

Lecture 7: Smoothing

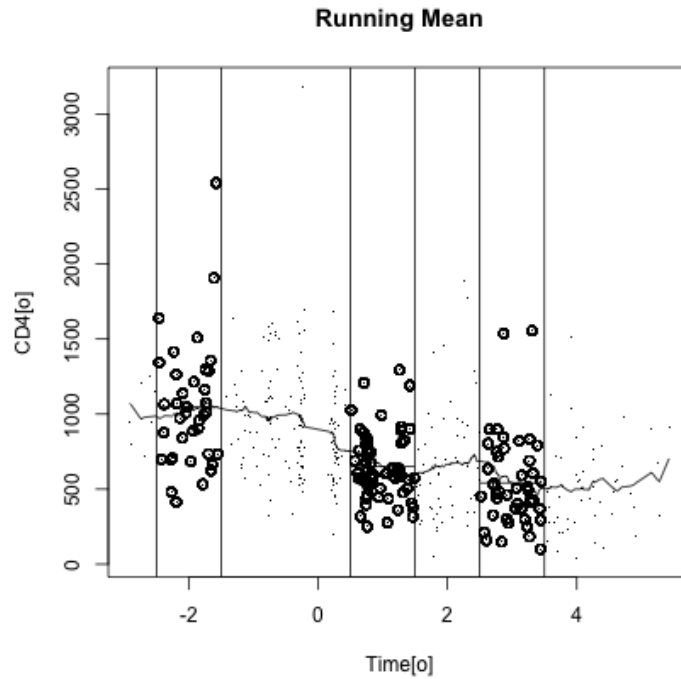
Rafael A. Irizarry and Hector Corrada Bravo

March, 2010

Kernel Methods

Below is the results of using running mean (K nearest neighbor) to estimate the effect of time to zero conversion on CD4 cell count.

Figure 1: Running mean estimate: CD4 cell count since zeroconversion for HIV infected men.



One of the reasons why the running mean (seen in Figure 1) is wiggly is because when we move from x_i to x_{i+1} two points are usually changed in the group

we average. If the new two points are very different then $s(x_i)$ and $s(x_{i+1})$ may be quite different. One way to try and fix this is by making the transition smoother. That's one of the main goals of kernel smoothers.

Kernel Smoothers

Generally speaking a kernel smoother defines a set of weights $\{W_i(x)\}_{i=1}^n$ for each x and defines

$$\hat{f}(x) = \sum_{i=1}^n W_i(x)y_i. \quad (1)$$

Most smoothers can be considered to be kernel smoothers in this very general definition. However, what is called a kernel smoother in practice has a simple approach to represent the weight sequence $\{W_i(x)\}_{i=1}^n$: by describing the shape of the weight function $W_i(x)$ via a density function with a scale parameter that adjusts the size and the form of the weights near x . It is common to refer to this shape function as a *kernel* K . The kernel is a continuous, bounded, and symmetric real function K which integrates to one:

$$\int K(u) du = 1.$$

For a given scale parameter h , the weight sequence is then defined by

$$W_{hi}(x) = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)}$$

Notice: $\sum_{i=1}^n W_{hi}(x_i) = 1$

The kernel smoother is then defined for any x as before by

$$\hat{f}(x) = \sum_{i=1}^n W_{hi}(x)Y_i.$$

Because we think points that are close together are similar, a kernel smoother usually defines weights that decrease in a smooth fashion as one moves away from the target point.

Running mean smoothers are kernel smoothers that use a “box” kernel. A natural candidate for K is the standard Gaussian density. (This is very inconvenient computationally because its never 0). This smooth is shown in Figure 2 for $h = 1$ year.

In Figure 3 we can see the weight sequence for the box and Gaussian kernels for three values of x .

Figure 2: CD4 cell count since zeroconversion for HIV infected men.

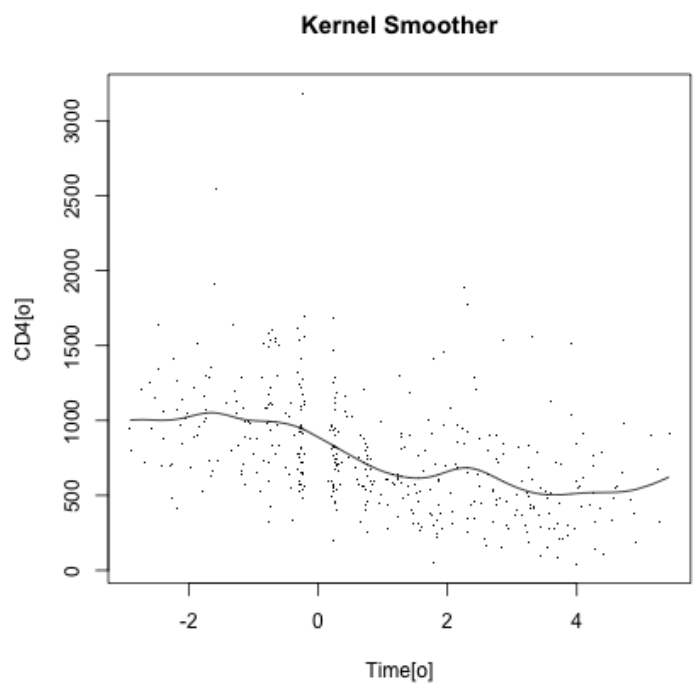
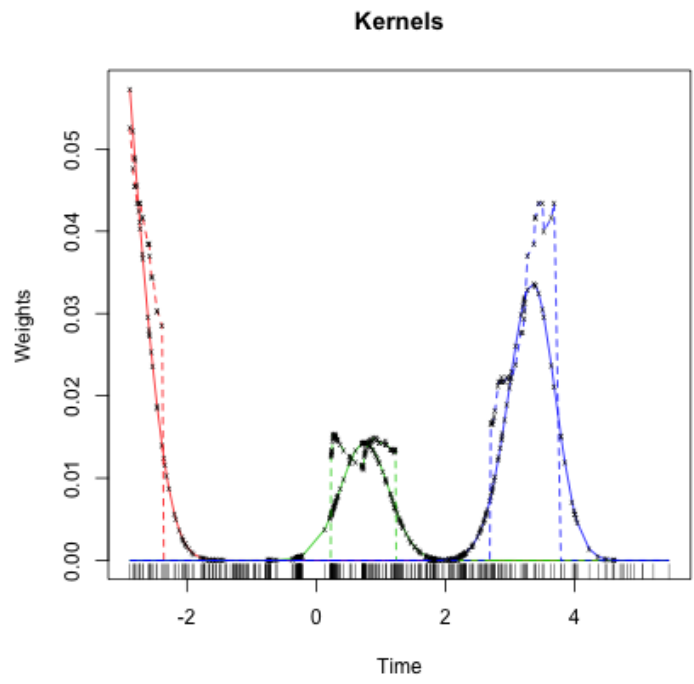


Figure 3: CD4 cell count since zeroconversion for HIV infected men.



Technical Note: An Asymptotic result

For the asymptotic theory presented here we will assume the stochastic design model with a one-dimensional covariate.

For the first time in this Chapter we will set down a specific stochastic model. Assume we have n IID observations of the random variables (X, Y) and that

$$Y_i = f(X_i) + \varepsilon_i, i = 1, \dots, n \quad (2)$$

where X has marginal distribution $f_X(x)$ and the ε_i IID errors independent of the X . A common extra assumption is that the errors are normally distributed. We are now going to let n go to infinity. . . What does that mean?

For each n we define an estimate for $f(x)$ using the kernel smoother with scale parameter h_n .

Theorem 1. *Under the following assumptions*

1. $\int |K(u)| du < \infty$
2. $\lim_{|u| \rightarrow \infty} uK(u) = 0$
3. $E(Y^2) \leq \infty$
4. $n \rightarrow \infty, h_n \rightarrow 0, nh_n \rightarrow \infty$

Then, at every point of continuity of $f(x)$ and $f_X(x)$ we have

$$\frac{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) y_i}{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)} \rightarrow f(x) \text{ in probability.}$$

Proof: Optional homework. Hint: Start by proving the fixed design model.

Local Regression

Local regression is used to model a relation between a predictor variable and response variable. To keep things simple we will consider the fixed design model. We assume a model of the form

$$Y_i = f(x_i) + \varepsilon_i$$

where $f(x)$ is an unknown function and ε_i is an error term, representing random errors in the observations or variability from sources not included in the x_i .

We assume the errors ε_i are IID with mean 0 and finite variance $\text{var}(\varepsilon_i) = \sigma^2$.

We make no global assumptions about the function f but assume that locally it can be well approximated with a member of a simple class of parametric function, e.g. a constant or straight line. Taylor's theorem says that any continuous function can be approximated with polynomial.

Technical note: Taylor's theorem

We are going to show three forms of Taylor's theorem.

- This is the original. Suppose f is a real function on $[a, b]$, $f^{(K-1)}$ is continuous on $[a, b]$, $f^{(K)}(t)$ is bounded for $t \in (a, b)$ then for any distinct points $x_0 < x_1$ in $[a, b]$ there exist a point x between $x_0 < x < x_1$ such that

$$f(x_1) = f(x_0) + \sum_{k=1}^{K-1} \frac{f^{(k)}(x_0)}{k!} (x_1 - x_0)^k + \frac{f^{(K)}(x)}{K!} (x_1 - x_0)^K.$$

Notice: if we view $f(x_0) + \sum_{k=1}^{K-1} \frac{f^{(k)}(x_0)}{k!} (x_1 - x_0)^k$ as function of x_1 , it's a polynomial in the family of polynomials

$$\mathcal{P}_{K+1} = \{f(x) = a_0 + a_1x + \dots + a_Kx^K, (a_0, \dots, a_K)' \in \mathbb{R}^{K+1}\}.$$

- Statistician sometimes use what is called Young's form of Taylor's Theorem:

Let f be such that $f^{(K)}(x_0)$ is bounded for x_0 then

$$f(x) = f(x_0) + \sum_{k=1}^K \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k + o(|x - x_0|^K), \text{ as } |x - x_0| \rightarrow 0.$$

Notice: again the first two term of the right hand side is in \mathcal{P}_{K+1} .

- In some of the asymptotic theory presented in this class we are going to use another refinement of Taylor's theorem called Jackson's Inequality:

Suppose f is a real function on $[a, b]$ with K is continuous derivatives then

$$\min_{g \in \mathcal{P}_k} \sup_{x \in [a, b]} |g(x) - f(x)| \leq C \left(\frac{b-a}{2k} \right)^K$$

with \mathcal{P}_k the linear space of polynomials of degree k .

Fitting local polynomials

Local weighter regression, or loess, or lowess, is one of the most popular smoothing procedures. It is a type of kernel smoother. The default algorithm for loess adds an extra step to avoid the negative effect of influential outliers.

We will now define the recipe to obtain a loess smooth for a target covariate x_0 .

The first step in loess is to define a weight function (similar to the kernel K we defined for kernel smoothers). For computational and theoretical purposes we

will define this weight function so that only values within a *smoothing window* $[x_0 + h(x_0), x_0 - h(x_0)]$ will be considered in the estimate of $f(x_0)$.

Notice: In local regression $h(x_0)$ is called the span or bandwidth. It is like the kernel smoother scale parameter h . As will be seen a bit later, in local regression, the span may depend on the target covariate x_0 .

This is easily achieved by considering weight functions that are 0 outside of $[-1, 1]$. For example Tukey's tri-weight function

$$W(u) = \begin{cases} (1 - |u|^3)^3 & |u| \leq 1 \\ 0 & |u| > 1. \end{cases}$$

The weight sequence is then easily defined by

$$w_i(x_0) = W\left(\frac{x_i - x_0}{h(x)}\right)$$

We define a window by a procedure similar to the k nearest points. We want to include $\alpha \times 100\%$ of the data.

Within the smoothing window, $f(x)$ is approximated by a polynomial. For example, a quadratic approximation

$$f(x) \approx \beta_0 + \beta_1(x - x_0) + \frac{1}{2}\beta_2(x - x_0)^2 \text{ for } x \in [x_0 - h(x_0), x_0 + h(x_0)].$$

For continuous function, Taylor's theorem tells us something about how good an approximation this is.

To obtain the local regression estimate $\hat{f}(x_0)$ we simply find the $\beta = (\beta_0, \beta_1, \beta_2)'$ that minimizes

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^3} \sum_{i=1}^n w_i(x_0) [Y_i - \{\beta_0 + \beta_1(x_i - x_0) + \frac{1}{2}\beta_2(x_i - x_0)\}]^2$$

and define $\hat{f}(x_0) = \hat{\beta}_0$.

Notice that the Kernel smoother is a special case of local regression. Proving this is a Homework problem.

Defining the span

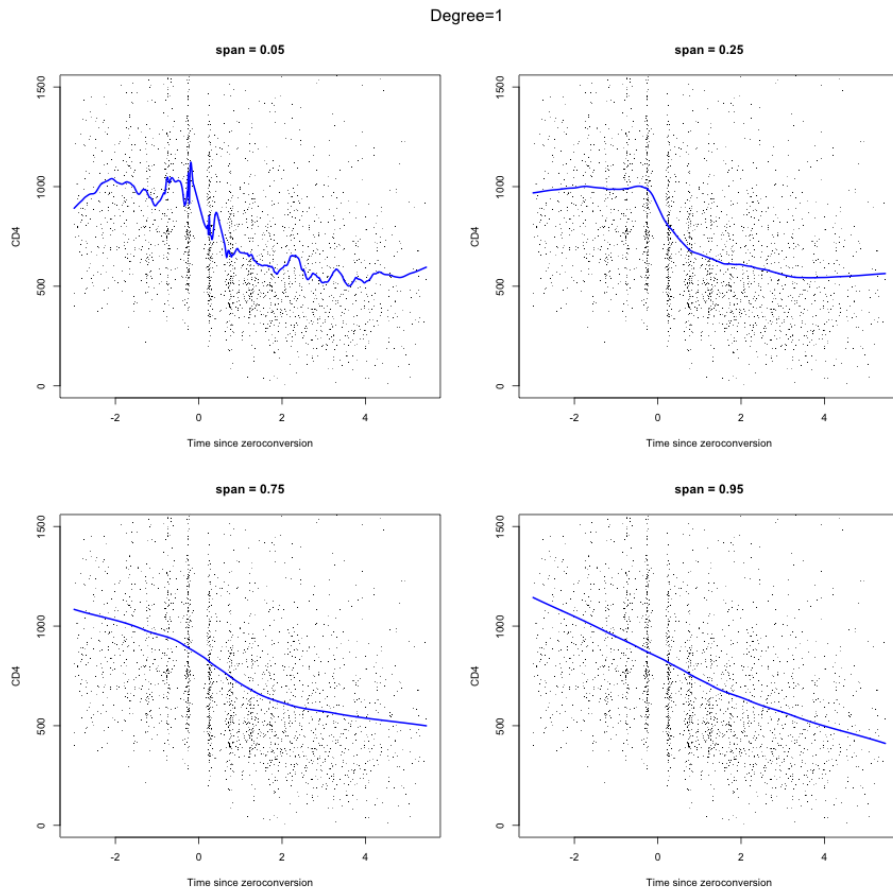
In practice, it is quite common to have the x_i irregularly spaced. If we have a fixed span h then one may have local estimates based on many points and others is very few. For this reason we may want to consider a nearest neighbor strategy to define a span for each target covariate x_0 .

Define $\Delta_i(x_0) = |x_0 - x_i|$, let $\Delta_{(i)}(x_0)$ be the ordered values of such distances. One of the arguments in the local regression function `loess()` is the `span`. A span of α means that for each local fit we want to use $\alpha \times 100\%$ of the data.

Let q be equal to αn truncated to an integer. Then we define the span $h(x_0) = \Delta_{(q)}(x_0)$. As α increases the estimate becomes smoother.

In Figures 4 — 6 we see loess smooths for the CD4 cell count data using spans of 0.05, 0.25, 0.75, and 0.95. The smooth presented in the Figures are fitting a constant, line, and parabola respectively.

Figure 4: CD4 cell count since zeroconversion for HIV infected men.



Symmetric errors and Robust fitting

If the errors have a symmetric distribution (with long tails), or if there appears to be outliers we can use robust loss.

Figure 5: CD4 cell count since zeroconversion for HIV infected men.

Degree=2, the default

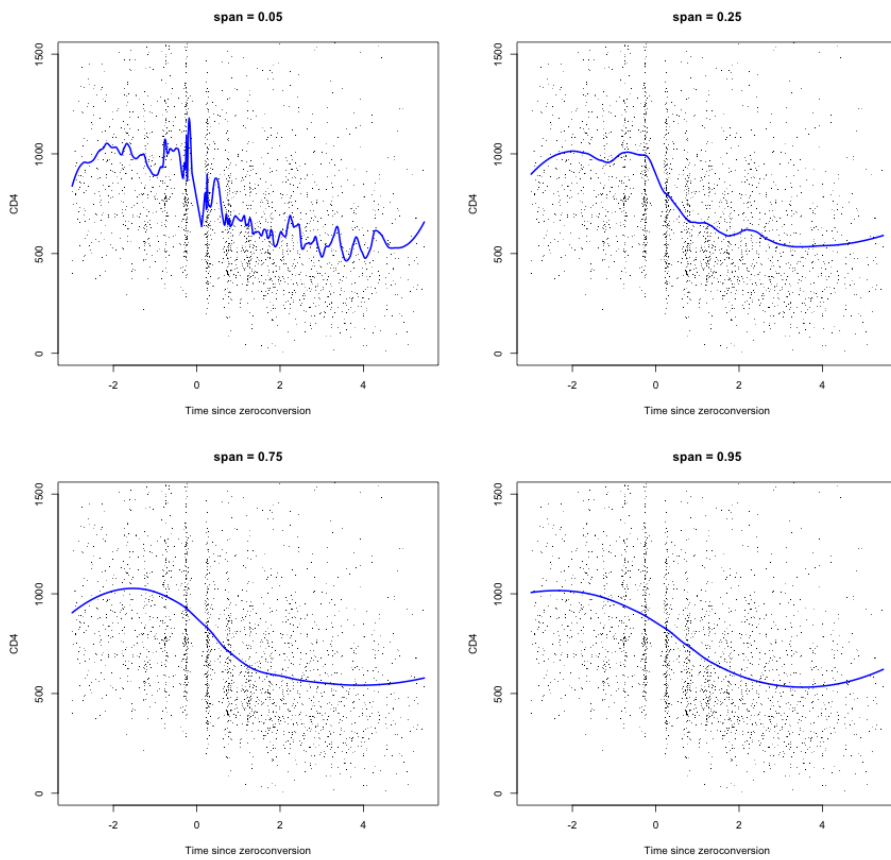
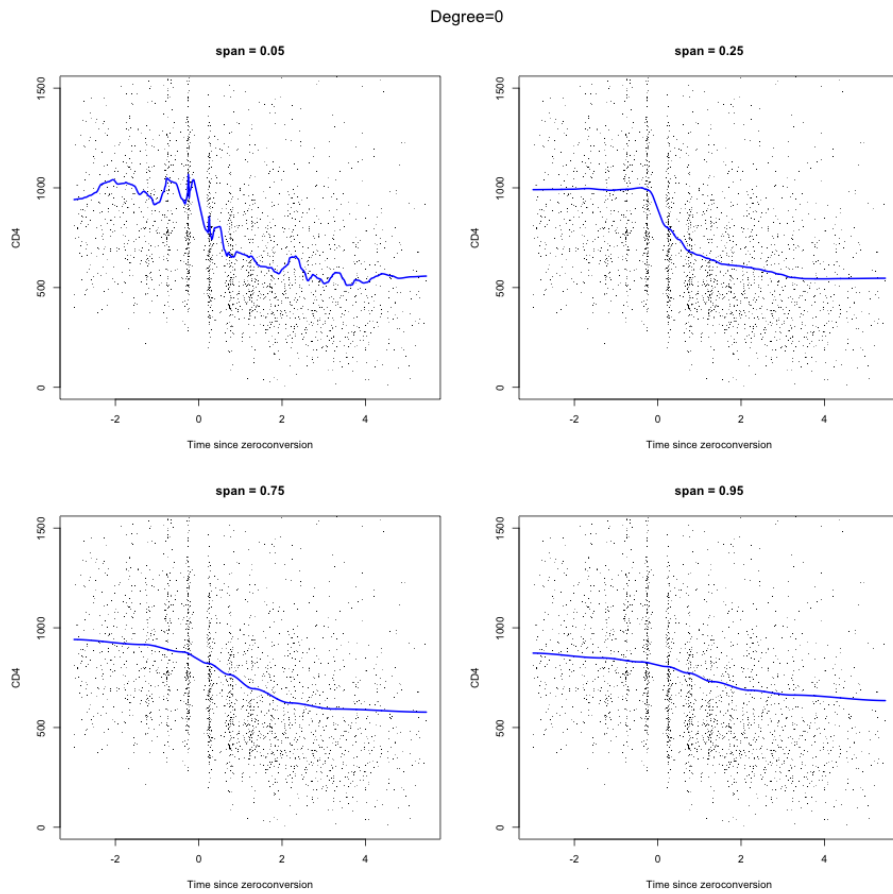


Figure 6: CD4 cell count since zeroconversion for HIV infected men.



We begin with the estimate described above $\hat{f}(x)$. The residuals

$$\hat{\varepsilon}_i = y_i - \hat{f}(x_i)$$

are computed.

Let

$$B(u; b) = \begin{cases} \{1 - (u/b)^2\}^2 & |u| < b \\ 0 & |u| \geq b \end{cases}$$

be the bisquare weight function. Let $m = \text{median}(|\hat{\varepsilon}_i|)$. The robust weights are

$$r_i = B(\hat{\varepsilon}_i; 6m)$$

The local regression is repeated but with new weights $r_i w_i(x)$. The robust estimate is the result of repeating the procedure several times.

If we believe the variance $\text{var}(\varepsilon_i) = a_i \sigma^2$ we could also use this double-weight procedure with $r_i = 1/a_i$.

Multivariate Local Regression

Because Taylor's theorems also applies to multidimensional functions it is relatively straight forward to extend local regression to cases where we have more than one covariate. For example if we have a regression model for two covariates

$$Y_i = f(x_{i1}, x_{i2}) + \varepsilon_i$$

with $f(x, y)$ unknown. Around a target point $\mathbf{x}_0 = (x_{01}, x_{02})$ a local quadratic approximation is now

$$f(x_1, x_2) \approx \beta_0 + \beta_1(x_1 - x_{01}) + \beta_2(x_2 - x_{02}) + \beta_3(x_1 - x_{01})(x_2 - x_{02}) + \frac{1}{2}\beta_4(x_1 - x_{01})^2 + \frac{1}{2}\beta_5(x_2 - x_{02})^2$$

Once we define a distance, between a point \mathbf{x} and \mathbf{x}_0 , and a span h we can define define weights as in the previous sections:

$$w_i(\mathbf{x}_0) = W\left(\frac{\|\mathbf{x}_i, \mathbf{x}_0\|}{h}\right).$$

It makes sense to re-scale x_1 and x_2 so we smooth the same way in both directions. This can be done through the distance function, for example by defining a distance for the space \mathbb{R}^d with

$$\|\mathbf{x}\|^2 = \sum_{j=1}^d (x_j/v_j)^2$$

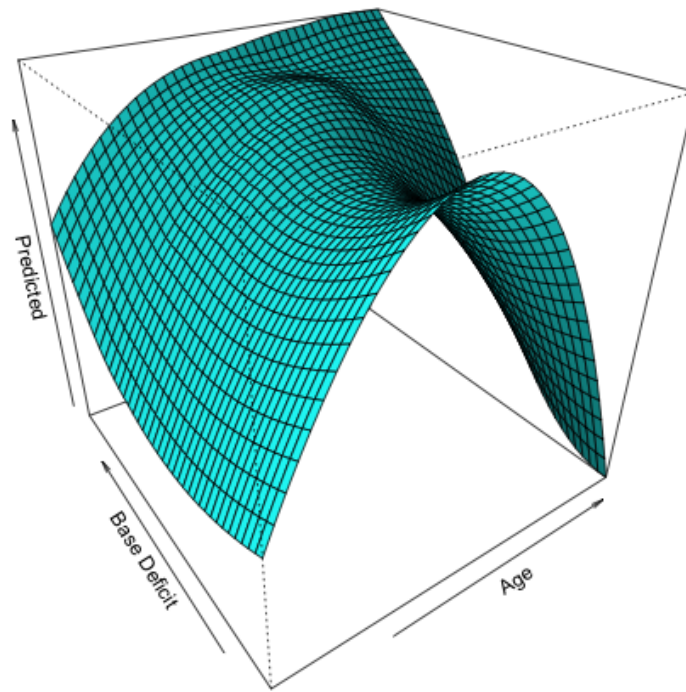
with v_j a scale for dimension j . A natural choice for these v_j are the standard deviation of the covariates.

Notice: We have not talked about k-nearest neighbors. As we will see in later the *curse of dimensionality* will make this hard.

Example

We look at part of the data obtained from a study by Socket et. al. (1987) on the factors affecting patterns of insulin-dependent diabetes mellitus in children. The objective was to investigate the dependence of the level of serum C-peptide on various other factors in order to understand the patterns of residual insulin secretion. The response measurement is the logarithm of C-peptide concentration (pmol/ml) at diagnosis, and the predictors are age and base deficit, a measure of acidity. In Figure 7 we show a loess two dimensional smooth. Notice that the effect of age is clearly non-linear.

Figure 7: Loess fit for predicting C.Peptide from Base.deficit and Age.



Linear Smoothers: Influence, Variance, and Degrees of Freedom

All the smoothers presented in this course are linear smoothers. This means that we can think of them as version of Kernel smoothers because every estimate $\hat{f}(x)$ is a linear combination of the data Y thus we can write it in the form of equation (1).

If we forget about estimating f at every possible x and consider only the observed (or design) points x_1, \dots, x_n , we can write equation (1) as

$$\hat{\mathbf{f}} = \mathbf{S}\mathbf{y}.$$

Here $\mathbf{f} = \{f(x_1), \dots, f(x_n)\}$ and \mathbf{S} is defined by the algorithm we are using.

Question: What is \mathbf{S} for linear regression? How about for the kernel smoother defined above?

How can we characterize the amount of smoothing being performed? The smoothing parameters provide a characterization, but it is not ideal because it does not permit us to compare between different smoothers and for smoothers like loess it does not take into account the shape of the weight function nor the degree of the polynomial being fit.

We now use the connections between smoothing and multivariate linear regression (they are both linear smoothers) to characterize pointwise criteria that characterize the amount of smoothing at a single point and global criteria that characterize the global amount of smoothing.

We will define variance reduction, influence, and degrees of freedom for linear smoothers.

The variance of the interpolation estimate is $\text{var}[Y_1] = \sigma^2$. The variance of our smooth estimate is

$$\text{var}[\hat{f}(x)] = \sigma^2 \sum_{i=1}^n W_i^2(x)$$

so we define $\sum_{i=1}^n W_i^2(x)$ as the variance reduction. Under mild conditions one can show that this is less than 1.

Because

$$\sum_{i=1}^n \text{var}[\hat{f}(x_i)] = \text{tr}(\mathbf{S}\mathbf{S}')\sigma^2,$$

the total variance reduction from $\sum_{i=1}^n \text{var}[Y_i]$ is $\text{tr}(\mathbf{S}\mathbf{S}')/n$.

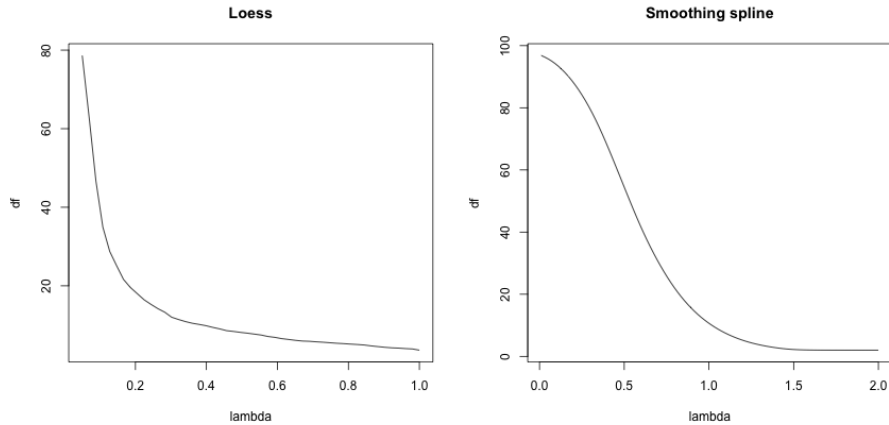
In linear regression the variance reduction is related to the degrees of freedom, or number of parameters. For linear regression, $\sum_{i=1}^n \text{var}[\hat{f}(x_i)] = p\sigma^2$. One widely used definition of degrees of freedom for smoothers is $df = \text{tr}(\mathbf{S}\mathbf{S}')$.

The sensitivity of the fitted value, say $\hat{f}(x_i)$, to the data point y_i can be measured by $W_i(x_i)/\sum_{i=1}^n W_n(x_i)$ or \mathbf{S}_{ii} (remember the denominator is usually 1).

The total influence or sensitivity is $\sum_{i=1}^n W_i(x_i) = \text{tr}(\mathbf{S})$.

In linear regression $\text{tr}(\mathbf{S}) = p$ is also equivalent to the degrees of freedom. This is also used as a definition of degrees of freedom.

Figure 8: Degrees of freedom for loess and smoothing splines as functions of the smoothing parameter. We define smoothing splines in a later lecture.



Finally we notice that

$$E[(\mathbf{y} - \hat{\mathbf{f}})'(\mathbf{y} - \hat{\mathbf{f}})] = \{n - 2\text{tr}(\mathbf{S}) + \text{tr}(\mathbf{S}\mathbf{S}')\}\sigma^2$$

In the linear regression case this is $(n - p)\sigma^2$. We therefore denote $n - 2\text{tr}(\mathbf{S}) + \text{tr}(\mathbf{S}\mathbf{S}')$ as the residual degrees of freedom. A third definition of degrees of freedom of a smoother is then $2\text{tr}(\mathbf{S}) - \text{tr}(\mathbf{S}\mathbf{S}')$.

Under relatively mild assumptions we can show that

$$1 \leq \text{tr}(\mathbf{S}\mathbf{S}') \leq \text{tr}(\mathbf{S}) \leq 2\text{tr}(\mathbf{S}) - \text{tr}(\mathbf{S}\mathbf{S}') \leq n$$

Splines and Friends: Basis Expansion and Regularization

Through-out this section, the regression function f will depend on a single, real-valued predictor X ranging over some possibly infinite interval of the real line, $I \subset \mathbb{R}$. Therefore, the (mean) dependence of Y on X is given by

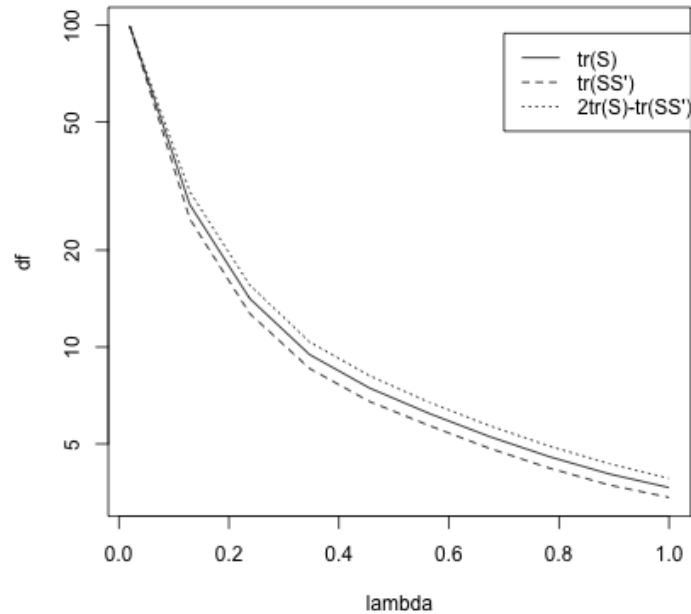
$$f(x) = E(Y|X = x), x \in I \subset \mathbb{R}. \quad (3)$$

For spline models, estimate definitions and their properties are more easily characterized in the context of linear spaces.

Linear Spaces

In this chapter our approach to estimating f involves the use of finite dimensional linear spaces.

Figure 9: Comparison of three definition of degrees of freedom



Remember what a linear space is? Remember definitions of dimension, linear subspace, orthogonal projection, etc...

Why use linear spaces?

- Makes estimation and statistical computations easy.
- Has nice geometrical interpretation.
- It actually can specify a broad range of models given we have discrete data.

Using linear spaces we can define many families of function f ; straight lines, polynomials, splines, and many other spaces (these are examples for the case where \mathbf{x} is a scalar). The point is: we have many options.

Notice that in most practical situation we will have observations $(\mathbf{X}_i, Y_i), i = 1, \dots, n$. In some situations we are only interested in estimating $f(\mathbf{X}_i), i = 1, \dots, n$. In fact, in many situations it is all that matters from a statistical point of view. We will write \mathbf{f} when referring to the this vector and $\hat{\mathbf{f}}$ when referring to an estimate. Think of how its different to know f and know \mathbf{f} .

Let's say we are interested in estimating \mathbf{f} . A common practice in statistics is to assume that \mathbf{f} lies in some *linear space*, or is well approximated by a \mathbf{g} that lies in some *linear space*.

For example for simple linear regression we assume that \mathbf{f} lies in the linear space of lines:

$$\alpha + \beta\mathbf{x}, (\alpha, \beta)' \in \mathbb{R}^2.$$

For linear regression in general we assume that \mathbf{f} lies in the linear space of linear combinations of the covariates or rows of the design matrix. How do we write it out?

Note: Through out this chapter f is used to denote the true regression function and g is used to denote an arbitrary function in a particular space of functions. It isn't necessarily true that f lies in this space of function. Similarly we use \mathbf{f} to denote the true function evaluated at the design points or observed covariates and \mathbf{g} to denote an arbitrary function evaluated at the design points or observed covariates.

Now we will see how and why it's useful to use linear models in a more general setting.

Technical note: A linear model of order p for the regression function (3) consists of a p -dimensional linear space \mathcal{G} , having as a basis the function

$$B_j(\mathbf{x}), j = 1, \dots, p$$

defined for $\mathbf{x} \in I$. Each member $g \in \mathcal{G}$ can be written uniquely as a linear combination

$$g(\mathbf{x}) = g(\mathbf{x}; \boldsymbol{\theta}) = \theta_1 B_1(\mathbf{x}) + \dots + \theta_p B_p(\mathbf{x})$$

for some value of the coefficient vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)' \in \mathbb{R}^p$.

Notice that $\boldsymbol{\theta}$ specifies the point $g \in \mathcal{G}$.

How would you write this out for linear regression?

Given observations $(\mathbf{X}_i, Y_i), i = 1, \dots, n$ the least squares estimate (LSE) of \mathbf{f} or equivalently $f(\mathbf{x})$ is defined by $\hat{f}(\mathbf{x}) = g(\mathbf{x}; \hat{\boldsymbol{\theta}})$, where

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \sum_{i=1}^n \{Y_i - g(\mathbf{X}_i, \boldsymbol{\theta})\}^2.$$

Define the vector $\mathbf{g} = \{g(x_1), \dots, g(x_n)\}'$. Then the distribution of the observations of $Y|X = x$ are in the family

$$\{N(\mathbf{g}, \sigma^2 \mathbf{I}_n); \mathbf{g} = [g(x_1), \dots, g(x_n)]', g \in \mathcal{G}\} \quad (4)$$

and if we assume the errors ε are IID normal and that $f \in \mathcal{G}$ we have that $\hat{\mathbf{f}} = [g(x_1; \hat{\boldsymbol{\theta}}), \dots, g(x_n; \hat{\boldsymbol{\theta}})]$ is the

maximum likelihood estimate. The estimand \mathbf{f} is an $n \times 1$ vector. But how many parameters are we really estimating?

Equivalently we can think of the distribution is in the family

$$\{N(\mathbf{B}\boldsymbol{\theta}, \sigma^2); \boldsymbol{\theta} \in \mathbb{R}^p\} \quad (5)$$

and the maximum likelihood estimate for $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}}$. Here \mathbf{B} is a matrix of basis elements defined soon...

Here we start seeing for the first time where the name *non-parametric* comes from. How are the approaches (4) and (5) different?

Notice that obtaining $\hat{\boldsymbol{\theta}}$ is easy because of the linear model set-up. The ordinary least square estimate is

$$(\mathbf{B}'\mathbf{B})\hat{\boldsymbol{\theta}} = \mathbf{B}'\mathbf{Y}$$

where \mathbf{B} is the $n \times p$ design matrix with elements $[\mathbf{B}]_{ij} = B_j(\mathbf{X}_i)$. When this solution is unique we refer to $g(x; \hat{\boldsymbol{\theta}})$ as the OLS projection of \mathbf{Y} into \mathcal{G} (as learned in the first term).

Parametric versus non-parametric

In some cases, we have reason to believe that the function f is actually a member of some linear space \mathcal{G} . Traditionally, inference for regression models depends on f being representable as some combination of known predictors. Under this assumption, f can be written as a combination of basis elements for some value of the coefficient vector $\boldsymbol{\theta}$. This provides a *parametric* specification for f . No matter how many observations we collect, there is no need to look outside the fixed, finite-dimensional, linear space \mathcal{G} when estimating f .

In practical situations, however, we would rarely believe such relationship to be exactly true. Model spaces \mathcal{G} are understood to provide (at best) approximations to f ; and as we collect more and more samples, we have the freedom to audition richer and richer classes of models. In such cases, all we might be willing to say about f is that it is *smooth* in some sense, a common assumption being that f have two bounded derivatives. Far from the assumption that f belong to a fixed, finite-dimensional linear space, we instead posit a *nonparametric* specification for f . In this context, model spaces are employed mainly in our approach to inference; first in the questions we pose about an estimate, and then in the tools we apply to address them. For example, we are less interested in the actual values of the coefficient $\boldsymbol{\theta}$, e.g. whether or not an element of $\boldsymbol{\theta}$ is significantly different from zero to the 0.05 level. Instead we concern ourselves with functional properties of $g(\mathbf{x}; \hat{\boldsymbol{\theta}})$, the estimated curve or surface, e.g. whether or not a peak is real.

To ascertain the local behavior of OLS projections onto approximation spaces \mathcal{G} , define the pointwise, mean squared error (MSE) of $\hat{g}(\mathbf{x}) = g(\mathbf{x}; \hat{\boldsymbol{\theta}})$ as

$$E\{f(\mathbf{x}) - \hat{g}(\mathbf{x})\}^2 = \text{bias}^2\{\hat{g}(\mathbf{x})\} + \text{var}\{\hat{g}(\mathbf{x})\}$$

where

$$\text{bias}\{\hat{g}(\mathbf{x})\} = f(x) - E\{\hat{g}(\mathbf{x})\} \tag{6}$$

and

$$\text{var}\{\hat{g}(\mathbf{x})\} = E\{\hat{g}(\mathbf{x}) - E[\hat{g}(\mathbf{x})]\}^2$$

When the input values $\{\mathbf{X}_i\}$ are deterministic the expectations above are with respect to the noisy observation Y_i . In practice, MSE is defined in this way even in the random design case, so we look at expectations conditioned on \mathbf{X} .

Note: The MSE and EPE are equivalent. The only difference is that we ignore the first σ^2 due to measurement error contained in the EPE. The reason I use MSE here is because it is what is used in the Spline and Wavelet literature.

When we do this, standard results in regression theory can be applied to derive an expression for the variance term

$$\text{var}\{\hat{g}(\mathbf{x})\} = \sigma^2 \mathbf{B}(\mathbf{x})' (\mathbf{B}' \mathbf{B})^{-1} \mathbf{B}(\mathbf{x})$$

where $\mathbf{B}(\mathbf{x}) = (B_1(\mathbf{x}), \dots, B_p(\mathbf{x}))'$, and the error variance is assumed constant.

Under the parametric specification that $f \in \mathcal{G}$, what is the bias?

This leads to classical t- and F-hypothesis tests and associated parametric confidence intervals for $\boldsymbol{\theta}$. Suppose on the other hand, that f is not a member of \mathcal{G} , but rather can be reasonably approximated by an element in \mathcal{G} . The bias (6) now reflects the ability of functions in \mathcal{G} to capture the essential features of f .

Local Polynomials

In practical situations, a statistician is rarely blessed with simple linear relationship between the predictor X and the observed output Y . That is, as a description of the regression function f , the model

$$g(x; \boldsymbol{\theta}) = \theta_1 + \theta_2 x, x \in I$$

typically ignores obvious features in the data. This is certainly the case for the values of ⁸⁷Sr.

The Strontium data set was collected to test several hypotheses about the catastrophic events that occurred approximately 65 million years ago. The data contains Age in million of years and the ratios described here. There is a division between two geological time periods, the Cretaceous (from 66.4 to 144 million years ago) and the Tertiary (spanning from about 1.6 to 66.4 million years ago). Earth scientist believe that the boundary between these periods is distinguished by tremendous changes in climate that accompanied a mass extinction of over half of the species inhabiting the planet at the time. Recently, the compositions of Strontium (Sr) isotopes in sea water has been used to evaluate several

hypotheses about the cause of these extreme events. The dependent variable of the data-set is related to the isotopic make up of Sr measured for the shells of marine organisms. The Cretaceous-Tertiary boundary is referred to as KTB. There data shows a peak is at this time and this is used as evidence that a meteor collided with earth.

The data presented in the Figure ?? represents standardized ratio of strontium-87 isotopes (^{87}Sr) to strontium-86 isotopes (^{86}Sr) contained in the shells of foraminifera fossils taken from cores collected by deep sea drilling. For each sample its time in history is computed and the standardized ratio is computed:

$$^{87}\delta\text{Sr} = \left(\frac{^{87}\text{Sr}/^{86}\text{Sr sample}}{^{87}\text{Sr}/^{86}\text{Sr sea water}} - 1 \right) \times 10^5.$$

Earth scientist expect that $^{87}\delta\text{Sr}$ is a smooth-varying function of time and that deviations from smoothness are mostly measurement error.

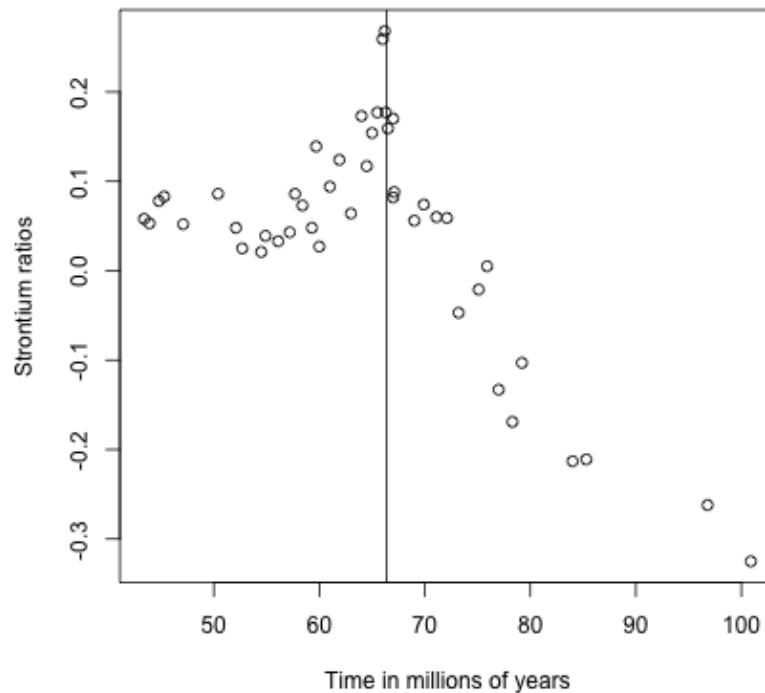


Figure 10: $^{87}\delta\text{Sr}$ data.

To overcome this deficiency, we might consider a more flexible polynomial model.

Let \mathcal{P}_k denote the linear space of polynomials in x of order at most k defined as

$$g(x; \boldsymbol{\theta}) = \theta_1 + \theta_2 x + \cdots + \theta_k x^{k-1}, x \in I$$

for the parameter vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k) \in \mathbb{R}^k$. Note that the space \mathcal{P}_k consists of polynomials having degree at most $k - 1$.

In exceptional cases, we have reasons to believe that the regression function f is in fact a high-order polynomial. This parametric assumption could be based on physical or physiological models describing how the data were generated.

For historical values of δSr we consider polynomials simply because our scientific intuition tells us that f should be smooth.

Recall Taylor's theorem: polynomials are good at approximating well-behaved functions in reasonably tight neighborhoods. If all we can say about f is that it is smooth in some sense, then either implicitly or explicitly we consider high-order polynomials because of their favorable approximation properties.

If f is not in \mathcal{P}_k then our estimates will be biased by an amount that reflects the approximation error incurred by a polynomial model.

Computational Issue: The basis of monomials

$$B_j(x) = x^{j-1} \text{ for } j = 1, \dots, k$$

is not well suited for numerical calculations (x^8 can be VERY BIG compared to x). While convenient for analytical manipulations (differentiation, integration), this basis is *ill-conditioned* for k larger than 8 or 9. Most statistical packages use the orthogonal Chebyshev polynomials (used by the R command `poly()`).

An alternative to polynomials is to consider the space $\mathcal{PP}_k(\mathbf{t})$ of piecewise polynomials with break points $\mathbf{t} = (t_0, \dots, t_{m+1})'$. Given a sequence $a = t_0 < t_1 < \cdots < t_m < t_{m+1} = b$, construct $m + 1$ (disjoint) intervals

$$I_l = [t_{l-1}, t_l], 1 \leq l \leq m \text{ and } I_{m+1} = [t_m, t_{m+1}],$$

whose union is $I = [a, b]$. Define the piecewise polynomials of order k

$$g(x) = \begin{cases} g_1(x) = \theta_{1,1} + \theta_{1,2}x + \cdots + \theta_{1,k}x^{k-1}, & x \in I_1 \\ \vdots & \vdots \\ g_{m+1}(x) = \theta_{m+1,1} + \theta_{m+1,2}x + \cdots + \theta_{m+1,k}x^{k-1}, & x \in I_{m+1}. \end{cases}$$

Splines

In many situations, breakpoints in the regression function do not make sense. Would forcing the piecewise polynomials to be continuous suffice? What about continuous first derivatives?

We start by consider the subspaces of the piecewise polynomial space. We will denote it with $\mathcal{PP}_k(\mathbf{t})$ with $\mathbf{t} = (t_1, \dots, t_m)'$ the break-points or interior knots. Different break points define different spaces.

We can put constrains on the behavior of the functions g at the break points. (We can construct tests to see if these constrains are suggested by the data but, will not go into this here)

Here is a trick for forcing the constrains and keeping the linear model set-up. We can write any function $g \in \mathcal{PP}_k(\mathbf{t})$ in *the truncated basis power*:

$$\begin{aligned} g(x) = & \theta_{0,1} + \theta_{0,2}x + \dots + \theta_{0,k}x^{k-1} + \\ & \theta_{1,1}(x - t_1)_+^0 + \theta_{1,2}(x - t_1)_+^1 + \dots + \theta_{1,k}(x - t_1)_+^{k-1} + \\ & \vdots \\ & \theta_{m,1}(x - t_m)_+^0 + \theta_{m,2}(x - t_m)_+^1 + \dots + \theta_{m,k}(x - t_m)_+^{k-1} \end{aligned}$$

where $(\cdot)_+ = \max(\cdot, 0)$. Written in this way the coefficients $\theta_{1,1}, \dots, \theta_{1,k}$ record the jumps in the different derivative from the first piece to the second.

Notice that the constrains reduce the number of parameters. This is in agreement with the fact that we are forcing more smoothness.

Now we can force constrains, such as continuity, by putting constrains like $\theta_{1,1} = 0$ etc...

We will concentrate on the cubic splines which are continuous and have continuous first and second derivatives. In this case we can write:

$$g(x) = \theta_{0,1} + \theta_{0,2}x + \dots + \theta_{0,4}x^3 + \theta_{1,k}(x - t_1)^3 + \dots + \theta_{m,k}(x - t_m)^3$$

How many “parameters” in this space?

Note: It is always possible to have less restrictions at knots where we believe the behavior is “less smooth”, e.g for the Sr ratios, we may have “unsmoothness” around KTB.

We can write this as a linear space. This setting is not computationally convenient. In S-Plus there is a function `bs()` that makes a basis that is convenient for computations.

There is asymptotic theory that goes along with all this but we will not go into the details. We will just notice that

$$E[f(x) - g(x)] = O(h_l^{2k} + 1/n_l)$$

where h_l is the size of the interval where x is in and n_l is the number of points in it. What does this say?

Splines in terms of Spaces and sub-spaces

The p -dimensional spaces described in Section 4.1 were defined through basis function $B_j(\mathbf{x}), j = 1, \dots, p$. So in general we defined for a given range $I \subset \mathbb{R}^k$

$$\mathcal{G} = \{g : g(\mathbf{x}) = \sum_{j=1}^p \theta_j \beta_j(\mathbf{x}), \mathbf{x} \in I, (\theta_1, \dots, \theta_p) \in \mathbb{R}^p\}$$

In the previous section we concentrated on $\mathbf{x} \in \mathbb{R}$.

In practice we have design points x_1, \dots, x_n and a vector of responses $\mathbf{y} = (y_1, \dots, y_n)$. We can think of \mathbf{y} as an element in the n -dimensional vector space \mathbb{R}^n . In fact we can go a step further and define a Hilbert space with the usual inner product definition that gives us the norm

$$\|\mathbf{y}\| = \sum_{i=1}^n y_i^2$$

Now we can think of least squares estimation as the projection of the data \mathbf{y} to the sub-space $\mathbf{G} \subset \mathbb{R}^n$ defined by \mathcal{G} in the following way

$$\mathbf{G} = \{\mathbf{g} \in \mathbb{R}^n : \mathbf{g} = [g(x_1), \dots, g(x_n)]', g \in \mathcal{G}\}$$

Because this space is spanned by the vectors $[B_1(x_1), \dots, B_p(x_n)]$ the projection of \mathbf{y} onto \mathbf{G} is

$$\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{y}$$

as learned in 751. Here $[\mathbf{B}]_{ij} = B_j(x_i)$.

Natural Smoothing Splines

Natural splines add the constrain that the function must be linear after the knots at the end points. This forces 2 more restrictions since f'' must be 0 at the end points, i.e the space has $k + 4 - 2$ parameters because of this extra 2 constrains.

So where do we put the knots? How many do we use? There are some data-driven procedures for doing this. Natural Smoothing Splines provide another approach.

What happens if the knots coincide with the dependent variables $\{X_i\}$. Then there is a function $g \in \mathcal{G}$, the space of cubic splines with knots at (x_1, \dots, x_n) , with $g(x_i) = y_i, i, \dots, n$, i.e. we haven't smoothed at all.

Consider the following problem: among all functions g with two continuous first two derivatives, find one that minimizes the penalized residual sum of squares

$$\sum_{i=1}^n \{y_i - g(x_i)\}^2 + \lambda \int_a^b \{g''(t)\}^2 dt$$

where λ is a fixed constant, and $a \leq x_1 \leq \dots \leq x_n \leq b$. It can be shown (Reinsch 1967) that the solution to this problem is a natural cubic spline with knots at the values of x_i (so there are $n - 2$ interior knots and $n - 1$ intervals). Here a and b are arbitrary as long as they contain the data.

It seems that this procedure is over-parameterized since a natural cubic spline as this one will have n degrees of freedom. However we will see that the penalty makes this go down.

Computational Aspects

We use the fact that the solution is a natural cubic spline and write the possible answers as

$$g(x) = \sum_{j=1}^n \theta_j B_j(x)$$

where θ_j are the coefficients and $B_j(x)$ are the basis functions. Notice that if these were cubic splines the functions lie in a $n + 2$ dimensional space, but the natural splines are an n dimensional subspace.

Let \mathbf{B} be the $n \times n$ matrix defined by

$$B_{ij} = B_j(x_i)$$

and a penalty matrix $\mathbf{\Omega}$ by

$$\Omega_{ij} = \int_a^b B_i''(t) B_j''(t) dt$$

now we can write the penalized criterion as

$$(\mathbf{y} - \mathbf{B}\boldsymbol{\theta})'(\mathbf{y} - \mathbf{B}\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}'\mathbf{\Omega}\boldsymbol{\theta}$$

It seems there are no boundary derivatives constraints but they are implicitly imposed by the penalty term.

Setting derivatives with respect to $\boldsymbol{\theta}$ equal to 0 gives the estimating equation:

$$(\mathbf{B}'\mathbf{B} + \lambda\mathbf{\Omega})\boldsymbol{\theta} = \mathbf{B}'\mathbf{y}.$$

The $\hat{\boldsymbol{\theta}}$ that solves this equation will give us the estimate $\hat{\mathbf{g}} = \mathbf{B}\hat{\boldsymbol{\theta}}$.

Is this a linear smoother?

Write:

$$\hat{\mathbf{g}} = \mathbf{B}\boldsymbol{\theta} = \mathbf{B}(\mathbf{B}'\mathbf{B} + \lambda\mathbf{\Omega})^{-1}\mathbf{B}'\mathbf{y} = (\mathbf{I} + \lambda\mathbf{K})^{-1}\mathbf{y}$$

where $\mathbf{K} = \mathbf{B}^{-1}\mathbf{\Omega}\mathbf{B}$. Notice we can write the criterion as

$$(\mathbf{y} - \mathbf{g})'(\mathbf{y} - \mathbf{g}) + \lambda\mathbf{g}'\mathbf{K}\mathbf{g}$$

If we look at the “kernel” of this linear smoother we will see that it is similar to the other smoothers presented in this class.

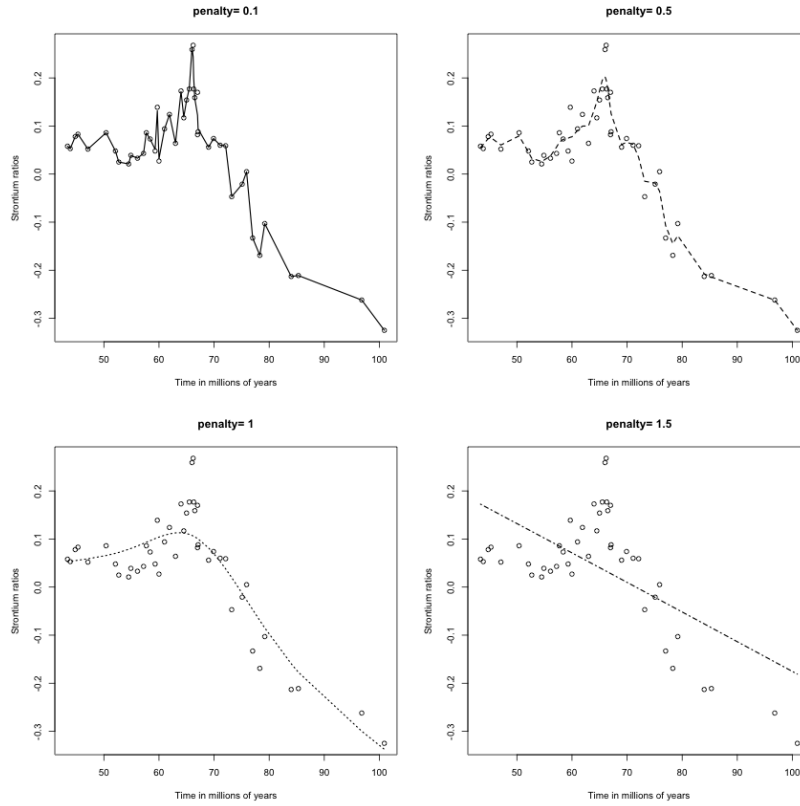


Figure 11: Smoothing spline fitted using different penalties.

Smoothing and Penalized Least Squares

In Section 4.4.1 we saw that the smoothing spline solution to a penalized least squares is a linear smoother.

Using the notation of Section 4.4.1, we can write the penalized criterion as

$$(\mathbf{y} - \mathbf{B}\boldsymbol{\theta})'(\mathbf{y} - \mathbf{B}\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}'\boldsymbol{\Omega}\boldsymbol{\theta}$$

Setting derivatives with respect to $\boldsymbol{\theta}$ equal to 0 gives the estimating equation:

$$(\mathbf{B}'\mathbf{B} + \lambda\boldsymbol{\Omega})\boldsymbol{\theta} = \mathbf{B}'\mathbf{y}$$

the $\hat{\boldsymbol{\theta}}$ that solves this equation will give us the estimate $\hat{\mathbf{g}} = \mathbf{B}\hat{\boldsymbol{\theta}}$.

Write:

$$\hat{\mathbf{g}} = \mathbf{B}\boldsymbol{\theta} = \mathbf{B}(\mathbf{B}'\mathbf{B} + \lambda\boldsymbol{\Omega})^{-1}\mathbf{B}'\mathbf{y} = (\mathbf{I} + \lambda\mathbf{K})^{-1}\mathbf{y}$$

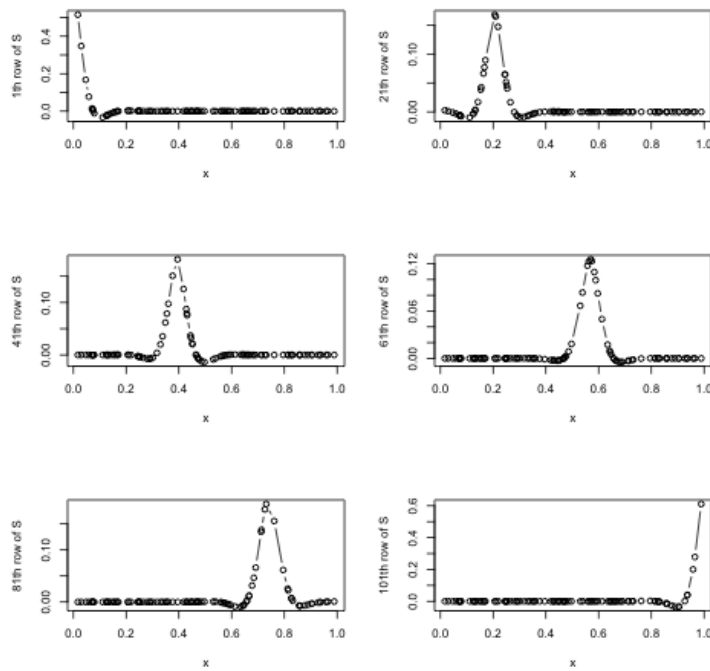
where $\mathbf{K} = \mathbf{B}'^{-1}\mathbf{\Omega}\mathbf{B}^{-1}$.

Notice we can write the penalized criterion as

$$(\mathbf{y} - \mathbf{g})'(\mathbf{y} - \mathbf{g}) + \lambda \mathbf{g}'\mathbf{K}\mathbf{g}$$

If we plot the rows of this linear smoother we will see that it is like a kernel smoother.

Figure 12: Kernels of a smoothing spline.



Notice that for any linear smoother with a symmetric and nonnegative definite \mathbf{S} , i.e. there \mathbf{S}^{-} exists, then we can argue in reverse: $\hat{\mathbf{f}} = \mathbf{S}\mathbf{y}$ is the value that minimizes the penalized least squares criteria of the form

$$(\mathbf{y} - \mathbf{f})'(\mathbf{y} - \mathbf{f}) + \mathbf{f}'(\mathbf{S}^{-} - \mathbf{I})\mathbf{f}.$$

Some of the smoothers presented in this class are not symmetrical but are close. In fact for many of them one can show that asymptotically they are symmetric.

Eigen analysis and spectral smoothing

For a smoother with symmetric smoother matrix \mathbf{S} , the eigendecomposition of \mathbf{S} can be used to describe its behavior.

Let $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ be an orthonormal basis of eigenvectors of \mathbf{S} with eigenvalues $\theta_1 \geq \theta_2 \geq \dots \geq \theta_n$:

$$\mathbf{S}\mathbf{u}_j = \theta_j \mathbf{u}_j, j = 1, \dots, n$$

or

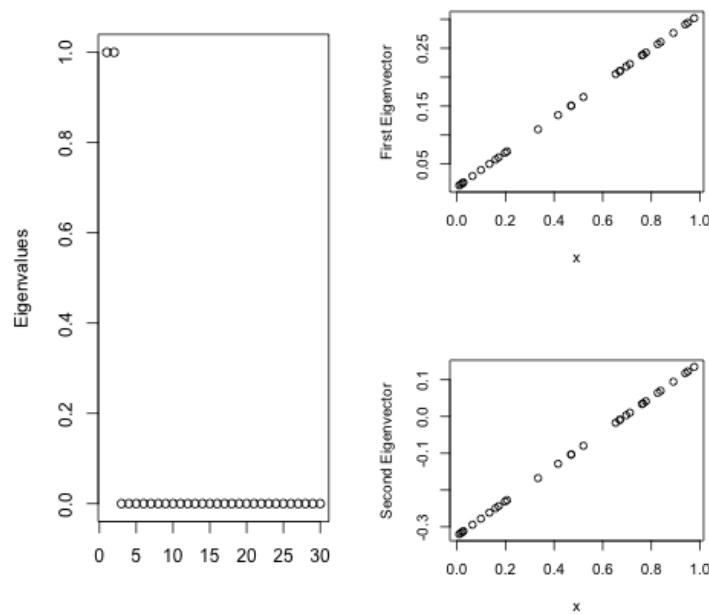
$$\mathbf{S} = \mathbf{U}\mathbf{D}\mathbf{U}' = \sum_{j=1}^n \theta_j \mathbf{u}_j \mathbf{u}_j'.$$

Here \mathbf{D} is a diagonal matrix with the eigenvalues as the entries.

For simple linear regression we only have two nonzero eigenvalues. Their eigenvectors are an orthonormal basis for lines.

Figure 13: Eigenvalues and eigenvectors of the hat matrix for linear regression.

Simple linear regression

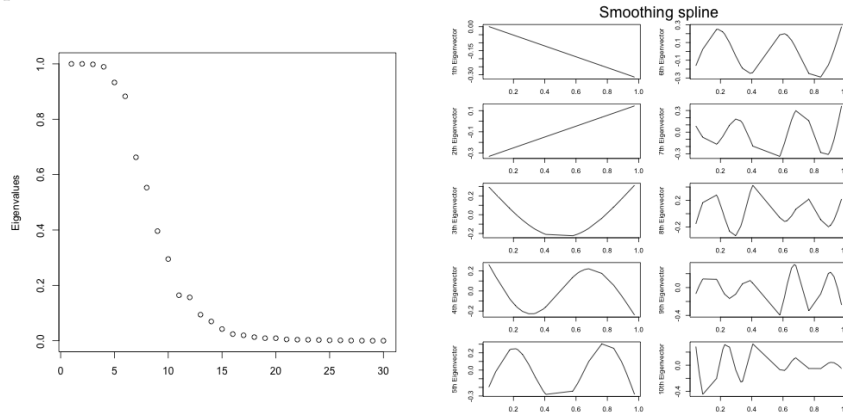


The cubic spline is an important example of a symmetric smoother, and its eigenvectors resemble polynomials of increasing degree.

It is easy to show that the first two eigenvalues are unity, with eigenvectors which correspond to linear functions of the predictor on which the smoother is based. One can also show that the other eigenvalues are all strictly between zero and one.

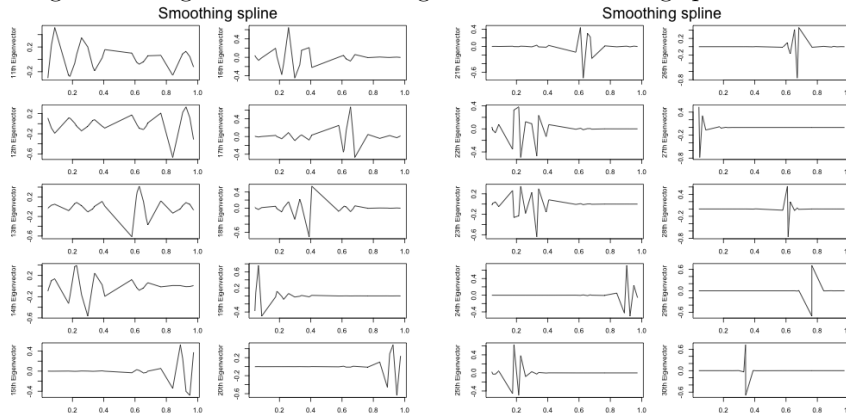
The action of the smoother is now transparent: if presented with a response $\mathbf{y} = \mathbf{u}_j$, it shrinks it by an amount θ_j as above.

Figure 14: Eigenvalues and eigenvectors 1 through 10 of \mathbf{S} for a smoothing spline.



Cubic smoothing splines, regression splines, linear regression, polynomial regression are all symmetric smoothers. However, loess and other “nearest neighbor” smoothers are not.

Figure 15: Eigen vectors 11 through 30 for a smoothing spline for $n = 30$.



If \mathbf{S} is not symmetric we have complex eigenvalues and the above decomposition is not as easy to interpret. However we can use the singular value decomposition

$$\mathbf{S} = \mathbf{UDV}'$$

One can think of smoothing as performing a basis transformation $\mathbf{z} = \mathbf{V}'\mathbf{y}$, shrinking with $\hat{\mathbf{z}} = \mathbf{D}\mathbf{z}$ the components that are related to “unsmooth components” and then transforming back to the basis $\hat{\mathbf{y}} = \mathbf{U}\hat{\mathbf{z}}$ we started out with... sort of.

In signal processing signals are “filtered” using linear transformations. The transfer function describes how the power of certain frequency components are reduced. A low-pass filter will reduce the power of the higher frequency components. We can view the eigen values of our smoother matrices as transfer functions.

Notice that the smoothing spline can be considered a low-pass filter. If we look at the eigenvectors of the smoothing spline we notice they are similar to sinusoidal components of increasing frequency. Figure 14 shows the “transfer function” defined by the smoothing splines.

Economical Bases: Wavelets and REACT estimators

If one consider the “equally spaced” Gaussian regression:

$$y_i = f(t_i) + \varepsilon_i, i = 1, \dots, n \quad (7)$$

$t_i = (i - 1)/n$ and the ε_i s IID $N(0, \sigma^2)$, many things simplify.

We can write this in matrix notation: the response vector \mathbf{y} is $N_n(\mathbf{f}, \sigma^2 \mathbf{I})$ with $\mathbf{f} = \{f(t_1), \dots, f(t_n)\}'$.

As usual we want to find an estimation procedure that minimizes risk:

$$n^{-1} \mathbf{E} \|\hat{\mathbf{f}} - \mathbf{f}\|^2 = n^{-1} \mathbf{E} \left[\sum_{i=1}^n \{\hat{f}(t_i) - f(t_i)\}^2 \right].$$

We have seen that the MLE is $\hat{f}_i = y_i$ which intuitively does not seem very useful. There is actually an important result in statistics that makes this more precise.

Stein (1956) noticed that the MLE is inadmissible: There is an estimation procedure producing estimates with smaller risk than the MLE for any \mathbf{f} .

To develop a non-trivial theory MLE won't do. A popular procedure is to specify some fixed class \mathcal{F} of functions where f lies and seek an estimator \hat{f} attaining minimax risk

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}} R(\hat{f}, f)$$

By restricting $f \in \mathcal{F}$ we make assumptions on the smoothness of f . For example, the L^2 Sobolev family makes an assumption on the number m of continuous derivatives and a limits the size of the m th derivative.

Useful transformations

Remember $\mathbf{f} \in \mathbb{R}^n$ and that there are many orthogonal bases for this space. Any orthogonal basis can be represented with an orthogonal transform \mathbf{U} that gives us the coefficients for any \mathbf{f} by multiplying $\boldsymbol{\xi} = \mathbf{U}'\mathbf{f}$. This means that we can represent any vector as $\mathbf{f} = \mathbf{U}\boldsymbol{\xi}$.

Remember that the eigen analysis of smoothing splines we can view the eigenvectors as such a transformation.

If we are smart, we can choose a transformation \mathbf{U} such that $\boldsymbol{\xi}$ has some useful interpretation. Furthermore, certain transformation may be more “economical” as we will see.

For **equally spaced data** a widely used transformation is the Discrete Fourier Transform (DFT). Fourier’s theorem says that any $\mathbf{f} \in \mathbb{R}^n$ can be re-written as

$$f_i = a_0 + \sum_{k=1}^{n/2-1} \left\{ a_k \cos\left(\frac{2\pi k}{n} i\right) + b_k \sin\left(\frac{2\pi k}{n} i\right) \right\} + a_{n/2} \cos(\pi i)$$

for $i = 1, \dots, n$. This defines a basis and the coefficients $\mathbf{a} = (a_0, a_1, b_1, \dots, \dots, a_{n/2})'$ can be obtained via $\mathbf{a} = \mathbf{U}'\mathbf{f}$ with \mathbf{U} having columns of sines and cosines:

$$\begin{aligned} U_1 &= [n^{-1/2} : 1 \leq i \leq n] \\ U_{2k} &= [(2/n)^{1/2} \sin\{2\pi ki/n\} : 1 \leq i \leq n], k = 1, \dots, n/2 \\ U_{2k+1} &= [(2/n)^{1/2} \cos\{2\pi ki/n\} : 1 \leq i \leq n], k = 1, \dots, n/2 - 1. \end{aligned}$$

Note: This can easily be changed to the case where n is odd by substituting $n/2$ by $\lfloor n/2 \rfloor$ and taking out the last term last term $a_{\lfloor n/2 \rfloor}$.

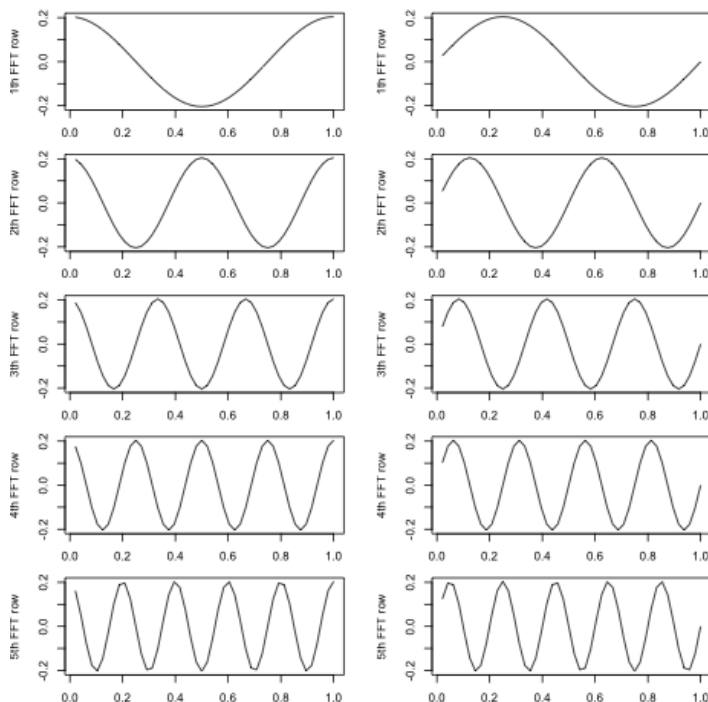
If a signal is close to a sine wave $f(t) = \cos(2\pi jt/n + \phi)$ for some integer $1 \leq j \leq n$, only two of the coefficients in \mathbf{a} will be big, namely the ones associated with the columns $2j - 1$ and $2j$, the rest will be close to 0.

This makes the basis associated with the DFT very economical (and the *periodogram* a good detector of hidden periodicities). Consider that if we were to transmit the signal, say using modems and a telephone line, it would be more “economical” to send \mathbf{a} instead of the \mathbf{f} . Once \mathbf{a} is received, $\mathbf{f} = \mathbf{U}\mathbf{a}$ is reconstructed. This is basically what data compression is all about.

Because we are dealing with equally spaced data, the coefficients of the DFT are also related to smoothness. Notice that the columns of U are increasing in frequency and thus decreasing in smoothness. This means that a “smooth” \mathbf{f} should have only the first $\mathbf{a} = \mathbf{U}'\mathbf{f}$ relatively different from 0.

A close relative of the DFT is the Discrete Cosine Transform (DCT).

$$\begin{aligned} U_1 &= [n^{-1/2} : 1 \leq i \leq n] \\ U_k &= [(2/n)^{1/2} \cos\{\pi(2i - 1)k/(2n)\} : 1 \leq i \leq n], k = 2, \dots, n \end{aligned}$$



Economical bases together with “shrinkage” ideas can be used to reduce risk and even to obtain estimates with minimax properties. We will see this through an example

An example

We consider body temperature data taken from a mouse every 30 minutes for a day, so we have $n = 48$. We believe measurements will have measurement error and maybe environmental variability so we use a stochastic model like (7). We expect body temperature to change “smoothly” through-out the day so we believe $f(x)$ is smooth. Under this assumption $\boldsymbol{\xi} = \mathbf{U}'\mathbf{f}$, with \mathbf{U} the DCT, should have only a few coefficients that are “big”.

Because the transformation is orthogonal we have that $\mathbf{z} = \mathbf{U}'\mathbf{y}$ is $N(\boldsymbol{\xi}, \sigma^2\mathbf{I})$. An idea we learn from Stein (1956) is to consider linear shrunken estimates $\hat{\boldsymbol{\xi}} = \{\mathbf{w}\mathbf{z}; \mathbf{w} \in [0, 1]^n\}$. Here the product $\mathbf{w}\mathbf{z}$ is taken component-wise like in S-plus.

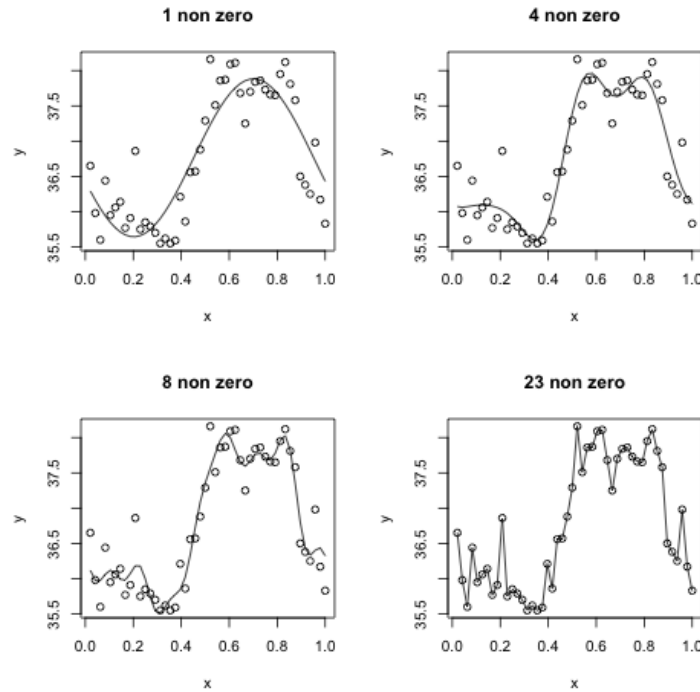
We can then choose the shrinkage coefficients that minimize the risk

$$E\|\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}\|^2 = E\|\mathbf{U}\hat{\boldsymbol{\xi}} - \mathbf{f}\|^2.$$

Remember that $\mathbf{U}\boldsymbol{\xi} = \mathbf{U}\mathbf{U}'\mathbf{f} = \mathbf{f}$.

Relatively simple calculations show that $\tilde{\mathbf{w}} = \boldsymbol{\xi}^2/(\boldsymbol{\xi}^2 + \sigma^2)$ minimizes the risk over all possible $\mathbf{w} \in \mathbb{R}^n$. The MLE obtained, with $\mathbf{w} = (1, \dots, 1)'$, minimizes the risk only if $\tilde{\mathbf{w}} = (1, \dots, 1)'$ which only happens when there is no variance!

Figure 16: Fitted curves obtained when using shrinkage coefficients of the form $\mathbf{w} = (1, 1, \dots, 1, 0, \dots, 0)$, with $2m + 1$ the number of 1s used.

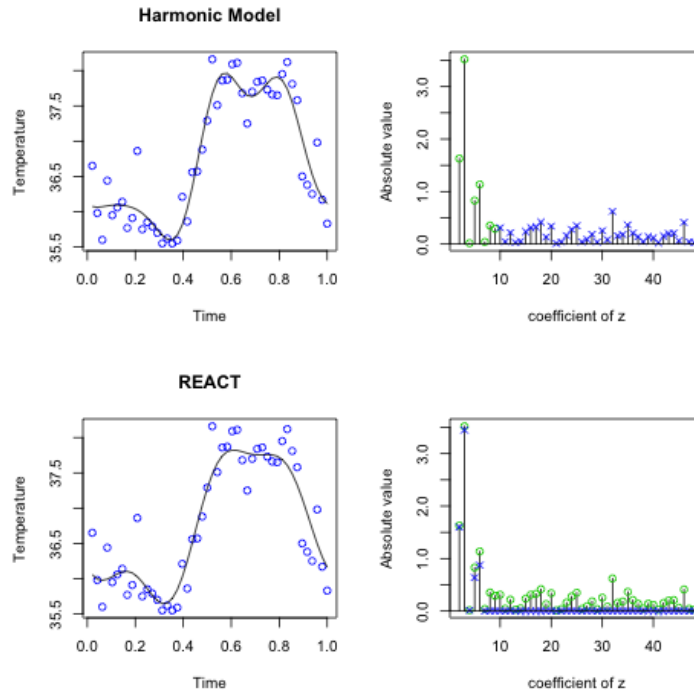


Notice that $\tilde{\mathbf{w}}$ makes sense because it shrinks coefficients with small signal to noise ratio. By shrinking small coefficients closer to 0 we reduce variance and the bias we add is not very large, thus reducing risk. However, we don't know $\boldsymbol{\xi}$ nor σ^2 so in practice we can't produce $\tilde{\mathbf{w}}$. Here is where having economical bases are helpful: we construct estimation procedures that shrink more aggressively the coefficients for which we have a-priori knowledge that they are "close to 0" i.e. have small signal to noise ratio. Two examples of such procedure are:

In Figure 16, we show for the body temperature data the the fitted curves obtained when using shrinkage coefficients of the form $\mathbf{w} = (1, 1, \dots, 1, 0, \dots, 0)$.

In Figure 17 we show the fitted curve obtained with $\mathbf{w} = (1, 1, \dots, 1, 0, \dots, 0)$ and using REACT. In the first plot we show the coefficients shrunk to 0 with crosses. In the second \mathbf{z} plot we show \mathbf{wz} with crosses. Notice that only the

Figure 17: Estimates obtained with harmonic model and with REACT. We also show the \mathbf{z} and how they have been shrunk.



first few coefficients of the transformation are “big”. Here are the same pictures for data obtained for 6 consecutive weekends.

Finally in Figure 18 we show the two fitted curves and compare them to the average obtained from observing many days of data.

Notice that using $\mathbf{w} = (1, 1, 1, 1, 0, \dots, 0)$ reduces to a parametric model that assumes f is a sum of 4 cosine functions.

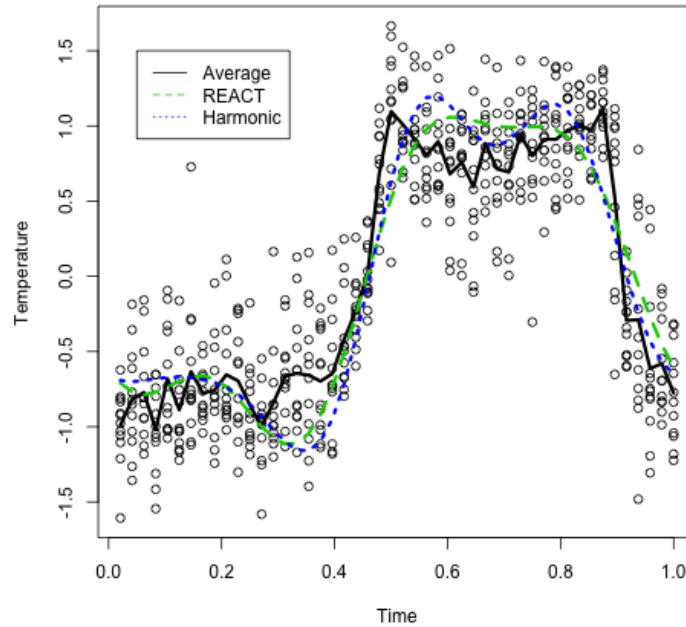
Any smoother with a smoothing matrix \mathbf{S} that is a projection, e.g. linear regression, splines, can be consider a special case of what we have described here.

Choosing the transformation \mathbf{U} is an important step in these procedure. The theory developed for Wavelets motivate a choice of \mathbf{U} that is especially good at handling functions \mathbf{f} that have “discontinuities”.

Wavelets

The following plot show a nuclear magnetic resonance (NMR) signal.

Figure 18: Comparison of two fitted curves to the average obtained from observing many days of data.



The signal does appear to have some added noise so we could use (7) to model the process. However, $f(x)$ appears to have a peak at around $x = 500$ making it not very smooth at that point.

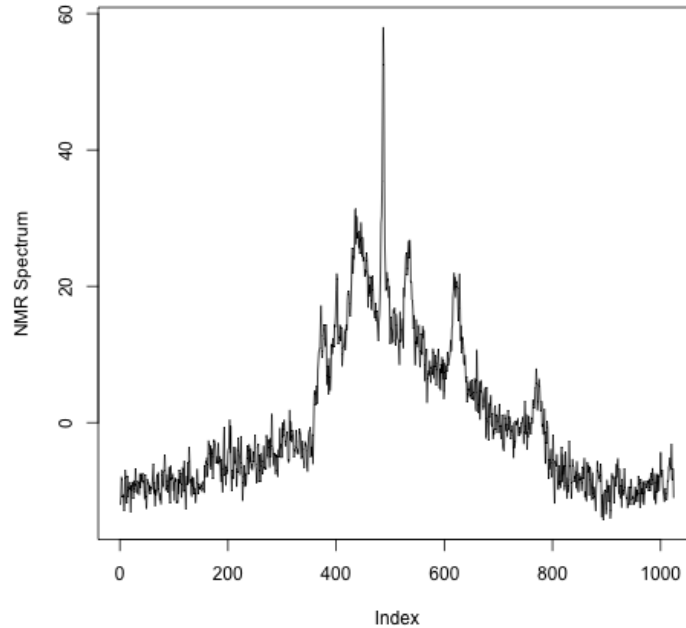
Situations like these are where wavelets analyses is especially useful for “smoothing”. Now a more appropriate word is “de-noising”.

The Discrete Wavelet Transform defines an orthogonal basis just like the DFT and DCT. However the columns of DWT are locally smooth. This means that the coefficients can be interpreted as local smoothness of the signal for different locations.

Here are the columns of the Haar DWT, the simplest wavelet.

Notice that these are step function. However, there are ways (they involve complicated math and no closed forms) to create “smoother” wavelets. The following are the columns of DWT using the Daubechies wavelets

The following plot shows the coefficients of the DWT by smoothness level and by location:

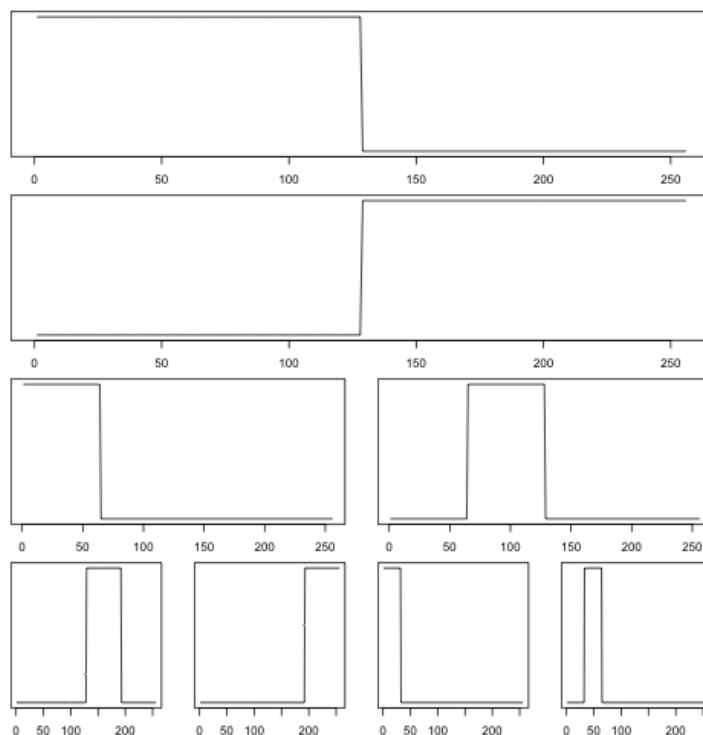


Using wavelet with shrinkage seems to perform better at de-noising than smoothing splines and loess as shown by the following figure.

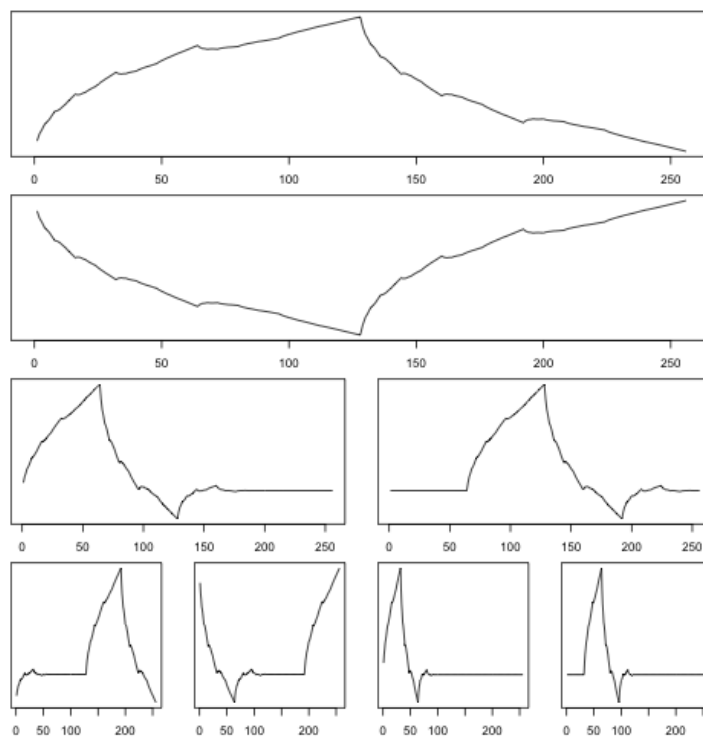
The last plot is what the wavelet estimate looks like for the temperature data

References

- [1] Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. (1990). Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6:3–33.
- [2] Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83:596–610.
- [3] Cleveland, W. S., Grosse, E., and Shyu, W. M. (1993). Local regression models. In Chambers, J. M. and Hastie, T. J., editors, *Statistical Models in S*, chapter 8, pages 309–376. Chapman & Hall, New York.
- [4] Loader, C. R. (1999), *Local Regression and Likelihood*, New York: Springer.



- [5] Socket, E.B., Daneman, D. Clarson, C., and Ehrich, R.M. (1987). Factors affecting and patterns of residual insulin secretion during the first year of type I (insulin dependent) diabetes mellitus in children. *Diabetes* 30, 453–459.
- [6] Eubank, R.L. (1988), *Smoothing Splines and Nonparametric Regression*, New York: Marcel Decker.
- [7] Reinsch, C.H. (1967) Smoothing by Spline Functions. *Numerische Mathematik*, 10: 177–183
- [8] Schoenberg, I.J. (1964), “Spline functions and the problem of graduation,” *Proceedings of the National Academy of Science, USA* 52, 947–950.
- [9] Silverman (1985) “Some Aspects of the spline smoothing approach to non-parametric regression curve fitting”. *Journal of the Royal Statistical Society B* 47: 1–52.
- [10] Wahba, G. (1990), *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series, Philadelphia: SIAM.
- [11] Beran, R. (2000). “REACT scatterplot smoothers: Superefficiency through basis economy”, *Journal of the American Statistical Association*, 95:155–171.



- [12] Brumback, B. and Rice, J. (1998). “Smoothing Spline Models for the Analysis of Nested and Crossed Samples of Curves”. *Journal of the American Statistical Association*. 93: 961–976.
- [13] Donoho, D.L. and Johnstone, I.M. (1995), “Adapting to Unknown Smoothness Via Wavelet Shrinkage” *Journal of the American Statistical Association*, 90: 1200–1224.
- [14] Donoho, D.L. and Johnstone, I.M. (1994), “Ideal Spatial Adaptation By Wavelet Shrinkage” *Biometrika*,81:425–455.
- [15] Robinson, G.K. (1991) “That BLUP Is a Good Thing: The Estimation of Random Effects”, *Statistical Science*, 6:15–32.
- [16] Speed, T. (1991). Comment on “That BLUP Is a Good Thing: The Estimation of Random Effects”, *Statistical Science*, 6:42–44.
- [17] Stein (1956). “Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution”. *Annals of Stat* 1:197-206.
- [18] Wahba, G. (1990), *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series, Philadelphia: SIAM.

Wavelet Decomposition Coefficients

