University of Illinois
Fall 2016

Department of Economics
Roger Koenker

Economics 536
Lecture 23
Introduction to Non-Parametrics

1. Introduction

In this lecture I would like to introduce some basic ideas about non-parametrics and also to reenforce some principal themes of the course. There are several meanings of "non-parametrics" in econometrics and statistics, I will use it to refer to function estimation. Instead of estimating models characterized by a finite dimensional parameter, e.g.

$$\begin{aligned} E(y|x) &= \beta_0 + \beta_1 x \\ E(y|x) &= \beta_0 + \beta_1 x + \beta_2 x^2 \end{aligned}$$

we are going to consider models with (in principle) an infinite dimensional parameter. Fear not, though, most of our effort will be devoted to making this look just like the finite dimensional case.

2. Splines, Sieves and Basis Expansion

Suppose we would like to have a somewhat more flexible specification of the effect of a covariate than is afforded by the simple quadratic specifications that we have discussed thus far. There are many options that can be formulated in the following way:

$$E(y|x) = \sum_{i=0}^{p} \beta_i \varphi_i(x)$$

where the φ_i denote known functions that represent, or span, the class of plausible candidate conditional mean functions. The traditional choice of monomials, $\varphi_i(x) = x^i$ turns out to be quite unsatisfactory beyond the conventional quadratic, but there are many other alternatives. To illustrate this consider fitting the scatter plot in the next figure with global polynomials of various orders. As illustrated, the global nature of the polynomial basis makes it very difficult to control fit, and small changes in the data in one region can exert a deleterious effect on the fit far away. Another classical choice is the Fourier expansion in which $E(y|x)$ is expressed as a sum of sine and cosine terms with increasing frequency, but this is most advantageous in special periodic, time series situations. There are many other orthogonal systems of functions that can be used: representations like this are often called sieves since higher order terms capture finer structure of the function. A leading example of this are splines.

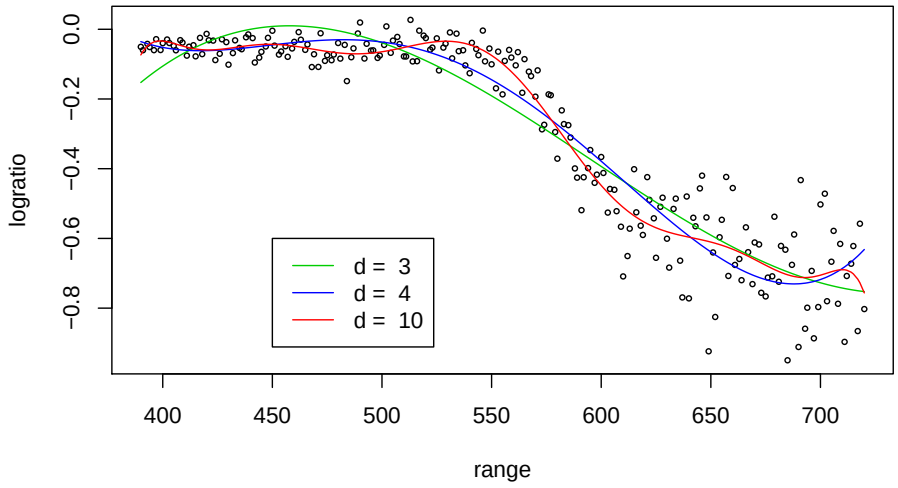

FIGURE 1. Polynomial Regression: The figure illustrates an example of global polynomial fitting. This is generally a bad idea since global polynomials are sensitive to perturbations of the data throughout the domain of x. This is especially true in the extremes of this domain. Local support as exemplified by the spline bases described below are much better in this respect.

2.1. **Splines:** A good way to generate flexible univariate functions is to consider piecewise polynomials. The simplest examples are linear splines. Denote the "positive part" function, $x_+ = \max\{0, x\}$ and consider functions represented as,

$$g(x) = \beta_0 + \sum_{i=1}^{p} \beta_i (x - \alpha_i)_+$$

where the α_i's denote fixed scalar values called knots, at which the function is allowed to bend. Between these knots the function is linear; it is continuous with piecewise constant derivative. When there are only two segments we have the socalled "broken stick" model illustrated in the left panel of Figure 1. While in the right panel we see that considerably more flexibility is possible when we introduce more knots. The fitted functions are simply linear combinations of the functions depicted in the lower panel, so it is natural to think of the collection of these functions as a linear vector space with these basis functions. Least squares fitting is always trying to find the element of this space that is closest to the observed response vector in the

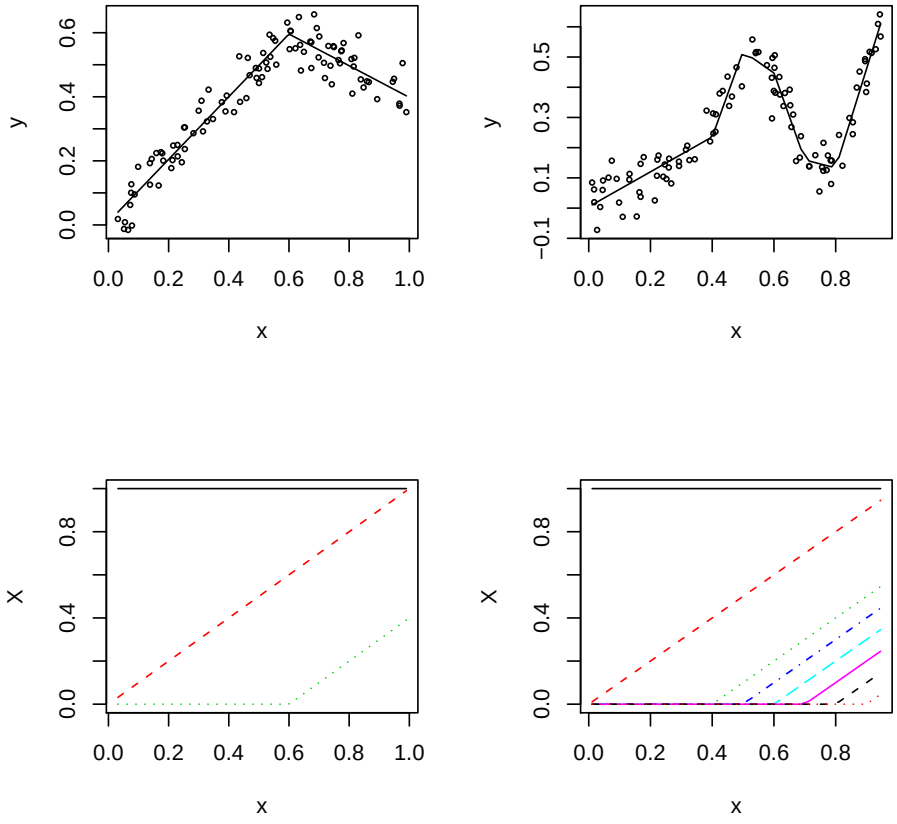

FIGURE 2. Linear Splines: The figure illustrates two examples of linear spline fitting. On the left we have a simple two segment example, on the right one sees 7 distinct segments. The underlying basis functions are shown below the respective scatterplots.

sense of solving,

$$\sum_{i=1}^{n}(y_i - g(x_i;\beta))^2.$$

Piecewise linear functions are simple and easy to interpret, but sometimes sharp kinks are undesirable. Piecewise cubic polynomials are often used as an alternative. In Figure 2 we depict basis functions, or B-splines, for a class of piecewise cubic functions with knots at the points indicated by points plotted on the $y = 0$ line. For cubic splines, not only do we have continuity of the function, but also of its first two derivatives. At the knots

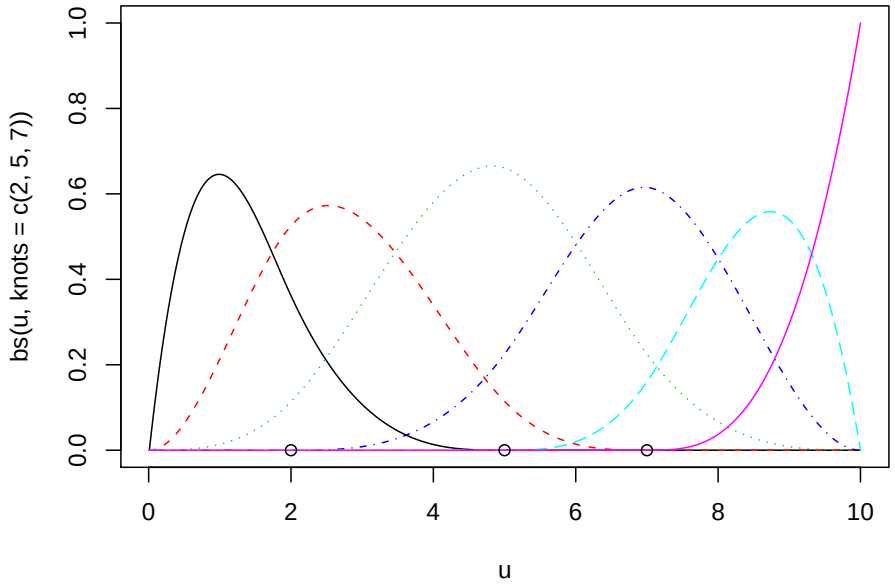

FIGURE 3. This figure illustrates some typical cubic B-spline basis functions. They are evaluated on an equally spaced grid from 0 to 10, interior knots are located at 2,5,7.

the third derivative is free to jump. This figure can be reproduced in R with the commands:

```
pdf("bspline.pdf",horizontal=FALSE,height = 5, width=6.5)
library(splines)
u <- 1:1000/100
matplot(u,bs(u,knots=c(2,5,7)),type="l")
points(c(2,5,7),rep(0,3))
dev.off()
```

A crucial specification decision for spline models is obviously the choice of the knot locations. Sometimes this is dictated by detailed knowledge of the functions to be fitted, but more frequently such knowledge is lacking and one has to rely on guesswork. Fortunately, it is typically the case that the fitted functions are not highly sensitive to the choice of knots. In the next subsection I describe methods that intentionally select too many knots and rely on "shrinkage" to control the variablity of the fitting.

We have focused thus far on what are sometimes called "scatterplot smoothers," that is methods for fitting a function to bivariate scatter plots. What if there is more than one covariate? A general answer to this question

is beyond the scope of this lecture, but a simple strategy can be described for some situations. If we are willing to assume that covariate effects are additive, then we can fit models of the form,

$$E(y|x) = \sum_{i=1}^{p} g_i(x_i)$$

where each component g_i takes the form of a spline. Or one can also fit partially linear models of the form,

$$E(y|x) = x^\top \beta + g(z)$$

where some covariates enter in a conventional linear fashion, and one or more enter as spline expansions. And finally as we have seen in PS 5, we can interact terms to obtain spline expansions that depend on discrete covariates, or spline expansions of interacted continuous variables.

3. Shrinkage and Bayesian Regression

One theme of the course has been the trade-off between bias and variance in model selection problems: too simple a model risks serious bias that may distort policy conclusions of the model, too complicated a model risks obscuring the important effects in a cloud of uncertainty. Until now we have tried to balance these risks by selecting a model that represents a compromise between our objectives. This requires a slightly schizophrenic viewpoint. On one hand we appear to believe that there are many possible models for our problem, but in the end only one will be taken seriously. (This is rationalized by Schwarz's 0-1 loss function, but often this all or nothing view of models isn't appropriate.)

In this lecture we will consider a new approach to reaching a compromise between simple and complex models. We will begin with a brief exposition of Bayesian methods for linear regression, in non-Bayesian statistical circles these methods are sometimes referred to as "shrinkage methods," or Stein-rule methods. After briefly discussing some connections to random effects models for panel data we will then consider non-parametric regression.

3.1. **Bayesian Regression.**

Consider (once more) the linear model,

$$y = X\beta + u$$

If we assume (as usual) that $u \sim \mathcal{N}(0, \sigma^2 I)$, then we have likelihood,

$$\mathcal{L}(b) = (2\pi)^{-n/2}\sigma^{-n}\exp\{-\frac{1}{2\sigma^2}(\hat{\beta} - b)'X'X(\hat{\beta} - b)\}$$

Suppose that we have a prior opinion that $\beta \sim \mathcal{N}(\beta_0, \Omega)$, i.e. that β has density

$$\pi(b) = (2\pi)^{-p/2}|\Omega|^{-1/2}\exp\{-\frac{1}{2}(b - \beta_0)'\Omega^{-1}(b - \beta_0)\}$$

Recall that Bayes rule says that we should update our prior opinion about β, to obtain,

$$p(b) = \frac{\mathcal{L}(b) \cdot \pi(b)}{\int \mathcal{L}(b)\pi(b)db}.$$

Focusing on the $\exp\{\cdot\}$ in the numerator we obtain, after some algebra,

$$p(b) = \kappa \exp\{\frac{1}{2}(b-\tilde{\beta})'(\sigma^{-2}X'X + \Omega^{-1})(b-\tilde{\beta})\}$$

where

$$\tilde{\beta} = (\sigma^{-2}(X'X) + \Omega^{-1})^{-1}(\sigma^{-2}(X'X)\hat{\beta} + \Omega^{-1}\beta_0)$$

The latter formula is yet another application of our general strategy for combining two estimates: we have $\hat{\beta}$ and β_0 and they have covariance matrices $\sigma^2(X'X)^{-1}$ and Ω respectively and they are combined accordingly – provided the normality assumptions are reasonable.

Some special cases to consider:

(1) When $\sigma^2 \to 0$ then the likelihood dominates the prior.
(2) When $\sigma^2 \to \infty$ then the prior dominates $\mathcal{L}$.
(3) As $n \to \infty$, typically we assume

$$n^{-1}X'X \to D \qquad \text{(positive definite)}$$

so $\sigma^{-2}(X'X) \to n\sigma^{-2}D$ and the factor n causes the $\mathcal{L}$ to dominate.
(4) Suppose

$$\Omega = \begin{bmatrix} w_0 I_p & 0 \\ 0 & w_1 I_q \end{bmatrix}$$

then by varying w's we get some interesting cases. E.G. $w_0 = \infty$ would express the view that we were clueless about the first p elements, but we might still want to shrink the last q elements towards their prior mean.

Some Practicalities

If we diagonalize $\Omega^{-1} = Q'Q$, then

$$\lambda Q\beta \sim \mathcal{N}(\lambda Q\beta_0, \lambda^2 Q\Omega Q') \sim \mathcal{N}(Q\beta_0, \lambda^2 I)$$

and we can now write the model augmented by the prior information as

$$\begin{pmatrix} y \\ \lambda Q\beta_0 \end{pmatrix} = \begin{pmatrix} X \\ \lambda Q \end{pmatrix} \beta + \begin{pmatrix} u \\ v \end{pmatrix}$$

where the vector $(u, v)' \sim \mathcal{N}(0, \tilde{\Omega})$ where

$$\tilde{\Omega} = \begin{bmatrix} \sigma^2 I & 0 \\ 0 & \lambda^2 I \end{bmatrix}.$$

and we can "run" this regression to get $\tilde{\beta}$ or "run" several of these regressions to get the whole "decolletage", $\tilde{\beta}(\lambda)$.

A special case is $\Omega = I$ so $Q = I$; this is often called ridge regression. All coefficients are shrunken toward zero.

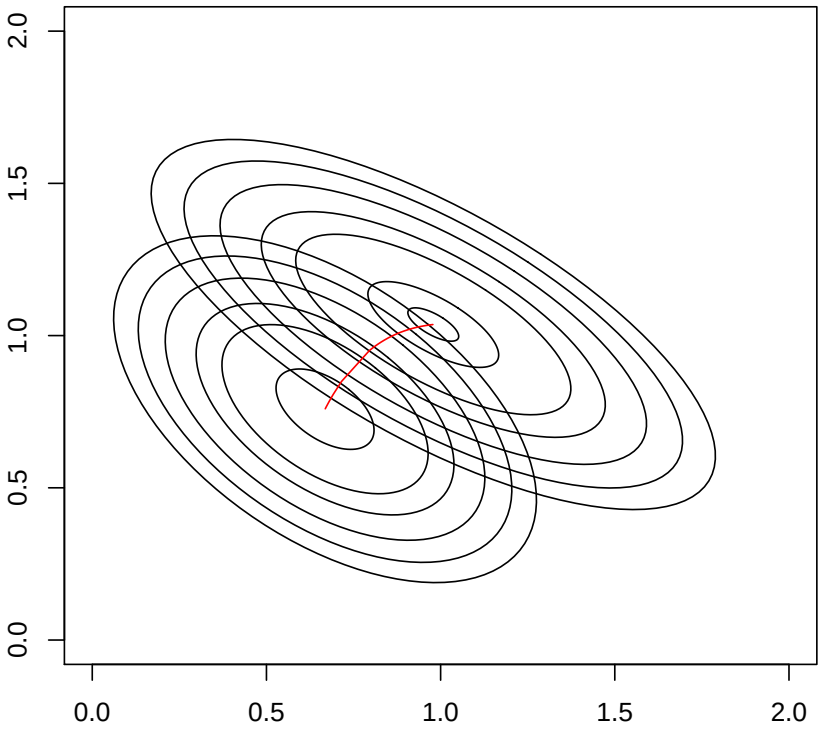

FIGURE 4. Bayesian Regression: Contours of the likelihood and prior are illustrated by the two families of concentric ellipses. The locus of tangencies between the contours of the prior and the contours of the likelihood represent a shrinkage path that connects the mle and the prior mean. Each point on the path can be interpreted as representing an intensity of belief in the prior.

3.2. **Panel Data.** Consider the panel data model of Lecture 13

$$y = X\beta + Z\alpha + u$$

where Z is a matrix of indicator variables representing the fixed effects. Suppose $u \sim \mathcal{N}(0, R)$ and $\alpha \sim \mathcal{N}(0, Q)$ then treating α as random we obtain,

$$\hat{\beta} = (X'(R + ZQZ')^{-1}X)^{-1}X'(R + ZQZ')^{-1}y$$

Why? Hint: Let $v = Z\alpha + u$, and show $Evv' = R + ZQZ'$. The following result connects this estimator to the earlier Bayesian shrinkage ideas.

Thm $\hat{\beta}$ solves $\min_{(\alpha,\beta)} \parallel y - X\beta - Z\alpha \parallel^2_{R^{-1}} + \parallel \alpha \parallel^2_{Q^{-1}}$.

Pf. Differentiating we have the normal equations

$$X'R^{-1}X\hat{\beta} + X'R^{-1}\hat{\alpha} = X'R^{-1}y$$
$$Z'R^{-1}X\hat{\beta} + (Z'R^{-1}Z + Q^{-1})\hat{\alpha} = Z'R^{-1}y$$

Now solve for $\hat{\alpha}$ in the first equation substitute into the second and then solve for $\hat{\beta}$ to get

$$\hat{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$$

where $\Omega^{-1} = R^{-1} - R^{-1}Z(Z'R^{-1}Z + Q^{-1})^{-1}Z'R^{-1}$. The result follows by verifying that $\Omega = R + ZQZ'$. □

This result has a long history. Goldberger (1962) introduced the phase best linear unbiased predictor, now sometimes abbreviated BLUP to refer to it. From our current viewpoint the interesting feature of $\hat{\beta}$ is that it shows that the random effects (GLS) estimator we have already studied can be viewed as a Bayesian estimator in which we estimate fixed effects, but shrink them toward a common value of 0. Again we can represent the "prior" term $\parallel \alpha \parallel^2_{R^{-1}}$ as a data augmentation.

4. Density Estimation

If we have a random sample $X_1, \ldots, X_n$ from a distribution F, with a smooth density $f = F'$ we might consider maximizing the log likelihood,

$$\max_f \sum \log f(X_i)$$

subject to the constraints that $f(x) \geq 0$ and $\int f(x)dx = 1$. Another way to write this problem ignoring for the moment the $f(x) \geq 0$ constraint is

$$\min - \int \log f(x) dF_n(x) + \mu \int f(x)dx$$

This problem has Euler function

$$-\frac{dF_n(x)}{f(x)} + \mu = 0$$

which has solution

$$\hat{f}(x) = dF_n(x).$$

This density has mass $1/n$ at each of the X_i.

A more reasonable estimator of f would introduce a prior for f that enabled us to shrink the very rough $\hat{f}$ just obtained toward something more reasonable. A proposal of Silverman's (1982) suggests the roughness penalty

$$R(f) = \int ((\log f)''')^2 dx.$$

The rationale for this is that at the normal model $(\log f)''' = (\log \phi)''' \equiv 0$ so this penalty shrinks toward the normal model.
Ex. Show that the penalty is zero for any normal density, not just the standard one, so this penalized mle

$$\min_f - \int \log f(x) dF_n(x) + \mu \int f(x) dx + \lambda R(f)$$

yields the $\eta(\hat{u}, \hat{\sigma}^2)$ density as $\lambda \to \infty$, where $\hat{\mu}, \hat{\sigma}^2$ are the usual mle estimates in the parametric normal model.

4.1. **Convolution and Kernel Density Estimation.** If we have a random variable X with distribution function F and density f and another r.v. Z with df G and density g, then what is the df of the r.v. $Y = X + Z$?

Recall that

$$\begin{aligned} P(Y = y) &= P(X + Z = y) \\ &= \int P(Z = y - X \big| X = x) P(X = x) dx \\ &= \int g(y - x) f(x) dx \end{aligned}$$

where the last line uses the fact that X and Z are $\perp\!\!\!\perp$ so conditional density of Z is equal to unconditional density.

Kernel density estimation smooths the empirical df by convolution.

We take X to be distributed as F_n

$$F_n(x) = n^{-1} \sum I(X_i \leq x)$$

then we smooth by convolution.

$$\begin{aligned} f_n(y) &= \int g(y - x) dF_n(x) \\ &= n^{-1} \sum_{i=1}^{n} g(y - X_i) \end{aligned}$$

so we are averaging the kernel density g evaluated at the points $y - X_i$. Clearly, the choice of the "bandwidth," the scale of the kernel function h is critical to this approach. There is a massive literature on this subject. To explicitly introduce the bandwidth we can suppose that g is a rescaled version of some standardized g_0, so that

$$g(x) = g_0(x/h)/h$$

Figure 2 illustrates several examples in which we have a small number of observations. The individual kernel functions are depicted in grey and the solid black line denotes the estimate f_n obtained by summation. There are nice ways to combine the foregoing approaches by penalization of the log likelihood of a kernel smoothed version of the density.

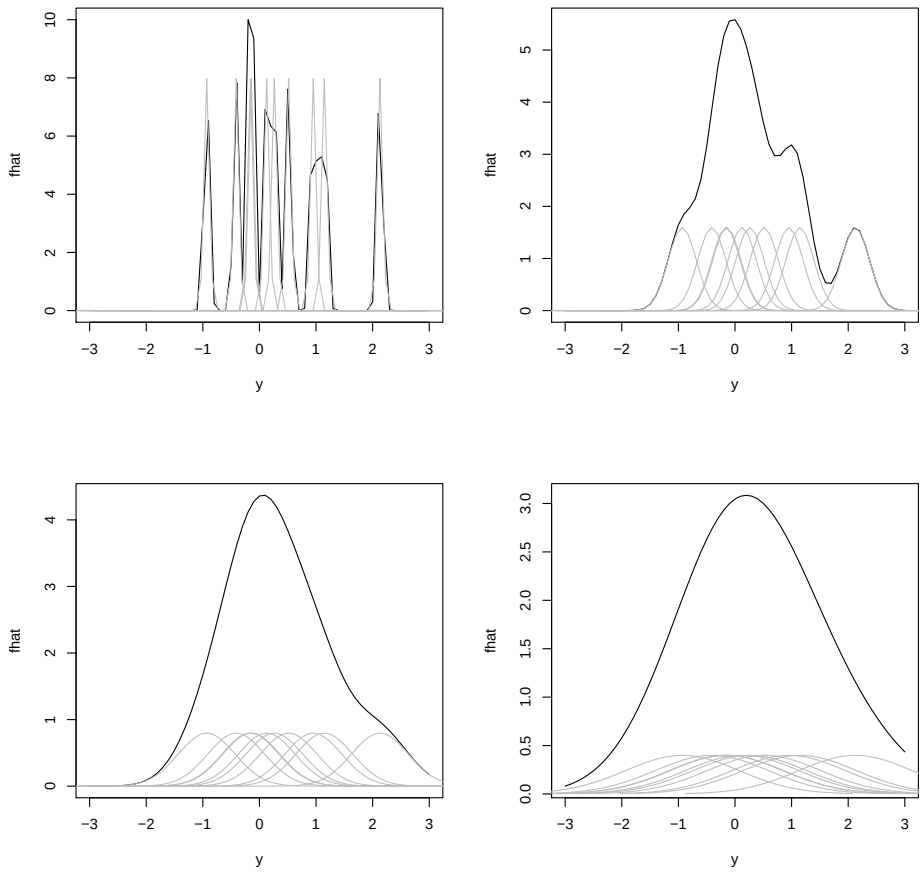

FIGURE 5. Kernel Density Estimation: The figure illustrates kernel density estimation for four different bandwidths applied to a small (n = 10) sample from the standard normal distribution. The grey constituent kernels are averaged to obtain the black estimate superimposed on the plot. The bandwidths for the four panels are $\{.05, .25, .5, 1\}$.

5. Nonparametric Regression

At this point you are probably wondering what does all this have to do with nonparametric regression estimation? There obviously isn't much time remaining to answer this question, but I'll try to sketch some basic ideas in one leading example.

5.1. **Kernel Regression and Local Polynomial Regression.** One can use kernel density estimation to estimate conditional mean regression functions. The basic idea was almost simultaneously proposed by Nadaraya and

Watson. Consider a general regression model

$$y_i = g(x_i) + u_i$$

where $g(x) = E(Y|X = x)$, in order to estimate g we may consider

$$E(Y|X = x) = \int y \frac{f(x, y)}{f(x)} dy$$

Now suppose we use Kernel density estimation to estimate both $f(x, y)$ and $f(x)$. This is the basic idea of Nadaraya-Watson estimator. In the numerator we have

$$\hat{f}(x, y) = n^{-1} \sum K_{h_1}(x - X_i) K_{h_2}(y - Y_i)$$

then

$$\begin{aligned}
\int y \hat{f}(x, y) dy &= n^{-1} \sum \int K_{h_1}(x - X_i) y K_{h_2}(y - Y_i) dy \\
&= n^{-1} \sum K_{h_1}(x - X_i) \int \frac{y}{h_2} K\left(\frac{y - Y_i}{h_2}\right) dy \\
&= n^{-1} \sum K_{h_1}(x - X_i) \int (s h_2 + Y_i) K(s) ds \\
&= n^{-1} \sum K_{h_1}(x - X_i) Y_i
\end{aligned}$$

Since $\int sK(s)ds = 0$ and $\int K(s)ds = 1$. Thus we have

$$\begin{aligned}
\hat{g}_h(x) &= n^{-1} \frac{\sum K_h(x - X_i) Y_i}{n^{-1} \sum K_h(x - X_i)} \\
&= \sum w_{h_i}(x) Y_i \\
\text{where} \quad w_{h_i}(x) &= \frac{(nh)^{-1} \sum K((x - X_i)/h)}{\hat{f}_h(x)}.
\end{aligned}$$

So at x the estimate of $E(Y|X = x)$ is a weighted average of the Y_i "near x". The problem with this idea is that we are really assuming a piecewise constant model for $E(y|x)$ and this is probably not very reasonable in most applications. Recent work has emphasized similar methods, but replacing the piecewise constant model with a piecewise linear, or piecewise polynomial model. We won't dwell on this, but instead briefly describe another approach that is more directly related to penalty methods.

5.2. **Smoothing Splines.** Consider the model

$$y_i = g_0(x_i) + u_i \qquad i = 1, \ldots, n$$

again we might begin by assuming that $u \sim \mathcal{N}(0, \sigma^2 I)$. We need to make some sort of assumptions about the form of g_0. Until now we say, "oh, lets' make it linear, or quadratic, or ...". Now we can assume that it might be linear, so we will take as a "prior for g",

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

This will play the role of $\| \alpha \|$ in the previous section. When g is linear then $P(g) = 0$ then g is very smooth, otherwise $P(g)$ is a measure of roughness and we choose g to minimize

$$(*) \qquad \sum (y_i - g(x_i))^2 + \lambda \int (g''(x))^2 dx$$

to balance fidelity to the data and fit.

A nice application of the usual Euler-Lagrange theory of the calculus of variations implies that our minimizer must satisfy the condition that

$$g''''(x) = 0$$

except at a finite set of points. This means $\hat{g}$ is a piecewise cubic *spline*.

The algebra turns out to be quite similar to the prior examples we can write $(*)$ as

$$\| y - a \|^2 + \lambda \| c \|_R^2$$

where $a = (g(x_i))$ and $c = (g''(x_i))$ but by continuity $Rc = Q'a$ for known R and Q depending only on the x's, so we can write

$$\| y - a \|^2 + \lambda a' Q R^{-1} Q' a$$

and this leads to an augmented data problem

$$\begin{pmatrix} y \\ 0 \end{pmatrix} = \begin{bmatrix} I \\ \lambda Q R^{-1} Q' \end{bmatrix} \; a \; + \begin{pmatrix} u \\ v \end{pmatrix}$$

so

$$\begin{aligned} \hat{a} &= (I - \lambda Q R^{-1} Q') y \\ &\equiv A(\lambda) y \end{aligned}$$

So $A(\lambda)$ plays a role something like $P = X(X'X)^{-1}X'$ in ordinary regression.

A simple example of this approach is the problem of smoothing equally spaced time series data, in macroeconomics this is sometimes referred to as the Hodrick-Prescott filter. For equally spaced x's the penalty may be written as

$$J(p) = \sum_{t=3}^{T} (\Delta^2 y_t)^2$$

and the computation becomes quite easy. A nice example of the power of the R language is the following program to implement this smoothing method

```
hpfilter <- function(y,lambda=1600){
eye <- diag(length(y))
solve(eye+lambda*crossprod(diff(eye,d=2)),y)
}
```

Try the following example:

```
x <- (1:1000)/50
y <- sin(x)+rnorm(1000)/5
plot(y)
```

```
lines(hpfilter(y))
lines(sin(x),col=‘‘red’’)
```

A very intriguing problem is how to choose λ. If $A(\lambda)$ were a projection matrix we could use Trace$A(\lambda) = p_\lambda$ as a measure of the dimension of the model. Recall that

$$\text{Trace}(P_X) = \text{Trace}((X'X)^{-1}X'X) = p$$

where p is the rank of X. Something similar can be done here even though $A(\lambda)$ isn't strictly a projection. Rather strangely we obtain a measure of the dimensionality of the fitted model that is real valued, not necessarily an integer, but this isn't really a problem and allows us to consider AIC, SIC, etc as model selection criteria for choosing λ.

5.3. **Quantile Smoothing Splines.** We have stressed estimation of conditional mean functions thus far, but similar methods can be used to estimate conditional quantile functions. In Koenker, Ng, and Portnoy (1994) it is proposed to estimate,

$$\min_g \sum \rho_\tau(y_i - g(x_i)) + \lambda \int |g''(x)|dx.$$

Note that the absolute value in the penalty term is less sensitive to sharp bends in he fitted function, This penalty leads to piecewise linear solutions with the tuning parameter, λ, controling the number of distinct segments. This approach is implemented in the `rqss` function in the R package `quantreg`. An advantage of this approach is that it is relatively easy to impose qualitative constraints such as monotonicity or convexity on the fitted functions. We illustrate an application of this approach in the final figure of the lecture.

References

Ruppert, D., M.P. Wand, and R.J. Carroll (2003), *Semiparametric Regression*, Cambridge U. Press.

Koenker, R. P. Ng, and S. Portnoy (1994) Quantile Smoothing Splines, *Biometrika* 81, 673–680.

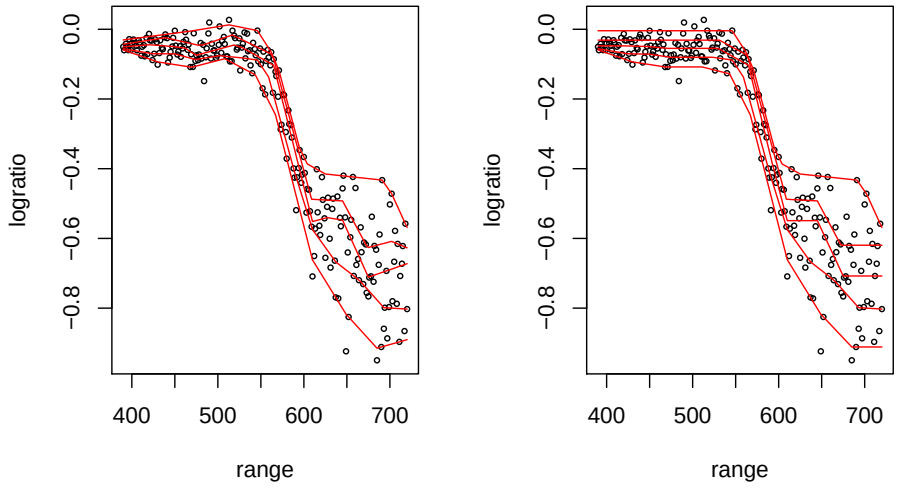

FIGURE 6. Quantile Smoothing Splines: The figure illustrates data from a light detection and ranging (lidar) experiment. See Ruppert, Wand, and Carroll (2003) for further details. We superimpose 5 conditional quantile function estimates on the scatterplot corresponding to the $\{0.05, 0.25, 0.50, 0.75, 0.95\}$ values of the parameter τ. The tuning parameter λ is fixed at the value 10 for all of the fitting. In the left panel the fitting is unconstrained, in the right panel the fitted functions are constrained to be monotone decreasing.