Chapter 6

Connections

6.1 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.

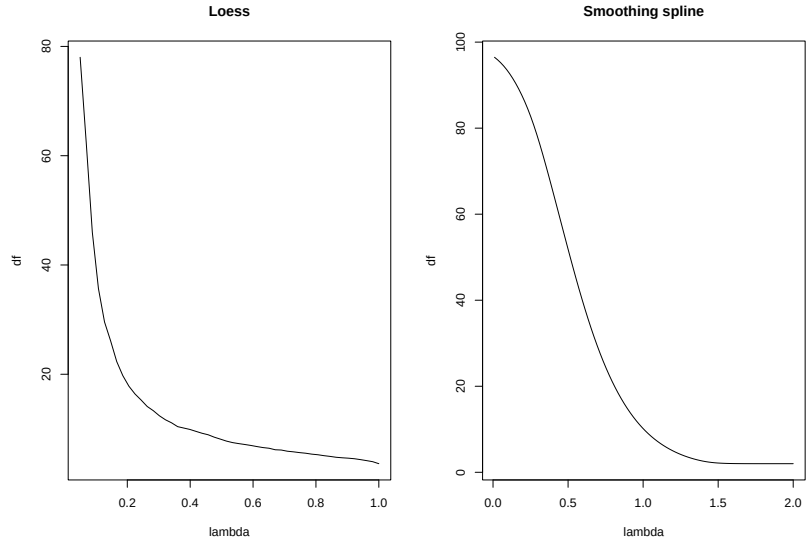
We will define variance reduction, influence, and degrees of freedom for linear smoothers.

The variance of the interpolation estimate is 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.

Figure 6.1: Degrees of freedom for loess and smoothing splines as functions of the smoothing parameter



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 $\text{tr}(\mathbf{S}\mathbf{S}')/n$.

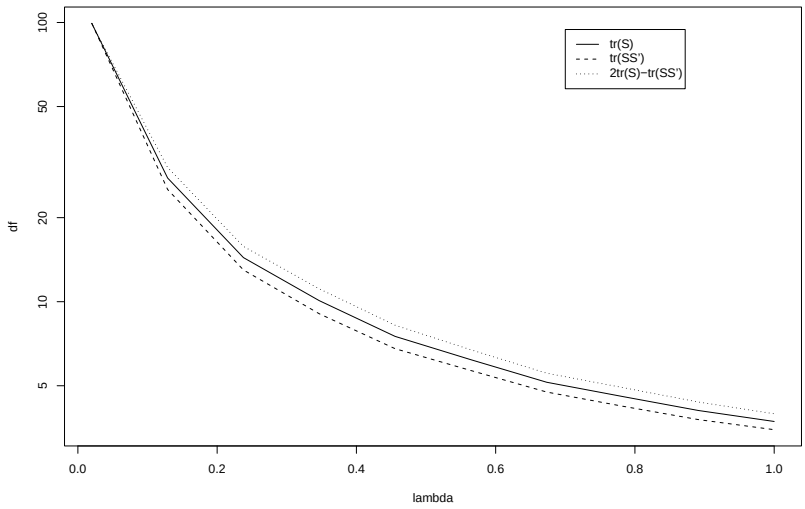
In linear regression the variance reduction is related to the degrees of freedom, or number of parameters. For linear regression, $\sum_{i=1}^{n} \text{var}[\hat{f}(x_i)] = p\sigma^2$. One widely used definition of degrees of freedoms for smoothers is $df = \text{tr}(\mathbf{S}\mathbf{S}')$.

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

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 6.2: Comparison of three definition of degrees of freedom



Finally we notice that

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

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

Under relatively mild assumptions we can show that

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

6.2 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:

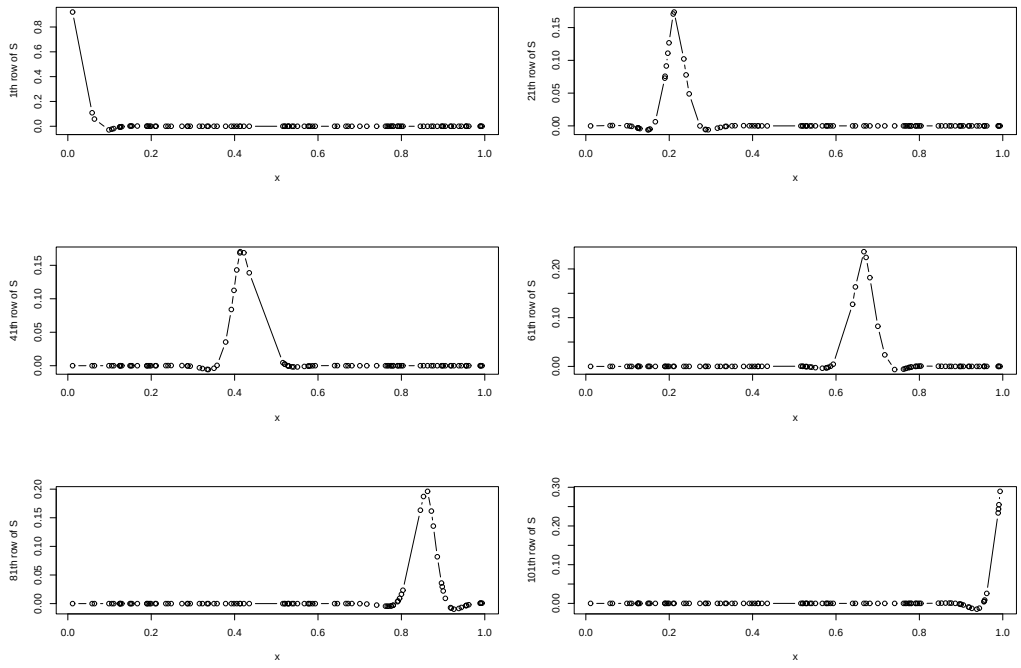
$$\hat{\mathbf{g}} = \mathbf{B}\boldsymbol{\theta} = \mathbf{B}(\mathbf{B}'\mathbf{B} + \lambda\boldsymbol{\Omega})^{-1}\mathbf{B}'\mathbf{y} = (\mathbf{I} + \lambda\mathbf{K})^{-1}\mathbf{y}$$

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

Notice we can write the penalized criterion as

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

Figure 6.3: Kernels of a smoothing spline.



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

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.

6.3 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 $\theta_1 \geq \theta_2 \ldots \geq \theta_n$:

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

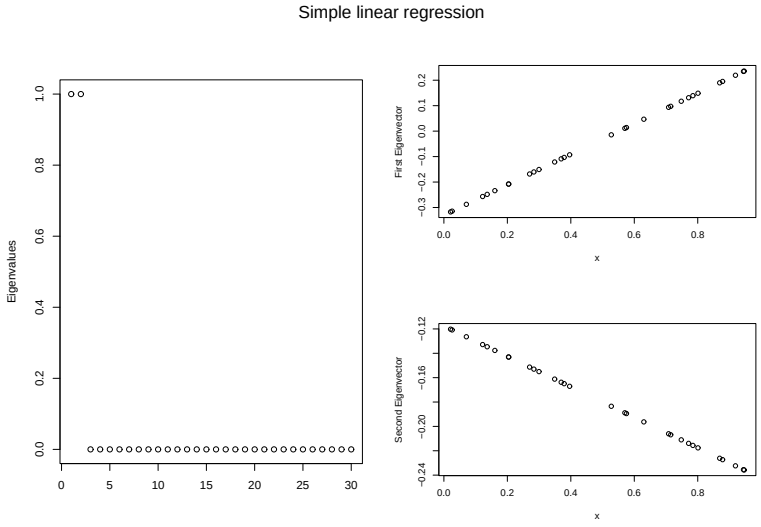
or

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

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

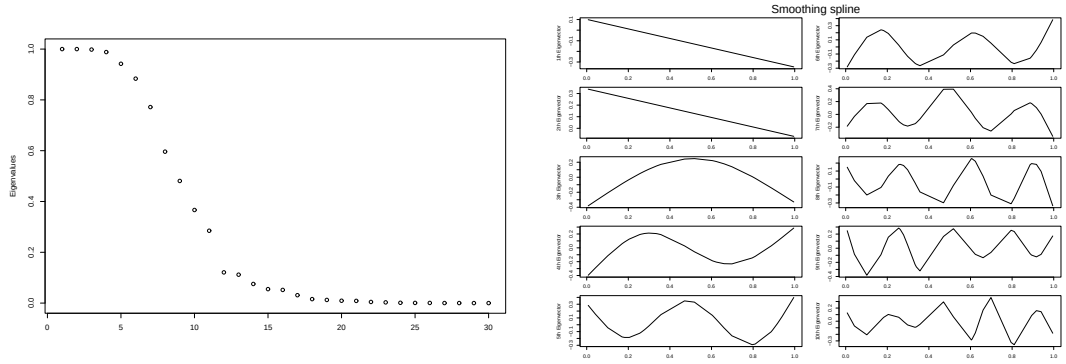


The cubic spline is an important example of a symmetric smoother, and its eigenvectors resemble polynomials of increasing degree.

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



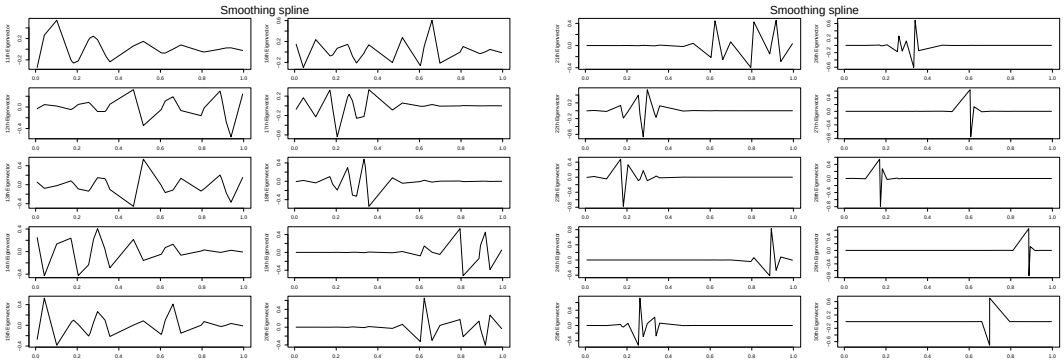
Cubic smoothing splines, regression splines, linear regression, polynomial regression are all symmetric smoothers. However, loess and other "nearest neighbor" smoothers are not.

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 $\hat{\mathbf{z}} = \mathbf{Dz}$ 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.

Figure 6.6: Eigen vectors 11 through 30 for a smoothing spline for $n = 30$.



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 6.5 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.

6.4 Economical Bases: Wavelets and REACT estimators

If one consider the "equally spaced" Gaussian regression:

$$y_i = f(t_i) + \varepsilon_i, i = 1, \ldots, n \tag{6.1}$$

$t_i = (i-1)/n$ 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.

6.4.1 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:

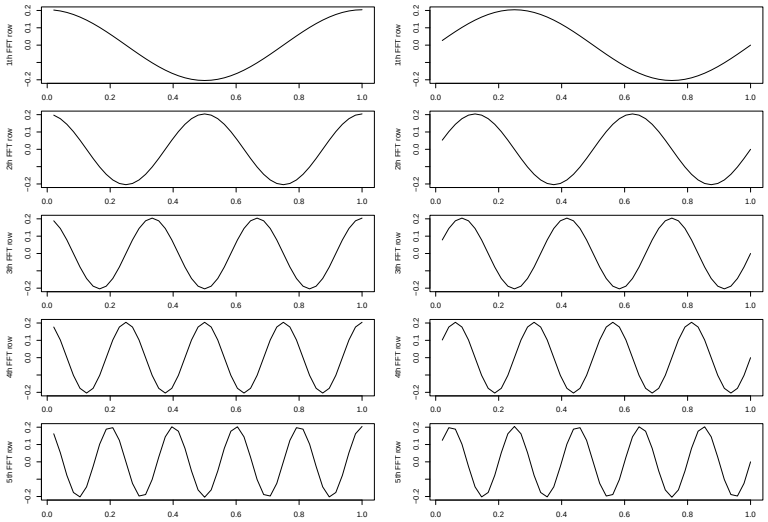
$$\begin{aligned} U_1 &= [n^{-1/2} : 1 \le i \le n] \\ U_{2k} &= [(2/n)^{1/2} \sin\{2\pi k i/n\} : 1 \le i \le n], k = 1, \ldots, n/2 \\ U_{2k+1} &= [(2/n)^{1/2} \cos\{2\pi k i/n\} : 1 \le i \le n], k = 1, \ldots, n/2 - 1. \end{aligned}$$

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).

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

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

6.4.2 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 (6.1). 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.

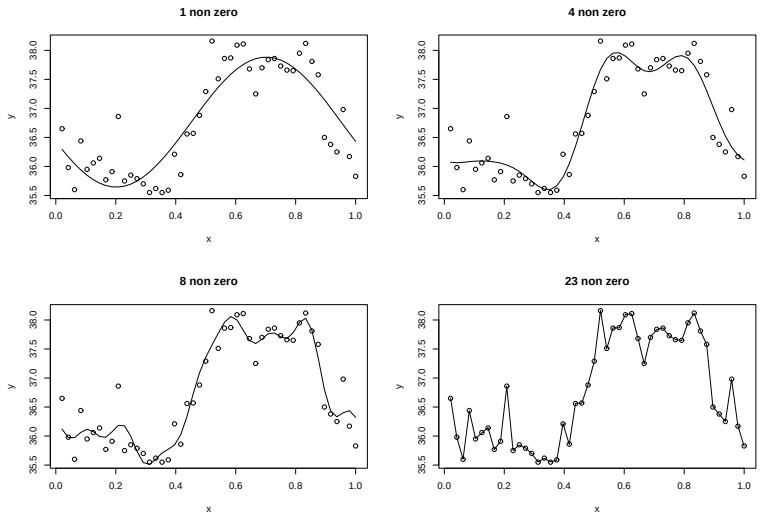
We can then choose the shrinkage coefficients that minimize the risk

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

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!

Figure 6.7: 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.

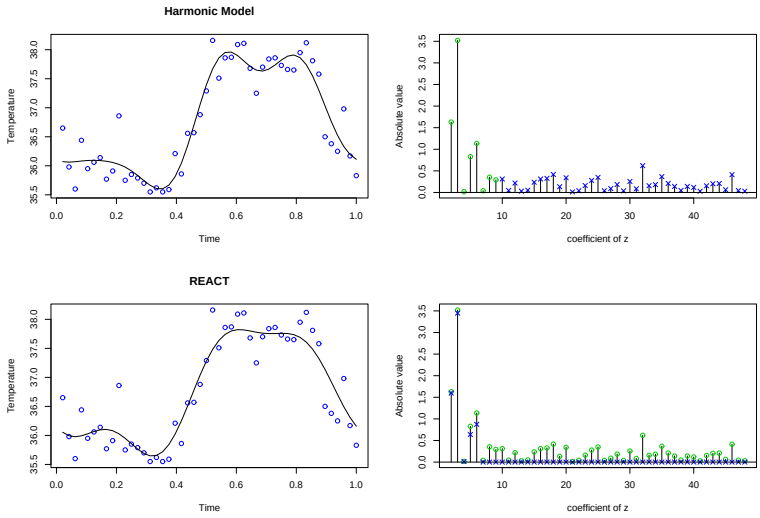


Notice that $\tilde{\mathbf{w}}$ makes sense because it shrinks coefficients with small signal to noise ratio. By shrinking small coefficients closer to 0 we reduce variance and the bias we add is not very large, thus reducing risk. However, we don't know $\boldsymbol{\xi}$ nor σ^2 so in practice we 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 **??**, we show for the body temperature data the the fitted curves obtained when using shrinkage coefficients of the from $\mathbf{w} = (1, 1, \ldots, 1, 0, \ldots, 0)$.

Figure 6.8: Estimates obtained with harmonic model and with REACT. We also show the $\mathbf{z}$ and how they have been shrunken.

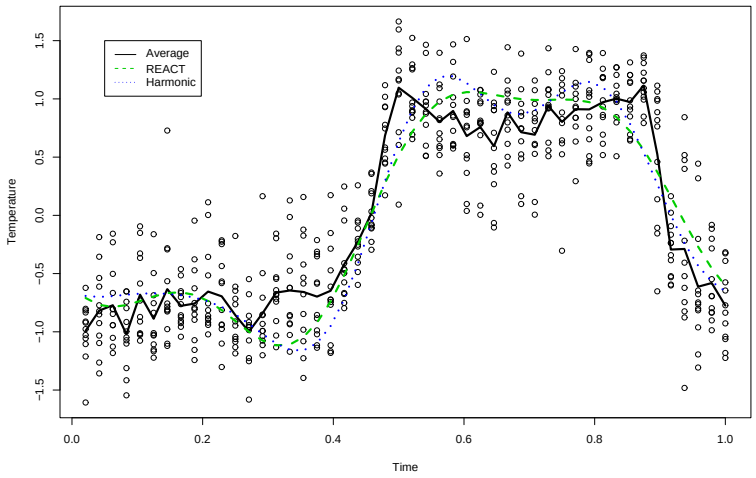


If Figure 6.8 we show the fitted curve obtained with $\mathbf{w} = (1, 1, \ldots, 1, 0, \ldots, 0)$ and using REACT. In the first plot we show the coefficients shrunken to 0 with crosses. In the second $\mathbf{z}$ plot we show $\mathbf{wz}$ with crosses. Notice that only the first few coefficients of the transformation are "big". Here are the same pictures for data obtained for 6 consecutive weekends.

Finally in Figure 6.9 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

Figure 6.9: Comparison of two fitted curves to the average obtained from observing many days of data.



assumes f is a sum of 4 cosine functions.

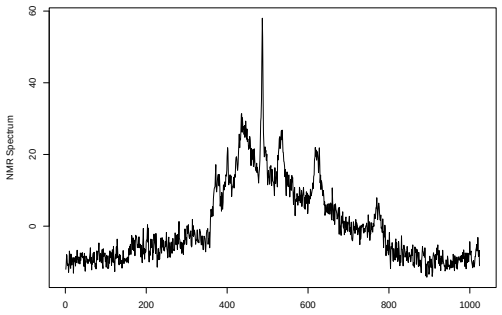
Any smoother with a smoothing matrix $\mathbf{S}$ that is a projection, e.g. linear regression, splines, can be consider a special case of what we have described here.

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".

6.4.3 Wavelets

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

The signal does appear to have some added noise so we could use (6.1) to model

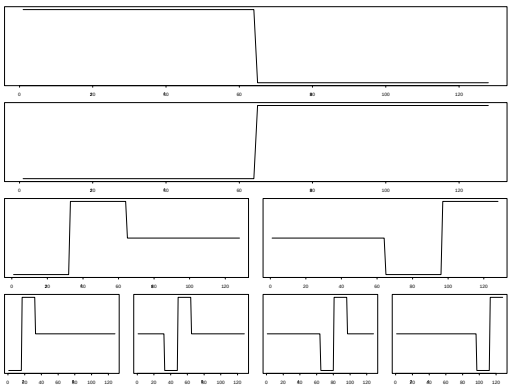


the process. However, $f(x)$ appears to have a peak at around $x = 500$ making it not very smooth at that point.

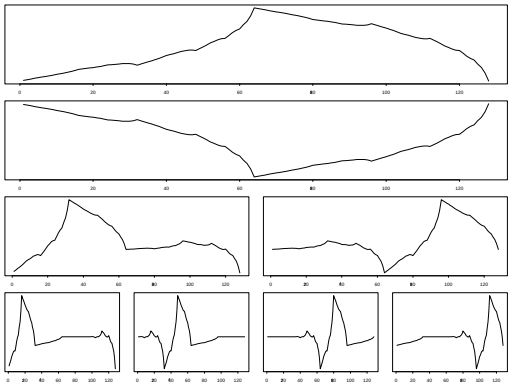
Situations like these are where wavelets analyses is especially useful for "smoothing". Now a more appropriate word is "de-noising".

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.

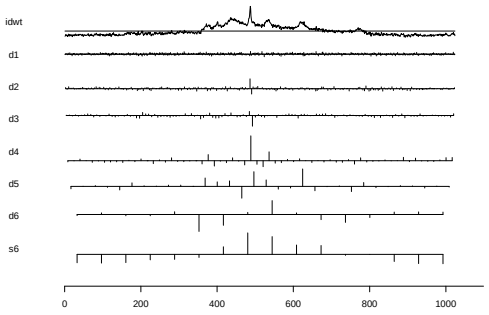
Here are the columns of the Haar DWT, the simplest wavelet.



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

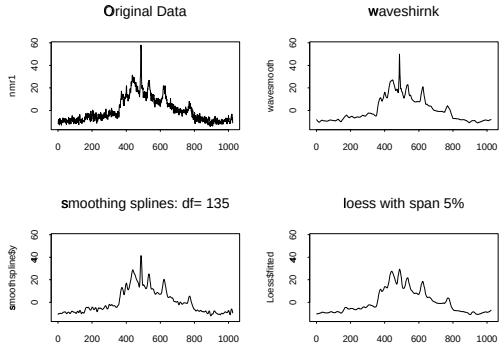


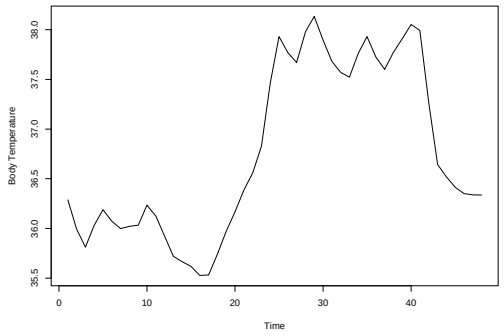
The following plot shows the coefficients of the DWT by smoothness level and by location:



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





6.5 Bayesian Model for Cubic Splines

Notice that the above definition can be easily extended to a Bayesian problem: If we define a distribution for f, say $dP(f)$ then we may consider the average risk

$$\int_f \mathrm{E}[||\hat{g} - f||^2 | f] \, dP(f)$$

as a criterion.

This follows section 3.6 in the book by Hastie and Tibshirani.

The cubic smoothing spline can be derived from a number of Bayesian models for

smoothing. The details are hard, but can be found in a book by Wahba.

Here we will demonstrate a fairly simple example.

Remember we can write any natural cubic spline as

$$g(x) = \mathbf{B}(x)\boldsymbol{\theta}.$$

Consider the following Bayesian set-up:

Model assumptions: Assume the data $\mathbf{y}$ follow a Gaussian distribution $N(\mathbf{B}\boldsymbol{\theta}, \sigma^2\mathbf{I}_n)$.

Prior assumptions: Assume $\boldsymbol{\theta}$ follows a multivariate Gaussian prior distribution with mean 0 and variance $\sigma^2/\lambda\boldsymbol{\Omega}^{-1}$.

it follows that the posterior distribution of $\boldsymbol{\theta}$ is multivariate Gaussian with mean

$$\mathrm{E}(\boldsymbol{\theta}|\mathbf{y}) = \mathbf{B}(\mathbf{B}'\mathbf{B} + \lambda\boldsymbol{\Omega})^{-1}\mathbf{B}'\mathbf{y}$$

which is the natural smoothing spline estimate of $\boldsymbol{\theta}$.

6.6 Mixed Models and Splines

Mixed models are defined by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

where $\mathbf{y}$ is a vector of n observable random variables, $\boldsymbol{\beta}$ is vector of p unknown parameters having fixed values (fixed effects), $\mathbf{X}$ and $\mathbf{Z}$ are known matrices and $\mathbf{u}$ and $\boldsymbol{\varepsilon}$ are vector, of length q and n, of unobservable variables (random effects) such that $\mathrm{E}(\mathbf{u}) = 0$ and $\mathrm{E}(\boldsymbol{\varepsilon}) = 0$ and

$$\mathrm{var}\begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}\sigma^2$$

. Here R and G are known positive definite matrices and σ^2 is a positive constant.

Random effects are especially useful to introduce correlation to the random part of the model.

Robinson (1991) describes how the best linear unbiased predictor of $\mathbf{y}$ is a "good thing". Speed (1991) notes that after defining an appropriate $\mathbf{X}$, $\mathbf{Z}$, and G we have that natural smoothing are BLUPs.

The smoothing parameter is included in the G and one can view it a "nuisance" parameter. Robinson (1991) suggests REML estimation as a way of "estimating" the smoothness parameter. Speed notes that this is equivalent to Wahba's Generalized Maximum Likelihood estimate of the smoothing parameter.

Furthermore, this idea permits us to model nested curves in a natural way. See Brumback and Rice (1998).

Bibliography

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

[2] 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.

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

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

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

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

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

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