \quad & K(x) = \frac{3}{4}(1 - x^2) I(x) \\
\textbf{the tricube kernel}: \quad & K(x) = \frac{70}{81}(1 - |x|^3)^3 I(x)
\end{aligned}$$

where

$$I(x) = \begin{cases} 1 & \text{if } |x| \leq 1 \\ 0 & \text{if } |x| > 1. \end{cases}$$

These kernels are plotted in Figure 1.

Let $h > 0$ be a positive number, called the **bandwidth**. The **Nadaraya–Watson kernel estimator** is defined by

$$\widehat{r}_n(x) = \sum_{i=1}^{n} \ell_i(x) Y_i \tag{1}$$

where K is a kernel and the weights $\ell_i(x)$ are given by

$$\ell_i(x) = \frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)}. \tag{2}$$

Think: What does this achieve? What happens to data "close to" versus "far away" from the evaluation point x? What are the differences between NW regression and k-NN regression?

Question: What's in the choice of kernel? Different kernels can give different results. But many of the common kernels tend to produce similar estimators; e.g., Gaussian vs. Epanechnikov, there's not a huge difference.

What does matter much more is the *choice of bandwidth h* which controls the amount of smoothing. What's the tradeoff when we vary h? Hint: as we've mentioned before, you should always keep two quantities in mind...

Bias and Variance of Kernel Smoothers

Conditional on $X = x$, recall how the generalization or prediction error decomposes (**review Lecture 3, Part I**):

$$\begin{aligned} R(x) = \mathbb{E}[\text{TestErr}(\widehat{r}(x))] &= \mathbb{E}\big[\big(Y - \widehat{r}(x)\big)^2 \big| X = x\big] \\ &= \sigma^2 + \text{Bias}(\widehat{r}(x))^2 + \text{Var}(\widehat{r}(x)). \end{aligned}$$

Questions: So what is the bias and variance of the kernel regression estimator? How do these terms depend on the smoothing parameter h_n and the sample size n? What is the optimal bandwidth h_n? How does the prediction error (for a kernel smoother with an optimal bandwidth h_n) depend on n?

Fortunately, these can roughly be worked out theoretically under some smoothness assumptions on r (and other assumptions). One can show that the bias at x is

$$\mathbb{E}[\widehat{r}(x) - r(x) | X_1 = x_1, \ldots X_n = x_n] = h^2 \left[\frac{1}{2} r''(x) + \frac{r'(x) f'(x)}{f(x)}\right] \sigma_K^2 + o(h^2) \tag{3}$$

where f is the density of x, and $\sigma_K^2 = \int u^2 K(u) du$ is the variance of the probability density corresponding to the kernel. One can also work out the variance of the kernel regression estimator,

$$\text{Var}[\widehat{r}(x) | X_1 = x_1, \ldots X_n = x_n] = \frac{\sigma^2 C(K)}{nhf(x)} + o((nh)^{-1}) \tag{4}$$

where $C(K) \equiv \int K^2(u) du$. Do these terms make sense? What happens to the bias and variance as h shrinks (i.e. less smoothing)? As h grows (i.e.

more smoothing)? Where does the sample size come in to the equation? What about the regression function itself?

Putting the bias together with the variance, we get an expression for the mean squared error of the kernel regression at x, MSE(x). Integrating the MSE, gives that the risk of the NW kernel estimator is

$$\begin{aligned} R(h_n) &= \frac{h_n^4}{4}\sigma_K^4 \int \left(r''(x) + 2r'(x)\frac{f'(x)}{f(x)} \right)^2 f(x)dx \\ &\quad + \frac{\sigma^2 \int K^2(x)dx}{nh_n} + o(nh_n^{-1}) + o(h_n^4) + \sigma^2 \end{aligned} \tag{5}$$

as $h_n \to 0$ and $nh_n \to \infty$. That is, using big-O order symbols, we have that:

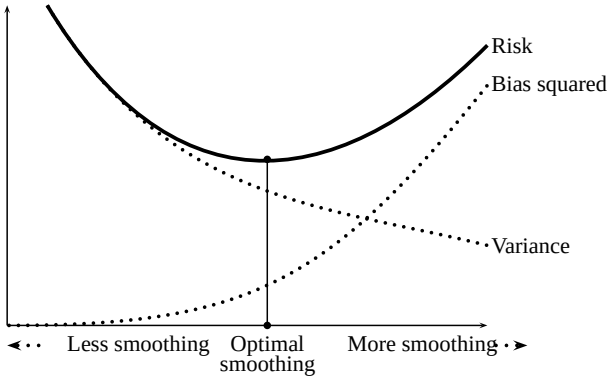

Figure 2: The bias–variance tradeoff. The bias increases and the variance decreases with the amount of smoothing. The optimal amount of smoothing, indicated by the vertical line, minimizes the risk = $\text{bias}^2 + \text{variance}$.

If we differentiate (5) and set the result equal to 0, we find that the **optimal bandwidth** h_* is

$$h_* = \left(\frac{1}{n}\right)^{1/5} \left(\frac{\sigma^2 \int K^2(x)dx}{\sigma_K^4 \int \left(r''(x) + 2r'(x)\frac{f'(x)}{f(x)}\right)^2 f(x)dx}\right)^{1/5}. \quad (6)$$

Thus, $h_* = O(n^{-1/5})$. Plugging h_* back into (5) we see that the risk de-

creases at rate $O(n^{-4/5})$. We can also derive the optimal bandwidth and the optimal rate "intuitively" by balancing the bias and variance terms:

In (most) parametric models, the risk of the maximum likelihood estimator decreases to 0 at rate $1/n$. The slower rate $n^{-4/5}$ is the price of using nonparametric methods. In practice, we cannot use the bandwidth given in (6) since h_* depends on the unknown function r. Instead, we use **cross-validation** as described in earlier lectures. To summarize (when to use parametric versus nonparametric models):

Multivariate Extension. The Curse of Dimensionality

In multiple dimensions, say, each $X_i \in \mathbb{R}^p$, we can easily use multivariate kernels: we just replace $X_i - x$ in the kernel argument by $\|X_i - x\|_2$, so that the (isotropic) multivariate kernel regression estimator is

$$\widehat{r}(x) = \frac{\sum_{i=1}^n K\left(\frac{\|x_i - x\|_2}{h}\right) \cdot Y_i}{\sum_{i=1}^n K\left(\frac{\|x_i - x\|_2}{h}\right)}$$

The same calculations as those that went into deriving the bias and variance bounds above can be done in this multivariate case. With some intuitive reasoning we can also figure out the order of the bias and variance terms:

Why is the variance so strongly affected by the dimension p? What is the optimal bandwidth h and the optimal rate of the kernel smoother in p dimensions?

Shalizi (Sec 9.3): *"For $p = 1$, the nonparametric rate is $O(n^{-4/5})$, which is of course slower than $O(n^{-1})$, but not all that much, and the improved bias usually more than makes up for it. But as p grows, the nonparametric rate gets slower and slower, and the fully nonparametric estimate more and more imprecise, yielding the infamous* **curse of dimensionality**. *For $p = 100$, say, we get a rate of $O(n^{-1/26})$, which is not very good at all. [....] Said another way, to get the same precision with p inputs that n data points gives us with one input takes $n^{(4+p)/5}$ data points. For $p = 100$, this is $n^{20.8}$, which tells us that matching the error of $n = 100$ one-dimensional observations requires $O(4 \times 10^{41})$ hundred-dimensional observations."*

In Lecture 8 we will use an alternative extension to higher dimensions that doesn't nearly suffer the same variance; this is called an *additive model*.