Chapter 6

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

$$f(x) = \mathrm{E}(Y|X = x), x \in I \subset \mathbb{R}. \tag{6.1}$$

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

6.1 Linear Spaces

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

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 $(\mathbf{X}_i, Y_i), i = 1, \ldots, n$. In some situations we are only interested in estimating $f(\mathbf{X}_i), i = 1, \ldots, n$. 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 (6.1) consists of a p-dimensional linear space $\mathcal{G}$, having as a basis the function

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

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

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

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

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

How would you write this out for linear regression?

Given observations $(\mathbf{X}_i, Y_i), i = 1, \ldots, n$ 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 $\mathbf{g} = \{g(x_1), \ldots, g(x_n)\}'$. Then the distribution of the observations of $Y|X = x$ are in the family

$$\{N(\mathbf{g}, \sigma^2 \mathbf{I}_n); \mathbf{g} = [g(x_1), \ldots, g(x_n)]',\ g \in \mathcal{G}\} \tag{6.2}$$

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

$$\{N(\mathbf{B}\boldsymbol{\theta}, \sigma^2); \boldsymbol{\theta} \in \mathbb{R}^p\} \tag{6.3}$$

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 (6.2) and (6.3) 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 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).

6.1.1 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

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

where

$$\mathrm{bias}\{\hat{g}(\mathbf{x})\} = f(x) - \mathrm{E}\{\hat{g}(\mathbf{x})\} \tag{6.4}$$

and

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

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 $\mathbf{X}$.

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

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

where $\mathbf{B}(\mathbf{x}) = (B_1(\mathbf{x}), \ldots, B_p(\mathbf{x}))'$, 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 (6.4) now reflects the ability of functions in $\mathcal{G}$ to capture the essential features of f.

6.2 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 ^{87}Sr.

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

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

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

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

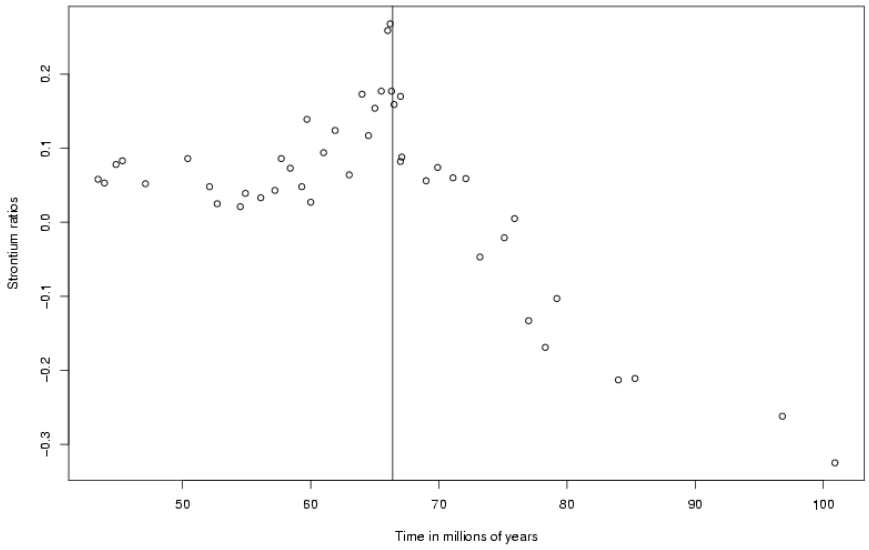


Figure 6.1: $^{87}\delta$Sr data.

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 $^{87}\delta Sr$ 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

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

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 $\mathbf{t} = (t_0, \ldots, t_{m+1})'$. Given a sequence $a = t_0 < t_1 < \ldots < t_m < t_{m+1} = b$, 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

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

In homework 2, we saw or will see that piecewise polynomials are a linear space that present an alternative to polynomials. However, it is hard to justify the breaks in the function $g(x; \hat{\boldsymbol{\theta}})$.

6.3 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 $\mathbf{t} = (t_1, \ldots, t_m)'$ 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*:

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

where $(\cdot)_+ = \max(\cdot, 0)$. Written in this way the coefficients $\theta_{1,1}, \ldots, \theta_{1,k}$ 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:

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

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

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

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?

6.3.1 Splines in terms of Spaces and sub-spaces

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

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

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

In practice we have design points $x_1, \ldots, x_n$ and a vector of responses $\mathbf{y} = (y_1, \ldots, y_n)$. 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

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

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

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

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

6.4 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 $(x_1, \ldots, x_n)$, with $g(x_i) = y_i, i, \ldots, n$, 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 $a \leq x_1 \leq \ldots \leq x_n \leq b$. 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.

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

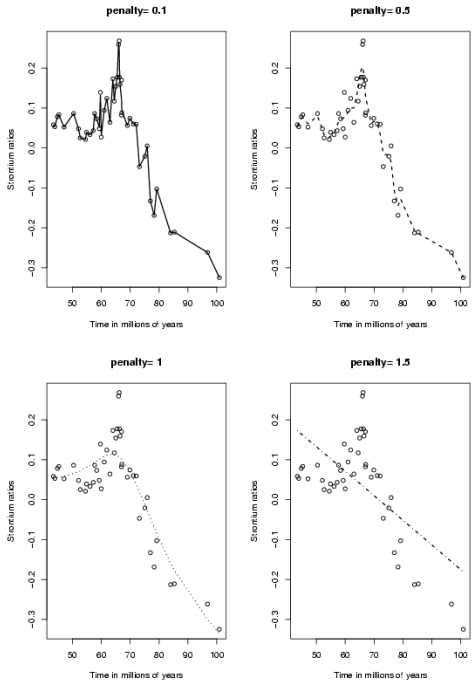


Figure 6.2: Smoothing spline fitted using different penalties.

6.5 Linear Smoothers: Influence, Variance, and Degrees of Freedom

All the smoothers we have discussed in this class are linear smoothers. The estimates of the regression function can be written as

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

For some of the smoothers we have defined we can define a weight sequence for any x and define

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

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

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

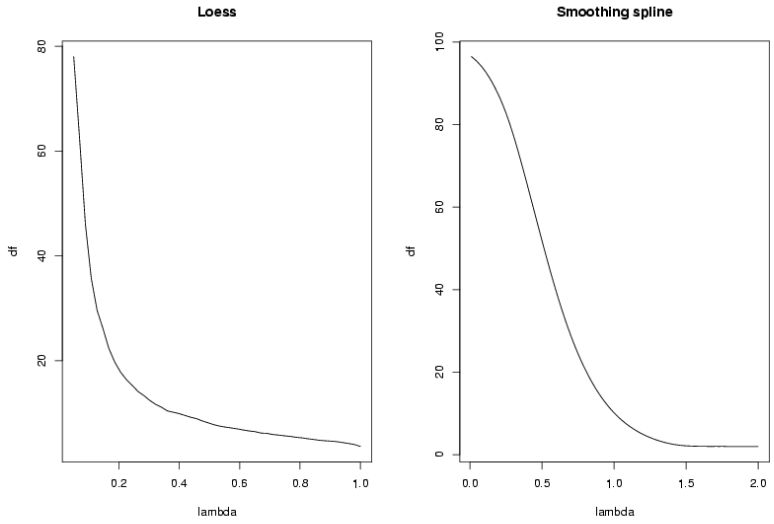
the total variance reduction from $\sum_{i=1}^{n} \text{var}[y_i]$ is $\text{tr}(\mathbf{S}\mathbf{S}')/n$.

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 freedoms 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 $\sum_{i=1}^{n} W_i(x_i) = \text{tr}(\mathbf{S})$.

Figure 6.3: Degrees of freedom for loess and smoothing splines as functions of the smoothing parameter



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.

Finally we notice that

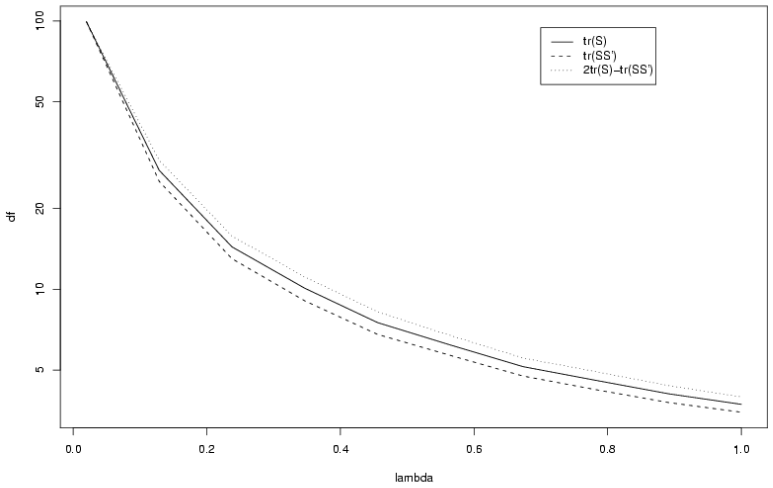
$$\text{E}[(\mathbf{y} - \hat{\mathbf{f}})'(\mathbf{y} - \hat{\mathbf{f}})] = \{n - 2\text{tr}(\mathbf{S}) + \text{tr}(\mathbf{S}\mathbf{S}')\}\sigma^2$$

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

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

Figure 6.4: Comparison of three definition of degrees of freedom



6.6 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:

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

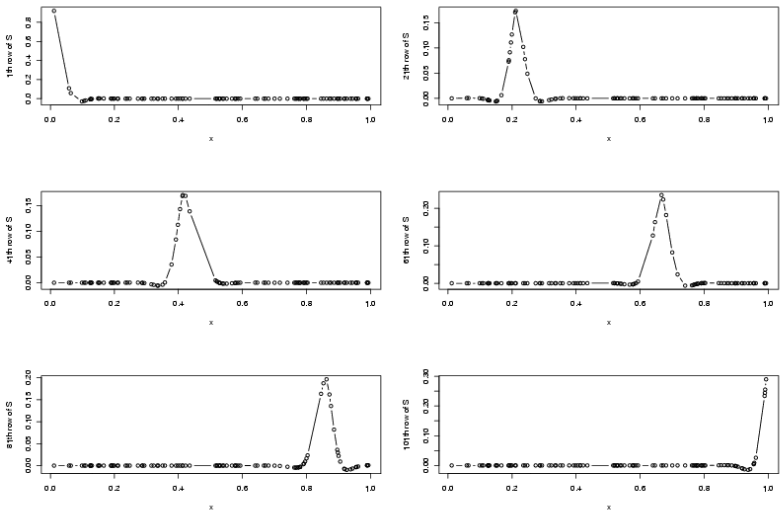
where $\mathbf{K} = \mathbf{B}'^{-}\boldsymbol{\Omega}\mathbf{B}^{-}$.

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 6.5: 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

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

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.

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

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

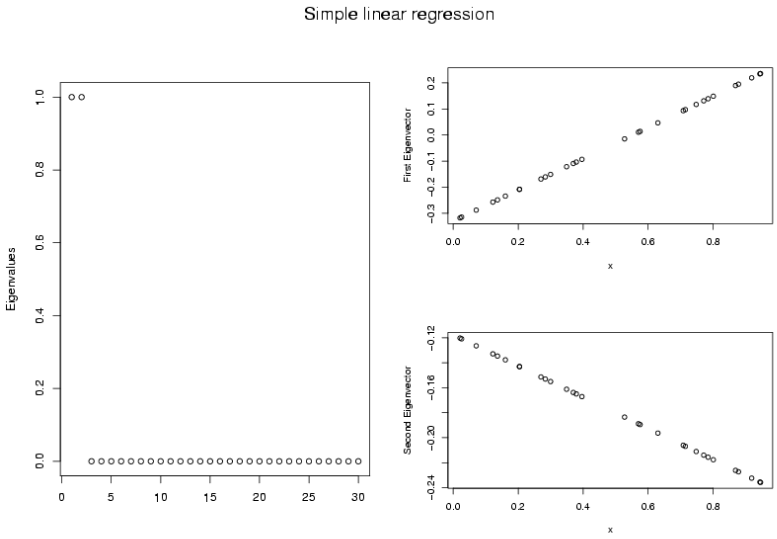
or

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

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

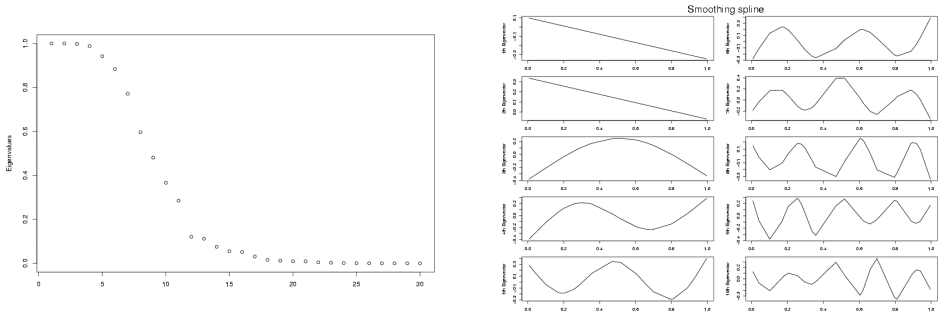


The cubic spline is an important example of a symmetric smoother, and its eigenvectors resemble polynomials of increasing degree.

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.

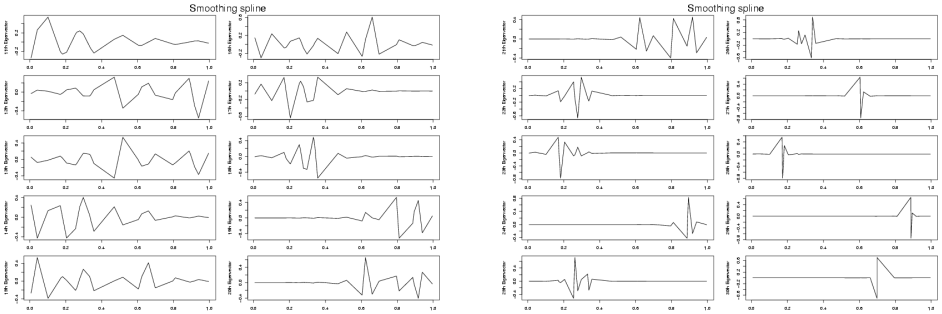
The action of the smoother is now transparent: if presented with a response $\mathbf{y} = \mathbf{u}_j$, it shrinks it by an amount θ_j as above.

Figure 6.7: 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 6.8: 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

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

On 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 6.7 shows the "transfer function" defined by the smoothing splines.

The change of basis idea described above has been explored by Donoho and Johnston 1994, 1995) and Beran (2000). In the following section we give a short introduction to these ideas.

6.8 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:

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

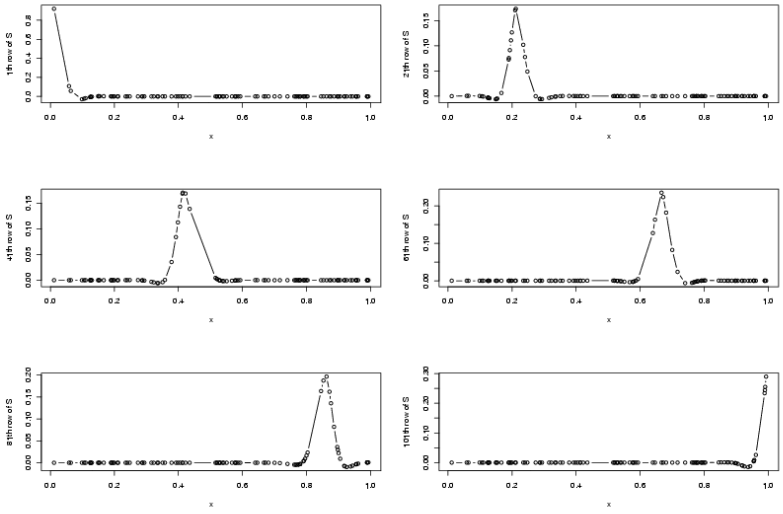
where $\mathbf{K} = \mathbf{B}'^{-}\boldsymbol{\Omega}\mathbf{B}^{-}$.

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 6.9: 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

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

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.

6.9 Economical Bases: Wavelets and REACT estimators

If one consider the "equally spaced" Gaussian regression:

$$y_i = f(t_i) + \varepsilon_i, i = 1, \ldots, n \tag{6.5}$$

$t_i = (i-1)/n$ and the ε_is 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 $\mathbf{f} = \{f(t_1), \ldots, f(t_n)\}'$.

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

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

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 that 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 mth derivative.

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

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

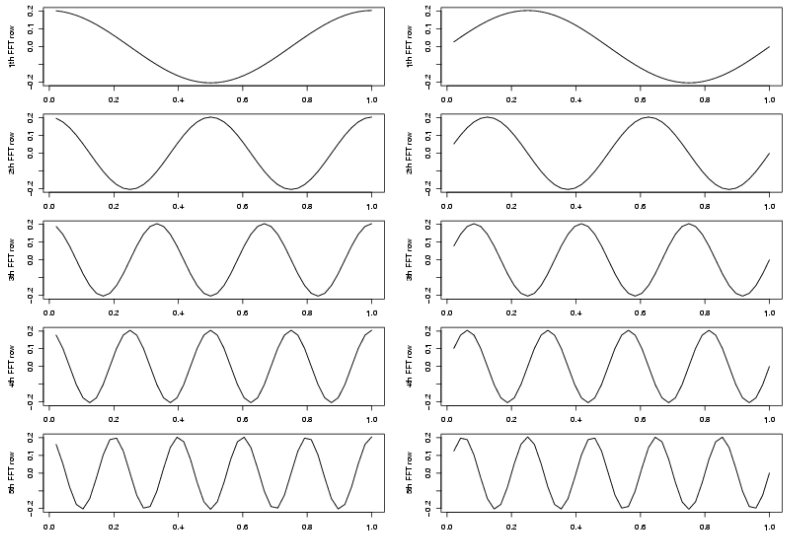
Note: This can easily be changed to the case where n is odd by substituting $n/2$ by $\lfloor n/2 \rfloor$ and taking out the last term last term $a_{\lceil n/2 \rceil}$.

If a signal is close to a sine wave $f(t) = \cos(2\pi j t/n + \phi)$ for some integer $1 \le j \le n$, 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 where 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).

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

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

6.9.2 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 (6.5). 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 $\hat{\boldsymbol{\xi}} = \{\mathbf{wz}; \mathbf{w} \in [0,1]^n\}$. Here the product $\mathbf{wz}$ 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 $\mathbf{U}\boldsymbol{\xi} = \mathbf{U}\mathbf{U}'\mathbf{f} = \mathbf{f}$.

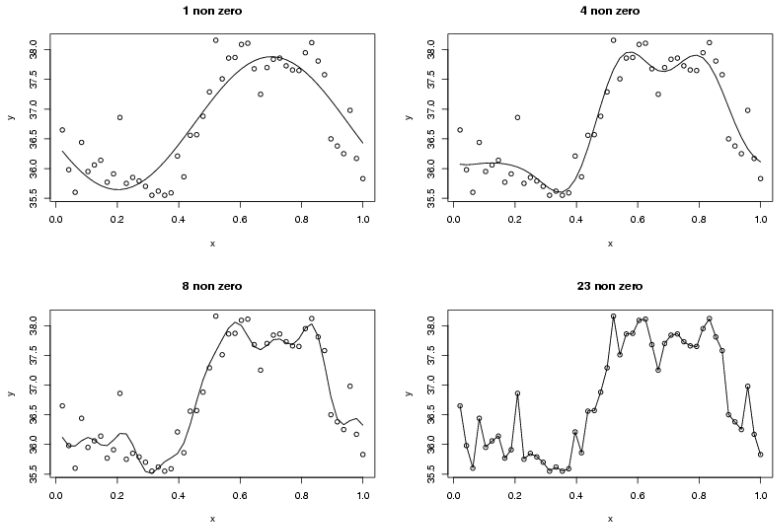
Relatively simple calculations show that $\tilde{\mathbf{w}} = \boldsymbol{\xi}^2/(\boldsymbol{\xi}^2 + \sigma^2)$ minimizes the risk over all possible $\mathbf{w} \in \mathbb{R}^n$. The MLE obtained, with $\mathbf{w} = (1, \ldots, 1)'$, minimizes the risk only if $\tilde{\mathbf{w}} = (1, \ldots, 1)'$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

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 6.10, we show for the body temperature data the the fitted curves obtained when using shrinkage coefficients of the from $\mathbf{w} = (1, 1, \ldots, 1, 0, \ldots, 0)$.

If Figure 6.11 we show the fitted curve obtained with $\mathbf{w} = (1, 1, \ldots, 1, 0, \ldots, 0)$ and using REACT. In the first plot we show the coefficients shrunken to 0 with

Figure 6.10: Fitted curves obtained when using shrinkage coefficients of the from $\mathbf{w} = (1, 1, \ldots, 1, 0, \ldots, 0)$, with $2m + 1$ the number of 1s used.



crosses. In the second $\mathbf{z}$ plot we show $\mathbf{wz}$ with crosses. Notice that only the first few coefficients of the transformation are "big". Here are the same pictures for data obtained for 6 consecutive weekends.

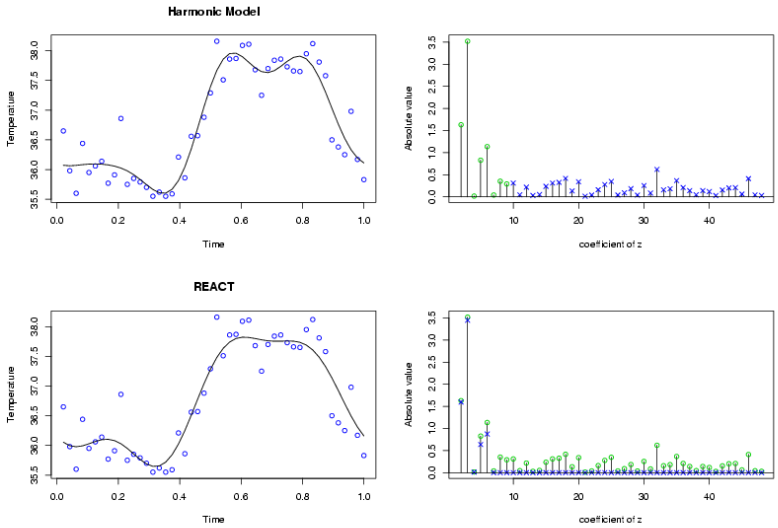
Finally in Figure 6.12 we show the two fitted curves and compare them to the average obtained from observing many days of data.

Notice that using $\mathbf{w} = (1, 1, 1, 1, 0, \ldots, 0)$ 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".

Figure 6.11: Estimates obtained with harmonic model and with REACT. We also show the **z** and how they have been shrunken.



6.9.3 Wavelets

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

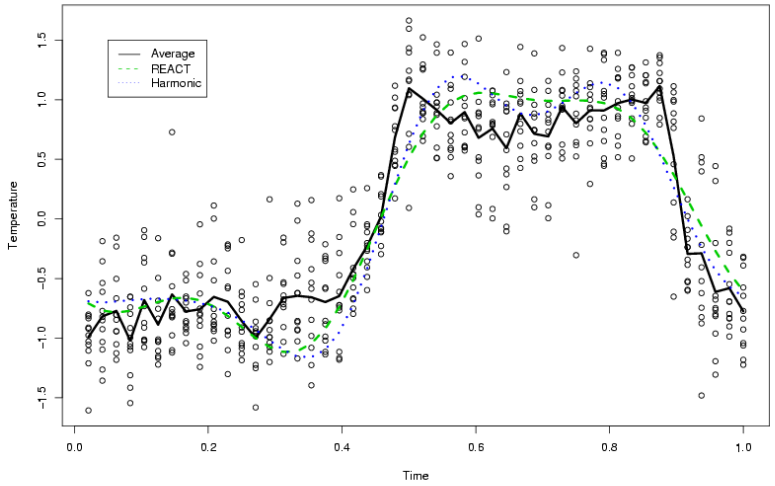
The signal does appear to have some added noise so we could use (6.5) to model the process. However, $f(x)$ appears to have a peak at around $x = 500$ making it not very smooth at that point.

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.

Figure 6.12: Comparison of two fitted curves to the average obtained from observing many days of data.

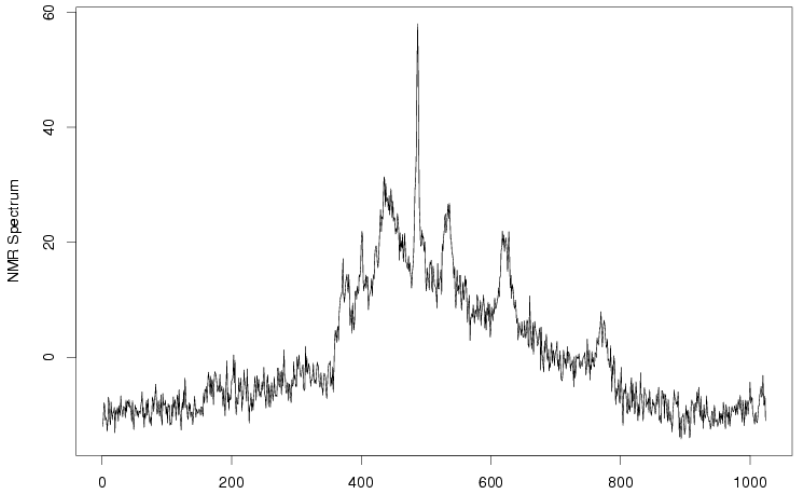


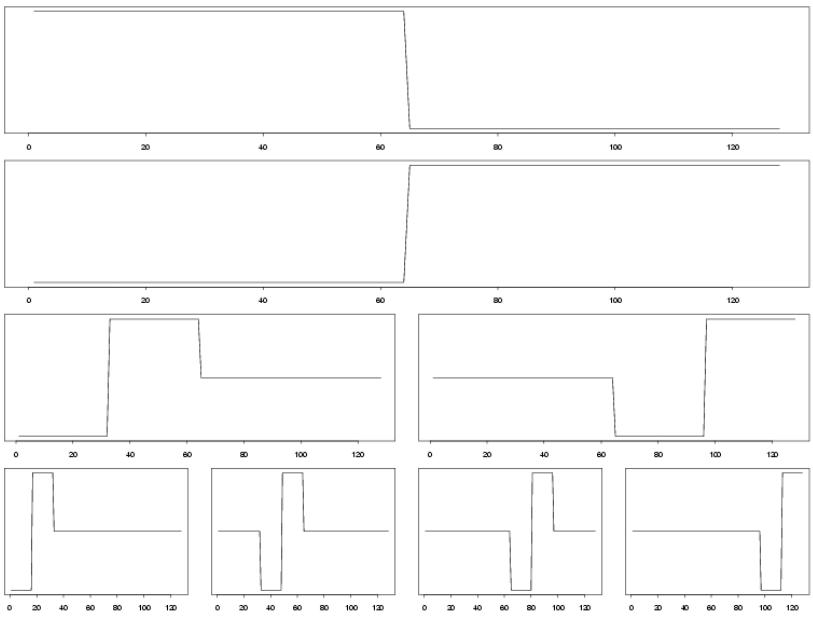
Notice that these are step function. However, there are ways (they involve complicated math and no closed forms) to create “smoother” wavelets. The following are the columns of DWT using the Daubechies wavelets

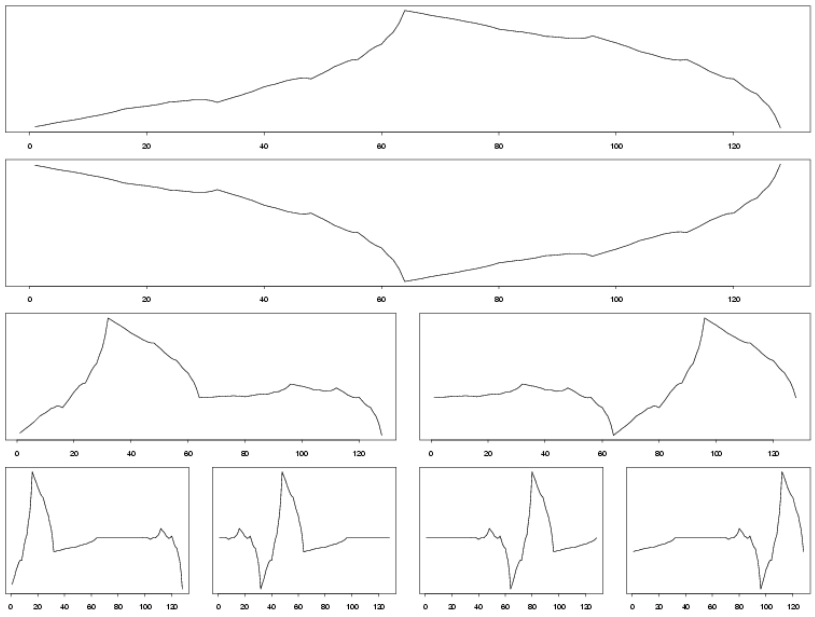
The following plot shows the coefficients of the DWT by smoothness level and by location:

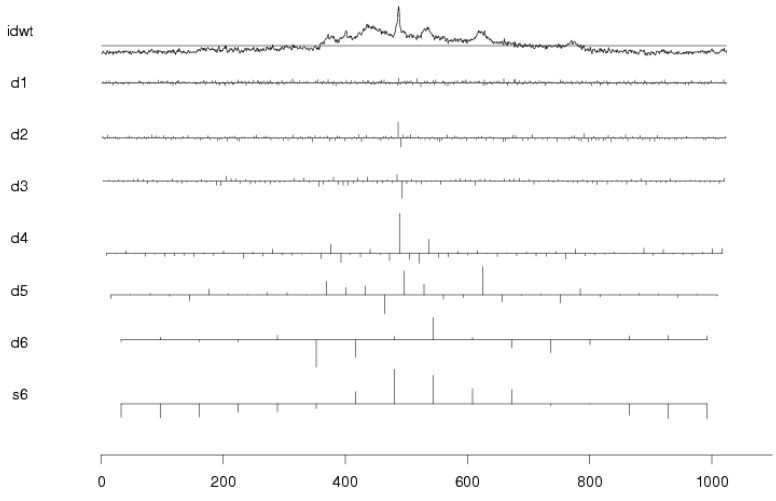
Using wavelet with shrinkage seems to perform better at de-noising than smoothing splines and loess as shown by the following figure.

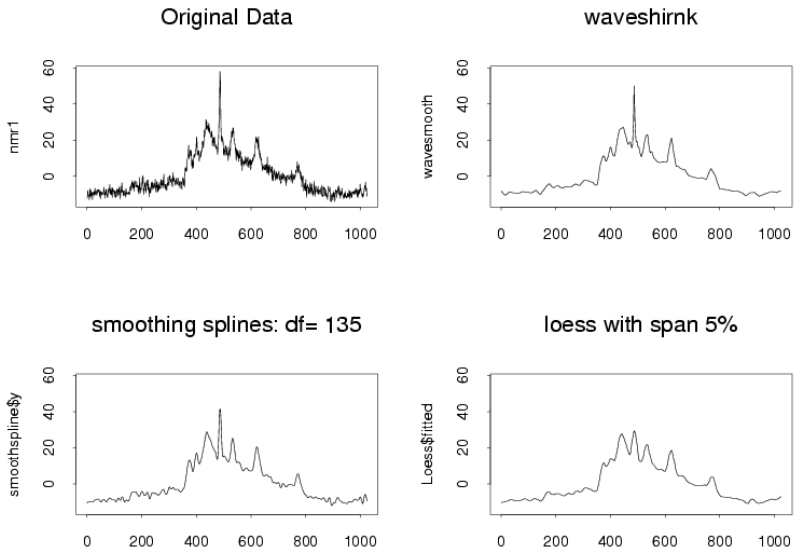
The last plot is what the wavelet estimate looks like for the temperature data

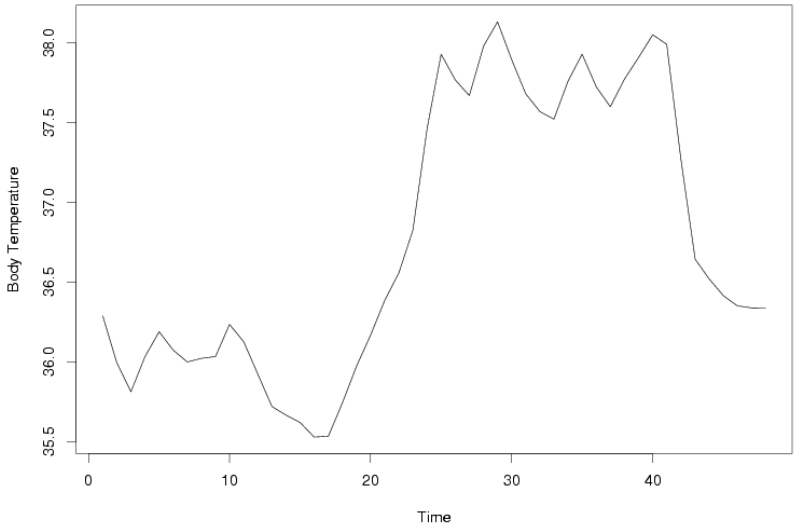

NMR Spectrum
0
20
40
60
0
200
400
600
800
1000




0
20
40
60
80
100
120


idwt
d1
d2
d3
d4
d5
d6
s6
0
200
400
600
800
1000


Original Data
waveshirnk
nmr1
wavesmooth
smoothing splines: df= 135
loess with span 5%
smoothspline$y
Loess$fitted
0
20
40
60
0
200
400
600
800
1000


Body Temperature
35.5
36.0
36.5
37.0
37.5
38.0
0
10
20
30
40
Time

Bibliography

[1] Eubank, R.L. (1988), *Smoothing Splines and Nonparametric Regression*, New York: Marcel Decker.

[2] Reinsch, C.H. (1967) Smoothing by Splins Functions. *Numerische Mathematik*, 10: 177–183

[3] Schoenberg, I.J. (1964), "Spline functions and the problem of graduation," *Proceedings of the National Academy of Science*, USA 52, 947–950.

[4] 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.

[5] Wahba, G. (1990), *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series, Philadelphia: SIAM.

[6] Beran, R. (2000). "REACT scatterplot smoothers: Superefficiency through basis economy", *Journal of the American Statistical Association*, 95:155–171.

[7] 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.

[8] Donoho, D.L. and Johnstone, I.M. (1995), "Adapting to Unknown Smoothness Via Wavelet Shrinkage" *Journal of the American Statistical Association*, 90: 1200–1224.

[9] Donoho, D.L. and Johnstone, I.M. (1994), "Ideal Spatial Adaptation By Wavelet Shrinkage" *Biometrika*,81:425–455.

[10] Robinson, G.K. (1991) "That BLUP Is a Good Thing: The Estimation of Random Effects", *Statistical Science*, 6:15–32.

[11] Speed, T. (1991). Comment on "That BLUP Is a Good Thing: The Estimation of Random Effects", *Statistical Science*, 6:42–44.

[12] Stein (1956). "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution". *Annals of Stat* 1:197-206.

[13] Wahba, G. (1990), *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series, Philadelphia: SIAM.