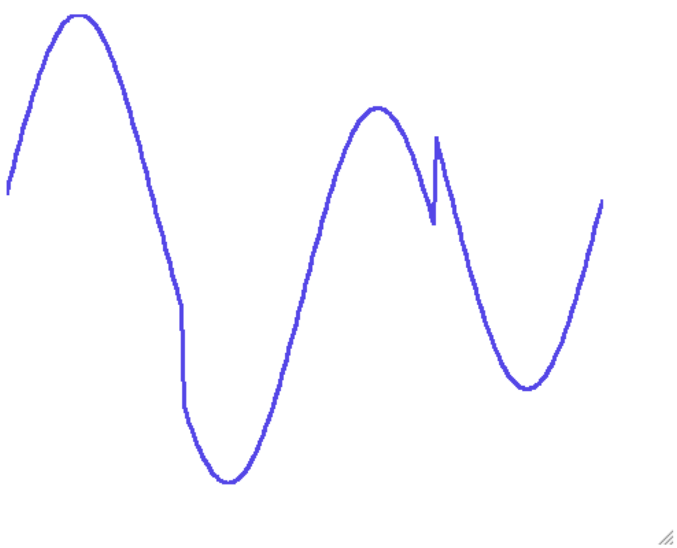
SPARSITY:
CORRECTING ERRORS IN DATA

Rebecca Willett

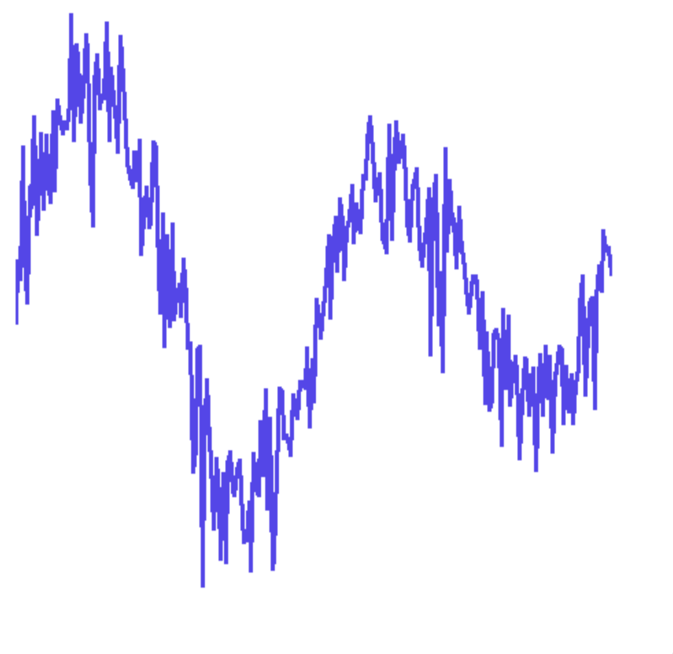
DUKE ENGINEERING
EDMUND T. PRATT, JR.
SCHOOL OF ENGINEERING

Noisy observations

In many settings, we do not directly observe the signal of interest; rather, our measurements are corrupted by “noise” and other errors.



What we want



What we get

NOISY OBSERVATIONS

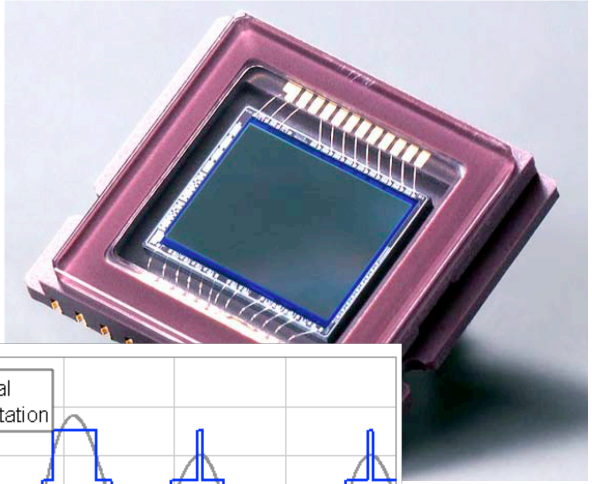
In many settings, we do not directly observe the signal of interest; rather, our measurements are corrupted by “noise” and other errors.

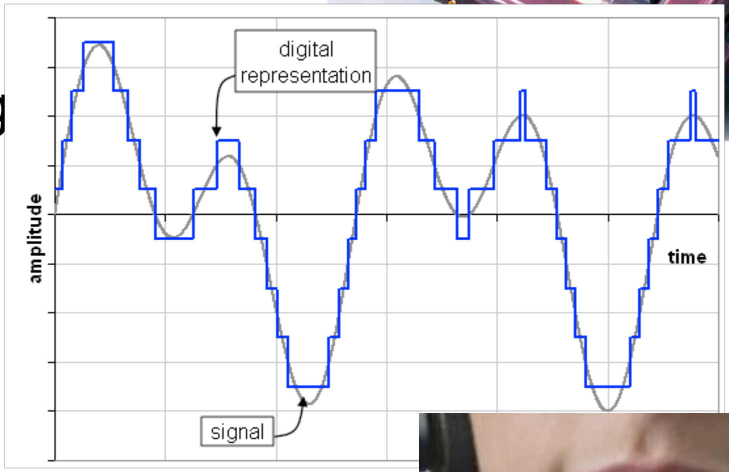
What we want

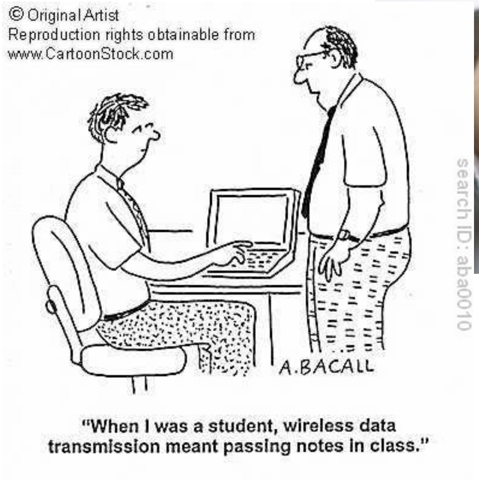
What we get

ORIGINS OF NOISE AND ERRORS

- The signal may be weak relative to the sensitivity of the sensor
- The true field being measured (e.g voltage or light intensity) gets quantized for storage on a digital system (e.g. computer)
- The field being sensed may be contaminated by the ambient environment (e.g. a microphone picks up not just a speaker, but also a little of the audience noise)
- Data may have been lost during storage and transmission







DENOISING

We model a noisy signal as

$$y_i = f_i + n_i$$

where f_i is the i^{th} element of the true signal, y_i is the corresponding observation, and n_i is the noise or error in that measurement.

Our goal is to estimate f from y without knowing n.

Without any assumptions about the structure of f and n, this task would be impossible. Thus we typically make two key assumptions:

- The noise has some known properties, such as
  - is stochastic with a known distribution
  - is bounded, so $\|n\|_2^2 < \epsilon$ where ϵ is known.
- The signal has some known properties, such as
  - is smooth or piecewise smooth
  - is sparse in some basis.

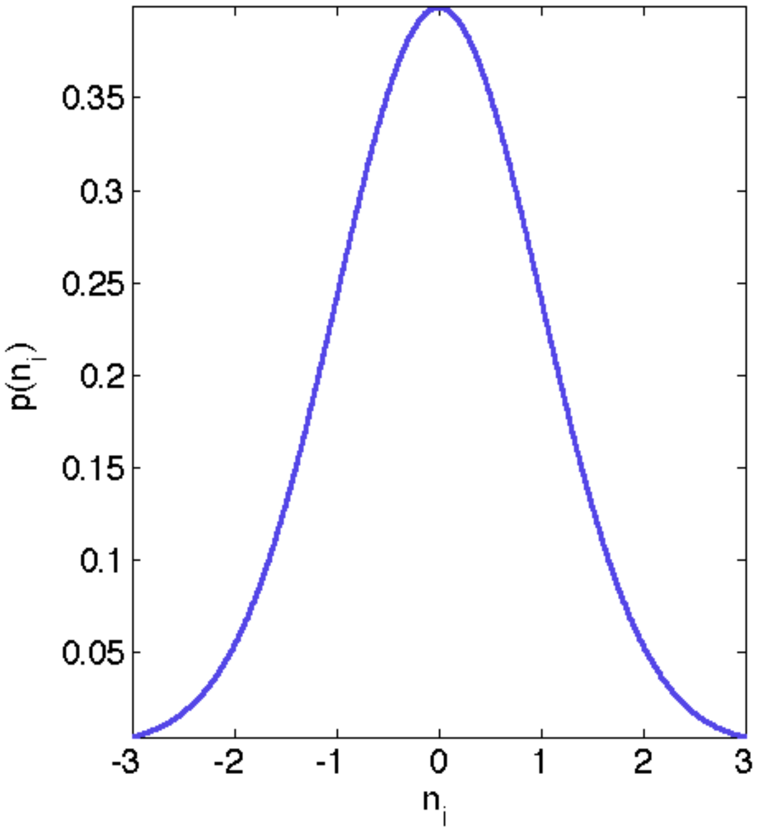
Gaussian Noise

We can often assume each noise element n_i is drawn independently from a Gaussian distribution, so that the probability distribution function underlying n_i is

$$p(n_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-n_i^2/2\sigma^2};$$

we write

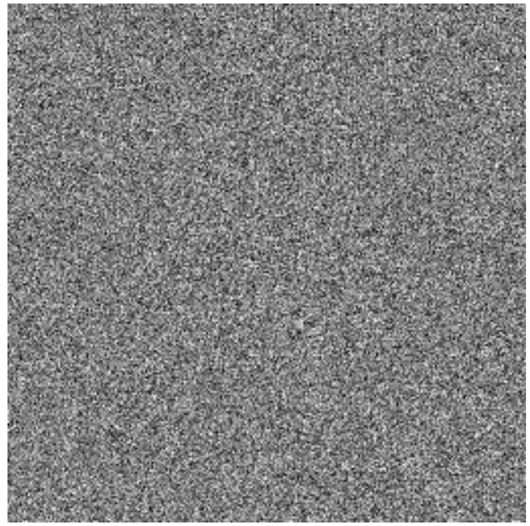
$$n_i \sim \mathcal{N}(0, \sigma^2).$$



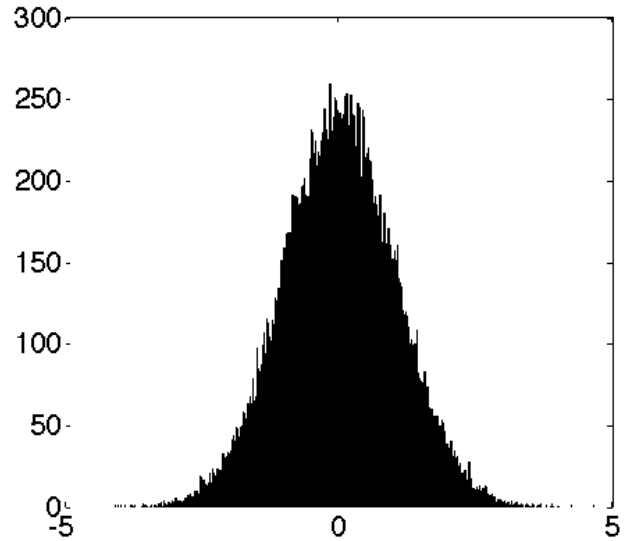
We typically assume that the n_i's are uncorrelated with the f_i's and independent of i, the sample index.

GAUSSIAN NOISE

f



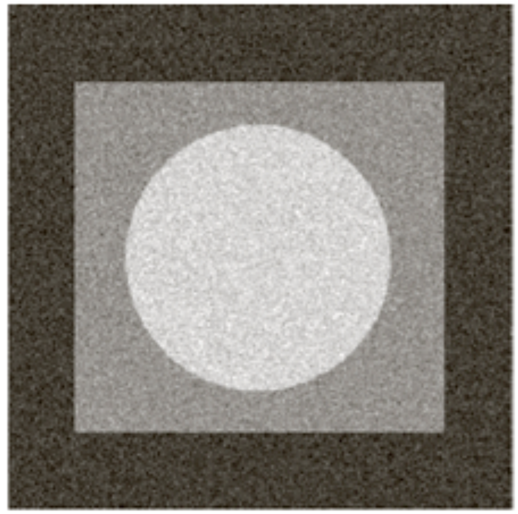
y



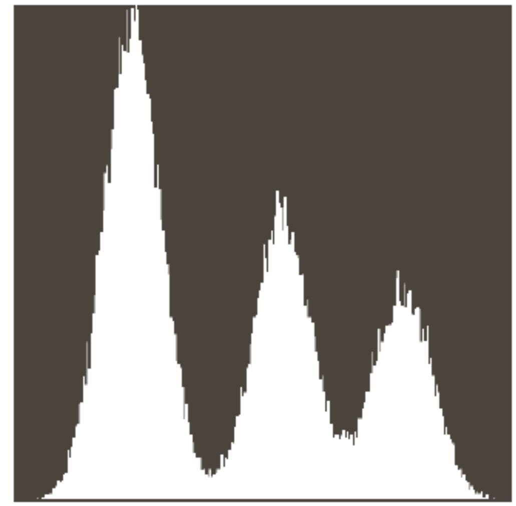
Histogram of y

Gaussian noise

f



y



Histogram of y

$$p(n_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(n_i-0)^2/2\sigma^2}; \quad n_i \sim \mathcal{N}(0, \sigma^2),$$

implies

$$p(y_i|f_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i-f_i)^2/2\sigma^2}; \quad y_i|f_i \sim \mathcal{N}(f_i, \sigma^2).$$

Multivariate Normal (Gaussian)

We can also consider the joint distribution of all the n_i's in a noisy signal of length N:

$$\begin{aligned} p(n|\mu,\Sigma) &\triangleq (2\pi)^{-N/2}|\Sigma|^{-1/2}e^{-\frac{1}{2}(n-\mu)^T\Sigma^{-1}(n-\mu)} \\ n &\sim \mathcal{N}(\mu,\Sigma) \\ \mu &\triangleq \mathbb{E}[n] \\ \Sigma &\triangleq \mathbb{E}\left[(n-\mu)(n-\mu)^T\right] \\ \Sigma_{i,j} &= \mathbb{E}\left[(n_i-\mu_i)(n_j-\mu_j)^T\right]. \end{aligned}$$

When $\Sigma = \sigma^2 I$, then all the elements of n are uncorrelated and

$$p(n|\mu,\Sigma) = \prod_{i=1}^{N} p(n_i|\mu_i,\sigma^2).$$

Linear Transformation of Gaussian Data

Suppose that we transform a multivariate normal vector n by applying a linear transformation (matrix) A:

$$m = An$$

(n, m random, A deterministic). Then

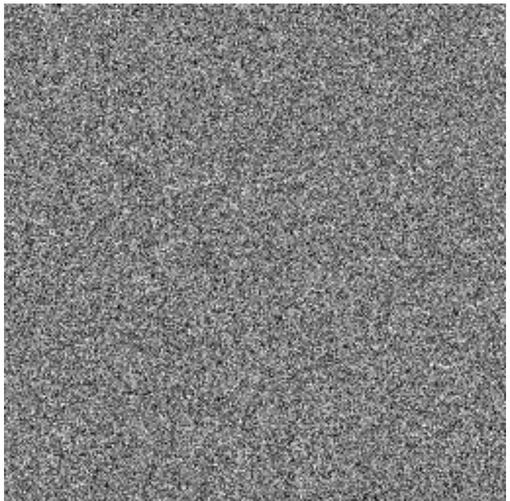
$$n \sim \mathcal{N}(\mu, \Sigma)$$

implies

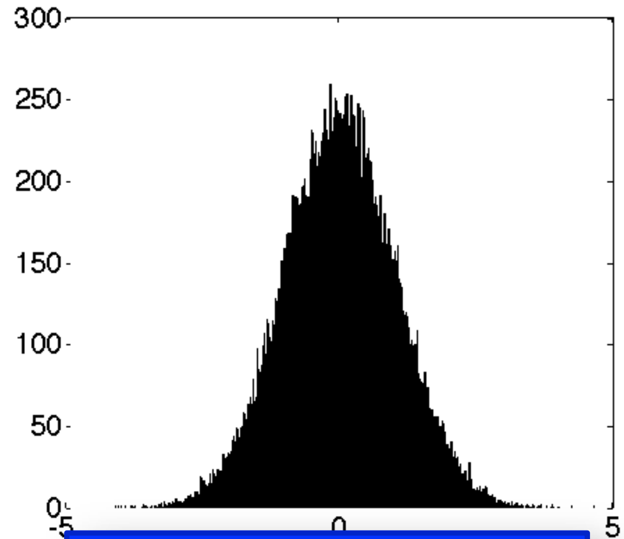
$$m \sim \mathcal{N}(A\mu, A\Sigma A^T).$$

Special case: When $\Sigma = \sigma^2 I$ and A is an orthonormal matrix (e.g. Fourier or wavelet transform), then m corresponds to independent noise in the transform coefficients.

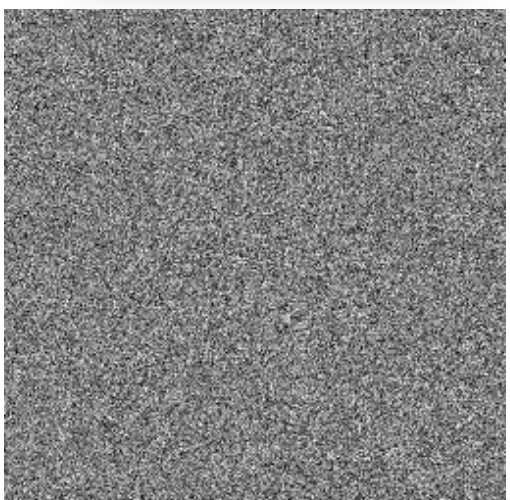
NOISE IN THE WAVELET DOMAIN



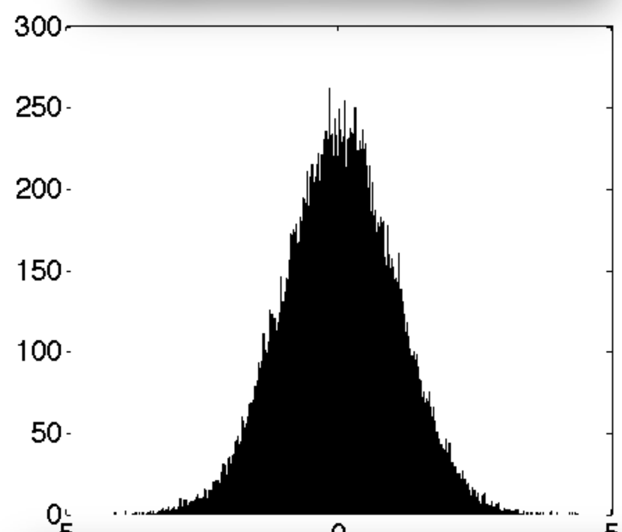
Signal noise



Signal histogram



Coefficient noise



Coefficient histogram

Statistical Estimation

Denoising is about developing a function of the data, $\widehat{f}(y)$, which, on average, will be close to f.

When the noise is stochastic, then $\widehat{f}(y)$ is also stochastic.

We can measure closeness via the "mean squared error":

$$\text{MSE} \triangleq \mathbb{E}\left[\|f - \widehat{f}(y)\|_2^2\right]$$

Bias-Variance Decomposition

$$\begin{aligned}
\text{MSE}\left(\widehat{f}\right) &= \mathbb{E}\left[\left\|f-\widehat{f}\right\|_2^2\right] \\
&= \mathbb{E}\left[\left\|\left(f-\mathbb{E}\left[\widehat{f}\right]\right)+\left(\mathbb{E}\left[\widehat{f}\right]-\widehat{f}\right)\right\|_2^2\right] \\
&= \underbrace{\left\|f-\mathbb{E}\left[\widehat{f}\right]\right\|_2^2}_{\text{Bias}^2\left(\widehat{f}\right)}+\underbrace{\mathbb{E}\left[\left\|\mathbb{E}\left[\widehat{f}\right]-\widehat{f}\right\|_2^2\right]}_{\text{Var}\left(\widehat{f}\right)}
\end{aligned}$$

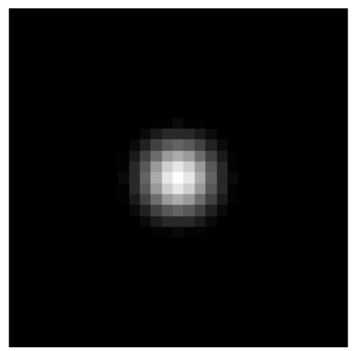
Hence, the MSE can be decomposed into two sources of error: bias and variance.

$\boxed{\text{Let's look at a simple example...}}$

IMAGE SMOOTHING

Consider removing noise by "smoothing" the image; i.e. convolve with a Gaussian blur.

What is the right blur radius?



vs.

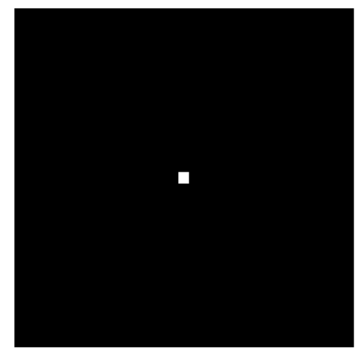


Image smoothing

$\mathbb{E}\left[\widehat{f}\right]$ $\widehat{f}$

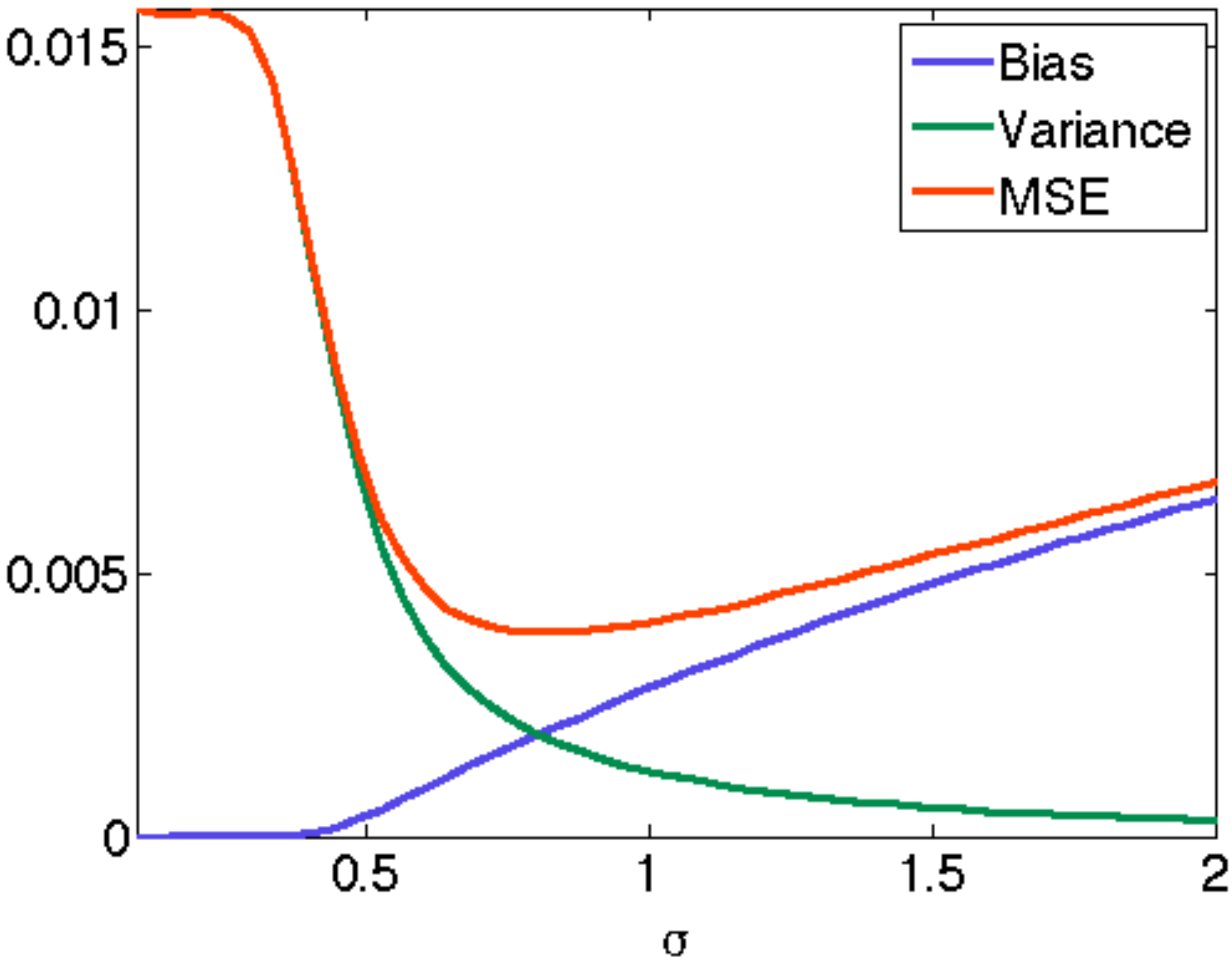
Zero bias, high variance

$\mathbb{E}\left[\widehat{f}\right]$ $\widehat{f}$

High bias, low variance

Low bias, medium variance

IMAGE SMOOTHING



Can sparsity give us low bias AND low variance?

Basis representation

Let us decompose our noisy signal

$$y = f + n$$

in an orthonormal basis $\{\psi_i\}_{i=1}^N$:

$$\boxed{\underbrace{\langle y, \psi_i \rangle}_{\zeta_i} = \underbrace{\langle f, \psi_i \rangle}_{\theta_i} + \underbrace{\langle n, \psi_i \rangle}_{\eta_i}}$$

So that

$$y = \sum_{i=1}^N \zeta_i \psi_i \qquad f = \sum_{i=1}^N \theta_i \psi_i \qquad n = \sum_{i=1}^N \eta_i \psi_i$$

What do we know about the distribution of the η_i's?

Coefficient Estimation

In the following discussion, we will estimate the signal f by estimating each coefficient $\theta_i = \langle f, \psi_i \rangle$ individually and computing the reconstruction

$$\boxed{\widehat{f} = \sum_{i=1}^{N} \widehat{\theta}_i \psi_i}$$

where $\widehat{\theta}_i = \widehat{\theta}(y)$ is our estimate of θ_i.

Ideal Coefficient Selection

Consider an estimator of the form

$$\widehat{f} = \sum_{i=1}^{N} \alpha_i \langle y, \psi_i \rangle \psi_i, \quad \alpha_i \in \mathbb{R};$$

i.e. $\widehat{\theta}_i = \alpha_i \langle y, \psi_i \rangle$. The MSE is

$$\mathbb{E}\left[\|f - \widehat{f}\|_2^2\right] = \mathbb{E}\left[\|\theta - \widehat{\theta}\|_2^2\right] = \sum_{i=1}^{N} \mathbb{E}\left[(\theta_i - \widehat{\theta}_i)^2\right].$$

Consider an estimator which selects the most important coefficients; i.e. restrict $\alpha_i \in \{0, 1\}$. Minimizing the MSE with respect to $\{\alpha_i\}_{i=1}^{N}$ yields the optimal coefficient selection

$$\alpha_i = \begin{cases} 1, & |\theta_i| \geq \sigma \\ 0, & |\theta_i| < \sigma \end{cases}$$

The problem is that θ is unknown, so this ideal coefficient attenuation is not practical.

Ideal coefficient selection

The MSE of the ideal coefficient selection estimator is

$$\begin{aligned}
\text{MSE}_s &= \sum_{i=1}^{N} \min(|\theta_i|^2, \sigma^2) \\
&= \sum_{i:|\theta_i|<\sigma} \min(|\theta_i|^2, \sigma^2) + \sum_{i:|\theta_i|\geq\sigma} \min(|\theta_i|^2, \sigma^2) \\
&= \sum_{i:|\theta_i|<\sigma} |\theta_i|^2 + \sum_{i:|\theta_i|\geq\sigma} \sigma^2
\end{aligned}$$

Let K be the number of coefficients satisfying $|\theta_i| \geq \sigma$, and recall our discussion of K-term approximations of the form

$$f_K = \sum_{i:|\theta_i|\geq\sigma} \theta_i \psi_i \qquad \text{so} \qquad f - f_K = \sum_{i:|\theta_i|<\sigma} \theta_i \psi_i.$$

MSE DECOMPOSITION

The MSE of the ideal selection rule is then

$$\text{MSE}_s = \mathbb{E}\left[\|f - \widehat{f}\|_2^2\right] = \underbrace{\|f - f_K\|_2^2}_{\substack{\text{Approximation error} \\ \approx \text{ bias}}} + \underbrace{K\sigma^2}_{\substack{\text{Estimation error} \\ \approx \text{ variance}}}.$$

This MSE is small if and only if both terms above are small – i.e. if K is small and f_K is a good approximation to f.

IDEAL COEFFICIENT SELECTION

Consider an estimator of the form

$$\widehat{f} = \sum_{i=1}^{N} \alpha_i \langle y, \psi_i \rangle \psi_i, \quad \alpha_i \in \mathbb{R};$$

i.e. $\widehat{\theta}_i = \alpha_i \langle y, \psi_i \rangle$. The MSE is

$$\mathbb{E}\left[\|f - \widehat{f}\|_2^2\right] = \mathbb{E}\left[\|\theta - \widehat{\theta}\|_2^2\right] = \sum_{i=1}^{N} \mathbb{E}\left[(\theta_i - \widehat{\theta}_i)^2\right].$$

Consider an estimator which selects the most important coefficients; i.e. restrict $\alpha_i \in \{0, 1\}$. Minimizing the MSE with respect to $\{\alpha_i\}_{i=1}^N$ yields the optimal coefficient selection

$$\alpha_i = \begin{cases} 1, & |\theta_i| \geq \sigma \\ 0, & |\theta_i| < \sigma \end{cases}$$

The problem is that θ is unknown, so this ideal coefficient attenuation is not practical.

Hard Thresholding

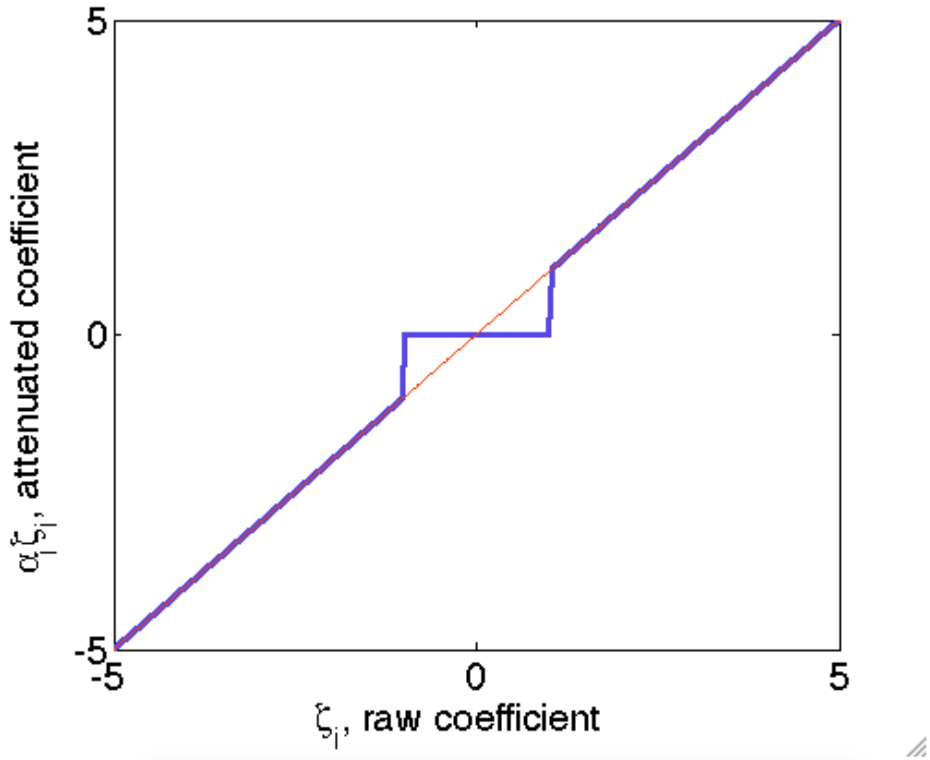
Ideal coefficient selection is also impractical. However, we can threshold the noisy coefficients to approximate the ideal coefficient selection. In particular, let

$$\widehat{f} = \sum_{i=1}^{N} \delta_T^{(H)}(\zeta_i)\psi_i$$

where $\delta_T^{(H)}$ is a "hard" threshold function

$$\delta_T^{(H)}(z) = \begin{cases} z, & |z| > T \\ 0, & |z| \leq T \end{cases}$$

and T is the threshold level.



Keep large coefficients, kill small coefficients

Sparsity regularization

This hard-threshold estimator is equivalent to performing the optimization

$$\widehat{\theta} = \arg\min_{\theta} \|\zeta - \theta\|_2^2 + \tau\|\theta\|_0$$

where τ is a "regularization parameter" which depends on T and

$$\|\theta\|_0 \triangleq \#\{i : |\theta|_i \neq 0\}.$$

One way to think of this optimization is that we want to find the vector θ which (a) is a good fit to the data and (b) is sparse.

Sparsity Regularization

To see this, first consider the following question: what value of θ minimizes

$$\min_{\theta:\|\theta\|_0=K} \|\zeta-\theta\|_2^2?$$

Now note that we can re-write our optimization as follows:

$$\widehat{K} = \arg\min_K \left[\min_{\theta:\|\theta\|_0=K} \|\zeta-\theta\|_2^2 + \tau K \right]$$

$$\widehat{\theta} = \arg\min_{\theta:\|\theta\|_0=\widehat{K}} \|\zeta-\theta\|_2^2$$

Thus solving the ℓ_0-optimization problem in this denoising context amounts to hard thresholding.

Theorem (Donoho and Johnstone)

If the threshold $T = \sigma\sqrt{2\log_e N}$, then the MSE of the hard thresholding estimator satisfies

$$\mathrm{MSE} \le (2\log_e N + 1)(\sigma^2 + \mathrm{MSE}_s).$$

That is, the performance of the practical hard thresholding estimator is within a $\log N$ factor of the ideal coefficient selection estimator.

So what does sparsity buy us?

We have seen the following:

$$\begin{aligned}\text{MSE} &\le (2\log_e N + 1)(\sigma^2 + \text{MSE}_s)\\ &\le (2\log_e N + 1)(\|f - f_K\|_2^2 + (K+1)\sigma^2).\end{aligned}$$

When f is K-sparse, we have

$$\boxed{\frac{\text{MSE}}{N} = \frac{\|f - \widehat{f}\|_2^2}{N} = O\left(\frac{K\log N}{N}\right).}$$

Recall that if we had a parametric signal with K parameters, our MSE would behave like K/N – so even though this is non-parametric estimation, sparsity leads to near-parametric performance!

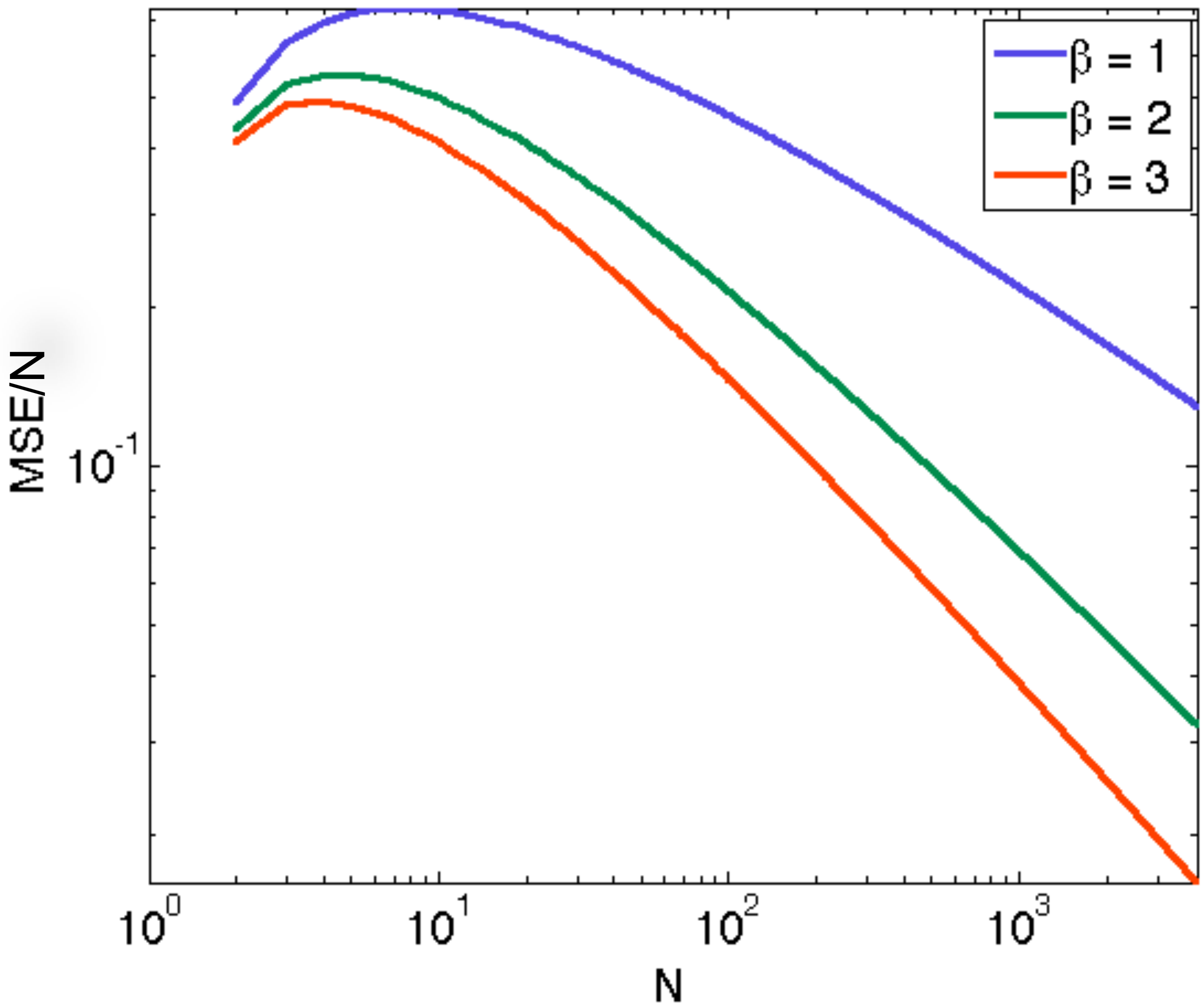
When f is compressible, so $(1/N)\|f - f_K\|_2^2 \leq K^{-\beta}$, we have

$$\frac{\|f - \widehat{f}\|_2^2}{N} = O\left([\log N]\left[K^{-\beta} + \frac{K\sigma^2}{N}\right]\right).$$

Given N and σ^2, the optimal sparsity level (which minimizes the MSE) is $K^* = (N/\sigma^2)^{1/(\beta+1)}$, yielding

$$\boxed{\frac{\|f - \widehat{f}\|_2^2}{N} = O\left([\log N]\left[\sigma^2/N\right]^{\beta/(\beta+1)}\right).}$$

For more compressible signals, the MSE decays more quickly with the amount of data or the signal-to-noise ratio.



Ideal coefficient attenuation

Hard-thresholding and ℓ_0-regularization work well for our noise removal problem, but the ℓ_0-regularizer creates computational problems in related settings (e.g. inverse problems).

Consider instead a coefficient attenuation estimator, where we let the α_i's take any values in $[0, 1]$.

Minimizing the MSE with respect to $\{\alpha_i\}_{i=1}^N$ yields the optimal coefficient attenuation

$$\alpha_i = \frac{|\theta_i|^2}{|\theta_i|^2 + \sigma^2} \qquad \Rightarrow \qquad \mathrm{MSE}_a = \sum_{i=1}^{N} \frac{|\theta_i|^2 \sigma^2}{|\theta_i|^2 + \sigma^2}.$$

Attenuation and Selection

$$\text{MSE}_s = \sum_{i=1}^{N} \min(|\theta_i|^2, \sigma^2) \qquad \text{MSE}_a = \sum_{i=1}^{N} \frac{|\theta_i|^2 \sigma^2}{|\theta_i|^2 + \sigma^2}.$$

$$\boxed{\text{MSE}_s \geq \text{MSE}_a \geq \frac{1}{2}\text{MSE}_s}$$

PRACTICAL COEFFICIENT ATTENUATION

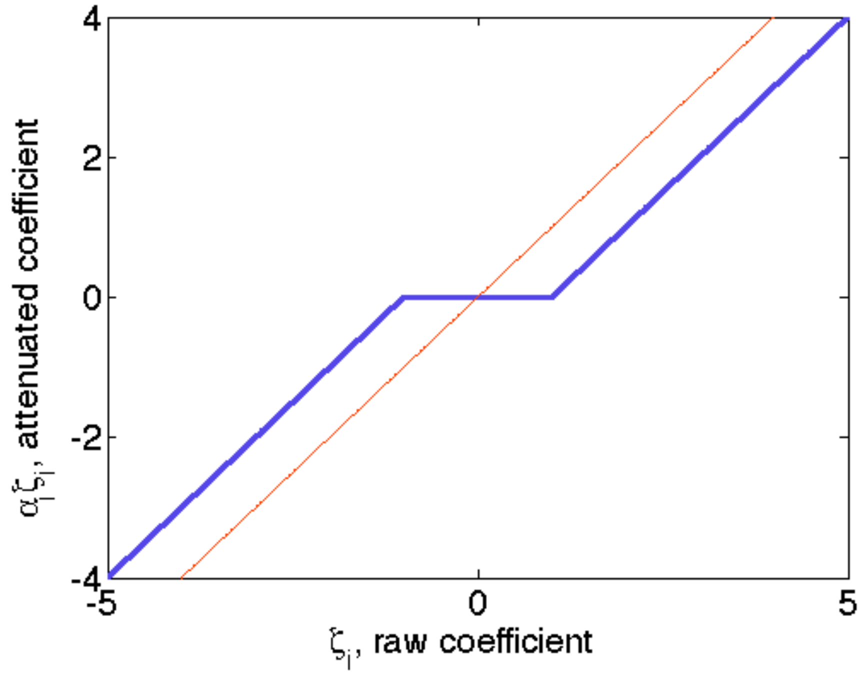
Ideal coefficient attenuation is also impractical. However, we can threshold the noisy coefficients to approximate the ideal coefficient attenuation. In particular, let

$$\widehat{f} = \sum_{i=1}^{N} \delta_T^{(S)}(\zeta_i)\psi_i$$

where $\delta_T^{(S)}$ is a "soft" threshold function

$$\delta_T^{(S)}(z) = \frac{(|z| - T)_+}{|z|} z$$

where $(z)_+ = \begin{cases} z & z \geq 0 \\ 0 & z < 0 \end{cases}$ and T is the threshold level.



Sparsity Regularization

This soft-threshold estimator is equivalent to performing the optimization

$$\widehat{\theta} = \arg\min_{\theta} \|\zeta - \theta\|_2^2 + \tau\|\theta\|_1$$

where τ is a "regularization parameter" which depends on σ and

$$\|\theta\|_1 \triangleq \sum_{i=1}^{N} |\theta|_1.$$

One way to think of this optimization is that we want to find the vector θ which (a) is a good fit to the data and (b) is *nearly sparse*.

This is a convex optimization problem that generalizes well to inverse problems.

Sparsity Regularization

To see this, first re-write our objective as

$$\|\zeta - \theta\|_2^2 + \tau\|\theta\|_1 \equiv \sum_{i=1}^{N} (\zeta_i - \theta_i)^2 + \tau|\theta_i|.$$

Thus we can solve this problem independently for each index i. Consider $\zeta_i \geq 0$; then we know $\widehat{\theta}_i \geq 0$. Compute the derivative and set it equal to zero:

$$\frac{d}{d\theta_i}(\zeta_i - \theta_i)^2 + \tau\theta_i = -2\zeta_i + 2\theta_i + \tau = 0$$

$$\Rightarrow \widehat{\theta}_i = (\zeta_i - \tau/2)_+$$

Now consider $\zeta_i \leq 0$, so that $\widehat{\theta}_i \leq 0$.

$$\frac{d}{d\theta_i}(\zeta_i - \theta_i)^2 - \tau\theta_i = -2\zeta_i + 2\theta_i - \tau = 0$$

$$\Rightarrow \widehat{\theta}_i = -(-\zeta_i - \tau/2)_+$$

Overall this gives us the soft thresholding function:

$$\widehat{\theta}_i = \frac{(|\zeta_i| - \tau/2)_+}{|\zeta_i|}\zeta_i.$$

Let’s see it in action!

WAVELET DENOISING

$\mathbb{E}\left[\widehat{f}\right]$ $\widehat{f}$

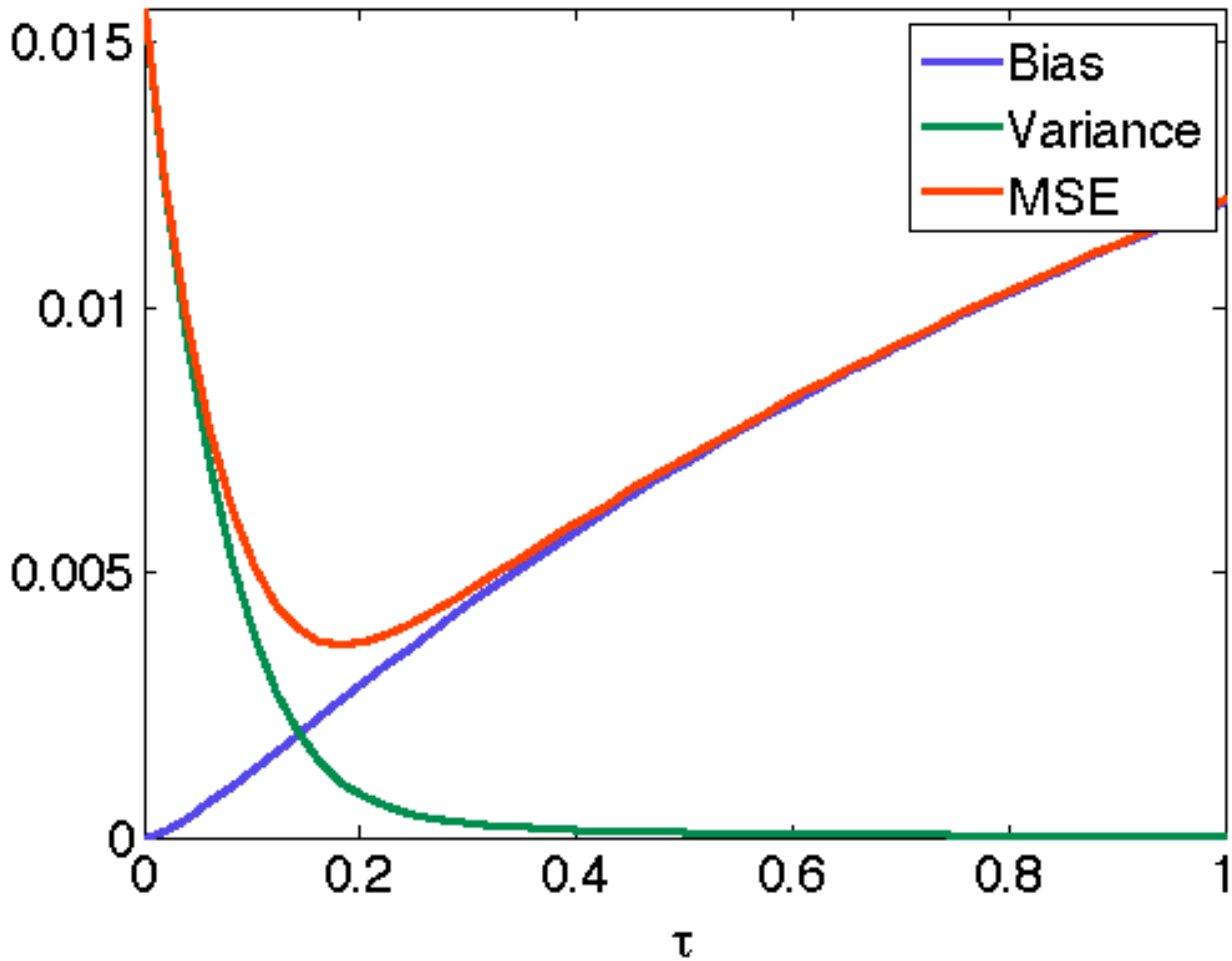
Zero bias, high variance

$\mathbb{E}\left[\widehat{f}\right]$ $\widehat{f}$

High bias, low variance

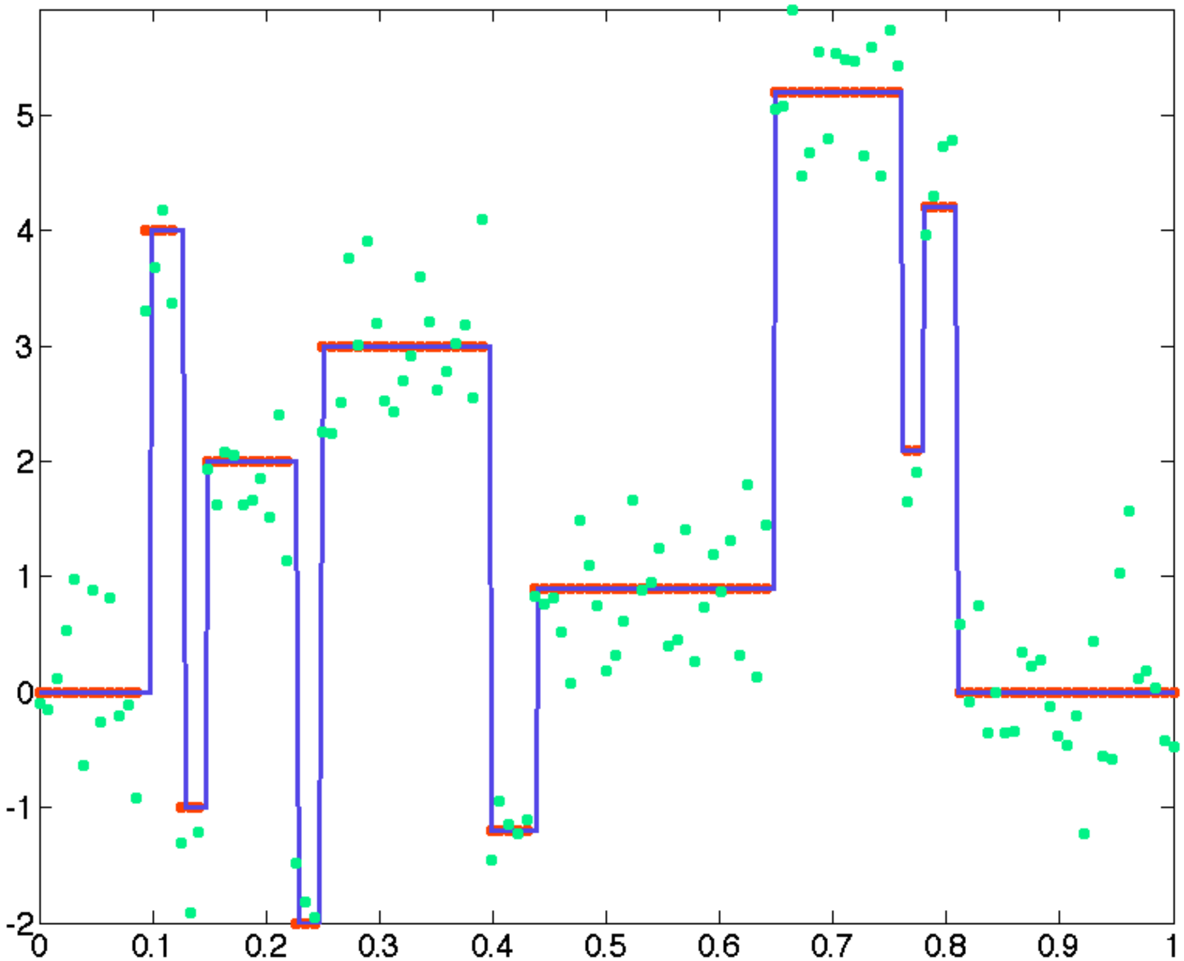
Low bias, low variance

WAVELET DENOISING



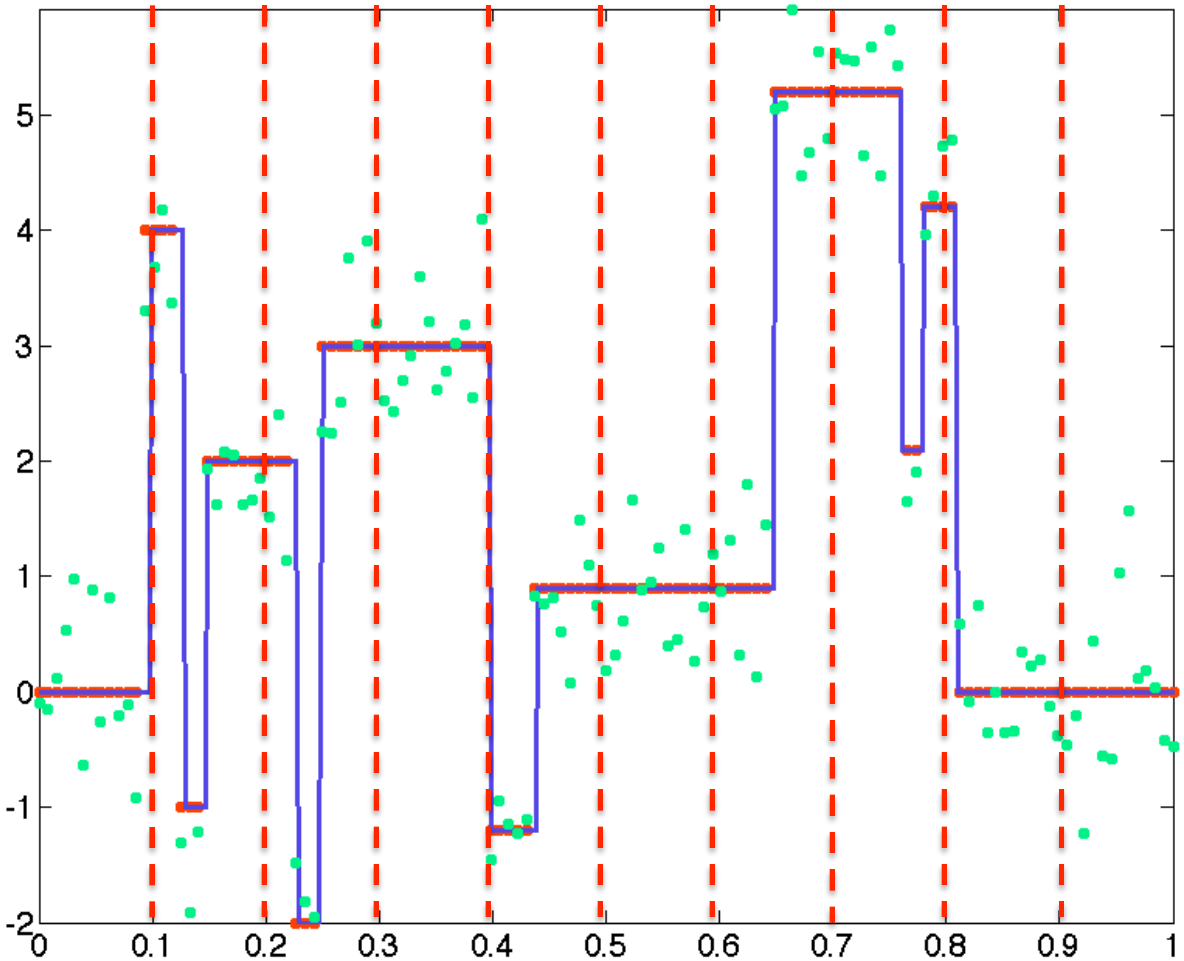
A cute example

A PIECEWISE CONSTANT SIGNAL



p breakpoints,
N samples

LINEAR ESTIMATOR



Break signal into m equi-sized pieces, each with N/m samples. Compute sample average on each piece.

Linear estimator

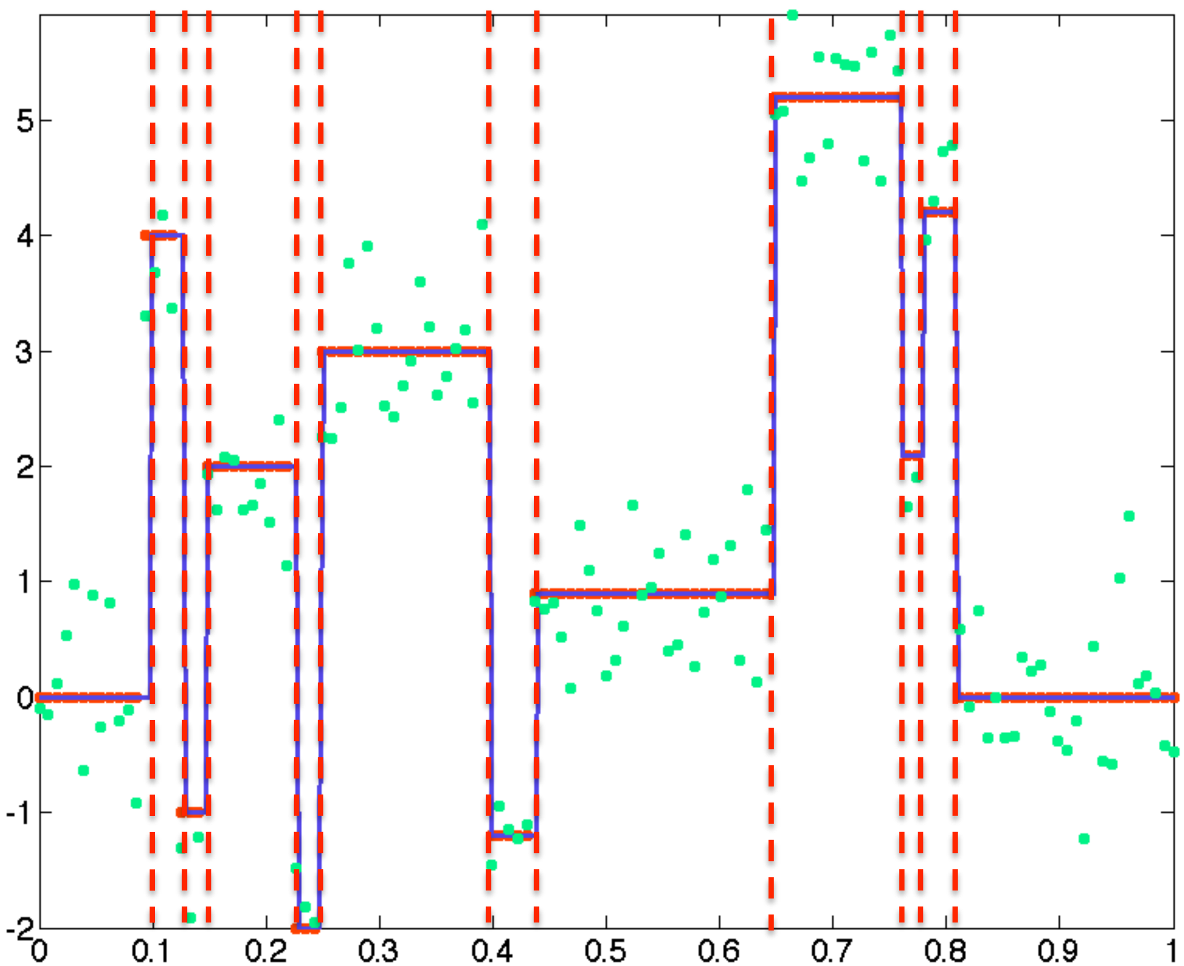
p of the m pieces will have a breakpoint. The error in these pieces will be $O(1/m)$ regardless of how much data we have.

The remaining $m-p$ pieces will not have a breakpoint, and the error of estimating a constant on each piece is $O(1/N)$.

The total error is then MSE $= O(p/m) + O((m-p)/N)$. Optimizing over m we find that $m^* \approx \sqrt{Np}$, giving us the total error

$$\boxed{\frac{\text{MSE}_{\text{linear}}}{N} = O\left(\sqrt{p/N}\right)}$$

ORACLE ESTIMATOR



An oracle tells us where the p break-points are, so we just have to estimate the constant level on each interval.

Oracle estimator

None of the $p+1$ pieces have a breakpoint. The error of estimating a constant on each piece is $O(1/N)$.

The total error is then

$$\boxed{\frac{\text{MSE}_{\text{oracle}}}{N} = O\left(\frac{p}{N}\right)}$$

SPARSE ESTIMATOR

A piecewise constant signal is
$p \log N$-sparse in the Haar wavelet basis.

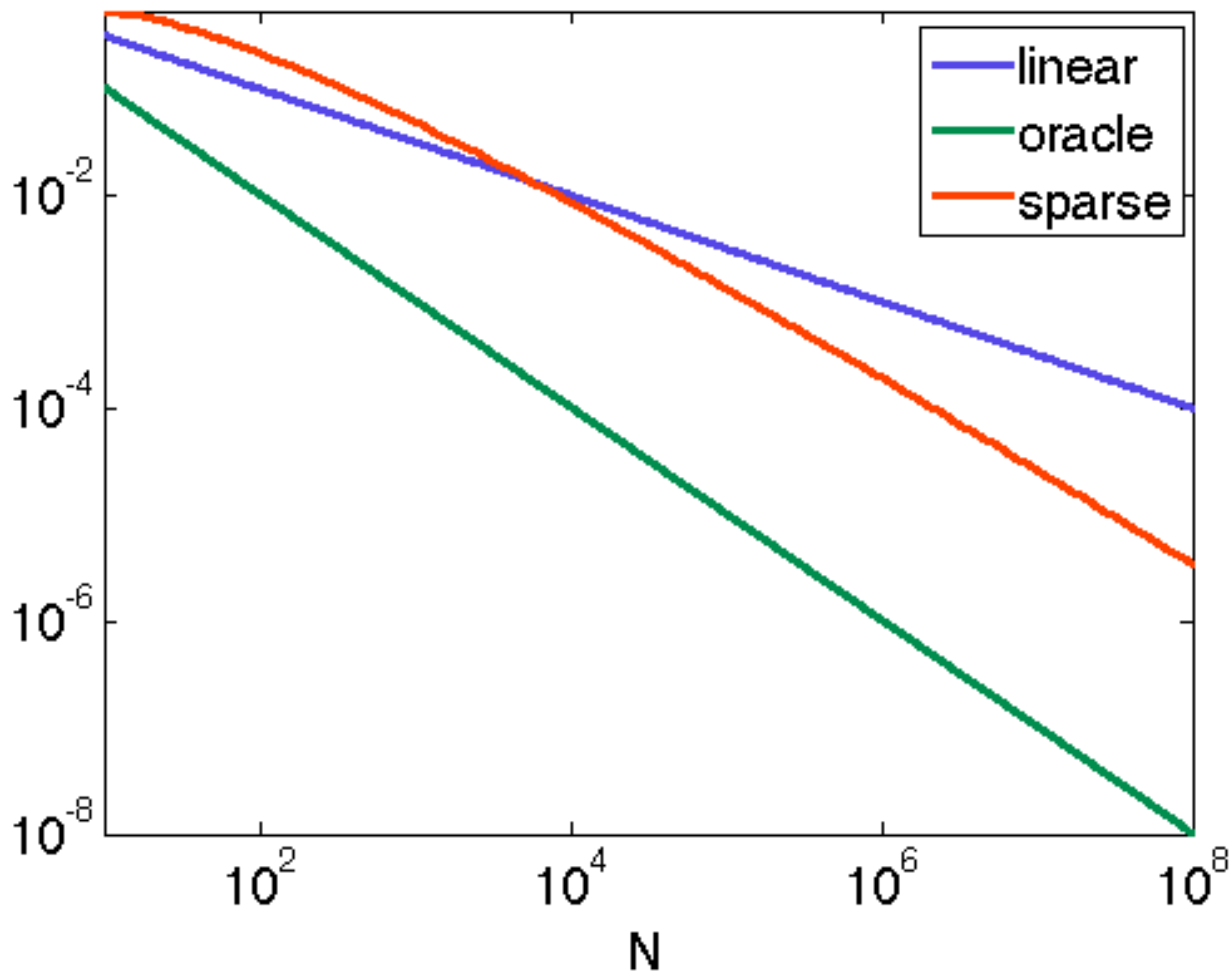
Ideal coefficient selection would give us

$$\text{MSE}_s = O(\sigma^2 p \log N)$$

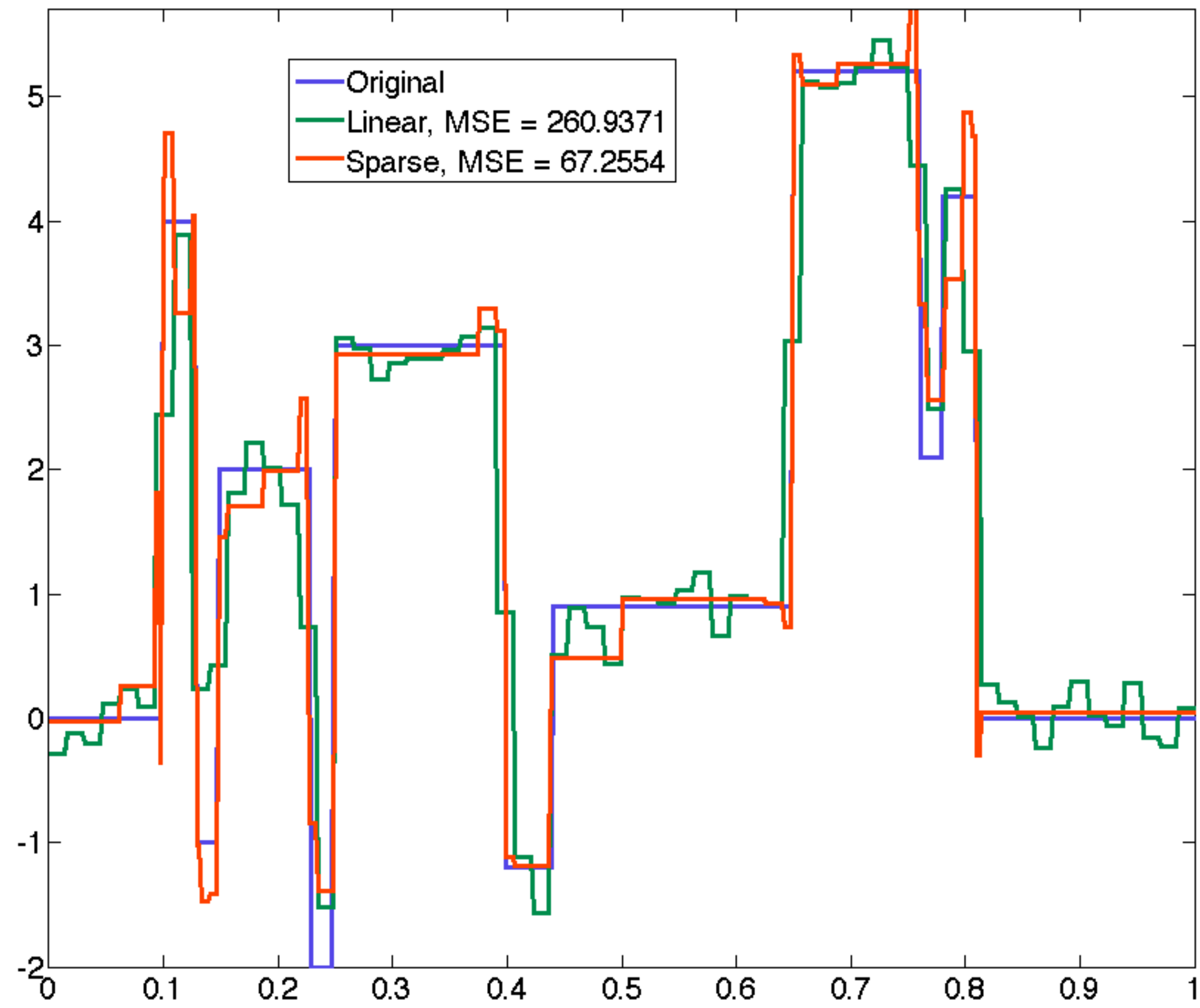
Practical coefficient selection would give us

$$\boxed{\frac{\text{MSE}_{\text{sparse}}}{N} = O\left(\frac{p \log^2 N}{N}\right)}$$

ERROR DECAY RATES



Example



Original
Linear, MSE = 260.9371
Sparse, MSE = 67.2554
5
4
3
2
1
0
-1
-2
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Generalizations

APPROXIMATION AND ESTIMATION

- We saw that our error had to main components: approximation error and estimation error
- We can think of this as follows: to estimate a sparse signal, we need to perform two tasks:
  - Figure out which coefficients are significant, giving ourselves an accurate sparse approximation
  - Computing the values of those coefficients from noisy or corrupted data.

A FUNDAMENTAL TRADEOFF

- Similar tradeoffs appear in many contexts.
  - Classification: find sparse representation of features, then do classification in space of significant coefficients
  - Compression: find sparse approximation, encode indices and values of sparse coefficients
  - Missing data: fill in missing values so result is sparse and fits data
  - Distributed processing: instead of communicating all observations, just communicate sparse coefficients
- This lets us sidestep the

"CURSE OF DIMENSIONALITY"

Examples

Estimation: Choose estimate $\widehat{\theta}$ where

$$\widehat{\theta} = \arg\min_{\theta} \quad \underbrace{\|y - \Psi\theta\|_2^2}_{\text{fit to data}} \quad + \quad \underbrace{\tau\|\theta\|_1}_{\text{sparsity}}$$

Compression: Encode approximation $\widehat{\theta}$ where

$$\widehat{\theta} = \arg\min_{\theta} \quad \underbrace{\|y - \Psi\theta\|_2^2}_{\text{fit to original}} \quad + \quad \underbrace{\tau\|\theta\|_1}_{\approx \text{ file size}}$$

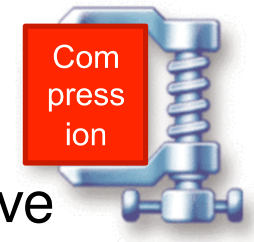
Distributed estimation: Transmit estimate $\widehat{\theta}$ where

$$\widehat{\theta} = \arg\min_{\theta} \quad \underbrace{\|y - \Psi\theta\|_2^2}_{\substack{\text{fit to data at} \\ \text{different sensors}}} \quad + \quad \underbrace{\tau\|\theta\|_1}_{\substack{\approx \text{ communication} \\ \text{power/bandwidth}}}$$

Inverse problems: Measure $y = Af + n$, choose estimate $\widehat{\theta}$ where

$$\widehat{\theta} = \arg\min_{\theta} \quad \underbrace{\|y - A\Psi\theta\|_2^2}_{\text{fit to data}} \quad + \quad \underbrace{\tau\|\theta\|_1}_{\text{sparsity}}$$

COMPRESSION



Similar concepts arise in data compression. Imagine we have an image with N pixels.

One option is to write each pixel value to a file. This is what a bitmap scheme does.

Another option is to transform the image into another domain (e.g. wavelet) in which it is sparse. Then we only need to store (a) the values and (b) the indices of the non-zero coefficients. This is what JPEG and JPEG-2000 do.

Since sparse images have few non-zero coefficients, part (a) requires relatively little storage. Determining methods for encoding part (b) can be more challenging, and significant research has been devoted to this topic.

COMING NEXT…

- We saw we can use sparsity to estimate signals in noise.
- This all assumed a very direct observation model, though.
- If we know our signal is sparse, are there better ways to sample it?
- Can we use sparsity to reduce the amount of data we need to collect?