APPLICATION: WAVELET DENOISING

The Discrete Wavelet Transform

The discrete wavelet transform (DWT) is a linear map

$$W^T : \mathbb{R}^N \longrightarrow \mathbb{R}^N, \quad \underline{x} \longmapsto \underline{y} = W^T \underline{x}$$

satisfying certain special properties.

Although a thorough and rigorous definition and treatment of the DWT is beyond our scope, we can understand it through analogy with the discrete Fourier transform (DFT).

DWT vs. DFT

- Both can be represented by orthogonal matrices
- Both have efficient implementations

  DFT : $O(N \log N)$

  DWT : $O(N)$

- Both are discretizations of continuous transforms

continuous and infinite	$\longrightarrow$	discrete and infinite	$\longrightarrow$	discrete and finite
CFT		DTFT		DFT
CWT		DTWT		DWT

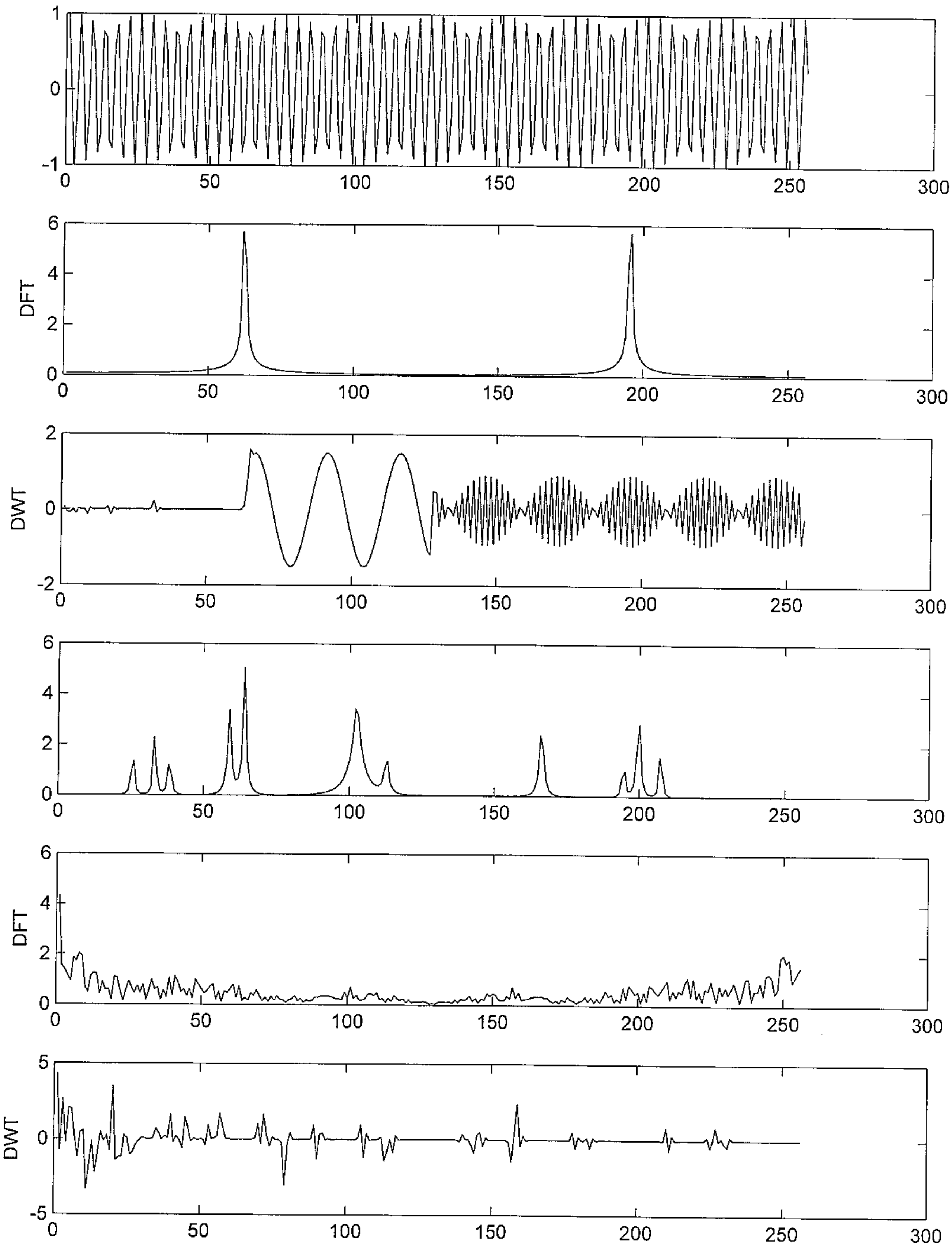
- Both are "change of basis" operators that compute the expansion coefficients of the signal in a different basis

DFT $\Longrightarrow$ Fourier basis

sparse representation of <u>sinusoidal</u> signals

DWT $\Longrightarrow$ Wavelet basis (colums of W)

sparse representation of <u>piecewise polynomial</u> signals



DFT
DWT
DFT
DWT

Haar Wavelet Transform

While there's only one DFT, there are in fact many different DWTs. The simplest is the <u>Haar</u> wavelet transform.

Consider a signal of length $N = 2^L$.

$$\underline{x} = [x(1) \quad x(2) \quad \cdots \quad x(N)]^T.$$

Define

$$\underline{y}_1 = \left[c_1(1) \ c_1(2) \cdots c_1\left(\tfrac{N}{2}\right) \ \middle| \ d_1(1) \cdots d_1\left(\tfrac{N}{2}\right) \right]^T$$

where

$$c_1(1) = \frac{x(1) + x(2)}{\sqrt{2}} \qquad d_1(1) = \frac{x(1) - x(2)}{\sqrt{2}}$$

$$c_1(2) = \frac{x(3) + x(4)}{\sqrt{2}} \qquad d_1(2) = \frac{x(3) - x(4)}{\sqrt{2}}$$

$$\vdots$$

Observe

- This transformation is invertible: we can recover $\underline{x}$ from $\underline{y}_1$
- The transformation is an orthogonal linear map

$$
\begin{bmatrix} c_1(1) \\ c_1(2) \\ c_1(3) \\ \vdots \\ d_1(1) \\ d_1(2) \\ d_1(3) \\ \vdots \end{bmatrix}
= \frac{1}{\sqrt{2}}
\begin{bmatrix}
1 & 1 & & & & & & \\
 & & 1 & 1 & & & & \\
 & & & & 1 & 1 & & \\
 & & & & & & \ddots & \\
1 & -1 & & & & & & \\
 & & 1 & -1 & & & & \\
 & & & & 1 & -1 & & \\
 & & & & & & & \ddots
\end{bmatrix}
\begin{bmatrix} x(1) \\ x(2) \\ \vdots \\ \\ \\ x(N) \end{bmatrix}
= W^T \underline{x}
$$

- The coefficients $c_1(n)$ are local <u>averages</u> and represent <u>coarse</u> information about the signal
- The coefficients $d_1(n)$ are local <u>differences</u> and represent <u>detailed</u> information

This operation is called the <u>level-1 Haar wavelet transform</u>. The idea behind the general Haar wavelet transform is to $\boxed{\text{recursively}}$ apply this operation to the $\boxed{\text{coarse}}$ coefficients.

$$\underline{y}_1 = \left[c_1(1) \quad \cdots \quad c_1\left(\tfrac{N}{2}\right) \;\middle|\; d_1(1) \quad \cdots \quad d_1\left(\tfrac{N}{2}\right) \right]^T$$

$$\underline{y}_2 = \left[c_2(1) \quad \cdots \quad c_2\left(\tfrac{N}{4}\right) \;\middle|\; d_2(1) \cdots d_2\left(\tfrac{N}{4}\right) \;\middle|\; d_1(1) \quad \cdots \quad d_1\left(\tfrac{N}{2}\right) \right]^T$$

$$\underline{y}_3 = \left[c_3(1) \cdots c_3\left(\tfrac{N}{8}\right) \middle| d_3(1) \cdots d_3\left(\tfrac{N}{8}\right) \middle| d_2(1) \cdots d_2\left(\tfrac{N}{4}\right) \middle| d_1(1) \cdots d_1\left(\tfrac{N}{2}\right) \right]^T$$

$$\vdots$$

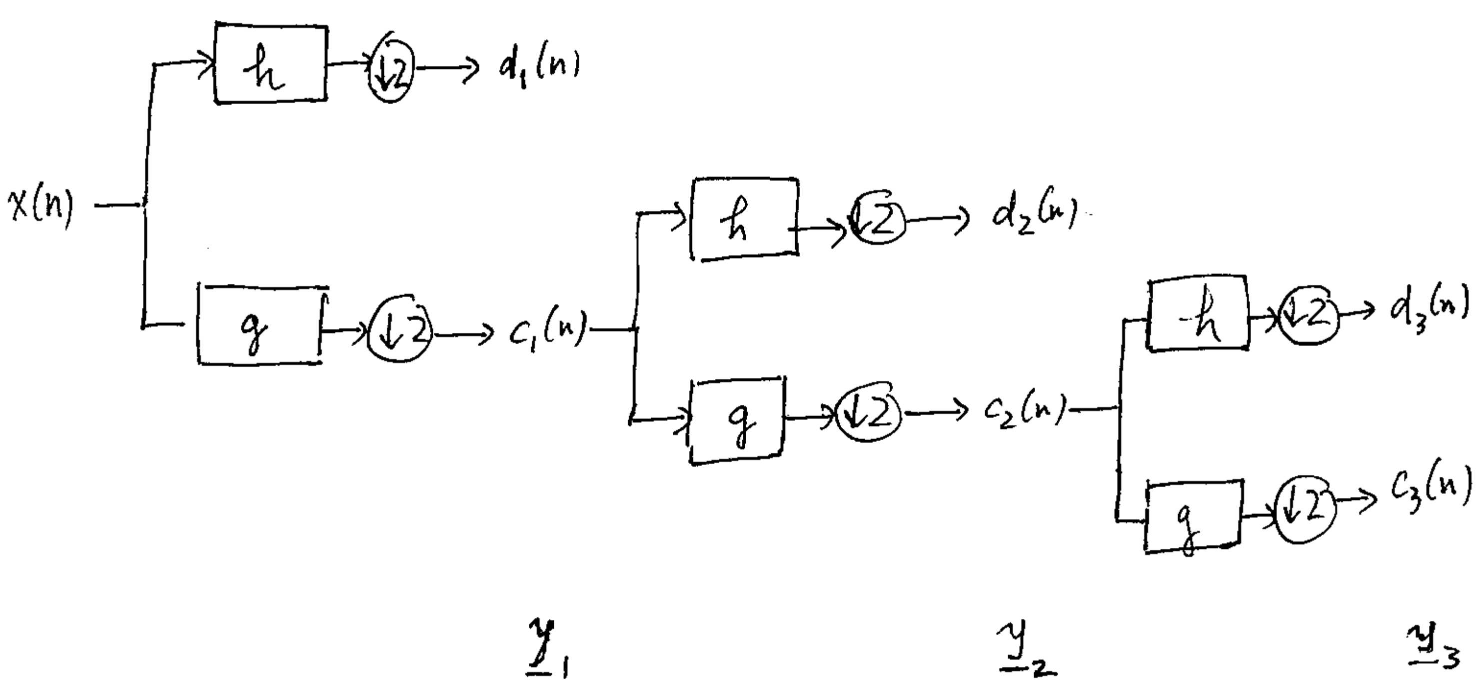
$$\underline{y}_L = \left[c_L(1) \middle| d_L(1) \middle| d_{L-1}(1) \quad d_{L-1}(2) \middle| d_{L-2}(1) \cdots d_{L-2}(4) \middle| \cdots \right]^T$$

We call $\underline{y}_l$ the <u>level-l Haar wavelet transform</u>

<u>Filter bank implementation</u>:

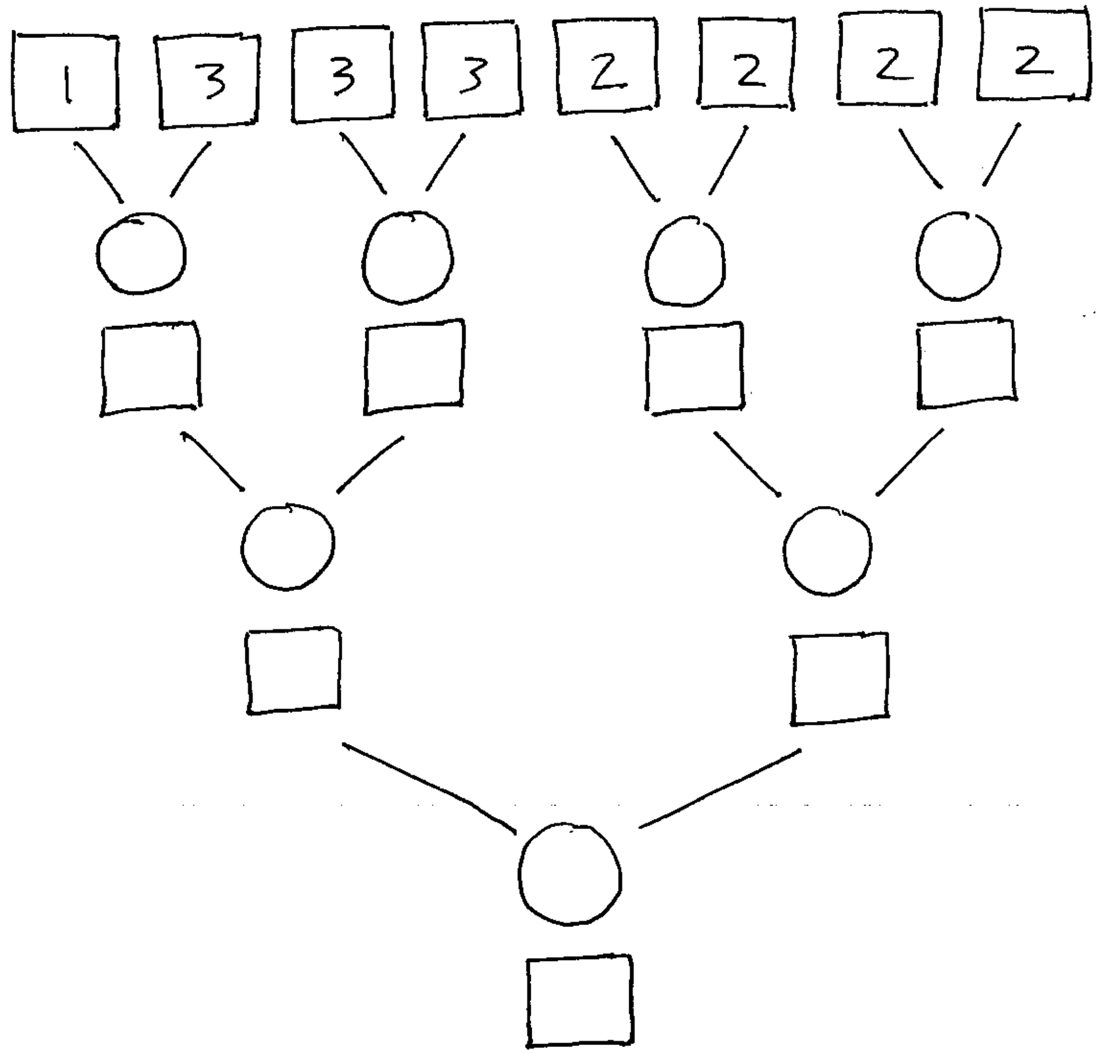
$$h = \left[\tfrac{1}{\sqrt{2}} \quad -\tfrac{1}{\sqrt{2}} \right] \quad \text{(high pass)}$$

$$g = \left[\tfrac{1}{\sqrt{2}} \quad \tfrac{1}{\sqrt{2}} \right] \quad \text{(low pass)}$$



Example

$\underline{x} =$



$$\underline{y}_1 = \left[\qquad\qquad\qquad\qquad \right]^T$$

$$\underline{y}_2 = \left[\qquad\qquad\qquad\qquad \right]^T$$

$$\underline{y}_3 = \left[\qquad\qquad\qquad\qquad \right]^T$$

Important things to notice:

- the detail coefficients are zero where the signal is constant. In particular, if $x(n)$ is constant on the interval

$$\left[k \cdot 2^{\ell}+1, k \cdot 2^{\ell}+2, \ldots, (k+1) 2^{\ell}\right], \quad \text{then } d_{\ell}(k)=0$$

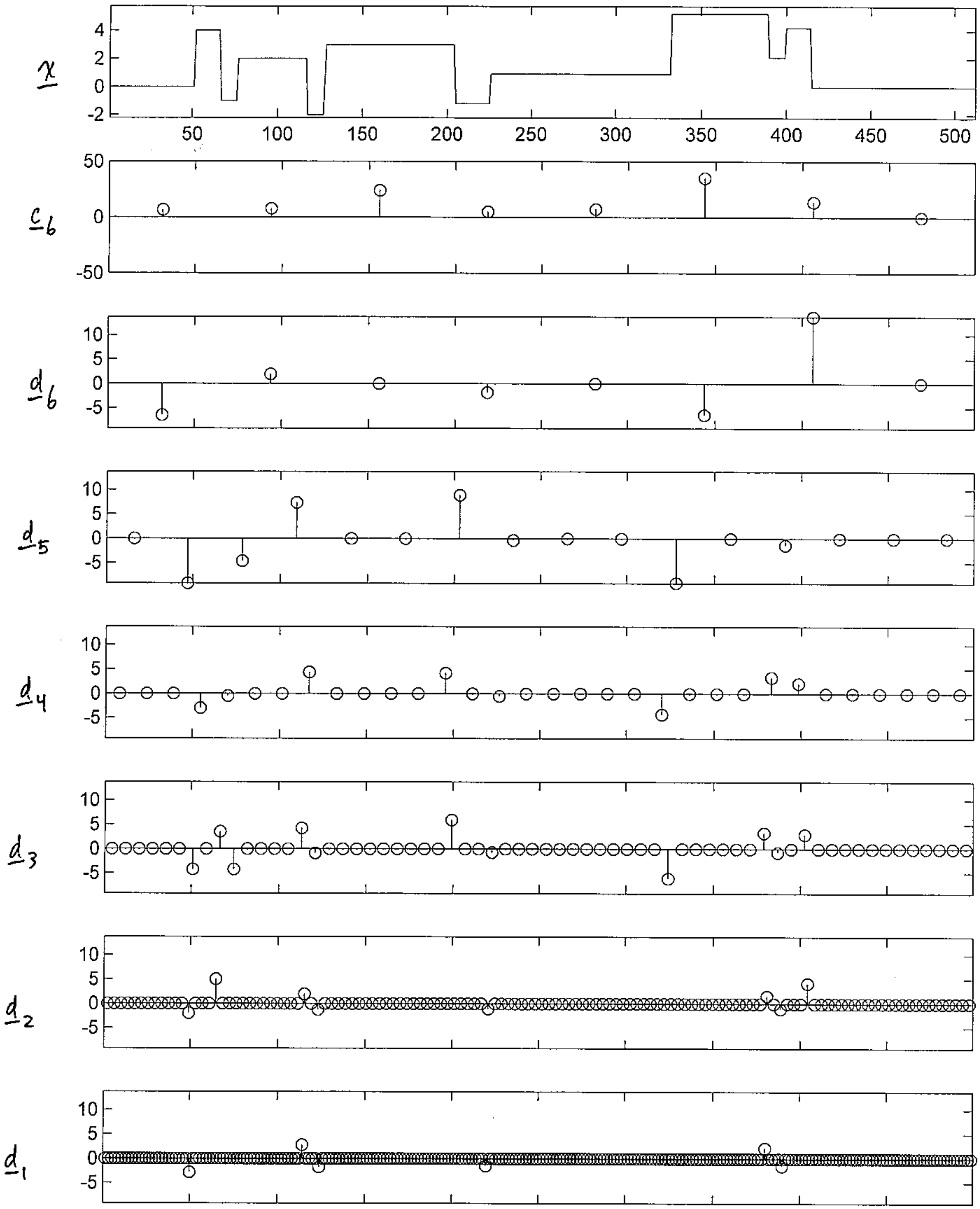
$\Longrightarrow$ $\boxed{\text{sparsity}}$

- the detail coefficients have a natural hierarchical (or tree-structured) arrangement; we can say that $d_\ell(k)$ is the <u>parent</u> of $d_{\ell-1}(2k-1)$ and $d_{\ell-1}(2k)$, who are its <u>children</u>

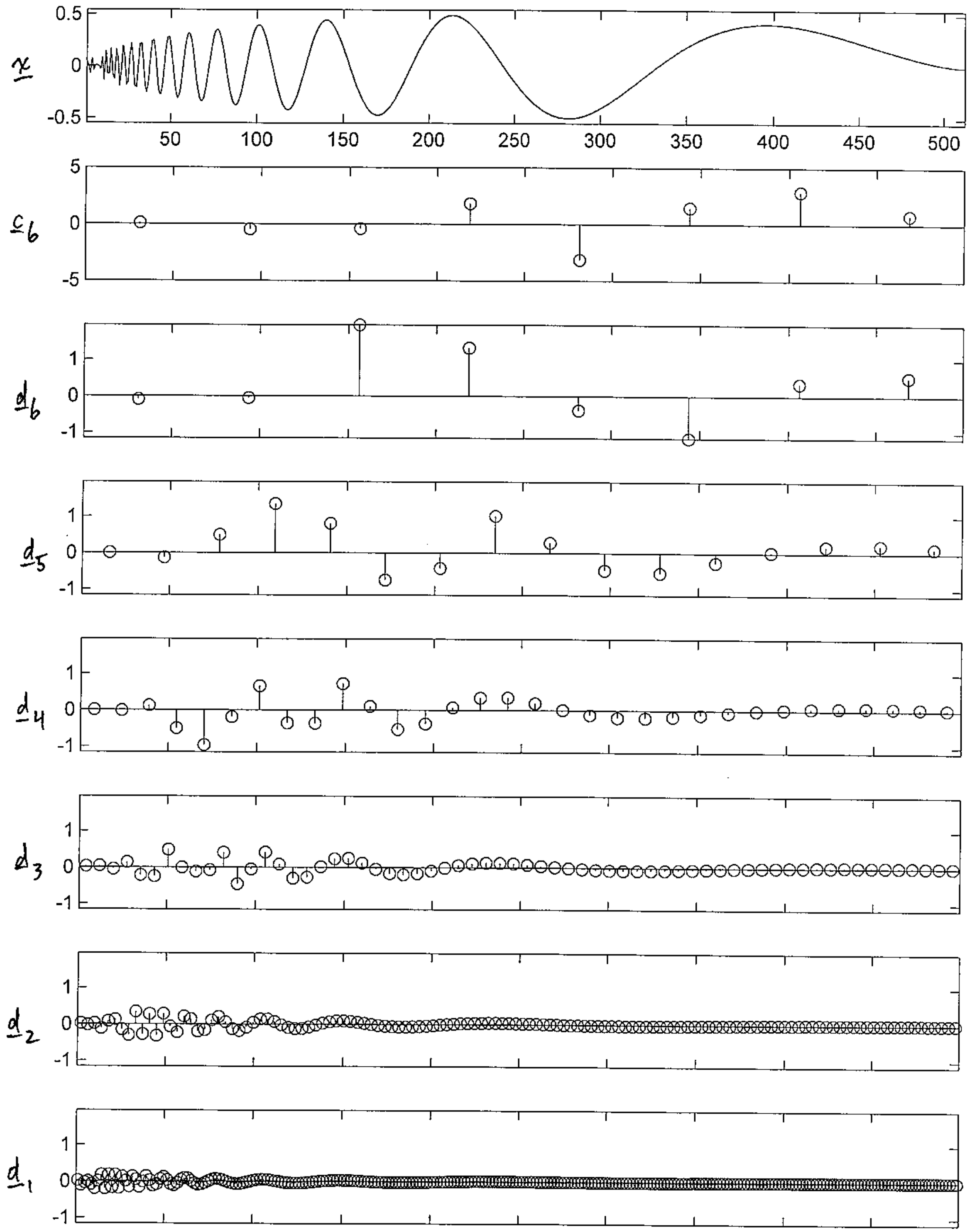
- $c_\ell(n)$ is a low resolution approximation to $x(n)$; it is the result of averaging and downsampling $x(n)$ ℓ times

- different levels capture different <u>resolutions</u> of detail:

$$\begin{array}{ccc}
\{d_1(k)\}_{k=1}^{2^{L-1}} & & \text{high frequency detail} \\
\downarrow & \Longleftrightarrow & \downarrow \\
\{d_{L-1}(k)\}_{k=1}^{2} & & \\
\{d_L(k)\}_{k=1}^{1} & & \text{low frequency detail}
\end{array}$$



$\Longrightarrow$ sparse representation



x
c6
d6
d5
d4
d3
d2
d1
0.5
0
-0.5
50
100
150
200
250
300
350
400
450
500
5
-5
1
-1

Other Wavelet Transforms

The Haar wavelet transform can be generalized by using different high-pass and low-pass filters h & g. These filters must satisfy certain properties for the resulting transform to be orthogonal (and qualify as a DWT).

The most important generalization are the Daubechies DWT. They are based on certain filters h_p, g_p such that

- the length of h_p and g_p is $2p$.
- if a signal behaves locally like a $(p-1)^{th}$ order polynomial, the corresponding detail coefficients are zero

Examples

$p = 1 \Rightarrow$ Haar

$$p = 2 \Rightarrow g = [.4830 \quad .8365 \quad .2241 \quad -.1294]$$
$$h = [.1294 \quad .2241 \quad -.8365 \quad .4830]$$

Wavelet Denoising

Suppose we measure a noisy signal

$$\underline{x} = \underline{s} + \underline{v} \qquad (*)$$

and assume

- $\underline{s} = [s_1 \; \dots \; s_N]^T$ has a sparse representation in a certain wavelet basis, e.g., $\underline{s}$ is piecewise constant / Haar basis
- $\underline{v} \sim N(\underline{0}, \sigma^2 I)$

Now take the wavelet transform of $(*)$:

$$\underline{y} = \underline{\theta} + \underline{z}$$

$$\boxed{\begin{aligned} \underline{y} &= W^T \underline{x} \\ \underline{\theta} &= W^T \underline{s} \\ \underline{z} &= W^T \underline{v} \end{aligned}}$$

Then we know

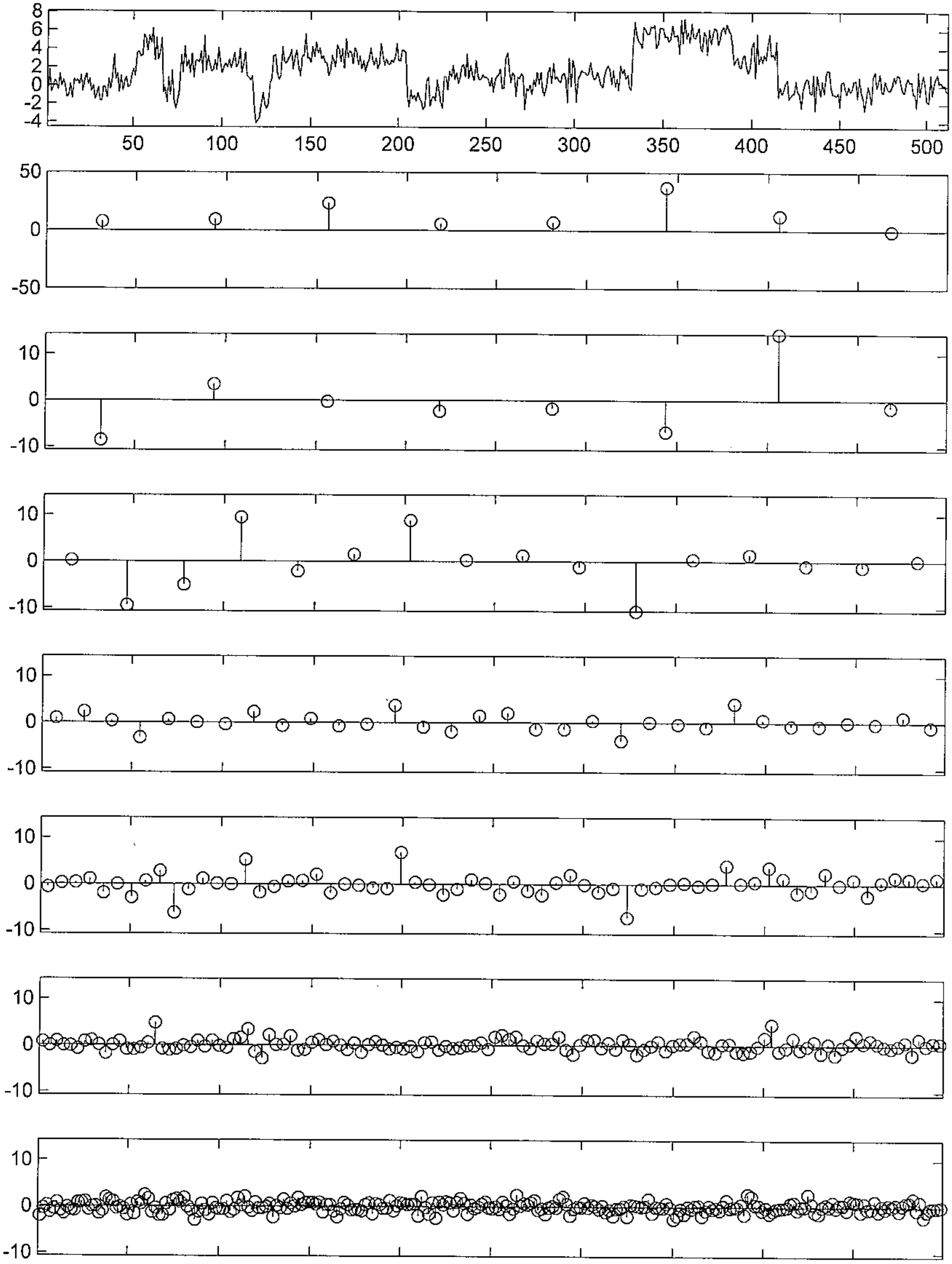
- most elements in

  $$\underline{\theta} = [\theta_1, \dots, \theta_N]^T$$

  are zero or very close to zero

- $\underline{z} \sim N(\underline{0}, \sigma^2 I)$

  because $W^T W = I$


8
6
4
2
0
-2
-4
50
100
150
200
250
300
350
400
450
500
50
0
-50
10
0
-10
10
0
-10
10
0
-10
10
0
-10
10
0
-10
10
0
-10

Since W is orthogonal, the estimation problem amounts to recovery of a signal in iid Gaussian noise, whether we treat the problem in the "time" domain or the "wavelet" domain

From a Bayesian perspective, we should tackle the problem in the domain for which it is easiest to specify a prior.

On one hand, the fact that $\underline{\theta}$ is sparse suggest a subspace model. Unfortunately we don't know a priori <u>which</u> detail coefficients will be zero.

<u>Q</u>: How can we take advantage of the prior knowledge that $\underline{\theta}$ is sparse? What statistical model captures this information?

<u>A</u>: One solution is to employ a ____________

Mixture Modeling

View the detail coefficients $\theta_2, \ldots, \theta_N$ as realizations of a single random variable θ.

We know

- Most θ_i are small (sparsity assumption)
- Some θ_i are large (W is orthogonal, so energy must be preserved)
- θ is zero mean, since θ_i are local differences
- θ_i are "approximately" independent, since the θ_i are local differences

This suggests the following prior:

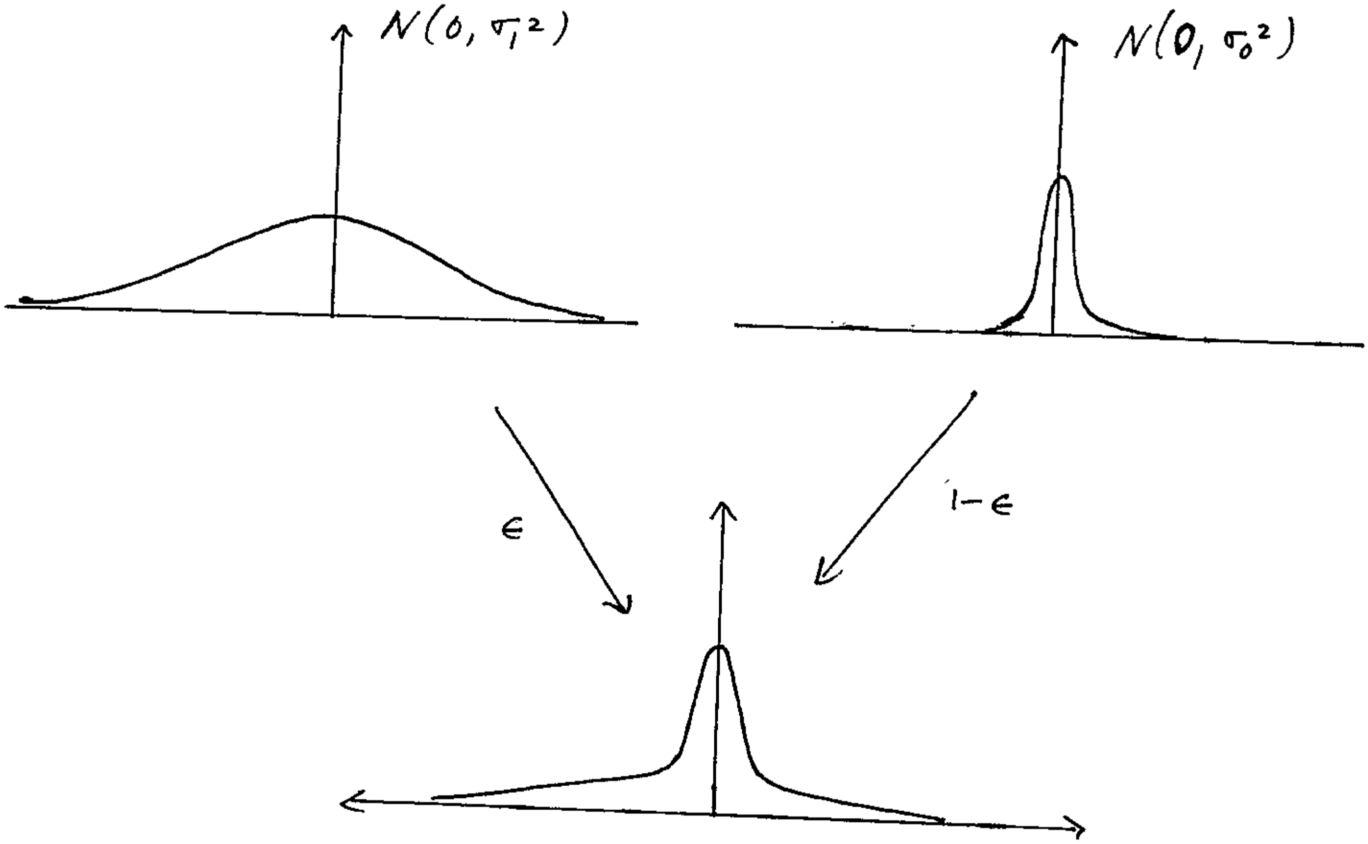
$$\theta_i \overset{iid}{\sim} \epsilon N(0, \sigma_1^2) + (1-\epsilon) N(0, \sigma_0^2).$$

where $\sigma_0^2 \ll \sigma_1^2$ and

ϵ = proportion of "significant" coefficients

σ_1^2 = variance of significant "

σ_0^2 = " " insignificant "



According to this prior, a detail coefficient θ is generated according to the following algorithm:

1. Flip an "ϵ-coin"
2. If heads,

$$\theta \sim N(0, \sigma_1^2)$$

Else

$$\theta \sim N(0, \sigma_0^2)$$

The Big Picture

- We observe $\underline{x} = \underline{s} + \underline{v}$, $\underline{v} \sim N(0, \sigma^2 I)$
- Compute $\underline{y} = \underline{\theta} + \underline{w}$ by taking wavelet transform
- View

$$y_i = \theta_i + w_i \quad , \quad w_i \sim N(0, \sigma^2)$$

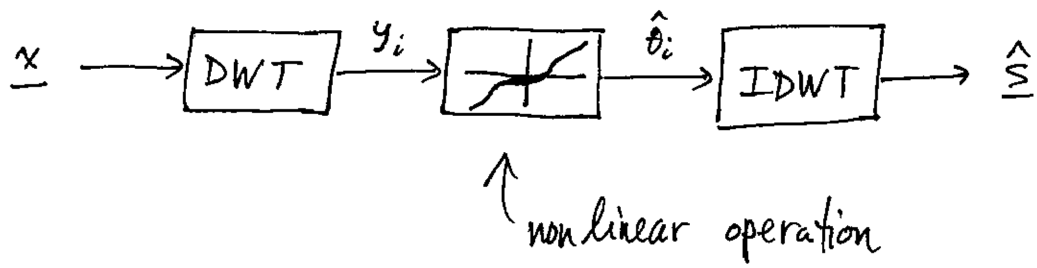
as independent estimation problems.

- Leave "coarse" coefficients unprocessed: noise will be "averaged out" since these are local averages
- Assume a mixture model prior on the detail coefficients and estimate

$$\hat{\theta}_i = E[\theta_i \,/\, y_i]$$

- Apply inverse wavelet transform to obtain

$$\underline{\hat{s}} = W \, \underline{\hat{\theta}}$$



Setting Parameters

Before this method is practical, we need ways to to set σ^2, ϵ, σ_1^2, and σ_0^2.

Donoho and Johnstone suggested the estimate

$$\hat{\sigma} = \frac{\underset{i > N/2}{\mathrm{MAD}}(y_i)}{.6745},$$

which takes the "median absolute deviation" of the wavelet coefficients at the "finest" level of detail, and .6745 makes the estimate unbiased <u>if</u> all of the θ_i are in fact 0.

The mixture model parameters ϵ, σ_1^2, σ_0^2 may be estimated via maximum likelihood:

$$(\hat{\epsilon}, \hat{\sigma}_1^2, \hat{\sigma}_0^2) = \underset{(\epsilon, \sigma_1^2, \sigma_0^2)}{\arg\max} \; \ell(\epsilon, \sigma_1^2, \sigma_0^2 ; \underline{y})$$

Exercise | Determine a formula for the likelihood of $\epsilon, \sigma_1^2, \sigma_0^2$ given the detail coefficients $\underline{y}_{detail} = [y_2, \ldots, y_N]^T$ (assuming a max-level wavelet transform)

Solution | Denote

$$\phi(y ; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}.$$

Since

$$y_i = \theta_i + w_i$$

$\boxed{\text{independent}}$

where

$$f(\theta_i) = \epsilon\,\phi(\theta_i ; 0, \sigma_1^2) + (1-\epsilon)\,\phi(\theta_i ; 0, \sigma_0^2)$$

$$f(w_i) = \phi(w_i ; 0, \sigma^2)$$

it follows that

$$f(y_i) = f(\theta_i) \star f(w_i)$$

$$= \epsilon\,\phi(y_i ; 0, \sigma^2+\sigma_1^2) + (1-\epsilon)\,\phi(y_i ; 0, \sigma^2+\sigma_0^2)$$

Hence the likelihood of $\epsilon, \sigma_1^2, \sigma_0^2$ is

$$l(\epsilon, \sigma_1^2, \sigma_0^2 ; \underline{y}_{detail}) = \prod_{i=2}^{N} f(y_i ; \epsilon, \sigma_1^2, \sigma_0^2)$$

$$= \prod_{i=2}^{N}\left[\epsilon\,\phi(y_i ; 0, \sigma^2+\sigma_1^2) + (1-\epsilon)\,\phi(y_i ; 0, \sigma^2+\sigma_0^2)\right]$$

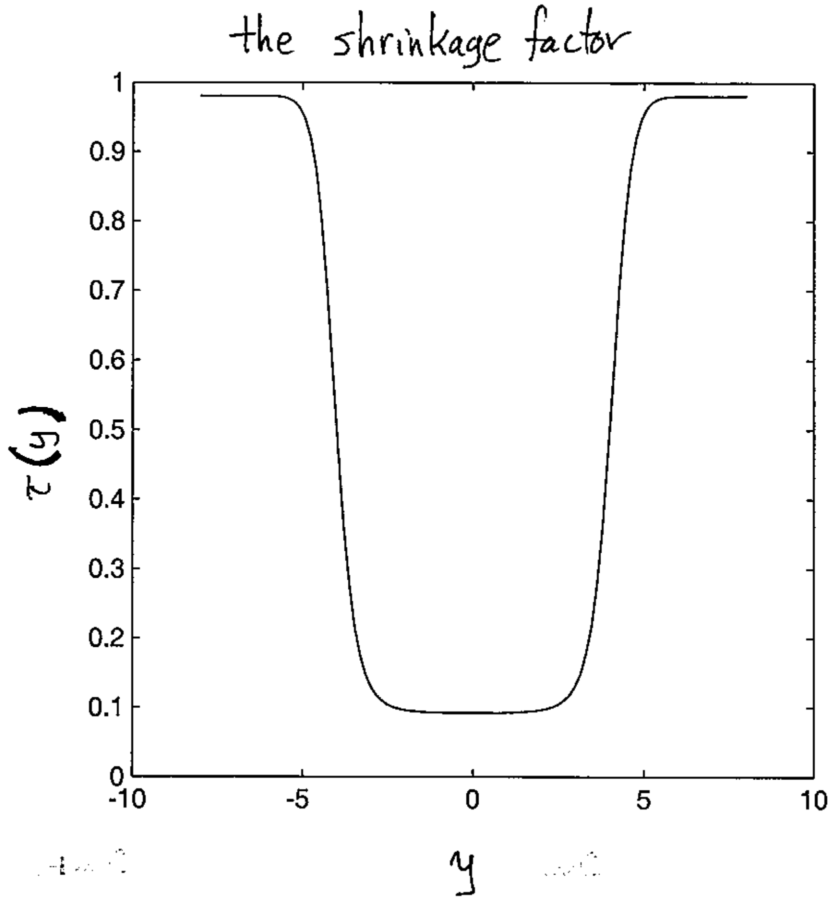
Typically one uses an iterative algorithm such as an EM algorithm to fit mixture models. However, because there are only 3 unknowns, we could also maximize the likelihood by more generic procedures.

Note that we are using the data to determine our prior - hence we are not strictly adhering to the Bayesian philosophy. This kind of procedure is called an <u>empirical Bayesian</u> method.

You have shown on the homework that

$$\hat{\theta} = E[\theta \mid y] = \tau(y) \cdot y$$

where $0 < \tau(y) < 1$ is called the <u>shrinkage</u> factor.

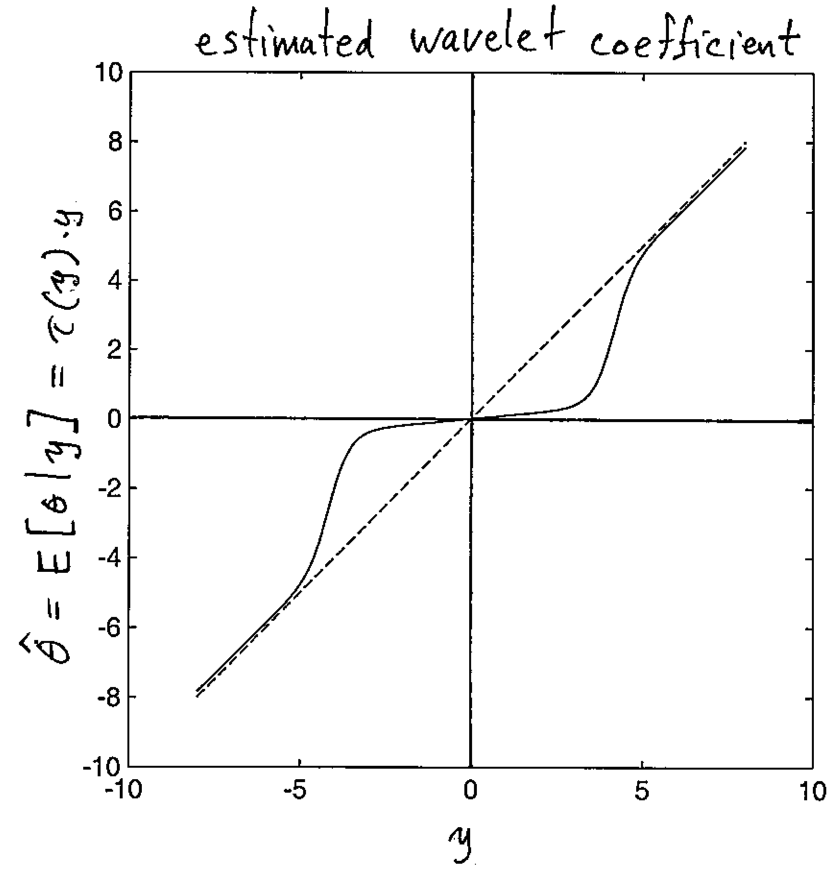


$\epsilon = 0.05$

$\sigma^2 = 1$

$\sigma_0^2 = .1$

$\sigma_1^2 = 50$



The effect of the shrinkage factor is that

- small coefficients are set nearly to zero
- large coefficients are virtually unaltered.

This property is consistent with our understanding

- small coefficients are mostly noise
- large coefficients contain actual signal

Extensions

- Image denoising
- More sophisticated priors

Summary

$$\left.\begin{array}{l}\text{Gaussian noise}\\ \text{Mixture of Gaussians prior}\end{array}\right\} \Rightarrow \text{Wavelet shrinkage}$$