Nonparametric Regression

10/36-702

Larry Wasserman

1 Introduction

Now we focus on the following problem: Given a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, where $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$, estimate the regression function

$$m(x) = \mathbb{E}(Y|X = x) \tag{1}$$

without making parametric assumptions (such as linearity) about the regression function $m(x)$. Estimating m is called *nonparametric regression* or *smoothing*. We can write

$$Y = m(X) + \epsilon$$

where $\mathbb{E}(\epsilon) = 0$. This follows since, $\epsilon = Y - m(X)$ and $\mathbb{E}(\epsilon) = \mathbb{E}(\mathbb{E}(\epsilon|X)) = \mathbb{E}(m(X) - m(X)) = 0$

A related problem is *nonparametric prediction*. Given a pair (X, Y), we want to predict Y from X. The optimal predictor (under squared error loss) is the regression function $m(X)$. Hence, estimating m is of interest for its own sake and for the purposes of prediction.

Example 1 *Figure 1 shows data on bone mineral density. The plots show the relative change in bone density over two consecutive visits, for men and women. The smooth estimates of the regression functions suggest that a growth spurt occurs two years earlier for females. In this example, Y is change in bone mineral density and X is age.*

Example 2 *Figure 2 shows an analysis of some diabetes data from Efron, Hastie, Johnstone and Tibshirani (2004). The outcome Y is a measure of disease progression after one year. We consider four covariates (ignoring for now, six other variables): age, bmi (body mass index), and two variables representing blood serum measurements. A nonparametric regression model in this case takes the form*

$$Y = m(x_1, x_2, x_3, x_4) + \epsilon. \tag{2}$$

A simpler, but less general model, is the additive model

$$Y = m_1(x_1) + m_2(x_2) + m_3(x_3) + m_4(x_4) + \epsilon. \tag{3}$$

Figure 2 shows the four estimated functions $\widehat{m}_1, \widehat{m}_2, \widehat{m}_3$ and $\widehat{m}_4$.

Notation. We use $m(x)$ to denote the regression function. Often we assume that X_i has a density denoted by $p(x)$. The support of the distribution of X_i is denoted by $\mathcal{X}$. We assume that $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$. Recall that the trace of a square matrix A is denoted by $\text{tr}(A)$ and is defined to be the sum of the diagonal elements of A.

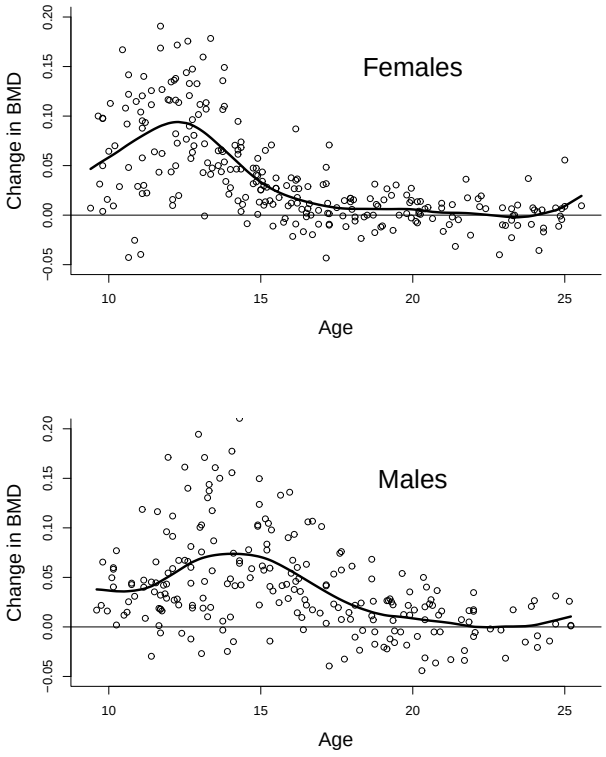

Figure 1: Bone Mineral Density Data

2 The Bias–Variance Tradeoff

Let $\widehat{m}(x)$ be an estimate of $m(x)$. The *pointwise risk* (or *pointwise mean squared error*) is

$$R(m(x), \widehat{m}(x)) = \mathbb{E}\big((\widehat{m}(x) - m(x))^2\big). \tag{4}$$

The *predictive risk* is

$$R(m, \widehat{m}) = \mathbb{E}((Y - \widehat{m}(X))^2) \tag{5}$$

where (X, Y) denotes a new observation. It follows that

$$\begin{aligned} R(m, \widehat{m}) &= \sigma^2 + \mathbb{E}\int (m(x) - \widehat{m}(x))^2 dP(x) && (6) \\ &= \sigma^2 + \int b_n^2(x) dP(x) + \int v_n(x) dP(x) && (7) \end{aligned}$$

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

The estimator $\widehat{m}$ typically involves smoothing the data in some way. The main challenge is to determine how much smoothing to do. When the data are oversmoothed, the bias term

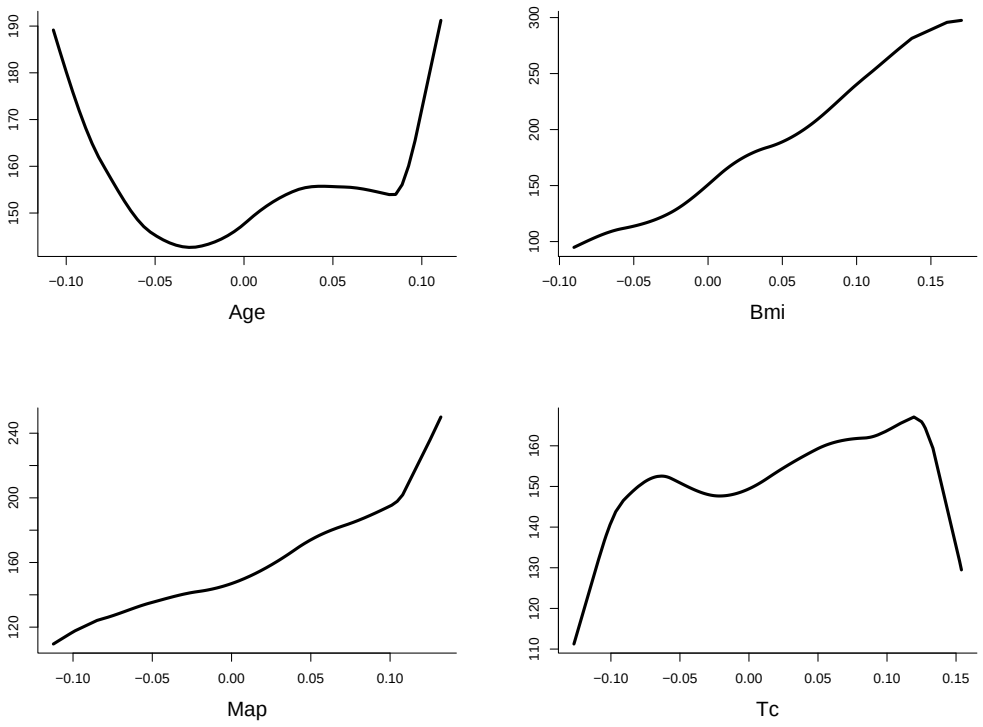

Figure 2: Diabetes Data

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.

An estimator $\widehat{m}$ is *consistent* if

$$\|\widehat{m} - m\| \xrightarrow{P} 0. \tag{8}$$

The *minimax risk* over a set of functions $\mathcal{M}$ is

$$R_n(\mathcal{M}) = \inf_{\widehat{m}} \sup_{m \in \mathcal{M}} R(m, \widehat{m}) \tag{9}$$

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

$$R(m, \widehat{m}) \asymp R_n(\mathcal{M}). \tag{10}$$

Typically the minimax rate is of the form $n^{-2\beta/(2\beta+d)}$ for some $\beta > 0$.

3 The Kernel Estimator

The simplest nonparametric estimator is the kernel estimator. The word "kernel" is often used in two different ways. Here are we referring to smoothing kernels. Later we will discuss *Mercer kernels* which are a distinct (but related) concept.

A one-dimensional *smoothing kernel* is any smooth, symmetric function K such that $K(x) \geq 0$ and

$$\int K(x)\,dx = 1, \quad \int xK(x)dx = 0 \quad \text{and} \quad \sigma_K^2 \equiv \int x^2 K(x)dx > 0. \tag{11}$$

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

$$\widehat{m}(x) \equiv \widehat{m}_h(x) = \frac{\sum_{i=1}^n Y_i\ K\left(\frac{\|x-X_i\|}{h}\right)}{\sum_{i=1}^n K\left(\frac{\|x-X_i\|}{h}\right)} = \sum_{i=1}^n Y_i \ell_i(x) \tag{12}$$

where $\ell_i(x) = K(\|x - X_i\|/h)/\sum_j K(\|x - X_j\|/h)$.

Thus $\widehat{m}(x)$ is a local average of the Y_i's. It can be shown that the optimal kernel is the Epanechnikov kernel. But, as with density estimation, the choice of kernel K is not too important. Estimates obtained by using different kernels are usually numerically very similar. This observation is confirmed by theoretical calculations which show that the risk is very insensitive to the choice of kernel. What does matter much more is the choice of bandwidth h which controls the amount of smoothing. Small bandwidths give very rough estimates while larger bandwidths give smoother estimates.

The kernel estimator can be derived by minimizing the localized squared error

$$\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)\ \left(c - Y_i\right)^2. \tag{13}$$

A simple calculation shows that this is minimized by the kernel estimator $c = \widehat{m}(x)$ as given in equation (12).

Kernel regression and kernel density estimation are related. Let $\widehat{p}(x, y)$ be the kernel density estimator and define

$$\widehat{m}(x) = \widehat{E}(Y|X = x) = \int y\widehat{p}(y|x)dy = \frac{\int y\widehat{p}(x, y)dy}{\widehat{p}(x)} \tag{14}$$

where $\widehat{p}(x) = \int \widehat{p}(x, y)dy$. Then $\widehat{m}(x)$ is the Nadaraya-Watson kernel regression estimator

Homework: Prove this.

4 Analysis of the Kernel Estimator

The bias-variance decomposition has some surprises with important practical consequences. Let's start with the one dimensional case.

Theorem 3 *Suppose that:*

1. $X \sim P_X$ *and* P_X *has density* p.
2. x *is in the interior of the support of* P_X.
3. $\lim_{|u|\to\infty} uK(u) = 0$.
4. $\mathbb{E}(Y^2) < \infty$.
5. $p(x) > 0$ *where* p *is the density of* P.
6. m *and* p *have two bounded, continuous derivatives.*
7. $h \to 0$ *and* $nh \to \infty$ *as* $n \to \infty$.

Then

$$\mathbb{E}(\widehat{m}(x) - m(x))^2 = \underbrace{\frac{h^4}{4}\left(m''(x) + 2\frac{m'(x)p'(x)}{p(x)}\right)\mu_2(K)}_{\text{bias}} + \underbrace{\frac{1}{nh}\frac{\sigma^2(x)}{p(x)}||K||^2}_{\text{variance}} + r(x) \tag{15}$$

where $\mu_2(K) = \int t^2 K(t)dt$, $\sigma^2(x) = \text{Var}(Y|X = x)$ *and* $r(x) = o(\text{bias} + \text{variance})$.

Proof Outline. We can write $\widehat{m}(x) = \widehat{a}(x)/\widehat{p}(x)$, $\widehat{a}(x) = n^{-1}\sum_i Y_i K_h(x, X_i)$, $\widehat{p}(x) = n^{-1}\sum_i K_h(x, X_i)$ where $K_h(x, y) = h^{-1}K((x - y)/h)$. Note that $\widehat{p}(x)$ is just the kernel density estimator of the true density $p(x)$. Recall (from 36-705) that $\mathbb{E}[\widehat{p}(x)] = p(x) + (h^2/2)p''(x)\mu_2(K)$. Then

$$\begin{aligned}
\widehat{m}(x) - m(x) &= \frac{\widehat{a}(x)}{\widehat{p}(x)} - m(x) \\
&= \left(\frac{\widehat{a}(x)}{\widehat{p}(x)} - m(x)\right)\left(\frac{\widehat{p}(x)}{p(x)} + \left(1 - \frac{\widehat{p}(x)}{p(x)}\right)\right) \\
&= \frac{\widehat{a}(x) - m(x)\widehat{p}(x)}{p(x)} + \frac{(\widehat{m}(x) - m(x))(p(x) - \widehat{p}(x))}{p(x)}.
\end{aligned}$$

It can be shown that the dominant term is the first term. Now, since $Y_i = m(X_i) + \epsilon_i$,

$$\mathbb{E}[\widehat{a}(x)] = \frac{1}{h}\mathbb{E}\left[Y_i K\left(\frac{X_i - x}{h}\right)\right] = h^{-1}\int m(u)K((x-u)/h)p(u)du = \int m(x+th)K(t)p(x+th)dt.$$

So

$$\mathbb{E}[\widehat{a}(x)] \approx \int \left(m(x) + thm'(x) + \frac{t^2h^2}{2}m''(x)\right)\left(p(x) + thp'(x) + \frac{t^2h^2}{2}p''(x)\right)K(t)dt.$$

Recall that $\int tK(t) = 0$ and $\int t^3 K(t) = 0$. So

$$\mathbb{E}[\widehat{a}(x)] \approx p(x)m(x) + m(x)p''(x)\frac{h^2}{2}\mu_2(K) + h^2 m'(x)p'(x)\mu_2(K) + m''(x)p(x)\frac{h^2}{2}$$

and $m(x)\widehat{p}(x) \approx m(x)p(x) + m(x)p''(x)(h^2/2)\mu_2(K)$ and so the mean of the first term is

$$\frac{h^2}{2}\left(\frac{2m'(x)p'(x)\mu_2(K)}{p(x)} + m''(x)\right).$$

Squaring this and keeping the dominant terms (of order $O(h^2)$) gives the stated bias. The variance term is calculated similarly. □

The bias term has two disturbing properties: it is large if $p'(x)$ is non-zero and it is large if $p(x)$ is small. The dependence of the bias on the density $p(x)$ is called **design bias**. It can also be shown that the bias is large if x is close to the boundary of the support of p. This is called **boundary bias**. Before we discuss how to fix these problems, let us state the multivariate version of this theorem.

Theorem 4 *Let $\widehat{m}$ be the multivariate kernel regression estimator with bandwidth matrix H. Thus*

$$\widehat{m}(x) = \frac{\sum_i Y_i K_H(X_i - x)}{\sum_i K_H(X_i - x)}$$

where H is a symmetric positive definite bandwidth matrix and $K_H(y) = [\det(H)]^{-1}K(H^{-1}y)$. Then

$$\mathbb{E}(\widehat{m}(x)) - m(x) \approx \mu_s(K)\frac{\nabla m(x)^T HH^T \nabla p(x)}{p(x)} + \frac{1}{2}\mu_2(K)\mathrm{tr}(H^T m''(x)H)$$

and

$$\mathrm{Var}(\widehat{m}(x)) \approx \frac{1}{n\det(H)}||K||_2^2\frac{\sigma(x)}{p(x)}.$$

The proof is the same except that we use multivariate Taylor expansions. Now suppose that $H = hI$. Then we see that the risk (mean squared error) is

$$C_1 h^4 + \frac{C_2}{nh^d}.$$

This is minimized by $h = \left(\frac{C_2}{n}\right)^{\frac{1}{4+d}}$. The resulting risk is $R_n \asymp \left(\frac{1}{n}\right)^{\frac{4}{4+d}}$. The main assumption was bounded, continuous second derivatives. If we repeat the analysis with bounded, continuous derivatives of order β (and we use an appropriate kernel), we get

$$R_n \asymp \left(\frac{1}{n}\right)^{\frac{2\beta}{2\beta+d}}. \tag{16}$$

Later we shall see that this is the minimax rate.

Remark: Note that this also gives the rate for the integrated risk $\mathbb{E}[\int(\widehat{m}(x)-m(x))^2dx]$.

How do we reduce the design bias and the boundary bias? The answer is to make a small modification to the kernel estimator. We discuss this in the next section.

5 Local Polynomials Estimators

Recall that the kernel estimator can be derived by minimizing the localized squared error

$$\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\ \left(c-Y_i\right)^2. \tag{17}$$

To reduce the design bias and the boundary bias we simply replace the constant c with a polynomial. In fact, it is enough to use a polynomial of order 1; in other words, we fit a local linear estimator instead of a local constant. The idea is that, for u near x, we can write, $m(u)\approx\beta_0(x)+\beta_1(x)(u-x)$. We define $\widehat{\beta}(x)=(\widehat{\beta}_0(x),\widehat{\beta}_1(x))$ to minimize

$$\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\ \left(Y_i-\beta_0(x)-\beta_1(x)(X_i-x)\right)^2.$$

Then $\widehat{m}(u)\approx\widehat{\beta}_0(x)+\widehat{\beta}_1(x)(u-x)$. In particular, $\widehat{m}(x)=\widehat{\beta}_0(x)$. The minimizer is easily seen to be

$$\widehat{\beta}(x)=(\widehat{\beta}_0(x),\widehat{\beta}_1(x))^T=(\mathbb{X}^TW\mathbb{X})^{-1}\mathbb{X}^TW\mathbb{Y}$$

where $\mathbb{Y}=(Y_1,\ldots,Y_n)$,

$$\mathbb{X}=\begin{pmatrix}1 & X_1-x\\ 1 & X_2-x\\ \vdots & \vdots\\ 1 & X_n-x\end{pmatrix},\quad W=\begin{pmatrix}K_h(x-X_1) & 0 & \cdots & 0\\ 0 & K_h(x-X_1) & \cdots & 0\\ \vdots & \vdots & \cdots & \cdots\\ 0 & 0 & \cdots & K_h(x-X_n)\end{pmatrix}.$$

Then $\widehat{m}(x)=\widehat{\beta}_0(x)$.

For the multivariate version we minimize

$$\sum_{i=1}^{n} K_H(x-X_i)\ (Y_i-\beta_0(x)-\beta_1^T(x)(X_i-x))^2.$$

The minimizer is

$$\widehat{\beta}(x)=(\widehat{\beta}_0(x),\widehat{\beta}_1(x))^T=(\mathbb{X}^TW\mathbb{X})^{-1}\mathbb{X}^TW\mathbb{Y}$$

where $\mathbb{Y} = (Y_1, \ldots, Y_n)$,

$$\mathbb{X} = \begin{pmatrix} 1 & (X_1 - x)^T \\ 1 & (X_2 - x)^T \\ \vdots & \vdots \\ 1 & (X_n - x)^T \end{pmatrix}, \quad W = \begin{pmatrix} K_H(x - X_1) & 0 & \cdots & 0 \\ 0 & K_H(x - X_1) & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots \\ 0 & 0 & \cdots & K_H(x - X_n) \end{pmatrix}.$$

Then $\widehat{m}(x) = \widehat{\beta}_0(x)$.

Now comes the amazing part.

Theorem 5 *The local linear estimator has risk*

$$\frac{1}{4}\mu_2^2(K)(\mathrm{tr}(H^T p''(x) H))^2 + \frac{1}{n\det(H)}||K||_2^2 \frac{\sigma(x)}{p(x)}. \tag{18}$$

The proof uses the same techniques as before but is much more involved. The key thing to note are that the rate is the same but the design bias (dependence of the bias on $p'(x)$) has disappeared.

It can also be shown that the boundary bias disappears. In fact, the kernel estimator has squared bias of order h^2 near the boundary of the support while the local linear estimator has squared bias of order h^4 near the boundary of the support. To the best of my knowledge, there is no simple way to get rid of these biases from RKHS estimators and other penalization-based estimators.

Remark: If you want to reduce bias at maxima and minima of the regression function, then use local quadratic regression.

Example 6 (LIDAR) *These are data from a light detection and ranging (LIDAR) experiment. LIDAR is used to monitor pollutants. The response is the log of the ratio of light received from two lasers. The frequency of one laser is the resonance frequency of mercury while the second has a different frequency. Figure 3 shows the 221 observations. The top left plot shows the data and the fitted function using local linear regression. The cross-validation curve (not shown) has a well-defined minimum at $h \approx 37$ corresponding to 9 effective degrees of freedom. The fitted function uses this bandwidth. The top right plot shows the residuals. There is clear heteroscedasticity (nonconstant variance). The bottom left plot shows the estimate of $\sigma(x) = \sqrt{\mathrm{Var}(Y|X = x)}$. This estimate was obtained by smoothing the residuals $(Y_i - \widehat{m}(X_i))^2$. The bands in the lower right plot are $\widehat{m}(x) \pm 2\sigma(x)$. As expected, there is much greater uncertainty for larger values of the covariate.*

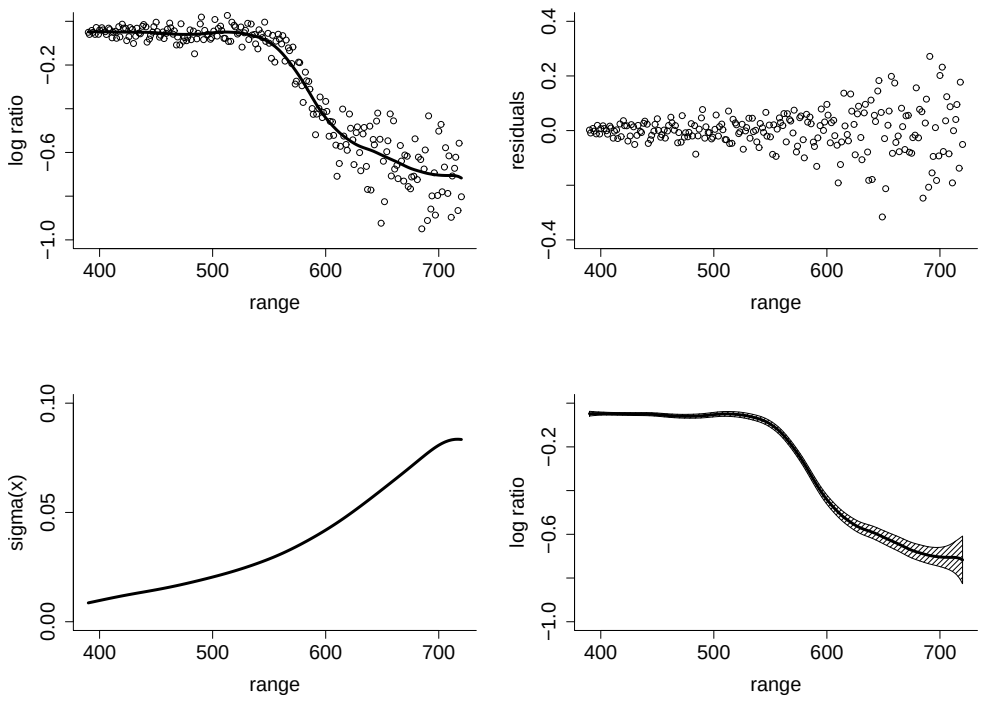

Figure 3: The LIDAR data from Example 6. Top left: data and the fitted function using local linear regression with $h \approx 37$ (chosen by cross-validation). Top right: the residuals. Bottom left: estimate of $\sigma(x)$. Bottom right: variability bands.

6 Linear Smoothers

Kernel estimators and local polynomial estimator are examples of *linear smoothers*.

Definition: An estimator $\widehat{m}$ of m is a *linear smoother* if, for each x, there is a vector $\ell(x) = (\ell_1(x), \ldots, \ell_n(x))^T$ such that

$$\widehat{m}(x) = \sum_{i=1}^{n} \ell_i(x) Y_i = \ell(x)^T Y \tag{19}$$

where $Y = (Y_1, \ldots, Y_n)^T$.

For kernel estimators, $\ell_i(x) = \frac{K(\|x - X_i\|/h)}{\sum_{j=1}^{n} K(\|x - X_j\|/h)}$. For local linear estimators, we can deduce the weights from the expression for $\widehat{\beta}(x)$. Here is an interesting fact: the following estimators are linear smoothers: Gaussian process regression, splines, RKHS estimators.

Example 7 *You should note confuse linear smoothers with linear regression. In linear regression we assume that $m(x) = x^T\beta$. In fact, least squares linear regression is a special*

case of linear smoothing. If $\widehat{\beta}$ *denotes the least squares estimator then* $\widehat{m}(x) = x^T\widehat{\beta} = x^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y = \ell(x)^T Y$ *where* $\ell(x) = x^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$.

Define the vector of *fitted values* $\widehat{Y} = (\widehat{m}(X_1), \ldots, \widehat{m}(X_n))^T$. It follows that $\widehat{Y} = LY$ where

$$L = \begin{pmatrix} \ell(X_1)^T \\ \ell(X_2)^T \\ \vdots \\ \ell(X_n)^T \end{pmatrix} = \begin{pmatrix} \ell_1(X_1) & \ell_2(X_1) & \cdots & \ell_n(X_1) \\ \ell_1(X_2) & \ell_2(X_2) & \cdots & \ell_n(X_2) \\ \vdots & \vdots & \vdots & \vdots \\ \ell_1(X_n) & \ell_2(X_n) & \cdots & \ell_n(X_n) \end{pmatrix}. \tag{20}$$

The matrix L defined in (20) is called the *smoothing matrix*. The i^{th} row of L is called the *effective kernel* for estimating $m(X_i)$. We define the *effective degrees of freedom* by

$$\nu = \text{tr}(L). \tag{21}$$

The effective degrees of freedom behave very much like the number of parameters in a linear regression model.

Remark. The weights in all the smoothers we will use have the property that, for all x, $\sum_{i=1}^n \ell_i(x) = 1$. This implies that the smoother preserves constants.

7 Penalized Regression

Another smoothing method is *penalized regression* (or *regularized regression*) where $\widehat{m}$ is defined to be the minimizer of

$$\sum_{i=1}^n (Y_i - \widehat{m}(X_i))^2 + \lambda J(\widehat{m}) \tag{22}$$

where $\lambda \geq 0$ and $J(\widehat{m})$ is a penalty (or regularization) term. A popular choice of J is

$$J(g) = \int (g''(x))^2 dx.$$

To find the minimizer of (22) we need to use cubic splines. Let (a, b) be an interval and let $x_1, \ldots, x_k$ be k points such that $a < x_1 < \cdots < x_k < b$. A continuous function f on (a, b) is a *cubic spline* with knots $\{x_1, \ldots, x_n\}$ if f is cubic polynomial over the intervals $(x_1, x_2), (x_2, x_2), \ldots$ and f has continuous first and second derivatives at the knots.

Theorem 8 *Let* $\widehat{m}$ *be the minimizer of (22) where* $J(g) = \int (g''(x))^2 dx$. *Then* $\widehat{m}$ *is a cubic spline with knots at the points* $X_1, \ldots, X_n$.

According to this result, the minimizer $\widehat{m}$ of (22) is contained in $\mathcal{M}_n$, the set of all cubic splines with knots at $\{X_1, \ldots, X_n\}$. However, we still have to find which function in $\mathcal{M}_n$ is the minimizer.

Define $B_1(x) = 1$, $B_2(x) = x$, $B_3(x) = x^2$, $B_4(x) = x^3$ and

$$B_j(x) = (x - X_{j-4})_+^3 \quad j = 5, \ldots, n+4.$$

It can be shown that $B_1, \ldots, B_{n+4}$ form a basis for the $\mathcal{M}_n$. (In practice, another basis for $\mathcal{M}$ called the B-spline basis is used since it has better numerical properties.) Thus, every $g \in \mathcal{M}_n$ can be written as $g(x) = \sum_{j=1}^N \beta_j B_j(x)$ for some coefficients $\beta_1, \ldots, \beta_N$. If we substitute $\widehat{m}(x) = \sum_{j=1}^N \beta_j B_j(x)$ into (22), the minimization problem becomes: find $\beta = (\beta_1, \ldots, \beta_N)$ to minimize

$$(Y - \mathbb{B}\beta)^T (Y - \mathbb{B}\beta) + \lambda \beta^T \Omega \beta \tag{23}$$

where $Y = (Y_1, \ldots, Y_n)$, $\mathbb{B}_{ij} = B_j(X_i)$ and $\Omega_{jk} = \int B_j''(x) B_k''(x) dx$. The solution is

$$\widehat{\beta} = (\mathbb{B}^T \mathbb{B} + \lambda \Omega)^{-1} \mathbb{B}^T Y$$

and hence

$$\widehat{m}(x) = \sum_j \widehat{\beta}_j B_j(x) = \ell(x)^T Y$$

where $\ell(x) = b(x)(\mathbb{B}^T \mathbb{B} + \lambda \Omega)^{-1} \mathbb{B}^T$ and $b(x) = (B_1(x), \ldots, B_N(x))^T$. Hence, the spline smoother is another example of a linear smoother.

The parameter λ is a smoothing parameter. As $\lambda \to 0$, $\widehat{m}$ tends to the interpolating function $\widehat{m}(X_i) = Y_i$. As $\lambda \to \infty$, $\widehat{m}$ tends to the least squares linear fit.

For RKHS regression, we minimize

$$\sum_{i=1}^n (Y_i - \widehat{m}(X_i))^2 + \lambda ||m||_K^2 \tag{24}$$

where $||m||_K$ is the RKHS norm for some Mercer kernel K. In the homework, you proved that the estimator is $\widehat{m}(x) = \sum_j \widehat{\alpha}_j K(x, X_i)$ where

$$\widehat{\alpha} = (\mathbb{K} + \lambda I)^{-1} \mathbb{Y}$$

and $\mathbb{K}(j,k) = K(j,k)$. Note that RKHS regression is just another linear smoother.

In my opinion, RKHS regression has several disadvantages:

1. There are two parameters to pick: the bandwidth in the kernel and λ.
2. It is very hard to derive the bias and variance of the estimator.
3. It is not obvious how to study and fix the design bias and boundary bias.
4. The kernel and local linear estimators are local averages. This makes them easy to understand. The RKHS estimators is the solution to a constrained minimization. This makes them much less intuitive.

8 Basis Functions and Dictionaries

Suppose that

$$m \in L_2(a,b) = \left\{ g : [a,b] \to \mathbb{R} : \quad \int_a^b g^2(x)dx < \infty \right\}.$$

Let $\phi_1, \phi_2, \ldots$ be an orthonormal basis for $L_2(a,b)$. This means that $\int \phi_j^2(x)dx = 1$, $\int \phi_j \phi_k(x)dx = 0$ for $j \neq k$ and the only function $b(x)$ such that $\int b(x)\phi_j(x)dx = 0$ for all j is $b(x) = 0$. It follows that any $m \in L_2(a,b)$ can be written as

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

where $\beta_j = \int m(x)\phi_j(x)dx$. For $[a,b] = [0,1]$, an example is the cosine basis

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

To use a basis for nonparametric regression, we regress Y on the first J basis functions and we treat J as a smoothing parameter. In other words we take $\widehat{m}(x) = \sum_{j=1}^{J} \widehat{\beta}_j \phi_j(x)$ where $\widehat{\beta} = (B^T B)^{-1} B^T Y$ and $B_{ij} = \phi_j(X_i)$. It follows that $\widehat{m}(x)$ is a linear smoother. See Chapters 7 and 8 of Wasserman (2006) for theoretical properties of orthogonal function smoothers.

It is not necessary to use orthogonal functions for smoothing. Let $\mathcal{D} = \{\psi_1, \ldots, \psi_N\}$ be any collection of functions, called a *dictionary*. The collection $\mathcal{D}$ could be very large. For example, $\mathcal{D}$ might be the union of several different bases. The smoothing problem is to decide which functions in $\mathcal{D}$ to use for approximating m. One way to approach this problem is to use the lasso: regress Y on $\mathcal{D}$ using an ℓ_1 penalty.

9 Choosing the Smoothing Parameter

The estimators depend on the bandwidth h. Let $R(h)$ denote the risk of $\widehat{m}_h$ when bandwidth h is used. We will estimate $R(h)$ and then choose h to minimize this estimate. As with linear regression, the *training error*

$$\widetilde{R}(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{m}_h(X_i))^2 \tag{25}$$

is biased downwards. We will estimate the risk using the cross-validation.

9.1 Leave-One-Out Cross-Validation

The *leave-one-out cross-validation score* is defined by

$$\widehat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{m}_{(-i)}(X_i))^2 \tag{26}$$

where $\widehat{m}_{(-i)}$ is the estimator obtained by omitting the i^{th} pair (X_i, Y_i), that is, $\widehat{m}_{(-i)}(x) = \sum_{j=1}^{n} Y_j \ell_{j,(-i)}(x)$ and

$$\ell_{j,(-i)}(x) = \begin{cases} 0 & \text{if } j = i \\ \frac{\ell_j(x)}{\sum_{k\neq i}\ell_k(x)} & \text{if } j \neq i. \end{cases} \tag{27}$$

Theorem 9 *Let $\widehat{m}$ be a linear smoother. Then the leave-one-out cross-validation score $\widehat{R}(h)$ can be written as*

$$\widehat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \widehat{m}_h(X_i)}{1 - L_{ii}}\right)^2 \tag{28}$$

where $L_{ii} = \ell_i(X_i)$ is the i^{th} diagonal element of the smoothing matrix L.

The smoothing parameter h can then be chosen by minimizing $\widehat{R}(h)$. An alternative is *generalized cross-validation* in which each L_{ii} in equation (28) is replaced with its average $n^{-1}\sum_{i=1}^{n} L_{ii} = \nu/n$ where $\nu = \text{tr}(L)$ is the effective degrees of freedom. (Note that ν depends on h.) Thus, we minimize

$$\text{GCV}(h) = \frac{\widetilde{R}}{(1 - \nu/n)^2}. \tag{29}$$

Usually, GCV and cross-validation are very similar. Using the approximation $(1-x)^{-2} \approx 1 + 2x$ we see that

$$\text{GCV}(h) \approx \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{m}(X_i))^2 + \frac{2\nu\widehat{\sigma}^2}{n} \equiv C_p \tag{30}$$

where $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(Y_i - \widehat{m}(X_i))^2$. Equation (30) is the nonparametric version of the C_p statistic that we saw in linear regression.

Example 10 (Doppler function) *Let*

$$m(x) = \sqrt{x(1-x)}\sin\left(\frac{2.1\pi}{x + .05}\right), \quad 0 \leq x \leq 1 \tag{31}$$

which is called the Doppler function. *This function is difficult to estimate and provides a good test case for nonparametric regression methods. The function is spatially inhomogeneous which means that its smoothness (second derivative) varies over x. The function is plotted*

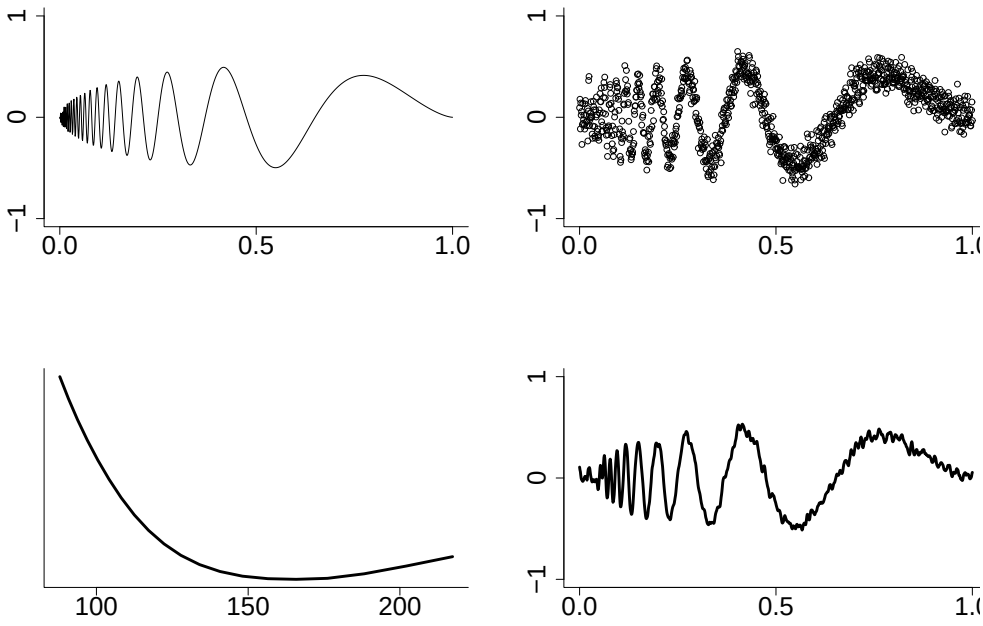

Figure 4: The Doppler function estimated by local linear regression. The function (top left), the data (top right), the cross-validation score versus effective degrees of freedom (bottom left), and the fitted function (bottom right).

in the top left plot of Figure 4. The top right plot shows 1000 data points simulated from $Y_i = m(i/n) + \sigma\epsilon_i$ *with* $\sigma = .1$ *and* $\epsilon_i \sim N(0,1)$. *The bottom left plot shows the cross-validation score versus the effective degrees of freedom using local linear regression. The minimum occurred at 166 degrees of freedom corresponding to a bandwidth of .005. The fitted function is shown in the bottom right plot. The fit has high effective degrees of freedom and hence the fitted function is very wiggly. This is because the estimate is trying to fit the rapid fluctuations of the function near* $x = 0$. *If we used more smoothing, the right-hand side of the fit would look better at the cost of missing the structure near* $x = 0$. *This is always a problem when estimating spatially inhomogeneous functions.*

9.2 Data Splitting

As with density estimation, stronger guarantees can be made using a *data splitting* version of cross-validation. Suppose the data are $(X_1, Y_1), \ldots, (X_{2n}, Y_{2n})$. Now randomly split the data into two halves that we denote by

$$\mathcal{D} = \Big\{(\widetilde{X}_1, \widetilde{Y}_1), \ldots, (\widetilde{X}_n, \widetilde{Y}_n)\Big\}$$

and

$$\mathcal{E} = \Big\{(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)\Big\}.$$

Construct regression estimators $\mathcal{M} = \{m_1, \ldots, m_N\}$ from $\mathcal{D}$. Define the risk estimator

$$\widehat{R}(m_j) = \frac{1}{n}\sum_{i=1}^{n} |Y_i^* - m_j(X_i^*)|^2.$$

Finally, let

$$\widehat{m} = \text{argmin}_{m\in\mathcal{M}}\widehat{R}(m).$$

Theorem 11 *Let $m_* \in \mathcal{M}$ minimize $\|m_j - m\|_P^2$. There exists $C > 0$ such that*

$$\mathbb{E}(\|\widehat{m} - m\|_P^2) \leq 2\,\mathbb{E}\|m_* - m\|_P^2 + \frac{C\log N}{n}.$$

10 Minimax Properties

Consider the nonparametric regression model

$$Y_i = m(X_i) + \epsilon_i$$

where $X_1, \ldots, X_n \sim P$ and $\mathbb{E}(\epsilon_i) = 0$.

Theorem 12 *Let $\mathcal{P}$ denote all distributions for X. Then*

$$\inf_{\widehat{m}} \sup_{m\in\Sigma(\beta,L), P\in\mathcal{P}} \mathbb{E}\|\widehat{m} - m\|_P^2 \geq \frac{C}{n^{2\beta/(2\beta+d)}}. \tag{32}$$

This theorem will be proved when we cover minimax theory. It follows that kernel regression is minimax for $\beta = 1$. If we were to impose stronger assumptions, then the rate of convergence for kernel regression would be $n^{-4/(4+d)}$ and hence would be minimax for $\beta = 2$.

11 Additive Models

Interpreting and visualizing a high-dimensional fit is difficult. As the number of covariates increases, the computational burden becomes prohibitive. A practical approach is to use an *additive model*. An additive model is a model of the form

$$Y = \alpha + \sum_{j=1}^{d} m_j(x_j) + \epsilon \tag{33}$$

The Backfitting Algorithm

Initialization: set $\widehat{\alpha} = \overline{Y}$ and set initial guesses for $\widehat{m}_1, \ldots, \widehat{m}_d$. Now iterate the following steps until convergence. For $j = 1, \ldots, d$ do:

- Compute $\widetilde{Y}_i = Y_i - \widehat{\alpha} - \sum_{k \neq j} \widehat{m}_k(X_i)$, $i = 1, \ldots, n$.
- Apply a smoother to $\widetilde{Y}$ on X_j to obtain $\widehat{m}_j$.
- Set $\widehat{m}_j(x) \longleftarrow \widehat{m}_j(x) - n^{-1} \sum_{i=1}^n \widehat{m}_j(X_i)$.
- end do.

Figure 5: Backfitting.

where $m_1, \ldots, m_d$ are smooth functions. The model (33) is not identifiable since we can add any constant to α and subtract the same constant from one of the m_j's without changing the regression function. This problem can be fixed in a number of ways; the simplest is to $\widehat{\alpha} = \overline{Y}$ and then regard the m_j's as deviations from $\overline{Y}$. In this case we require that $\sum_{i=1}^n \widehat{m}_j(X_i) = 0$ for each j.

There is a simple algorithm called *backfitting* for turning any one-dimensional regression smoother into a method for fitting additive models. This is essentially a coordinate descent, Gauss-Seidel algorithm. See Figure 5.

Example 13 *This example involves three covariates and one response variable. The data are 48 rock samples from a petroleum reservoir, the response is permeability (in milli-Darcies) and the covariates are: the area of pores (in pixels out of 256 by 256), perimeter in pixels and shape (perimeter/$\sqrt{\text{area}}$). The goal is to predict permeability from the three covariates. We fit the additive model*

$$\text{permeability} = m_1(\text{area}) + m_2(\text{perimeter}) + m_3(\text{shape}) + \epsilon.$$

We could scale each covariate to have the same variance and then use a common bandwidth for each covariate. Instead, we perform cross-validation to choose a bandwidth h_j for covariate x_j during each iteration of backfitting. The bandwidths and the functions estimates converged rapidly. The estimates of m_1, m_2 and m_3 are shown in Figure 6. $\overline{Y}$ was added to each function before plotting it.

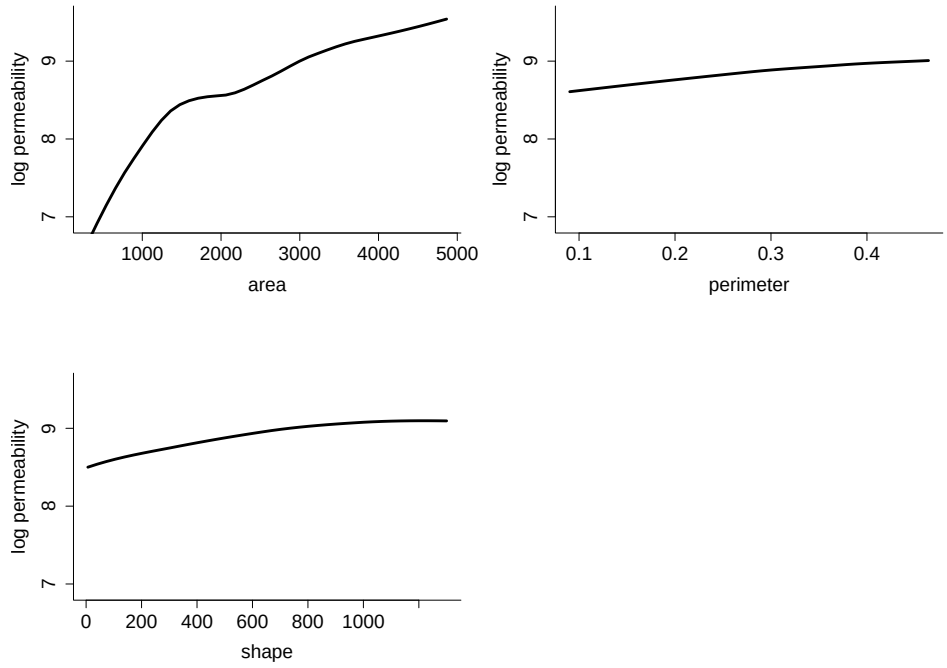

Figure 6: The rock data. The plots show $\widehat{m}_1$, $\widehat{m}_2$, and $\widehat{m}_3$ for the additive model $Y = \widehat{m}_1(X_1) + \widehat{m}_2(X_2) + \widehat{m}_3(x_3) + \epsilon$.

12 SpAM

Ravikumar, Lafferty, Liu and Wasserman (2007) introduced a sparse version of additive models called SpAM (Sparse Additive Models). This is a functional version of the grouped lasso (Yuan and Lin 2006) and is closely related to the COSSO (Lin and Zhang 2006).

We form an additive model

$$Y_i = \alpha + \sum_{j=1}^{d} \beta_j g_j(X_{ij}) + \epsilon_i \tag{34}$$

with the identifiability conditions $\int g_j(x_j)dP(x_j) = 0$ and $\int g_j^2(x_j)dP(x_j) = 1$. Further, we impose the sparsity condition $\sum_{j=1}^{d} |\beta_j| \leq L_n$ and the smoothness condition $g_j \in \mathcal{S}_j$ where $\mathcal{S}_j$ is some class of smooth functions. While this optimization problem makes plain the role ℓ_1 regularization of β to achieve sparsity, it is convenient to reexpress the model as

$$\min_{m_j \in \mathcal{H}_j} \mathbb{E}\left(Y - \textstyle\sum_{j=1}^{d} m_j(X_j)\right)^2$$

subject to

$$\sum_{j=1}^{d} \sqrt{\mathbb{E}(m_j^2(X_j))} \ \leq \ L, \quad \mathbb{E}(m_j) = 0, \ j = 1, \ldots, d.$$

The Lagrangian for the optimization problem is

$$L(f,\lambda,\mu) = \frac{1}{2}\mathbb{E}\left(Y - \sum_{j=1}^{d} m_j(X_j)\right)^2 + \lambda \sum_{j=1}^{d} \sqrt{\mathbb{E}(m_j^2(X_j))} + \sum_j \mu_j \mathbb{E}(m_j). \tag{35}$$

Theorem 14 *The minimizers $m_1, \ldots, m_p$ of (35) satisfy*

$$m_j = \left[1 - \frac{\lambda}{\sqrt{\mathbb{E}(P_j^2)}}\right]_+ P_j \quad \text{a.e.} \tag{36}$$

where $[\cdot]_+$ denotes the positive part, and $P_j = \mathbb{E}[R_j \mid X_j]$ denotes the projection of the residual $R_j = Y - \sum_{k \neq j} m_k(X_k)$ onto $\mathcal{H}_j$.

To solve this problem, we insert sample estimates into the population algorithm, as in standard backfitting. We estimate the projection P_j by smoothing the residuals:

$$\widehat{P}_j = \mathcal{S}_j R_j \tag{37}$$

where $\mathcal{S}_j$ is a linear smoother, such as a local linear or kernel smoother. Let

$$\widehat{s}_j = \frac{1}{\sqrt{n}} \|\widehat{P}_j\|_2 = \sqrt{\text{mean}(\widehat{P}_j^2)}. \tag{38}$$

be the estimate of $\sqrt{\mathbb{E}(P_j^2)}$. We have thus derived the SpAM backfitting algorithm given in Figure 7.

We choose λ by minimizing an estimate of the risk. Let ν_j be the effective degrees of freedom for the smoother on the j^{th} variable, that is, $\nu_j = \text{trace}(\mathcal{S}_j)$ where $\mathcal{S}_j$ is the smoothing matrix for the j-th dimension. Also let $\widehat{\sigma}^2$ be an estimate of the variance. Define the total effective degrees of freedom

$$\text{df}(\lambda) = \sum_j \nu_j I\left(\|\widehat{m}_j\| \neq 0\right). \tag{39}$$

Two estimates of risk are

$$C_p = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{d}\widehat{m}_j(X_j)\right)^2 + \frac{2\widehat{\sigma}^2}{n}\,\text{df}(\lambda) \tag{40}$$

and

$$\text{GCV}(\lambda) = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \sum_j \widehat{m}_j(X_{ij}))^2}{(1 - \text{df}(\lambda)/n)^2}. \tag{41}$$

SpAM Backfitting Algorithm

Input: Data (X_i, Y_i), regularization parameter λ.

Initialize $\widehat{m}_j = 0$, for $j = 1, \ldots, p$.

Iterate until convergence:

For each $j = 1, \ldots, p$:

(1) Compute the residual: $R_j = Y - \sum_{k \neq j} \widehat{f}_k(X_k)$;

(2) Estimate $P_j = \mathbb{E}[R_j \mid X_j]$ by smoothing: $\widehat{P}_j = \mathcal{S}_j R_j$;

(3) Estimate norm: $\widehat{s}_j^2 = \frac{1}{n} \sum_{i=1}^n \widehat{P}_j^2(i)$;

(4) Soft-threshold: $\widehat{m}_j = [1 - \lambda/\widehat{s}_j]_+ \widehat{P}_j$;

(5) Center: $\widehat{m}_j \leftarrow \widehat{m}_j - \text{mean}(\widehat{m}_j)$.

Output: Component functions $\widehat{m}_j$ and estimator $\widehat{m}(X_i) = \sum_j \widehat{m}_j(X_{ij})$.

Figure 7: The SpAM backfitting algorithm. The first two steps in the iterative algorithm are the usual backfitting procedure; the remaining steps carry out functional soft thresholding.

The first is C_p and the second is generalized cross validation but with degrees of freedom defined by $\text{df}(\lambda)$. A proof that these are valid estimates of risk is not currently available. Thus, these should be regarded as heuristics.

Synthetic Data. We generated $n = 150$ observations from the following 200-dimensional additive model:

$$Y_i = m_1(x_{i1}) + m_2(x_{i2}) + m_3(x_{i3}) + m_4(x_{i4}) + \epsilon_i \tag{42}$$
$$m_1(x) = -2\sin(2x),\ m_2(x) = x^2 - \frac{1}{3},\ m_3(x) = x - \frac{1}{2},\ m_4(x) = e^{-x} + e^{-1} - 1$$

and $m_j(x) = 0$ for $j \geq 5$ with noise $\epsilon_i \sim \mathcal{N}(0, 1)$. These data therefore have 196 irrelevant dimensions.

The results of applying SpAM with the plug-in bandwidths are summarized in Figure 8. The top-left plot in Figure 8 shows regularization paths as a function of the parameter λ; each curve is a plot of $\|\widehat{m}_j\|_2$ versus λ for a particular variable X_j. The estimates are generated efficiently over a sequence of λ values by "warm starting" $\widehat{m}_j(\lambda_t)$ at the previous value $\widehat{m}_j(\lambda_{t-1})$. The top-center plot shows the C_p statistic as a function of λ. The top-right plot compares the empirical probability of correctly selecting the true four variables as a function of sample size n, for $p = 128$ and $p = 256$. This behavior suggests the same threshold phenomenon that was proven for the lasso.

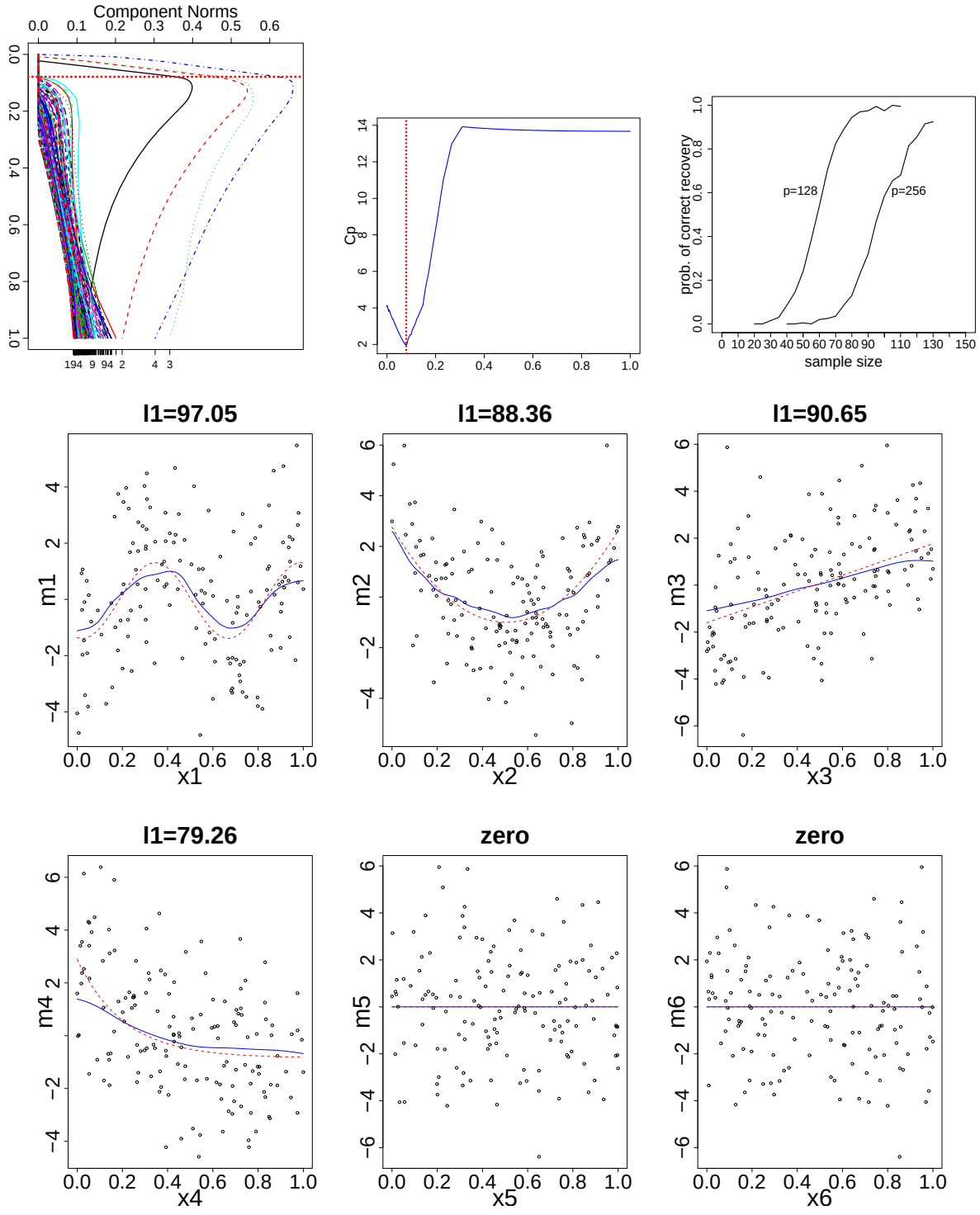

Figure 8: (Simulated data) Upper left: The empirical ℓ_2 norm of the estimated components as plotted against the regularization parameter λ; the value on the x-axis is proportional to $\sum_j \|\widehat{m}_j\|_2$. Upper center: The C_p scores against the regularization parameter λ; the dashed vertical line corresponds to the value of λ which has the smallest C_p score. Upper right: The proportion of 200 trials where the correct relevant variables are selected, as a function of sample size n. Lower (from left to right): Estimated (solid lines) versus true additive component functions (dashed lines) for the first 6 dimensions; the remaining components are zero.

Boston Housing. The Boston housing data were collected to study house values in the suburbs of Boston. There are 506 observations with 10 covariates. The dataset has been studied by many authors with various transformations proposed for different covariates. To explore the sparsistency properties of our method, we added 20 irrelevant variables. Ten of them are randomly drawn from Uniform$(0,1)$, the remaining ten are a random permutation of the original ten covariates. The model is

$$\begin{aligned} Y &= \alpha + m_1(\texttt{crim}) + m_2(\texttt{indus}) + m_3(\texttt{nox}) + m_4(\texttt{rm}) + m_5(\texttt{age}) \\ &\quad + m_6(\texttt{dis}) + m_7(\texttt{tax}) + m_8(\texttt{ptratio}) + m_9(\texttt{b}) + m_{10}(\texttt{lstat}) + \epsilon. \end{aligned} \tag{43}$$

The result of applying SpAM to this 30 dimensional dataset is shown in Figure 9. SpAM identifies 6 nonzero components. It correctly zeros out both types of irrelevant variables. From the full solution path, the important variables are seen to be `rm`, `lstat`, `ptratio`, and `crim`. The importance of variables `nox` and `b` are borderline. These results are basically consistent with those obtained by other authors. However, using C_p as the selection criterion, the variables `indux`, `age`, `dist`, and `tax` are estimated to be irrelevant, a result not seen in other studies.

13 Partitions and Trees

Simple and interpretable estimators can be derived by partitioning the range of X. Let $\Pi_n = \{A_1, \ldots, A_N\}$ be a partition of $\mathcal{X}$ and define

$$\widehat{m}(x) = \sum_{j=1}^{N} \overline{Y}_j \, I(x \in A_j)$$

where $\overline{Y}_j = n_j^{-1} \sum_{i=1}^{n} Y_i I(X_i \in A_j)$ is the average of the Y_i's in A_j and $n_j = \#\{X_i \in A_j\}$. (We define $\overline{Y}_j$ to be 0 if $n_j = 0$.)

The simplest partition is based on cubes. Suppose that $\mathcal{X} = [0,1]^d$. Then we can partition $\mathcal{X}$ into $N = k^d$ cubes with lengths of size $h = 1/k$. Thus, $N = (1/h)^d$. The smoothing parameter is h.

Theorem 15 *Let $\widehat{m}(x)$ be the partition estimator. Suppose that*

$$m \in \mathcal{M} = \left\{ m : \ |m(x) - m(x)| \leq L\|x - z\|, \quad x, z, \in \mathbb{R}^d \right\} \tag{44}$$

and that $\mathrm{Var}(Y|X = x) \leq \sigma^2 < \infty$ *for all x. Then*

$$\mathbb{E}\|\widehat{m} - m\|_P^2 \leq c_1 h^2 + \frac{c_2}{nh^d}. \tag{45}$$

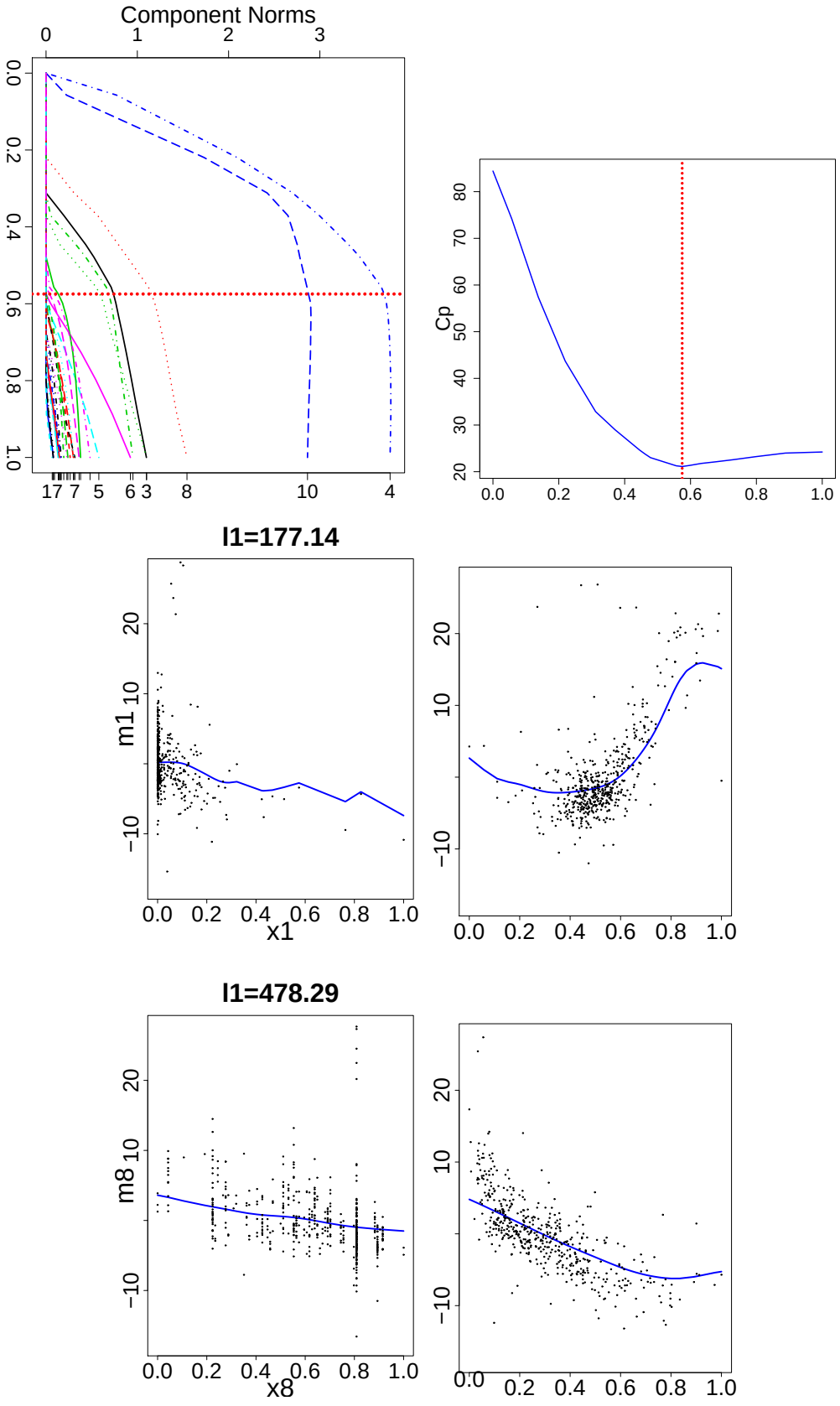

Figure 9: (Boston housing) Left: The empirical ℓ_2 norm of the estimated components versus the regularization parameter λ. Center: The C_p scores against λ; the dashed vertical line corresponds to best C_p score. Right: Additive fits for four relevant variables.

Hence, if $h \asymp n^{-1/(d+2)}$ *then*

$$\mathbb{E}\|\widehat{m} - m\|_P^2 \leq \frac{c}{n^{2/(d+2)}}. \tag{46}$$

The proof is virtually identical to the proof of Theorem 17.

A *regression tree* is a partition estimator of the form

$$m(x) = \sum_{m=1}^{M} c_m I(x \in R_m) \tag{47}$$

where $c_1, \ldots, c_M$ are constants and $R_1, \ldots, R_M$ are disjoint rectangles that partition the space of covariates and whose sides are parallel to the coordinate axes. The model is fit in a greedy, recursive manner that can be represented as a tree; hence the name.

Denote a generic covariate value by $x = (x_1, \ldots, x_j, \ldots, x_d)$. The covariate for the i^{th} observation is $X_i = (X_{i1}, \ldots, X_{ij}, \ldots, X_{id})$. Given a covariate j and a split point s we define the rectangles $R_1 = R_1(j,s) = \{x : \ x_j \leq s\}$ and $R_2 = R_2(j,s) = \{x : \ x_j > s\}$ where, in this expression, x_j refers the the j^{th} covariate not the j^{th} observation. Then we take c_1 to be the average of all the Y_i's such that $X_i \in R_1$ and c_2 to be the average of all the Y_i's such that $X_i \in R_2$. Notice that c_1 and c_2 minimize the sums of squares $\sum_{X_i \in R_1}(Y_i - c_1)^2$ and $\sum_{X_i \in R_2}(Y_i - c_2)^2$. The choice of which covariate x_j to split on and which split point s to use is based on minimizing the residual sums if squares. The splitting process is on repeated on each rectangle R_1 and R_2.

Figure 10 shows a simple example of a regression tree; also shown are the corresponding rectangles. The function estimate $\widehat{m}$ is constant over the rectangles.

Generally one first grows a very large tree, then the tree is pruned to form a subtree by collapsing regions together. The size of the tree is a tuning parameter and is usually chosen by cross-validation.

Example 16 *Figure 11 shows a tree for the rock data. Notice that the variable shape does not appear in the tree. This means that the shape variable was never the optimal covariate to split on in the algorithm. The result is that tree only depends on area and peri. This illustrates an important feature of tree regression: it automatically performs variable selection in the sense that a covariate* x_j *will not appear in the tree if the algorithm finds that the variable is not important.*

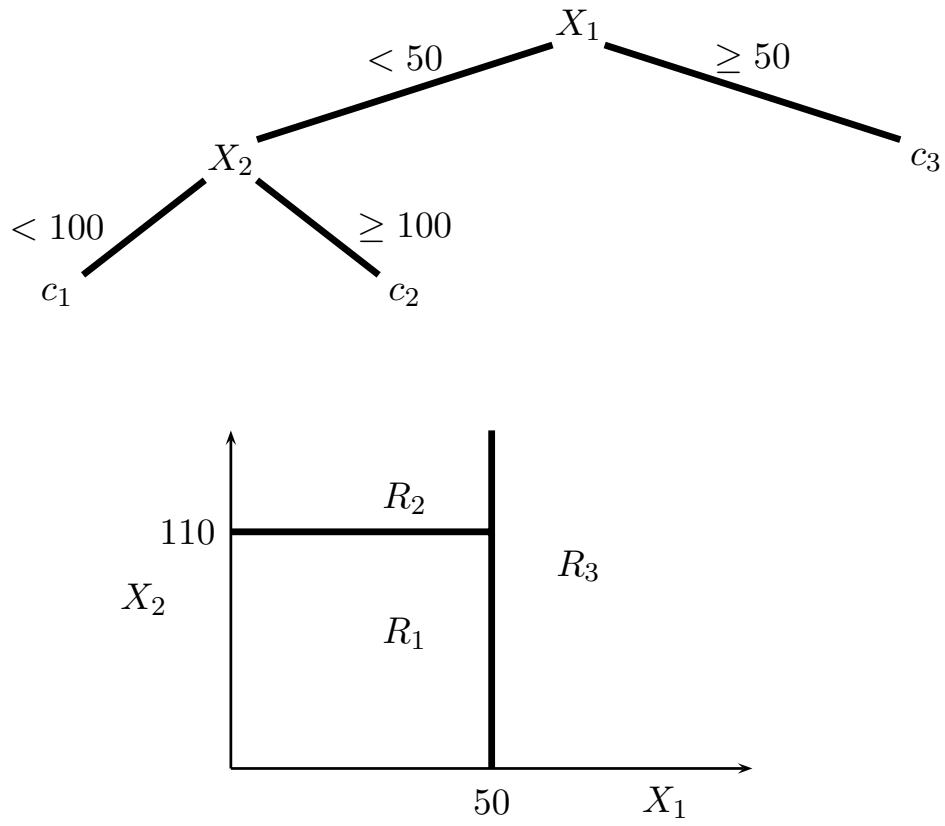

Figure 10: A regression tree for two covariates X_1 and X_2. The function estimate is $\widehat{m}(x) = c_1 I(x \in R_1) + c_2 I(x \in R_2) + c_3 I(x \in R_3)$ where R_1, R_2 and R_3 are the rectangles shown in the lower plot.

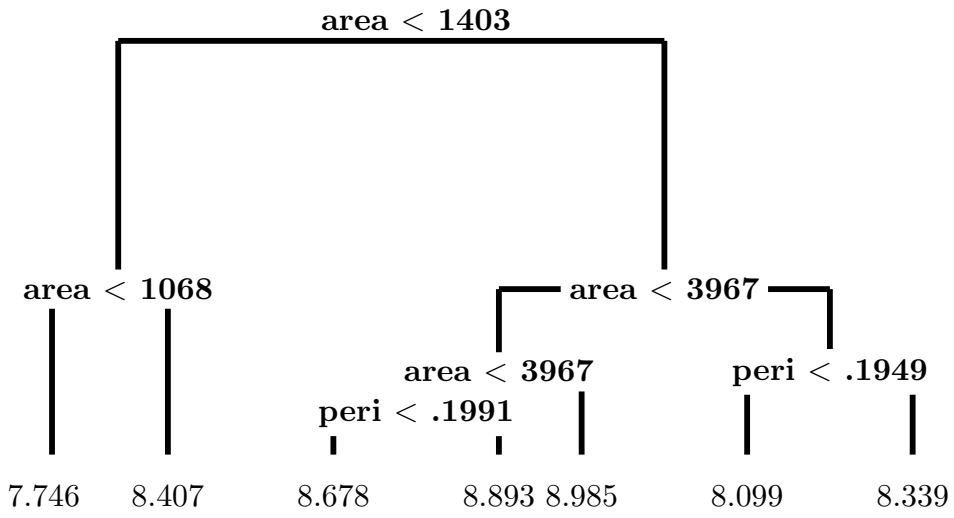

Figure 11: Regression tree for the rock data.

14 Confidence Bands

To get an idea of the precision of the estimate $\widehat{m}(x)$, one may compute a confidence band for $m(x)$. First, we get the squared residuals $r_i = (Y_i - \widehat{m}(X_i))^2$. If we regress the r_i's on the X_i's (using kernel regression for example) then we get an estimate of $\widehat{\sigma}^2(x)$. We then call $\widehat{m}(x) \pm 2\widehat{\sigma}(x)$ a *variability band*. It is not really a confidence band but it does give an idea of the variability of the estimator.

There are two reasons why this is not a confidence band. First, we have to account for uniformity over x. This can be done using the bootstrap. A more serious issue is bias. Note that

$$\frac{\sqrt{n}(\widehat{m}(x) - m(x))}{\widehat{\sigma}(x)} = \frac{\sqrt{n}(\widehat{m}(x) - \overline{m}(x))}{\widehat{\sigma}(x)} + \frac{\sqrt{n}(\overline{m}(x) - m(x))}{\widehat{\sigma}(x)}$$

where $\overline{m}(x) = \mathbb{E}[\widehat{m}(x)]$. Under weak conditions, the first term satisfies a central limit theorem. But the second term is the ratio of bias and variance. If we choose the bandwidth optimally, these terms are balanced. Hence, the second term does not go to 0 as $n \to \infty$. This means that the band is not centered at $m(x)$. To deal with this, we can estimate the bias and subtract the bias estimate. The resulting estimator *undersmooths*: it reduces bias but increases variance. The result will be an asymptotically valid confidence interval centered around a sub-optimal estimator.

The bottom line is this: in nonparametric estimation you can have optimal prediction or valid inferences but not both.

Appendix: Finite Sample Bounds for Kernel Regression

This analysis is from Györfi, Kohler, Krzyżak and Walk (2002). For simplicity, we will use the spherical kernel $K(\|x\|) = I(\|x\| \leq 1)$; the results can be extended to other kernels. Hence,

$$\widehat{m}(x) = \frac{\sum_{i=1}^n Y_i\, I(\|X_i - x\| \leq h)}{\sum_{i=1}^n I(\|X_i - x\| \leq h)} = \frac{\sum_{i=1}^n Y_i\, I(\|X_i - x\| \leq h)}{n\, P_n(B(x,h))}$$

where P_n is the empirical measure and $B(x,h) = \{u :\ \|x-u\| \leq h\}$. If the denominator is 0 we define $\widehat{m}(x) = 0$. Let

$$\mathcal{M} = \left\{ m :\ |m(x) - m(x)| \leq L\|x - z\|, \quad x, z, \in \mathbb{R}^d \right\}. \tag{48}$$

The proof of the following theorem is from Chapter 5 of Györfi, Kohler, Krzyżak and Walk (2002).

Theorem 17 *Suppose that the distribution of X has compact support and that* $\mathrm{Var}(Y|X = x) \leq \sigma^2 < \infty$ *for all x. Then*

$$\mathbb{E}\|\widehat{m} - m\|_P^2 \leq c_1 h^2 + \frac{c_2}{nh^d}. \tag{49}$$

Hence, if $h \asymp n^{-1/(d+2)}$ then

$$\mathbb{E}\|\widehat{m} - m\|_P^2 \leq \frac{c}{n^{2/(d+2)}}. \tag{50}$$

Note that the rate $n^{-2/(d+2)}$ is slower than the pointwise rate $n^{-4/(d+2)}$ because we have made weaker assumptions.

Proof. Let

$$m_h(x) = \frac{\sum_{i=1}^n m(X_i) I(\|X_i - x\| \leq h)}{nP_n(B(x,h))}.$$

Let $A_n = \{P_n(B(x,h)) > 0\}$. When A_n is true,

$$\mathbb{E}\left((\widehat{m}_h(x) - m_h(x))^2 \,\middle|\, X_1, \ldots, X_n \right) = \frac{\sum_{i=1}^n \mathrm{Var}(Y_i|X_i) I(\|X_i - x\| \leq h)}{n^2 P_n^2(B(x,h))} \leq \frac{\sigma^2}{nP_n(B(x,h))}.$$

Since $m \in \mathcal{M}$, we have that $|m(X_i) - m(x)| \leq L\|X_i - x\| \leq Lh$ for $X_i \in B(x,h)$ and hence

$$|m_h(x) - m(x)|^2 \leq L^2 h^2 + m^2(x) I_{A_n(x)^c}.$$

Therefore,

$$\begin{aligned}\mathbb{E}\int(\widehat{m}_h(x)-m(x))^2dP(x) &= \mathbb{E}\int(\widehat{m}_h(x)-m_h(x))^2dP(x)+\mathbb{E}\int(m_h(x)-m(x))^2dP(x)\\ &\leq \mathbb{E}\int\frac{\sigma^2}{nP_n(B(x,h))}I_{A_n(x)}dP(x)+L^2h^2+\int m^2(x)\mathbb{E}(I_{A_n(x)^c})dP(x). \quad (51)\end{aligned}$$

To bound the first term, let $Y = nP_n(B(x,h))$. Note that $Y \sim$ Binomial(n,q) where $q = \mathbb{P}(X \in B(x,h))$. Now,

$$\begin{aligned}\mathbb{E}\left(\frac{I(Y>0)}{Y}\right) &\leq \mathbb{E}\left(\frac{2}{1+Y}\right)=\sum_{k=0}^{n}\frac{2}{k+1}\binom{n}{k}q^k(1-q)^{n-k}\\ &= \frac{2}{(n+1)q}\sum_{k=0}^{n}\binom{n+1}{k+1}q^{k+1}(1-q)^{n-k}\\ &\leq \frac{2}{(n+1)q}\sum_{k=0}^{n+1}\binom{n+1}{k}q^{k}(1-q)^{n-k+1}\\ &= \frac{2}{(n+1)q}(q+(1-q))^{n+1}=\frac{2}{(n+1)q}\leq\frac{2}{nq}.\end{aligned}$$

Therefore,

$$\mathbb{E}\int\frac{\sigma^2 I_{A_n(x)}}{nP_n(B(x,h))}dP(x)\leq 2\sigma^2\int\frac{dP(x)}{nP(B(x,h))}.$$

We may choose points $z_1,\ldots,z_M$ such that the support of P is covered by $\bigcup_{j=1}^{M}B(z_j,h/2)$ where $M \leq c_2/(nh^d)$. Thus,

$$\begin{aligned}\int\frac{dP(x)}{nP(B(x,h))} &\leq \sum_{j=1}^{M}\int\frac{I(z\in B(z_j,h/2))}{nP(B(x,h))}dP(x)\leq\sum_{j=1}^{M}\int\frac{I(z\in B(z_j,h/2))}{nP(B(z_j,h/2))}dP(x)\\ &\leq \frac{M}{n}\leq\frac{c_1}{nh^d}.\end{aligned}$$

The third term in (51) is bounded by

$$\begin{aligned}\int m^2(x)\mathbb{E}(I_{A_n(x)^c})dP(x) &\leq \sup_x m^2(x)\int(1-P(B(x,h)))^n dP(x)\\ &\leq \sup_x m^2(x)\int e^{-nP(B(x,h))}dP(x)\\ &= \sup_x m^2(x)\int e^{-nP(B(x,h))}\frac{nP(B(x,h))}{nP(B(x,h))}dP(x)\\ &\leq \sup_x m^2(x)\sup_u(ue^{-u})\int\frac{1}{nP(B(x,h))}dP(x)\\ &\leq \sup_x m^2(x)\sup_u(ue^{-u})\frac{c_1}{nh^d}=\frac{c_2}{nh^d}.\end{aligned}$$

□

15 References

Györfi, Kohler, Krzyżak and Walk (2002). *A Distribution Free Theory of Nonparametric Regression.* Springer.

Hardle, Muller, Sperlich and Werwatz (2004). *Nonparametric and Semiparametric Models.* Springer.

Wasserman, L. (2006). *All of Nonparametric Statistics.* Springer.