MLISP: Machine Learning in Signal Processing

Lecture 11

Prof. V. I. Morgenshtern

Scribe: M. Elminshawi

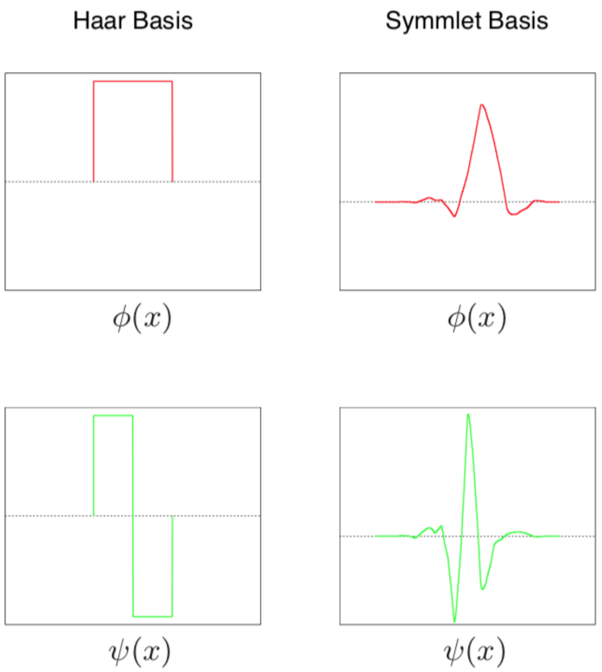
Illustrations: The elements of statistical learning, Hastie, Tibshirani, Friedman

Agenda

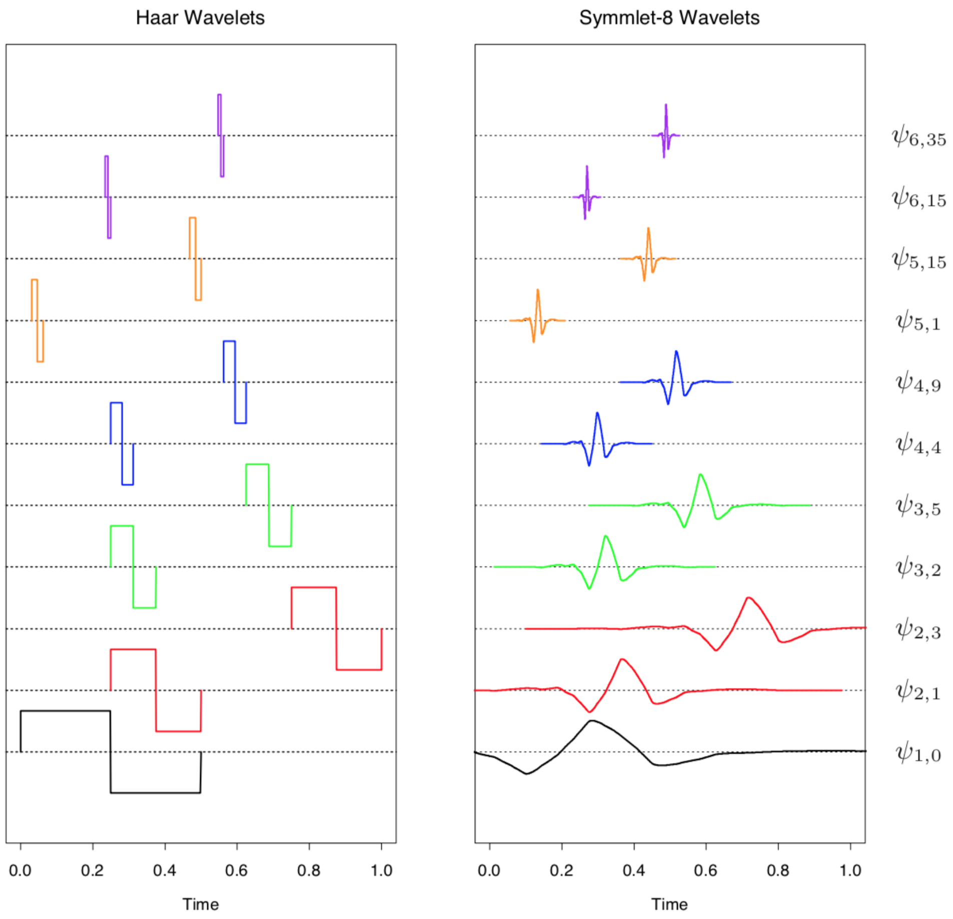
1. Wavelets with good frequency localization
2. Wavelet transform of NMR signal
3. Denoising via wavelet shrinkage

1 Wavelets with good frequency localization

Haar wavelets are simple to understand, but they are not smooth enough for most purposes. The Daubechies symmlet-8 wavelets, constructed in a similar way, have the same orthonormal properties as Haar wavelets, but are smoother. Here is the comparison of father functions for the two systems:



and here are the shifts:



Let's compare the two systems:

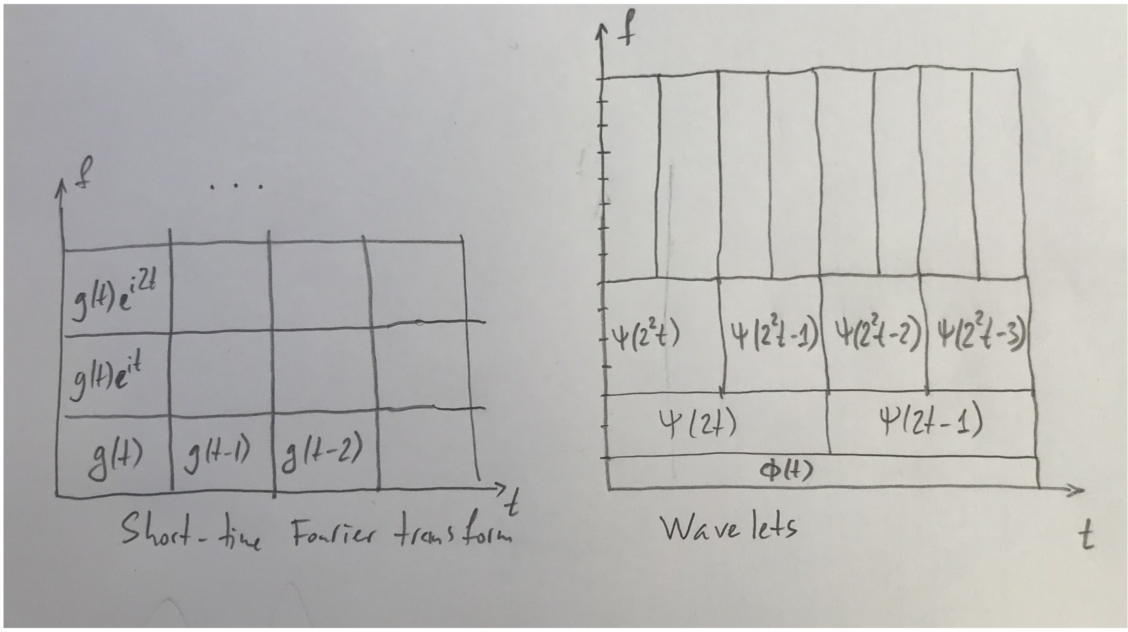
- Each symmlet-8 wavelet has a support covering 15 consecutive time intervals rather than one for the Haar basis. More generally, the symmlet-p family has the support of $2p - 1$ consecutive intervals. The wider support, the more time the wavelet has two die to zero, and so can achieve this more smoothly. Note that the effective support seems to be much narrower than the theoretical one.

- The symmlet-p wavelet $\psi(\cdot)$ has p vanishing moments:

  $$\int \psi(t) t^j dt = 0, \quad j = 0, ..., p-1. \tag{1}$$

  It follows that any order-p polynomial over $N = 2^J$ time points is reproduced exactly in V_0. The Haar wavelets have one vanishing moment and V_0 can reproduce any constant function.

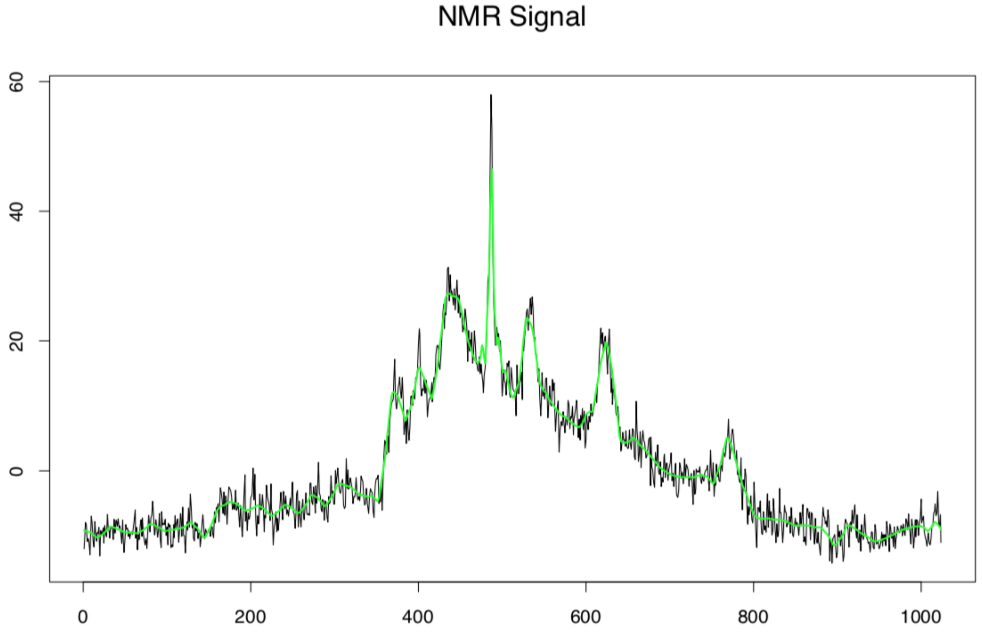
Consider the time-frequency plane. It is instructive to compare how wavelets tile this plane to how short time Fourier transform functions tile it:



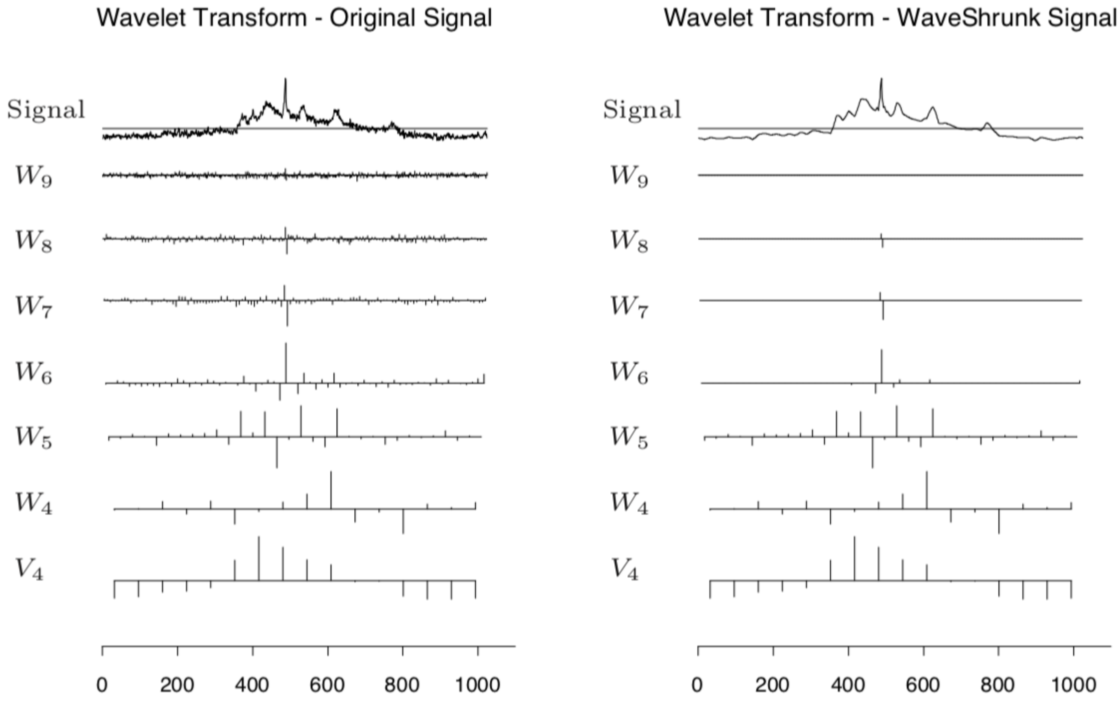
We observe that in the case of wavelets every time we go up the hierarchy stack, the number of functions doubles, each function becomes more localized (more narrow), and therefore it's frequency content doubles.

2 Wavelet transform of NMR signal

In the following figure we see an NMR (nuclear magnetic resonance) signal, which appears to be composed of smooth components and isolated spikes, plus some noise:



The wavelet transform using Symmlet basis, is shown in the lower left panel:



The wavelet coefficients are arranged in rows, from lowest scale at the bottom to highest scale at the top. The length of each line segment indicates the size of the coefficients. These rows correspond to the decomposition:

$$\text{signal span} = V_4 \oplus W_4 \oplus W_5 \oplus W_6 \oplus W_7 \oplus W_8 \oplus W_9 \tag{2}$$

that we studied before.

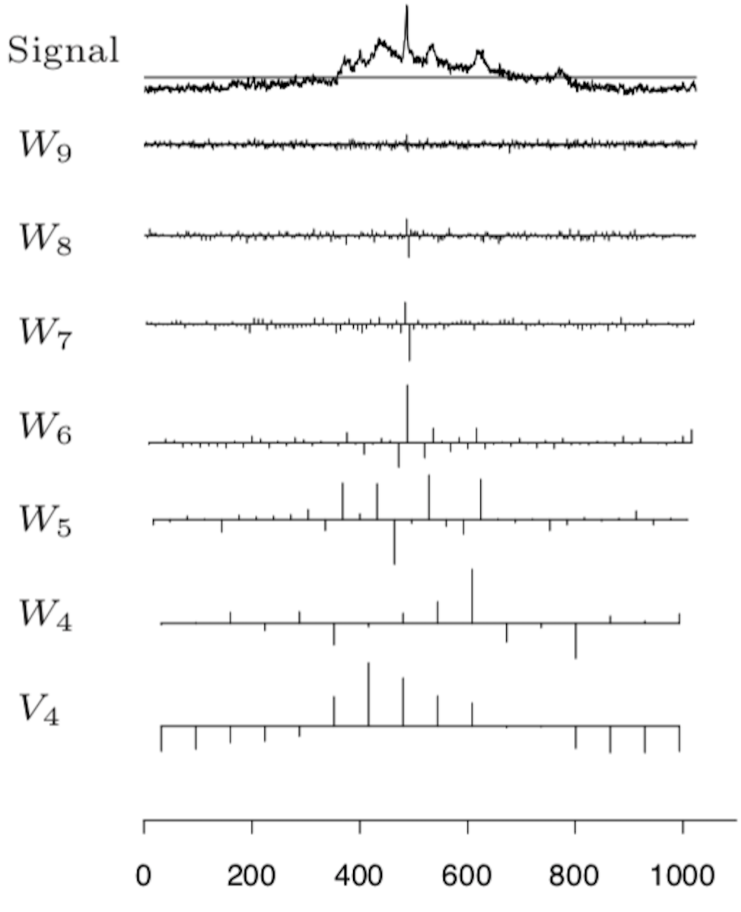
The bottom right panel shows the wavelet coefficients after they have been thresholded via soft-thresholding (will study next). Note that many of the smaller coefficients have been set to zero. The green curve in the top panel shows inverse transform of the thresholded coefficients. This is the denoised version of the original signal.

3 Denoising via wavelet shrinkage

Let's now give a mathematically precise method to obtain the wavelet smoothing. Suppose our NMR signal (rescaled to live on the interval $[0, 1]$), is sampled at $N = 2^J$ lattice points; call the discrete signal $\mathbf{y}$. Since wavelets form an orthonormal basis, we can stack the wavelet basis functions into an orthogonal matrix $\mathbf{W}^{\mathsf{T}}$. Each row of $\mathbf{W}^{\mathsf{T}}$ is one wavelet basis function, sampled at the lattice of N points $\mathbf{t} = [t_1, \ldots, t_N]$:

$$\mathbf{W}^\mathsf{T} = \begin{bmatrix} \phi(t_1) & \phi(t_2) & \ldots & \phi(t_N) \\ \psi(t_1) & \psi(t_2) & \ldots & \psi(t_N) \\ \sqrt{2}\psi(2t_1) & \sqrt{2}\psi(2t_2) & \ldots & \sqrt{2}\psi(2t_N) \\ \sqrt{2}\psi(2t_1 - 1) & \sqrt{2}\psi(2t_2 - 1) & \ldots & \sqrt{2}\psi(2t_N - 1) \\ 2\psi(2^2 t_1) & 2\psi(2^2 t_2) & \ldots & 2\psi(2^2 t_N) \\ 2\psi(2^2 t_1 - 1) & 2\psi(2^2 t_2 - 1) & \ldots & 2\psi(2^2 t_N - 1) \\ 2\psi(2^2 t_1 - 2) & 2\psi(2^2 t_2 - 2) & \ldots & 2\psi(2^2 t_N - 2) \\ \vdots & & & \\ 2^{J/2}\psi(2^J t_1 - 2^J + 1) & 2^{J/2}\psi(2^J t_2 - 2^J + 1) & \ldots & 2^{J/2}\psi(2^J t_N - 2^J + 1) \end{bmatrix}$$

Note that the matrix $\mathbf{W}$ is a square matrix of size $2^J \times 2^J$. Then $\mathbf{y}^* = \mathbf{W}^\mathsf{T}\mathbf{y}$ is called the discrete wavelet transform of $\mathbf{y}$. Here is a signal $\mathbf{y}$ at the top and it's wavelet transform coefficients arranged by scale in the remaining rows:



Notice that in this representation, the coefficients do not descent all way to V_0 but stop at V_4 which has 16 basis functions. As we ascend to each new level of detail, the coefficients get smaller, except in the locations where spiky behavior is present.

We would like to somehow capture the fact that the true signal should have few nonzero coefficients only those that correspond to spiky behavior.

One approach to do this is to solve:

$$\min_{\boldsymbol{\theta}} \quad \underbrace{\|\mathbf{y} - \mathbf{W}\boldsymbol{\theta}\|_2^2}_{\text{least squares regression term}} \quad + \quad \underbrace{2\lambda \cdot (\#\text{ of nonzero elements in } \boldsymbol{\theta})}_{\text{regularization term}}.$$

Above, $\mathbf{y}$ is the noisy data, $\mathbf{W}$ is the transformation matrix from the wavelet domain to the signal domain, $\boldsymbol{\theta}$ is the vector of estimated wavelet coefficients, λ is the parameter that controls the trade-off between data fidelity and sparsity.

Unfortunately, the above optimization problem is computationally infeasible! It is non-convex and even non-smooth; we cannot use the gradient descent. Instead we solve this:

$$\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{W}\boldsymbol{\theta}\|_2^2 + 2\lambda \|\boldsymbol{\theta}\|_1 . \tag{3}$$

where the $\|\boldsymbol{\theta}\|_1$ is the term that promotes sparsity as we will shortly see.

In general, the optimization problem of form (3) is convex, and therefore the minimum can easily be found numerically. The case we are considering now is special, because $\mathbf{W}$ is an orthogonal matrix. In this case, the solution can be found analytically as follows. First note that since $\mathbf{W}$ is orthogonal, it does not change the norm of a vector so that:

$$\|\mathbf{y} - \mathbf{W}\boldsymbol{\theta}\|_2^2 = \left\|\mathbf{W}^\mathsf{T}(\mathbf{y} - \mathbf{W}\boldsymbol{\theta})\right\|_2^2 = \left\|\mathbf{W}^\mathsf{T}\mathbf{y} - \underbrace{\mathbf{W}^\mathsf{T}\mathbf{W}}_{\mathbf{I}}\boldsymbol{\theta}\right\|_2^2 = \left\|\mathbf{W}^\mathsf{T}\mathbf{y} - \boldsymbol{\theta}\right\|_2^2 = \|\mathbf{y}^* - \boldsymbol{\theta}\|_2^2 .$$

Hence, (3) is equivalent to:

$$\min_{\boldsymbol{\theta}} \|\mathbf{y}^* - \boldsymbol{\theta}\|_2^2 + 2\lambda \|\boldsymbol{\theta}\|_1$$

which can be written as

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} (y_i^* - \theta_i)^2 + 2\lambda \sum_{i=1}^{N} |\theta_i|.$$

Therefore, the problem decouples and we just need to solve:

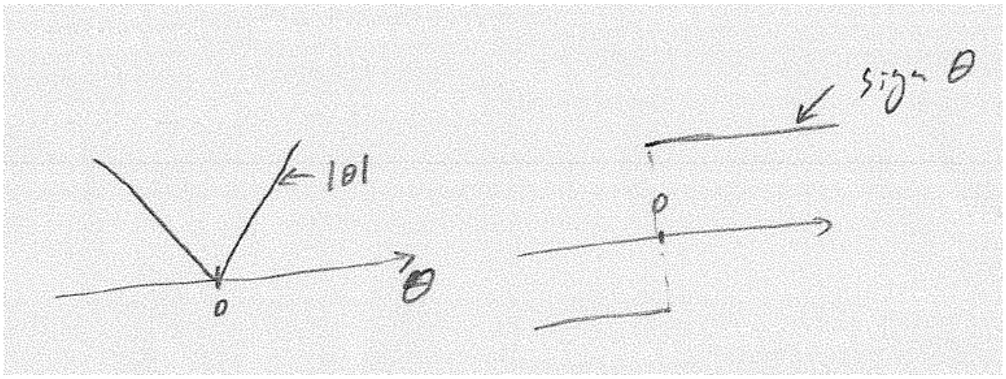
$$\min_{\theta_i} (y_i^* - \theta_i)^2 + 2\lambda|\theta_i|$$

for each i separately.

Let $f(\theta) = (y^* - \theta)^2 + 2\lambda|\theta|$. What is the minimum of $f(\theta)$? To find the minimum of this function, take the derive and set it to zero:

$$f'(\theta) = -2(y^* - \theta) + 2\lambda \operatorname{sign}\theta$$

where we used $|\theta|' = \operatorname{sign}(\theta)$ as can be seen from the plot:



We find, $f'(\theta) = 0$ iff

$$-y^* + \theta + \lambda \operatorname{sign} \theta = 0. \tag{4}$$

Claim: The solution to (4) is

$$\theta^* = \operatorname{sign}(y^*)(|y^*| - \lambda)_+, \tag{5}$$

where

$$(u)_+ = \begin{cases} u, & u > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Proof. If $\theta > 0$, then (4) is equivalent to:

$$\begin{aligned} &- y^* + \theta + \lambda = 0 \\ &\Leftrightarrow \theta = y^* - \lambda \\ &\Leftrightarrow \theta = (y^* - \lambda)_+ \\ &\Leftrightarrow \theta = \operatorname{sign}(y^*)(|y^*| - \lambda)_+ \end{aligned}$$

where the last step follows because $\theta = y^* - \lambda > 0 \Rightarrow y^* > \lambda > 0 \Rightarrow \operatorname{sign}(y^*) = 1$.

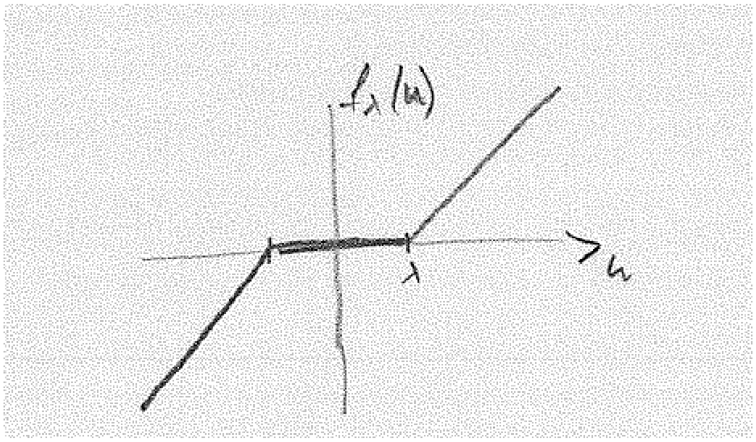
If $\theta < 0$, then (4) is equivalent to:

$$\begin{aligned} &- y^* + \theta - \lambda = 0 \\ &\Leftrightarrow \theta = y^* + \lambda \\ &\Leftrightarrow \theta = \operatorname{sign}(y^*)(|y^*| - \lambda)_+ \end{aligned}$$

where the last step follows because $\theta = y^* + \lambda < 0 \Rightarrow y^* < -\lambda < 0 \Rightarrow \operatorname{sign}(y^*) = -1$ and also $|y^*| = -y^*$. □

What is the meaning of the formula (5)?

The function $f_\lambda(u) = \operatorname{sign}(u)(|u| - \lambda)_+$ is called the *soft-thresholding* function. It looks like this:



Therefore, all the coefficients that are smaller than λ are set to zero. All other coefficients are reduced by λ in absolute value.

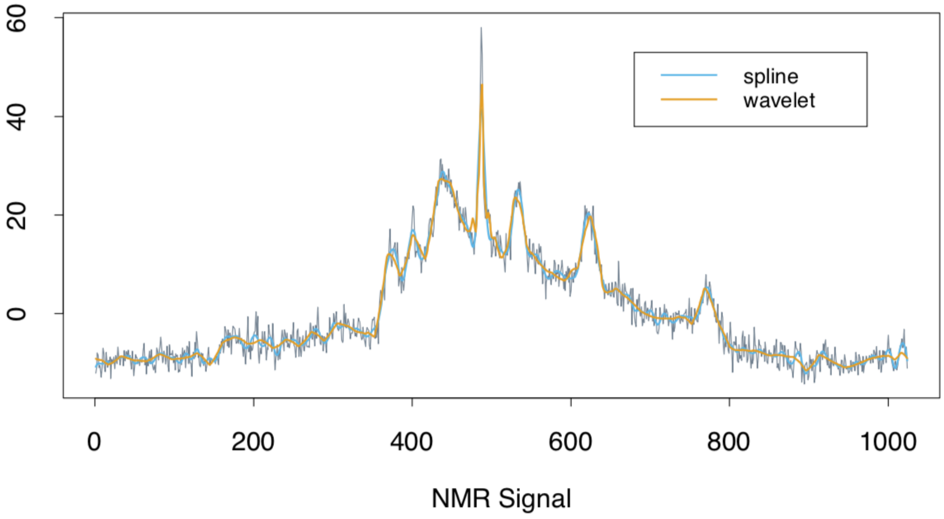
What is the good choice for λ? The answer depends on the amount of noise in the signal. Suppose $\mathbf{y} = \mathbf{s} + \mathbf{z}$, where $\mathbf{s}$ is the true signal and $\mathbf{z} = [z_1, \ldots, z_N]^\mathsf{T}$ is the noise vector. Assume $z_i \sim \mathcal{N}(0, \sigma^2)$ and independent over i.

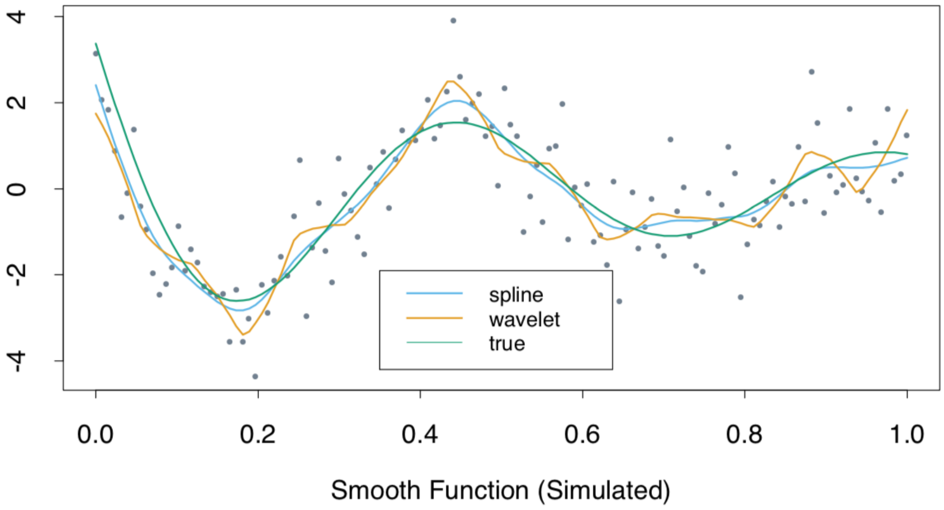
Now

$$\mathbf{y}^* = \mathbf{W}^\mathsf{T}\mathbf{y} = \mathbf{W}^\mathsf{T}\mathbf{s} + \mathbf{W}^\mathsf{T}\mathbf{z}.$$

Because $\mathbf{W}$ is an orthogonal matrix, the elements z_i^* of $\mathbf{z}^* = \mathbf{W}\mathbf{z} = [z_1^*, \ldots, z_N^*]^\mathsf{T}$ are again independent over i and distributed as $z_i^* \sim \mathcal{N}(0, \sigma^2)$.

Let $z^* = \max_i z_i^*$. It is not difficult to calculate that $\mathbb{E}z^* \approx \sigma\sqrt{2\log N}$. Hence, if we set all coefficients that are smaller than $\sigma\sqrt{2\log N}$ to zero, we are likely to remove all the noise from the signal. This means that a principled choice is $\lambda := \sigma\sqrt{2\log N}$. With this choice we obtain the denoising result as follows (orange line, top plot):





Let's compare the spline and the wavelet smoothing for two different signals as in the figure above. In NMR example, splines introduce many unnecessary features. Wavelets fit nicely localized spikes. In the second plot, the true function is smooth and the noise is high. The wavelet fit has left unnecessary wiggles - a price it pays in variance for additional adaptivity.