COURANT INSTITUTE OF MATHEMATICAL SCIENCES

Sparse linear models

Optimization-Based Data Analysis

http://www.cims.nyu.edu/~cfgranda/pages/OBDA_spring16

Carlos Fernandez-Granda

2/22/2016

Linear model

A linear model for a signal $x \in \mathbb{R}^n$ is a representation of the form

$$x = \sum_{i=1}^{m} c_i \, \phi_i$$

$\{\phi_1, \ldots, \phi_m\}$ is a family of atoms in $\mathbb{R}^n$

$c \in \mathbb{R}^m$ is an alternative representation or transform of x

Sparse linear model

A sparse linear model contains a small number of coefficients

$$x = \sum_{i \in \mathcal{S}} c_i \, \phi_i \qquad |\mathcal{S}| \ll m$$

How do we choose a transform that sparsifies a class of signals?

- Intuition / Domain knowledge (this lecture)
- Learning it from the data (later on)

Why?

Signals of interest (speech, natural images, biomedical activity, etc.) are often highly structured

Sparse linear models are able to exploit this structure to enhance data analysis and processing

Applications:

- Compression
- Denoising
- Inverse problems

Sparse representation in an orthonormal basis

If the atoms $\{\phi_1, \ldots, \phi_n\}$ form an orthonormal basis, then

$$U := \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_n \end{bmatrix} \qquad UU^T = \mathrm{I}$$

Coefficients are obtained by computing inner products with the atoms

$$x = UU^T x = \sum_{i=1}^{m} \langle \phi_i, x \rangle \, \phi_i$$

Sparse representation in a basis

If the atoms $\{\phi_1, \ldots, \phi_n\}$ form a basis, then

$$B := \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_n \end{bmatrix} \qquad BB^{-1} = I$$

Coefficients are obtained by computing inner products with dual atoms

$$x = BB^{-1}x = \sum_{i=1}^{m} \langle \theta_i, x \rangle \, \phi_i$$

where

$$B^{-1} = \begin{bmatrix} \theta_1^T \\ \theta_2^T \\ \cdots \\ \theta_n^T \end{bmatrix}$$

Overcomplete dictionaries

If the atoms $\{\phi_1, \ldots, \phi_m\}$ are linearly independent and $m > n$

$$D := \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_m \end{bmatrix}$$

Two alternative sparse models

1. Synthesis sparse model

   $$x = Dc \quad \text{where } c \text{ is sparse}$$

   Problem: Given x find a sparse c

2. Analysis sparse model:

   $$D^T x \text{ is sparse}$$

Both are equivalent if D is an orthonormal basis

Introduction

Linear transforms

- Frequency representation
- Short-time Fourier transform (STFT)
- Wavelets

Overcomplete sparse models

Denoising

- The denoising problem
- Thresholding
- Synthesis model
- Analysis model

Fourier series

Fourier series coefficients of $f : [0,1] \to \mathbb{C}$

$$c_k := \int_0^1 f(t)\, e^{-i2\pi kt}\, \mathrm{d}t$$

Fourier series

$$S_n(t) := \sum_{k=-n}^{n} c_k e^{i2\pi kt} = \sum_{k=-n}^{n} \langle \phi_k, f \rangle\, \phi_k$$

Sinusoidal atoms $\phi_k(t) = e^{i2\pi kt} = \cos(2\pi kt) + i \sin(2\pi kt)$

Orthonormal basis of $\mathbb{L}_2$ under the usual inner product

$$\lim_{n\to\infty} ||f(t) - S_n(t)|| = 0 \qquad \text{for all } f \in \mathbb{L}_2$$

Discrete Fourier transform (DFT)

Discretized frequency representation for vectors in $\mathbb{C}^n$

Sinusoidal atoms are a basis for $\mathbb{C}^n$

$$\phi_k = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 \\ e^{\frac{i2\pi k}{n}} \\ e^{\frac{i2\pi k2}{n}} \\ \cdots \\ e^{\frac{i2\pi k(n-1)}{n}} \end{bmatrix} \qquad F := \begin{bmatrix} \phi_0 & \phi_1 & \cdots & \phi_{n-1} \end{bmatrix}$$

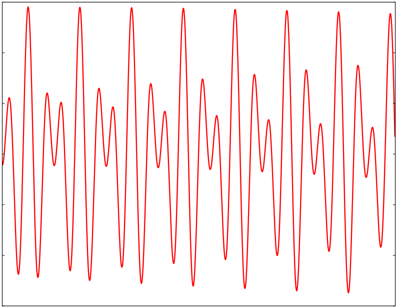
$$\text{DFT}\left\{x\right\}_k := (Fx)_k = \langle \phi_k, x \rangle$$

The fast Fourier transform (FFT) computes the DFT in $\mathcal{O}(n \log n)$

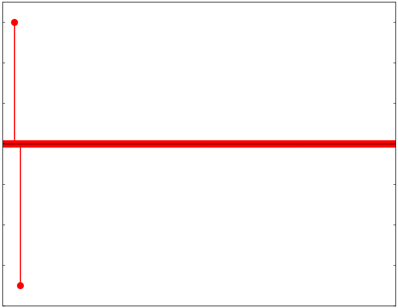
The discrete cosine transform (DCT) is a related transformation designed for real vectors

Discrete cosine transform

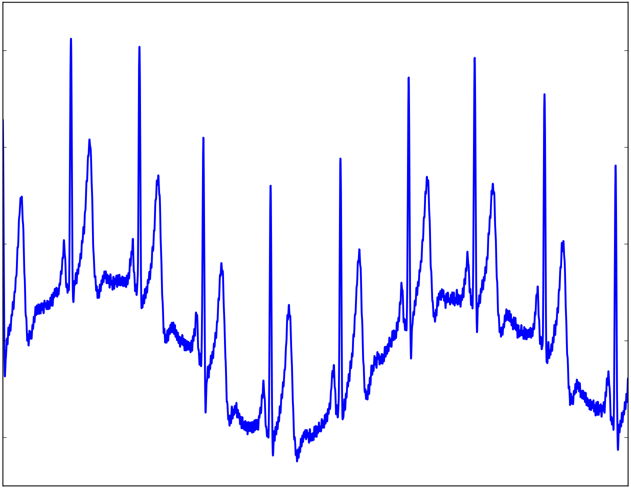
Signal



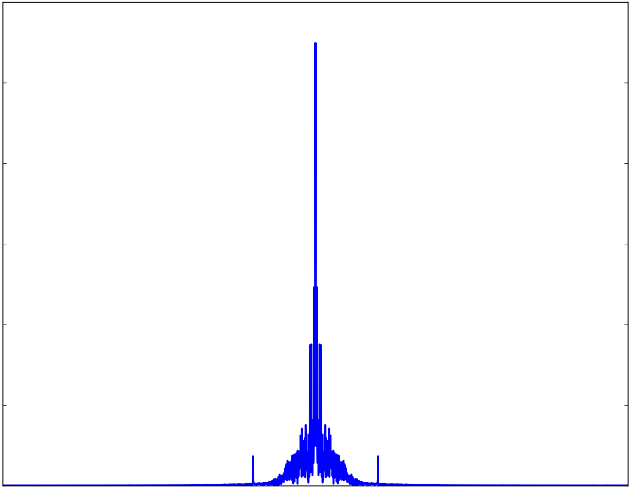
DCT coefficients



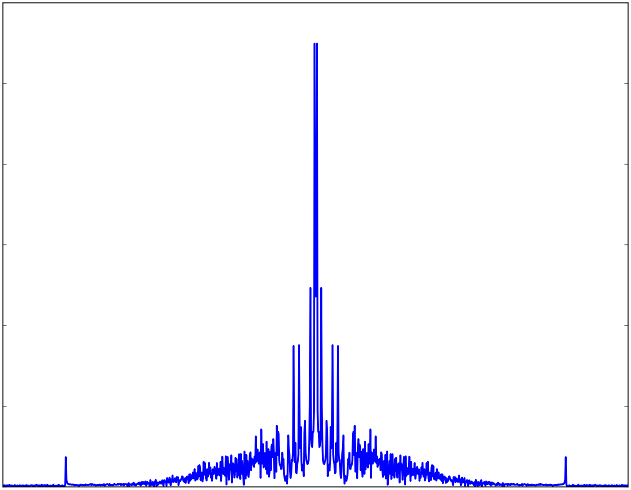
Electrocardiogram



Electrocardiogram (spectrum)



Electrocardiogram (spectrum)



2D DFT

Discretized frequency representation for 2D arrays in $\mathbb{C}^{n\times n}$

Sinusoidal atoms are a basis for $\mathbb{C}^{n\times n}$

$$\phi^{\text{2D}}_{k_1,k_2} = \frac{1}{n}\begin{bmatrix} 1 & e^{\frac{i2\pi k_2}{n}} & \cdots & e^{\frac{i2\pi k_2(n-1)}{n}} \\ e^{\frac{i2\pi k_1}{n}} & e^{\frac{i2\pi(k_1+k_2)}{n}} & \cdots & e^{\frac{i2\pi(k_1+k_2(n-1))}{n}} \\ & & \cdots & \\ e^{\frac{i2\pi k_1(n-1)}{n}} & e^{\frac{i2\pi(k_1(n-1)+k_2)}{n}} & \cdots & e^{\frac{i2\pi(k_1(n-1)+k_2(n-1))}{n}} \end{bmatrix}$$

$$= \phi^{\text{1D}}_{k_1}\left(\phi^{\text{1D}}_{k_2}\right)^T$$

$$\text{DFT}^{\text{2D}}\{X\} := FXF = \sum_{k_1=0}^{n-1}\sum_{k_2=0}^{n-1}\left\langle \phi^{\text{2D}}_{k_1,k_2}, X\right\rangle \phi^{\text{2D}}_{k_1,k_2}$$

Generalizes to $\mathbb{C}^{m\times n}$, $m \neq n$, and to higher dimensions

Compression via frequency representation

The 2D frequency representation of images tends to be sparse

Thresholding the coefficients yields a compressed representation

The JPEG compression standard is based on the 2D DCT

High-frequency coefficients are discarded according to a perceptual model

Compression via frequency representation

Original

Compression via frequency representation

10 % largest DCT coeffs

Compression via frequency representation

2 % largest DCT coeffs

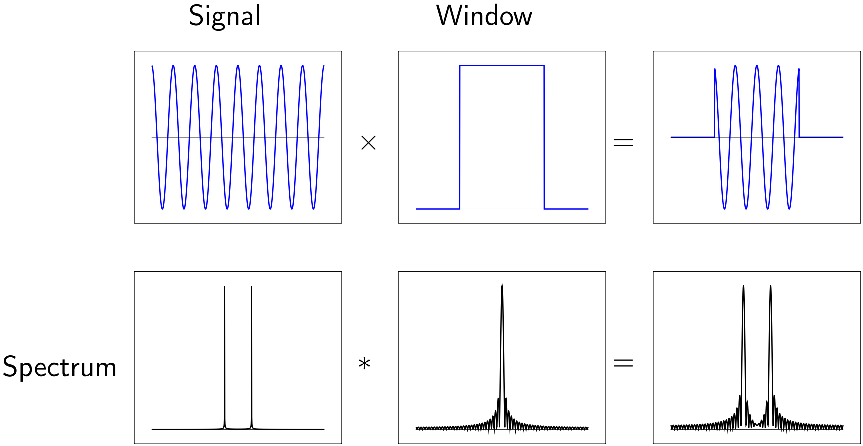
Motivation

Spectrum of speech, music, etc. varies over time

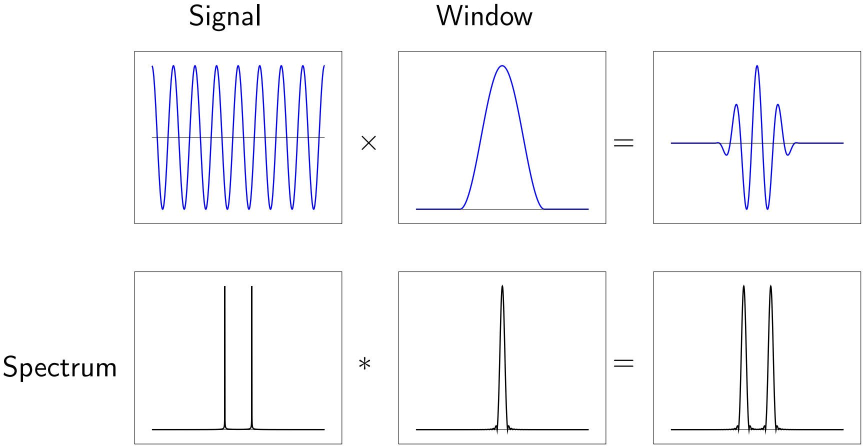
Idea: Compute frequency representation of time segments of the signal

We must use a window to avoid introducing spurious high frequencies

The need for windowing



The need for windowing



Short-time Fourier transform

Let $w : [0,1] \to \mathbb{C}$ be a window function localized in time and frequency

$$\mathrm{STFT}\left\{f\right\}(k,\tau) := \int_0^1 f(t)\,\overline{w(t-\tau)}e^{-i2\pi kt}\,\mathrm{d}t = \left\langle \phi_{k,\tau}, f\right\rangle$$

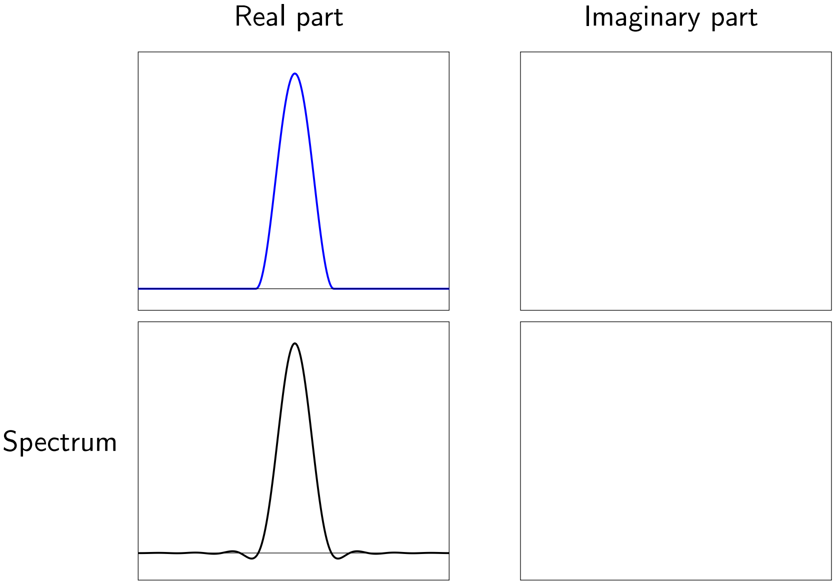
Each atom $\phi_{k,\tau}(t) := w(t-\tau)\,e^{i2\pi kt}$ corresponds to w shifted by τ in time and by k in frequency

In discrete time, pointwise multiplication by a shifted window followed by a DFT, equivalent to $D^T x$ where $D \in \mathbb{C}^{n\times m}$, $m > n$

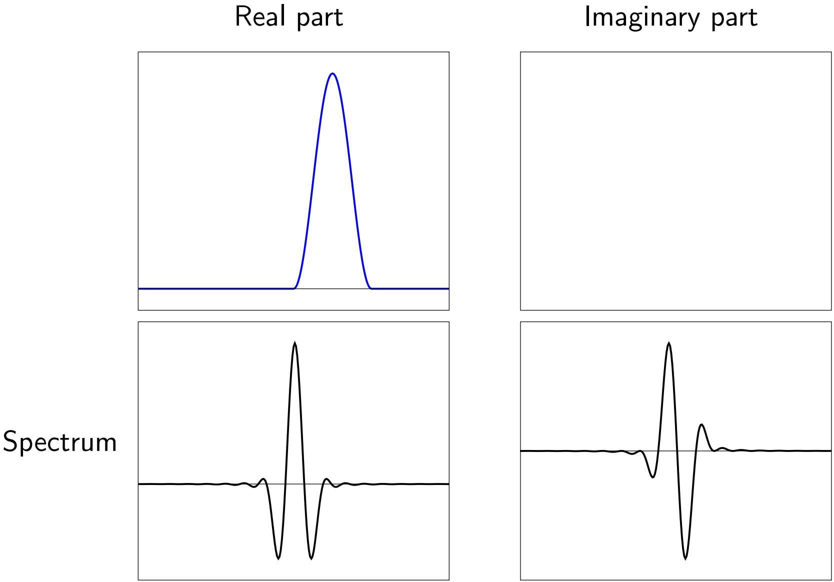
The STFT of speech tends to be sparse (analysis sparse model)

Including dilations of w (in addition to time and frequency translations) yields a dictionary of Gabor atoms

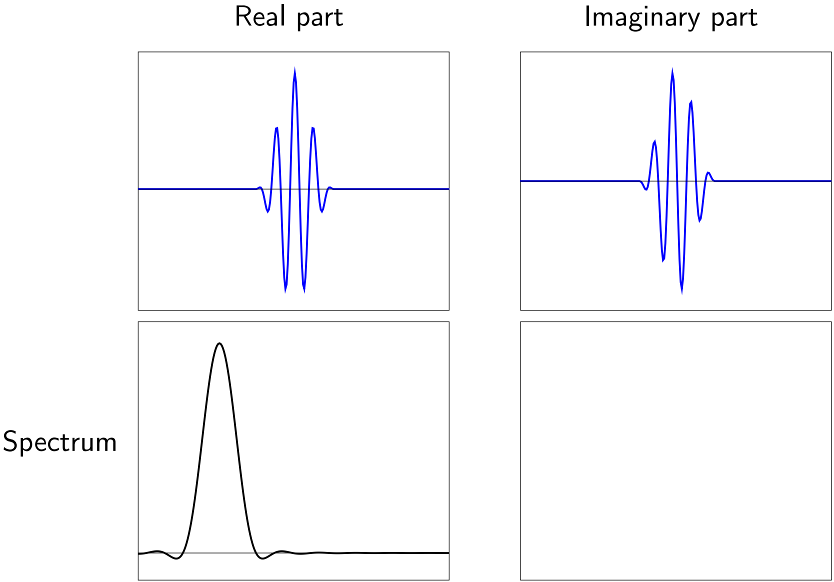
Atom $\tau = 0$, $k = 0$



Atom $\tau = 1/32$, $k = 0$



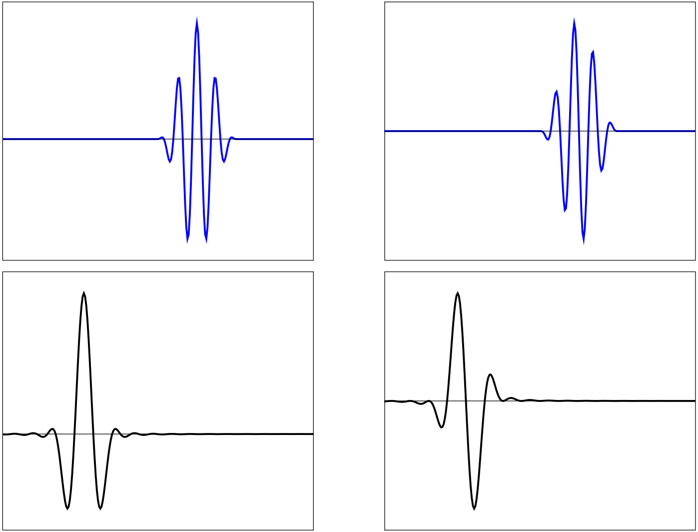
Atom $\tau = 0$, $k = 64$



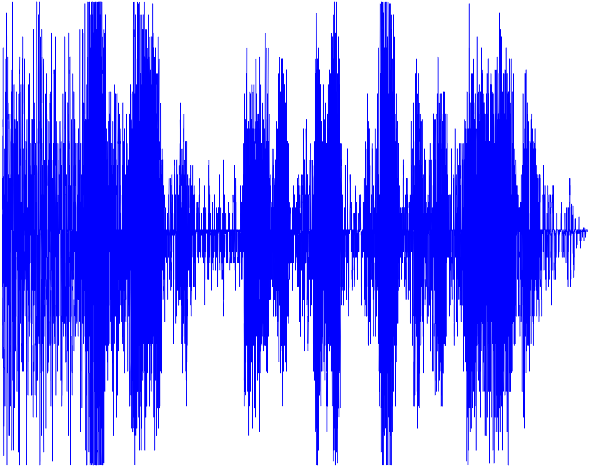
Atom $\tau = 1/32$, $k = 64$

Real part | Imaginary part

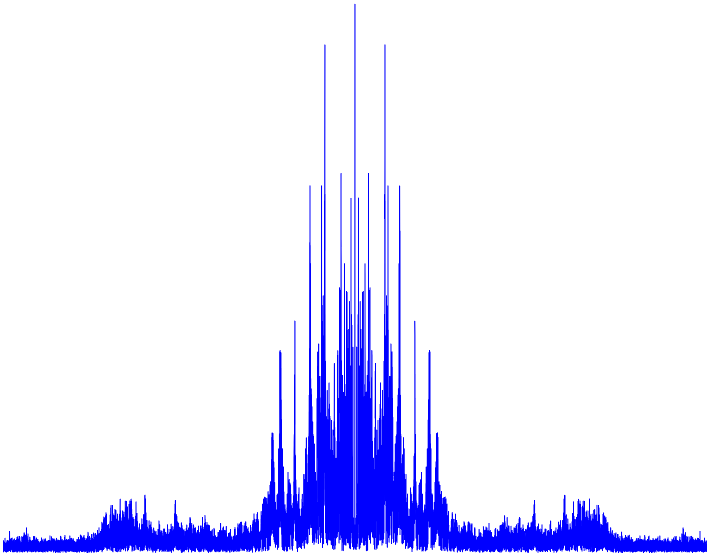
Spectrum



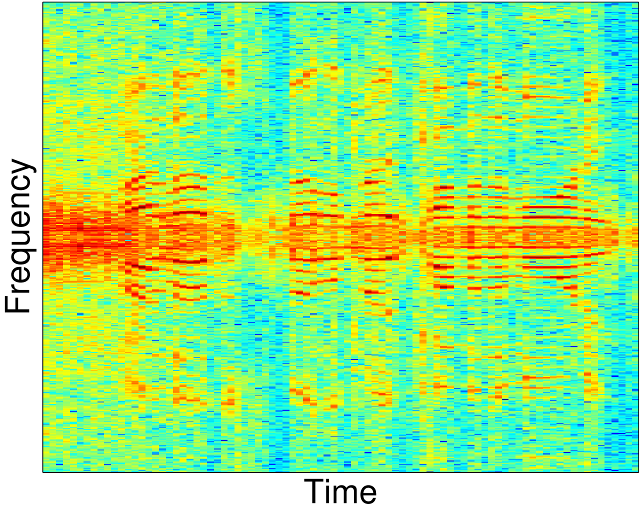
Speech signal



Spectrum



Spectrogram (log magnitude of STFT coefficients)



Wavelets

Aim: Approximate signals at different scales

A wavelet ψ is a unit-norm, zero-mean function in $\mathbb{L}_2$

Wavelet transform

$$\mathrm{W}\{f\}(s,\tau) := \frac{1}{\sqrt{s}} \int_0^1 f(t) \overline{\psi\left(\frac{t-\tau}{s}\right)} \, \mathrm{d}t = \langle \phi_{s,\tau}, f \rangle$$

Atoms are dilations and translations of the *mother* wavelet

$$\phi_{s,\tau}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-\tau}{s}\right)$$

We can build orthonormal basis for $\mathbb{L}_2$ using wavelets

Multiresolution approximation

Sequence $\{\mathcal{V}_j, j \in \mathbb{Z}\}$ of closed subspaces of $\mathbb{L}_2(\mathbb{R})$ such that $\mathcal{P}_{\mathcal{V}_j}(f)$ is an approximation of f at scale 2^j

Conditions:

- Dilating functions in $\mathcal{V}_j$ by 2 yields functions in $\mathcal{V}_{j+1}$

$$f(t) \in \mathcal{V}_j \iff f\left(\frac{t}{2}\right) \in \mathcal{V}_{j+1}$$

- Approximations at a scale 2^j are always better than at 2^{j+1}

$$\mathcal{V}_{j+1} \subset \mathcal{V}_j$$

Multiresolution approximation

- $\mathcal{V}_j$ is invariant to translations at the scale 2^j

$$f(t) \in \mathcal{V}_j \iff f\left(t - 2^j k\right) \in \mathcal{V}_j \qquad \text{for all } k \in \mathbb{Z}$$

- As $j \to \infty$ the approximation loses all information

$$\lim_{j \to \infty} \mathcal{V}_j = \{0\}$$

- As $j \to -\infty$ the approximation is perfect

$$\lim_{j \to -\infty} \mathcal{V}_j = \mathbb{L}_2$$

- There exists a scaling function $\zeta \in \mathcal{V}_0$ such that $\left\{\zeta_{0,k}(t) := \zeta(t-k),\ k \in \mathbb{Z}\right\}$ is an orthonormal basis for $\mathcal{V}_0$

Wavelet basis

Mallat and Meyer prove that there exists a wavelet ψ such that

$$\mathcal{P}_{\mathcal{V}_j}(f) = \mathcal{P}_{\mathcal{V}_{j+1}}(f) + \sum_{k \in \mathbb{Z}} \langle \psi_{2^j,k}, f \rangle \, \psi_{2^j,k}.$$

$\{\psi_{2^j,k},\ k \in \mathbb{Z}\}$ is an orthonormal basis for $\mathcal{V}_j \cap \mathcal{V}_j^{\perp}$

$\{\zeta_{0,k}(t), \psi_{2^1,k}, \psi_{2^2,k}, \ldots, \psi_{2^j,k},\ k \in \mathbb{Z}\}$ is an orthonormal basis for $\mathcal{V}_j$

Many different wavelet bases: Meyer, Daubechies, Battle-Lemarie, ...

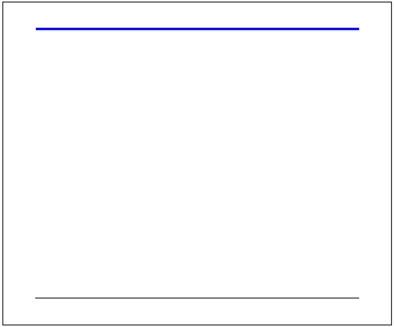
Discrete wavelet transform can be computed in $\mathcal{O}(n)$

Signal processing interpretation:

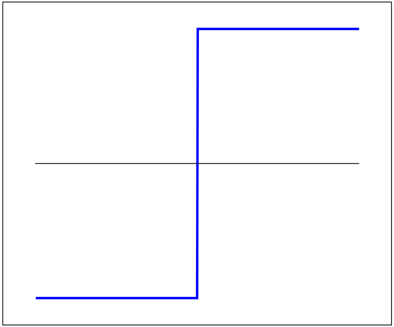
Wavelets act as band-pass filters, scaling functions act as low-pass filters

Haar wavelet

Scaling function

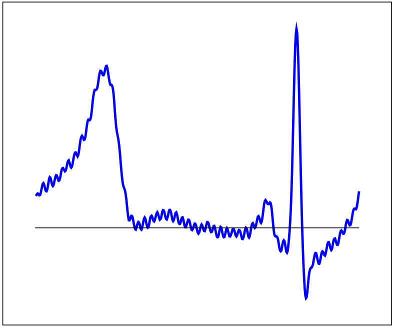


Mother wavelet

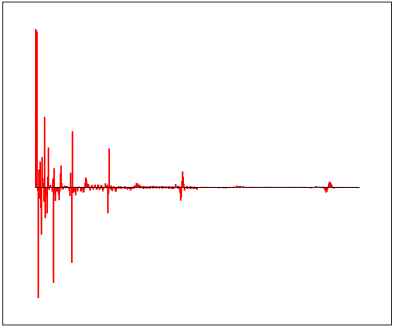


Electrocardiogram

Signal

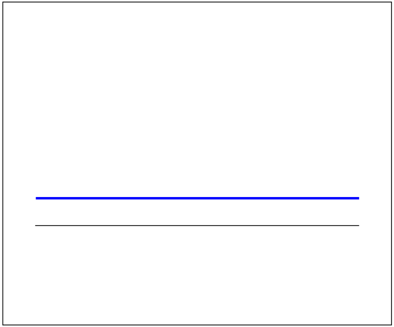


Haar transform

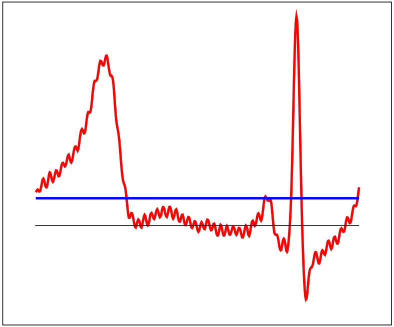


Scale 2^9

Contribution

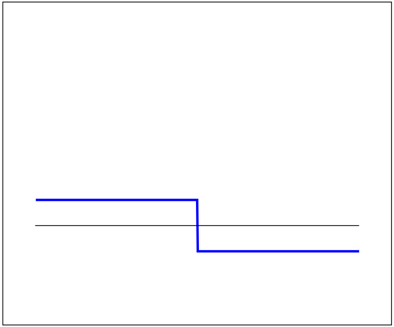


Approximation

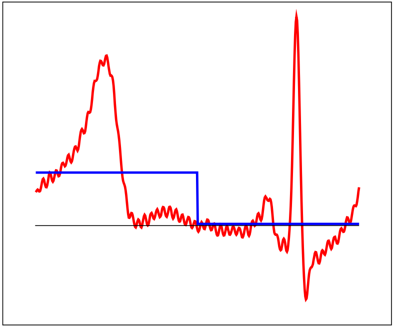


Scale 2^8

Contribution

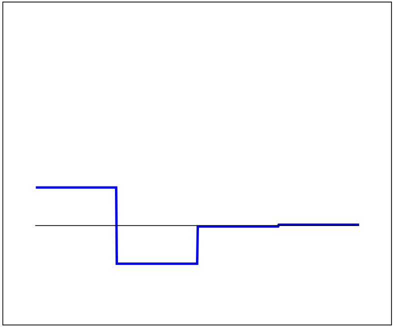


Approximation

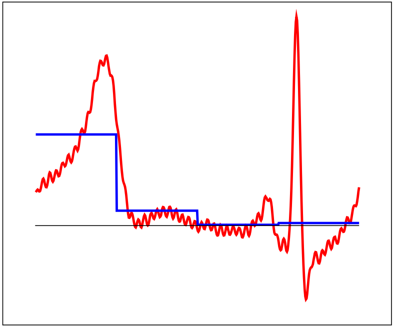


Scale 2^7

Contribution

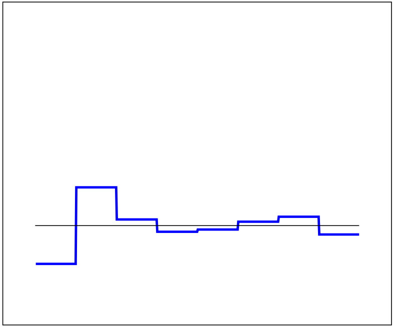


Approximation

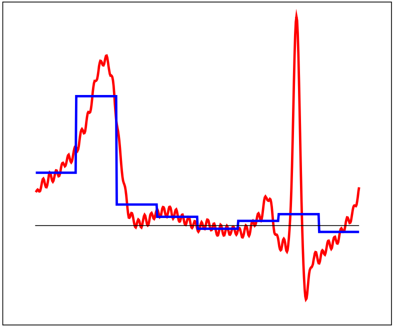


Scale 2^6

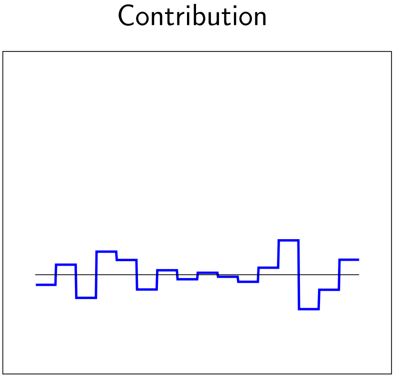
Contribution

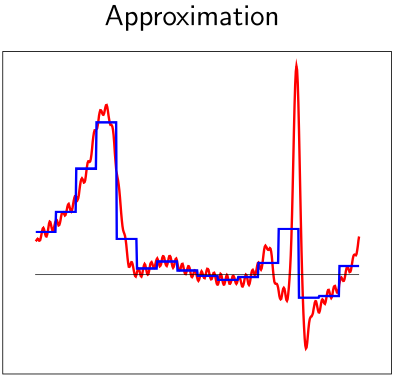


Approximation



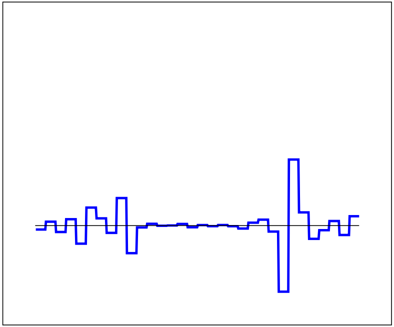
Scale 2^5



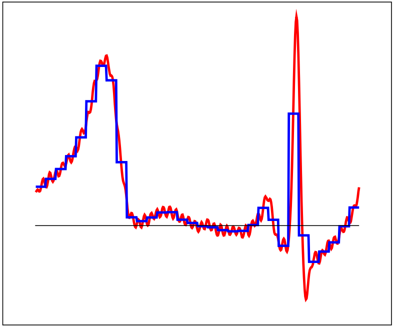


Scale 2^4

Contribution

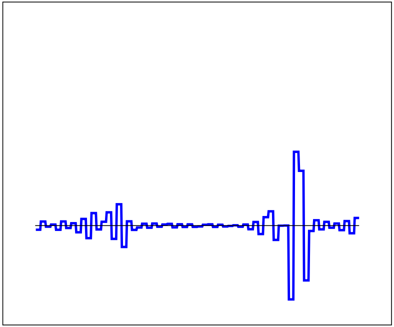


Approximation

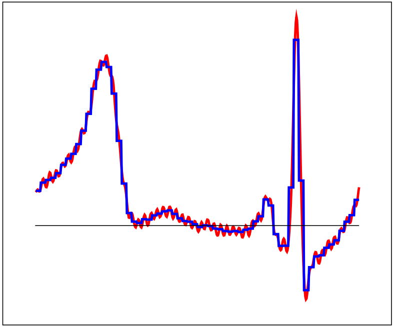


Scale 2^3

Contribution

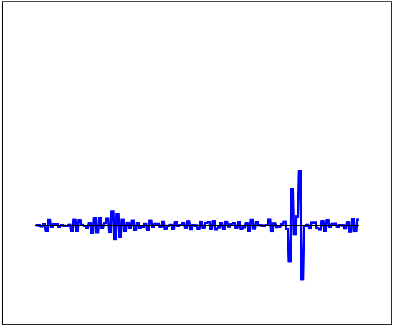


Approximation

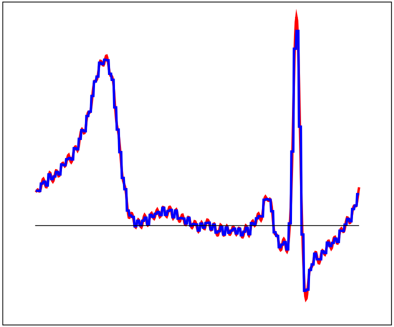


Scale 2^2

Contribution

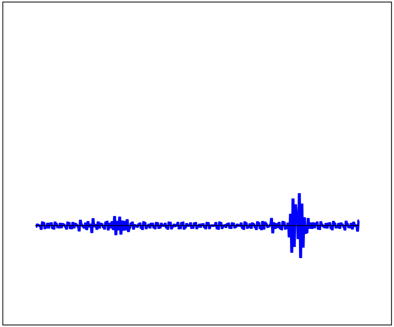


Approximation

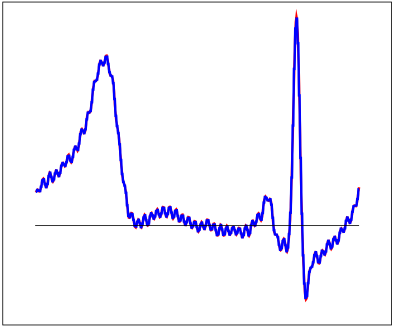


Scale 2^1

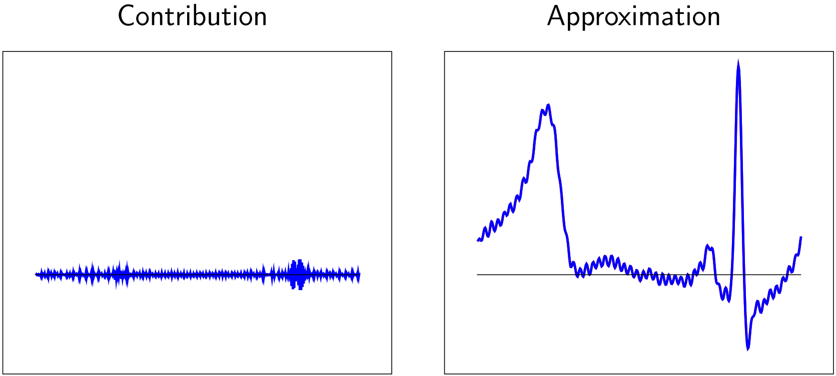
Contribution



Approximation



Scale 2^0



2D Wavelets

Extension to 2D by using outer products of 1D atoms

$$\phi^{2D}_{s_1,s_2,k_1,k_2} := \phi^{1D}_{s_1,k_1} \left(\phi^{1D}_{s_2,k_2}\right)^T$$

Yields sparse representation of natural images

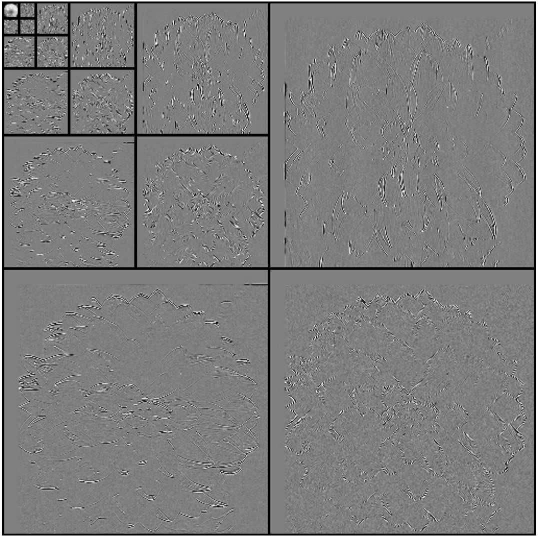
The JPEG 2000 compression standard is based on 2D wavelets

Many extensions:

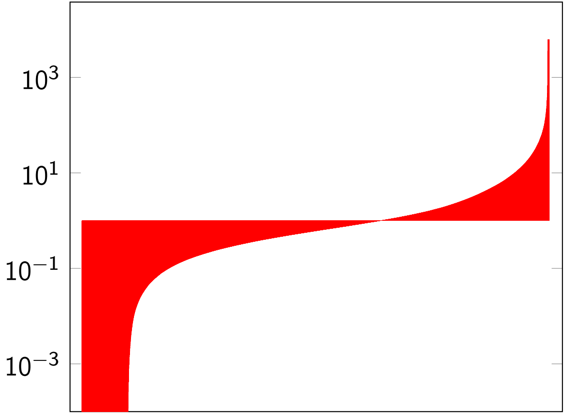
Steerable pyramid, ridgelets, curvelets, bandlets, ...

2D wavelet transform

2D wavelet transform



Sorted coefficients



Overcomplete dictionaries

Atoms $\{\phi_1, \ldots, \phi_m\}$ are linearly independent and $m > n$

$$D := \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_m \end{bmatrix}$$

Synthesis sparse model:

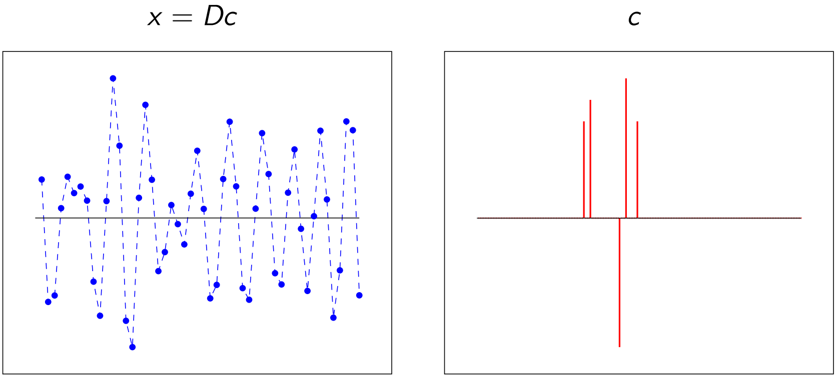
$$x = Dc \quad \text{where } c \text{ is sparse}$$

Problem: There are infinite choices of c such that $x = Dc$

Some may not be sparse at all!

Example: Dictionary of sinusoids

Dictionary of sinusoids



First idea

Apply pseudoinverse

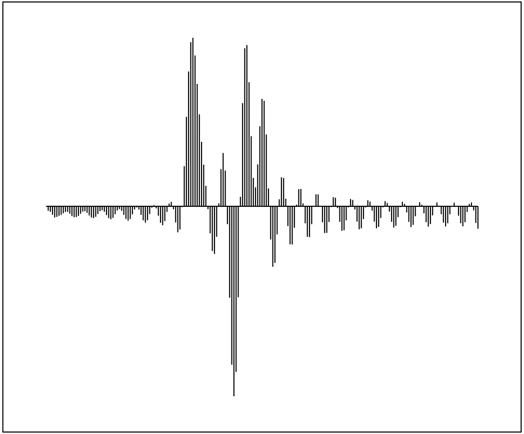
$$\widehat{c} := D^{\dagger} x = D^T \left(D D^T \right)^{-1} x$$

Interpretations:

- Projection of c onto row space of D
- Solution to

$$\begin{aligned} \text{minimize} \quad & ||\tilde{c}||_2 \\ \text{subject to} \quad & x = D\,\tilde{c}, \end{aligned}$$

Dictionary of sinusoids: Minimum ℓ_2-norm coefficients



Computing sparse representations

We would like to solve

$$\begin{aligned} &\text{minimize} \quad && ||\tilde{c}||_0 \\ &\text{subject to} \quad && x = D\,\tilde{c} \end{aligned}$$

Computationally intractable

Two possibilities:

- Greedy methods: Select atoms one by one
- ℓ_1-norm minimization:

$$\min_{\tilde{c} \in \mathbb{R}^m} ||\tilde{c}||_1 \quad \text{such that } x = Dc$$

Matching pursuit (MP)

Iteratively choose atoms that are most correlated with the signal

Initialization:

$$r^{(0)} = x$$
$$\hat{x}^{(0)} = 0$$

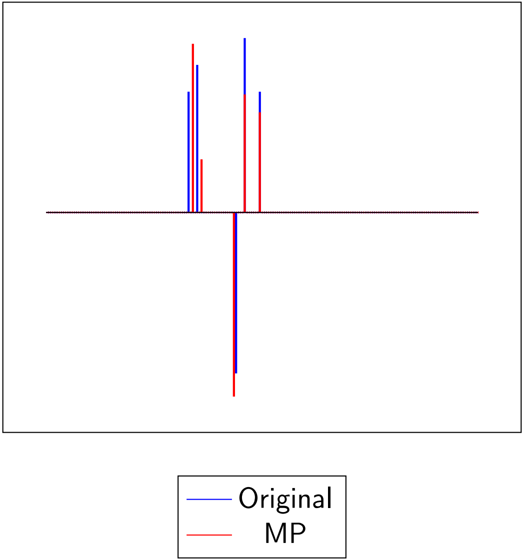
Iterations: $k = 1, 2, \ldots$

$$\phi^{(k)} = \arg\max_j \left| \left\langle r^{(k-1)}, \phi_k \right\rangle \right|$$

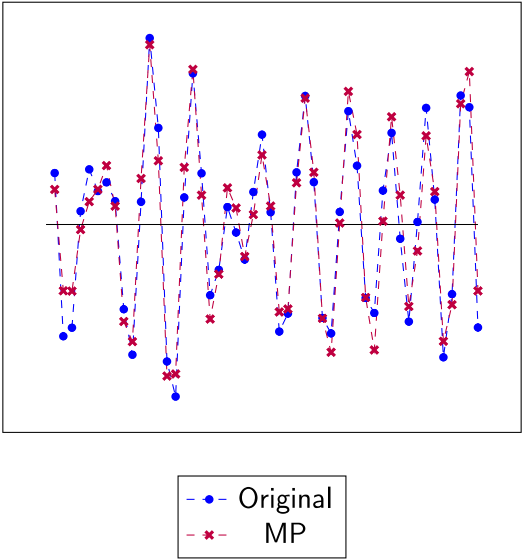
$$\hat{x}^{(k)} = \hat{x}^{(k-1)} + \left\langle r^{(k-1)}, \phi^{(k)} \right\rangle \phi^{(k)}$$

$$r^{(k)} = r^{(k-1)} - \left\langle r^{(k-1)}, \phi^{(k)} \right\rangle \phi^{(k)}$$

Dictionary of sinusoids: Coefficients



Dictionary of sinusoids: Approximation



Orthogonal matching pursuit (OMP)

Makes sure approximation is orthogonal to residual at

Initialization:

$$r^{(0)} = x$$

Iterations: $k = 1, 2, \ldots$

$$\phi^{(k)} = \arg\max_{j} \left| \left\langle r^{(k-1)}, \phi_k \right\rangle \right|$$

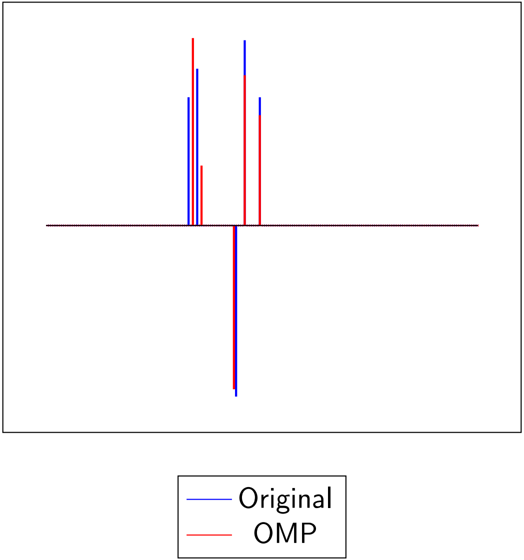
$$A^{(k)} = \begin{bmatrix} \phi^{(1)} & \phi^{(2)} & \cdots & \phi^{(k)} \end{bmatrix}$$

$$\hat{c}^{(k)} = A^{(k)\dagger} x = \left(A^{(k)} A^{(k)\,T} \right)^{-1} A^{(k)\,T} x$$

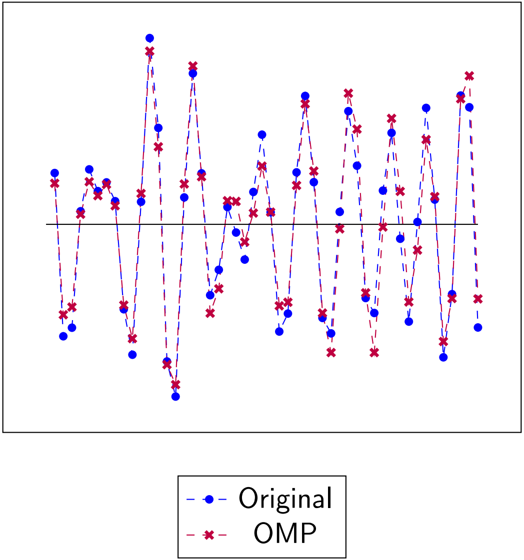
$$\hat{x}^{(k)} = A^{(k)} \hat{c}^{(k)}$$

$$r^{(k)} = x - \hat{x}^{(k)}$$

Dictionary of sinusoids: Coefficients



Dictionary of sinusoids: Approximation



ℓ_1-norm minimization

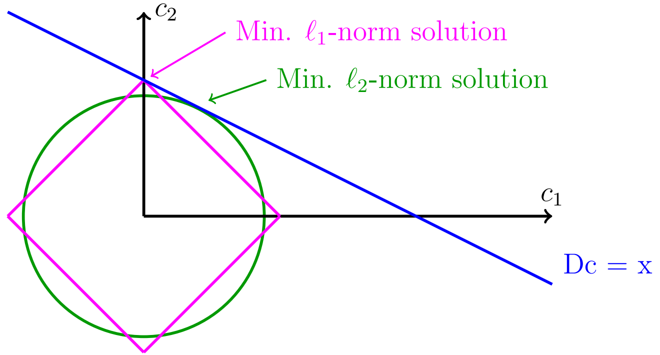
Estimate coefficients by solving

$$\begin{aligned} \text{minimize} \quad & \|\tilde{c}\|_1 \\ \text{subject to} \quad & x = D\,\tilde{c} \end{aligned}$$

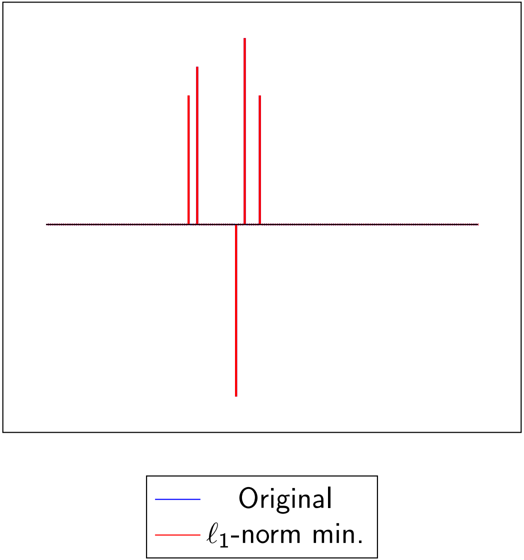
Computationally tractable (convex program)

Known as basis pursuit in the literature

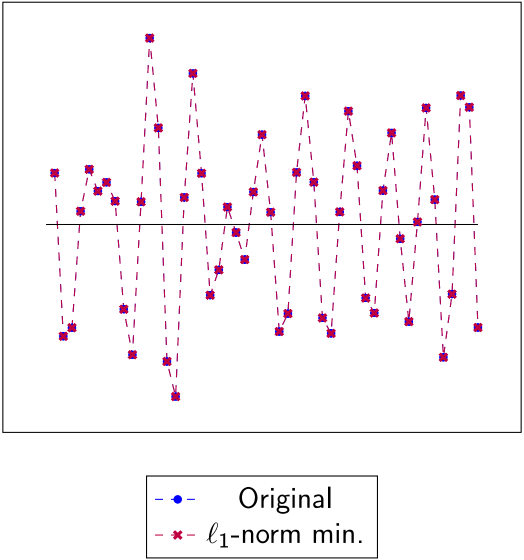
Geometric intuition



Dictionary of sinusoids: Coefficients



Dictionary of sinusoids: Approximation



Denoising

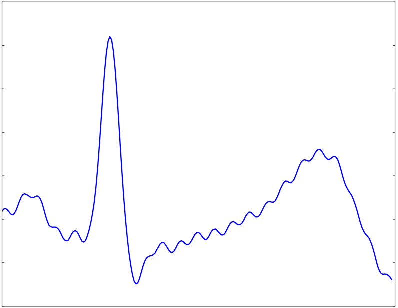
Aim: Extracting information (signal) from data in the presence of uninformative perturbations (noise)

Additive noise model

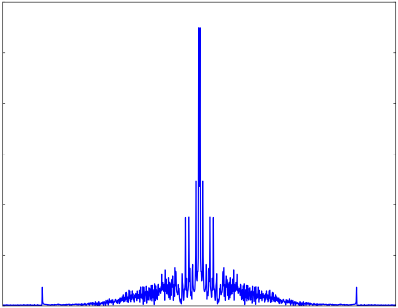
$$\text{data} = \text{signal} + \text{noise}$$
$$y = x + z$$

Prior knowledge about structure of signal vs structure of noise is required

Electrocardiogram

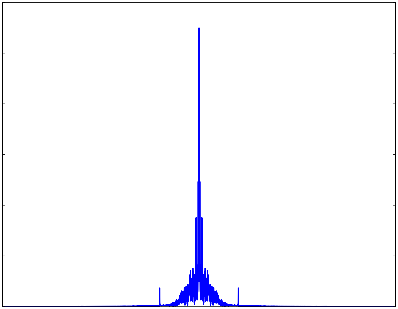


Spectrum

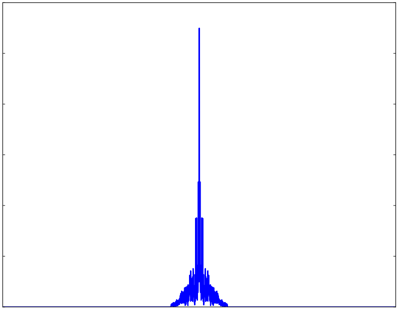


Electrocardiogram: High-frequency noise (power line hum)

Original spectrum

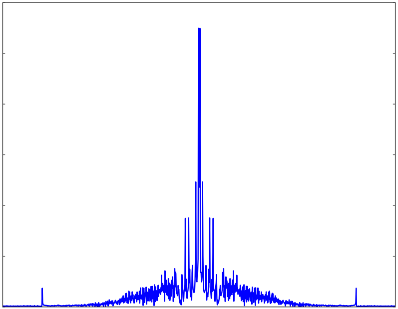


Low-pass filtered spectrum

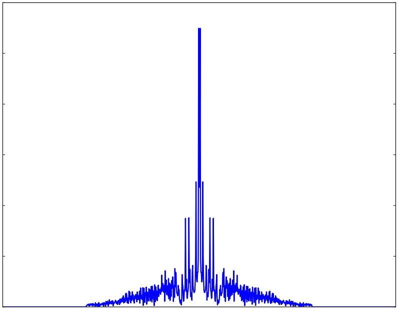


Electrocardiogram: High-frequency noise (power line hum)

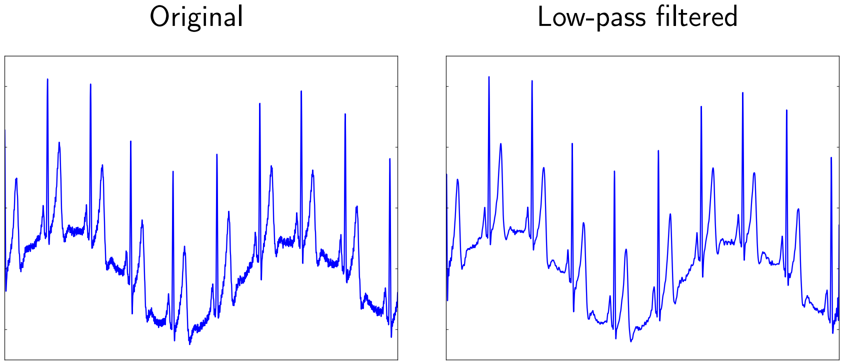
Original spectrum



Low-pass filtered spectrum

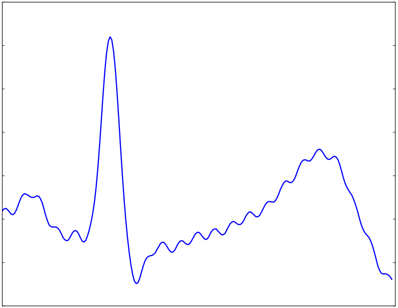


Electrocardiogram: High-frequency noise (power line hum)

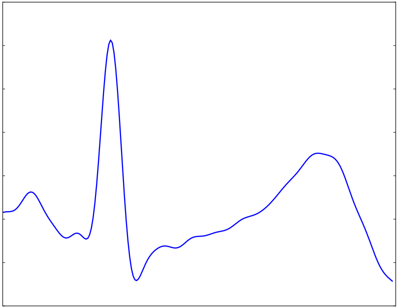


Electrocardiogram: High-frequency noise (power line hum)

Original



Low-pass filtered



Thresholding

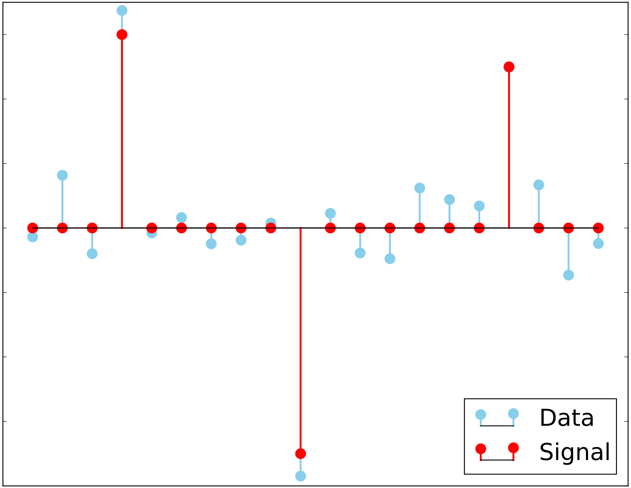
Prior knowledge:

- Signal is a sparse superposition of dictionary atoms
- Noise is not (incoherence between atoms and noise)

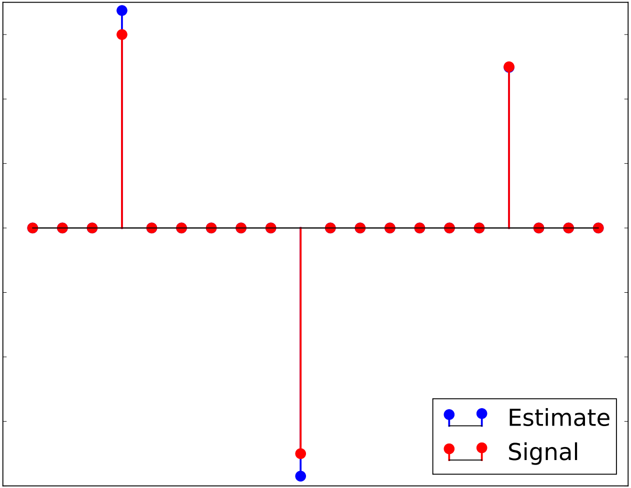
Hard-thresholding operator

$$\mathcal{H}_\eta(x)_i := \begin{cases} x_i & \text{if } |x_i| > \eta, \\ 0 & \text{otherwise} \end{cases}$$

Denoising via thresholding



Denoising via thresholding



Sparsity in a basis

Assumption: $x = Bc$, where c is sparse

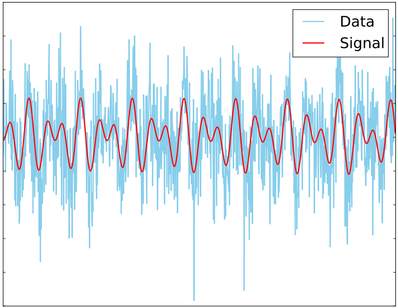
Threshold $B^{-1}y$

$$\begin{aligned} \hat{c} &= \mathcal{H}_\eta\left(B^{-1}y\right) \\ &= \mathcal{H}_\eta\left(c + B^{-1}z\right) \\ \hat{y} &= B\hat{c} \end{aligned}$$

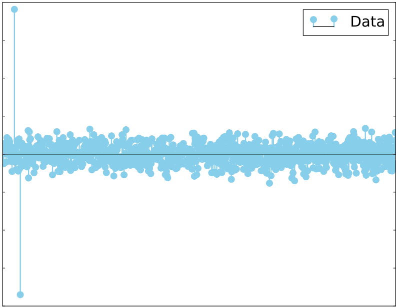
Noise and sparsifying atoms should be incoherent, i.e. $B^{-1}z$ is not sparse

Example: Orthogonal sparsifying basis and Gaussian noise

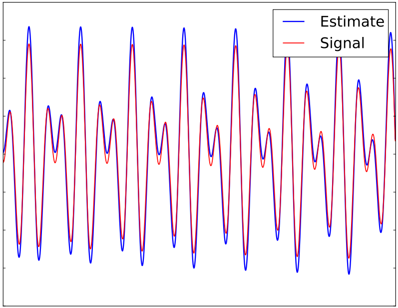
Denoising via thresholding in DCT basis



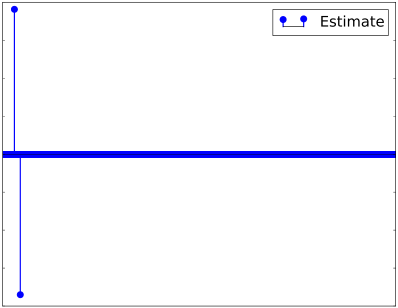
DCT coefficients



Denoising via thresholding in DCT basis

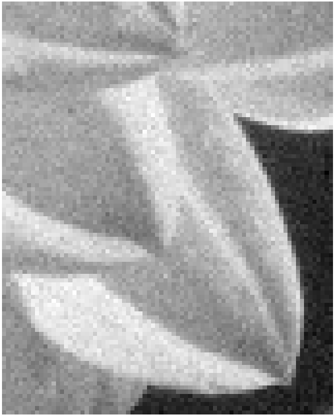


DCT coefficients

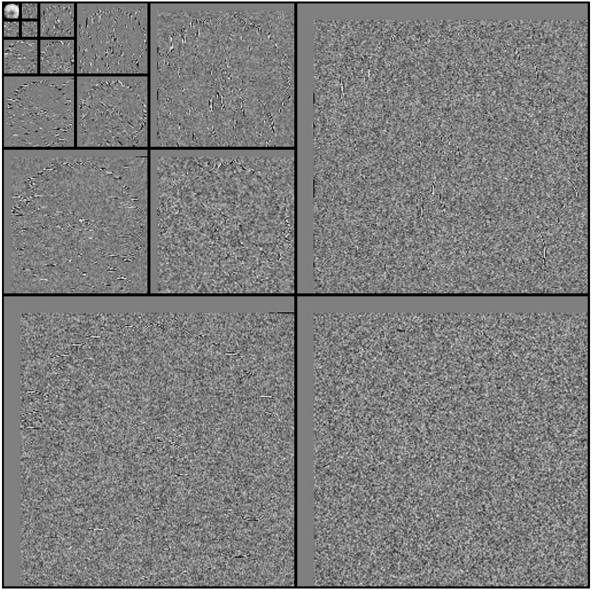


Denoising via thresholding in a wavelet basis

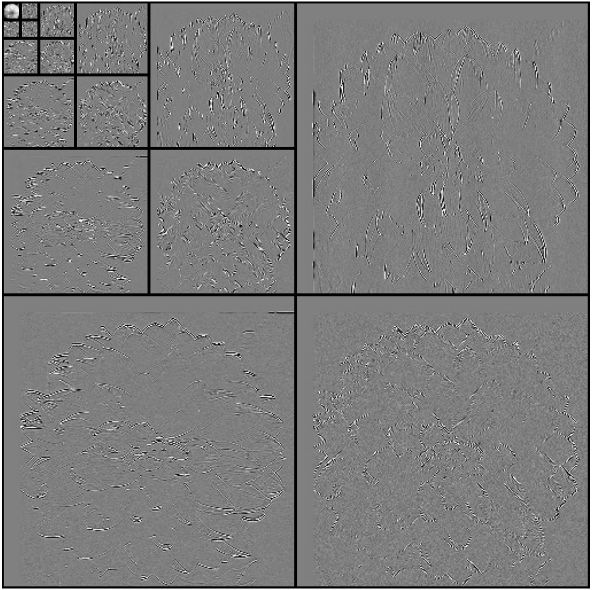
Denoising via thresholding in a wavelet basis



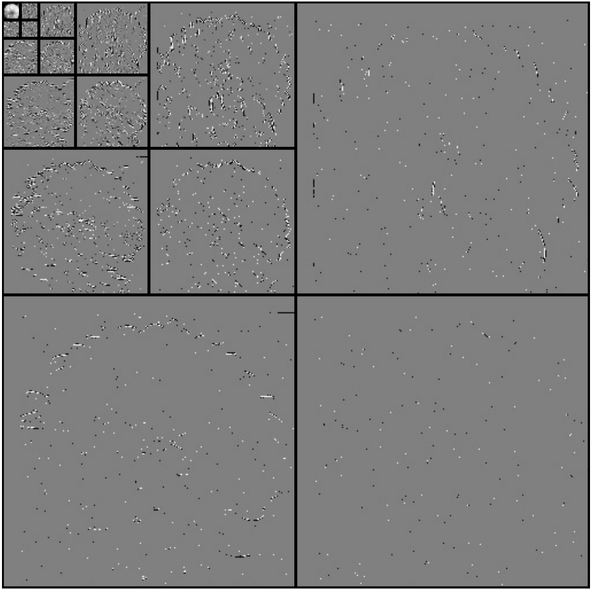
2D wavelet coefficients



Original coefficients



Thresholded coefficients



Estimate

Estimate

Denoising via thresholding in a wavelet basis

Original

Noisy

Estimate

Analysis model

Assumption: $D^T x$ is sparse

Threshold, then use left inverse of D^T L

$$\begin{aligned}
\hat{c} &= \mathcal{H}_\eta \left(D^T y \right) \\
&= \mathcal{H}_\eta \left(D^T x + D^T z \right) \\
\hat{y} &= L \hat{c}
\end{aligned}$$

Example: Thresholding STFT coefficients for speech denoising

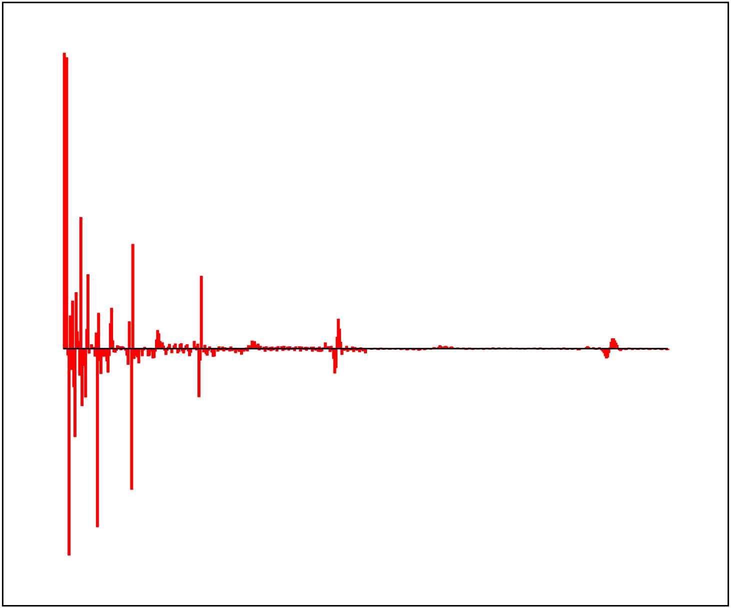
Block thresholding

Assumption: Coefficients are group sparse, nonzero coefficients cluster together

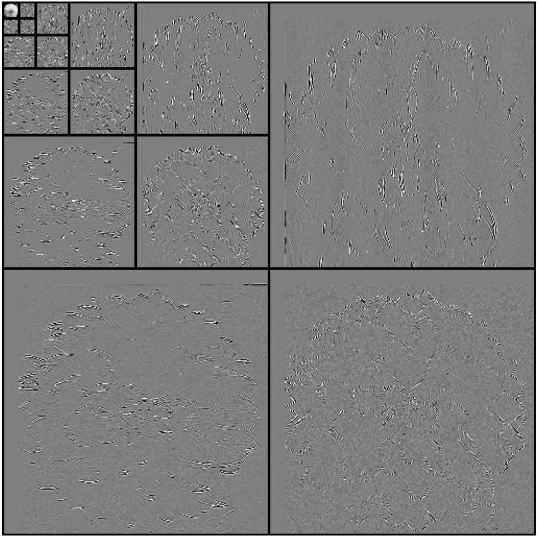
Block thresholding: Partition coefficients into blocks $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_k$ and threshold whole blocks

$$\mathcal{B}_\eta(x)_i := \begin{cases} x_i & \text{if } i \in \mathcal{I}_j \quad \text{such that } \left|\left|x_{\mathcal{I}_j}\right|\right|_2 > \eta, \\ 0 & \text{otherwise} \end{cases}$$

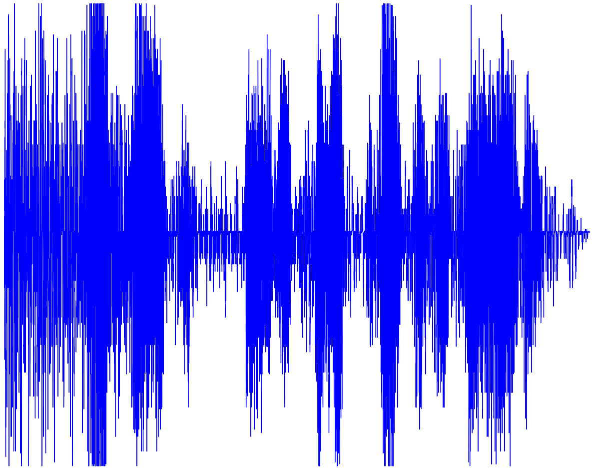
Haar transform



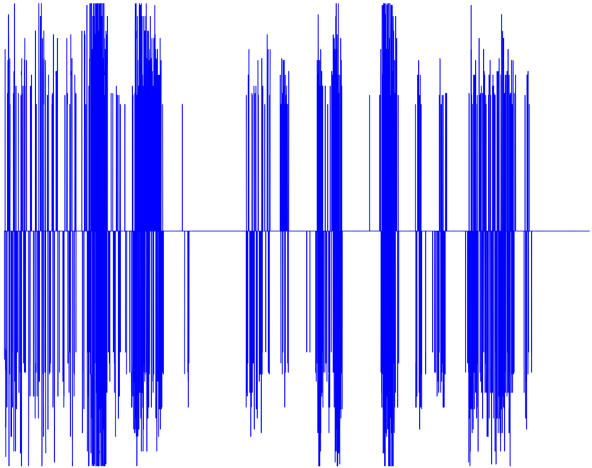
2D wavelet transform



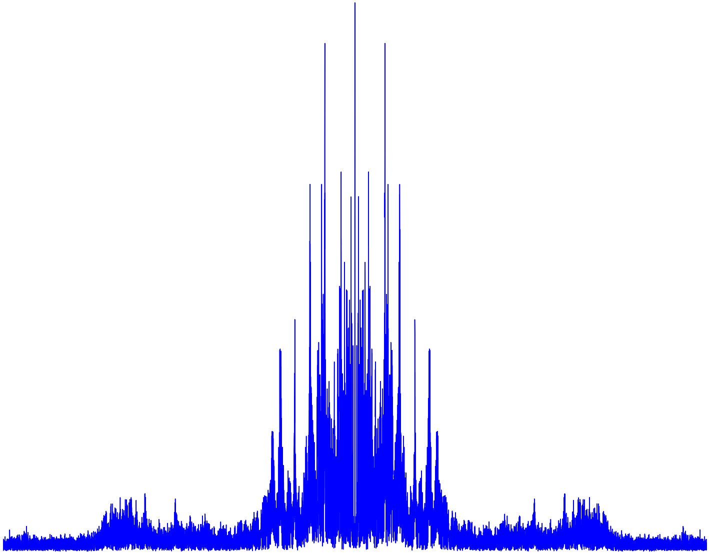
Speech denoising



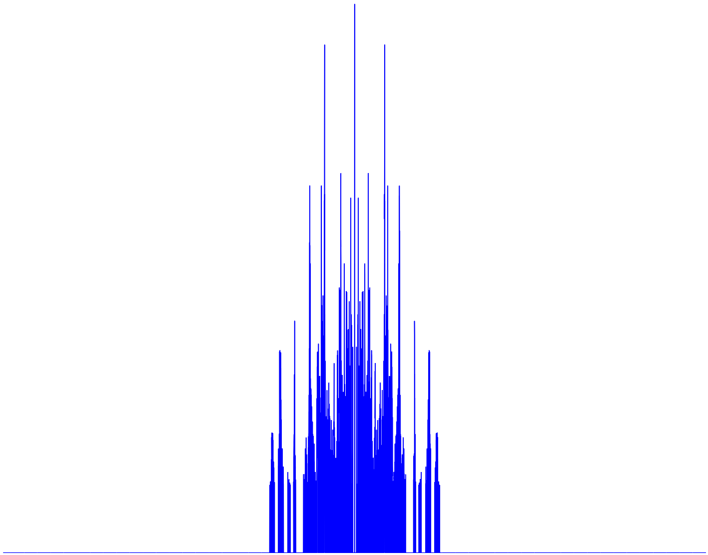
Time thresholding



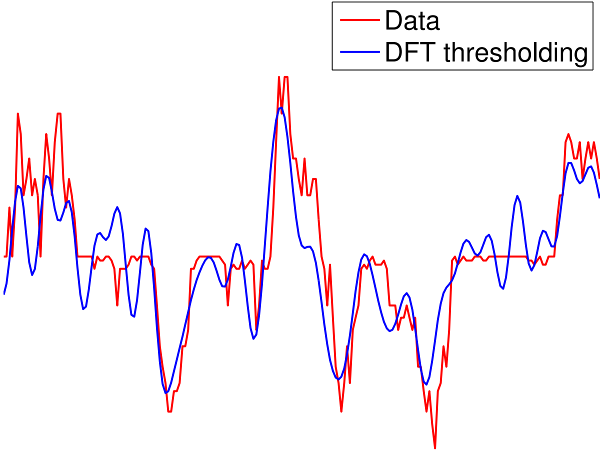
Spectrum



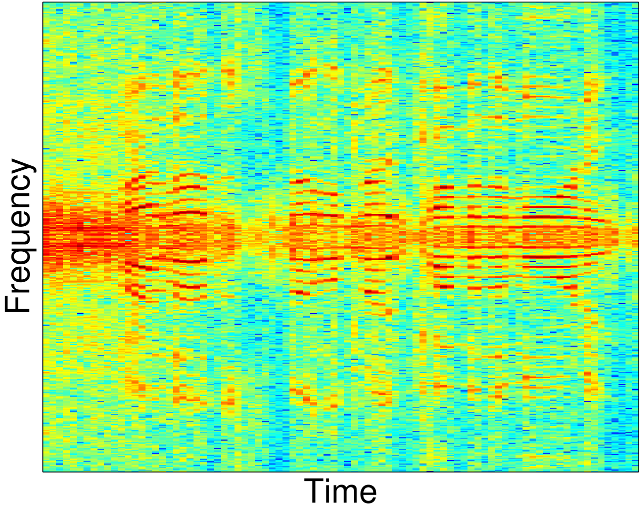
Frequency thresholding



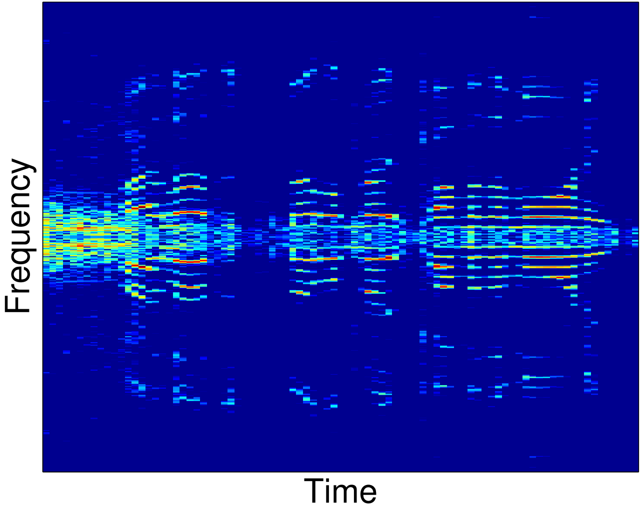
Frequency thresholding



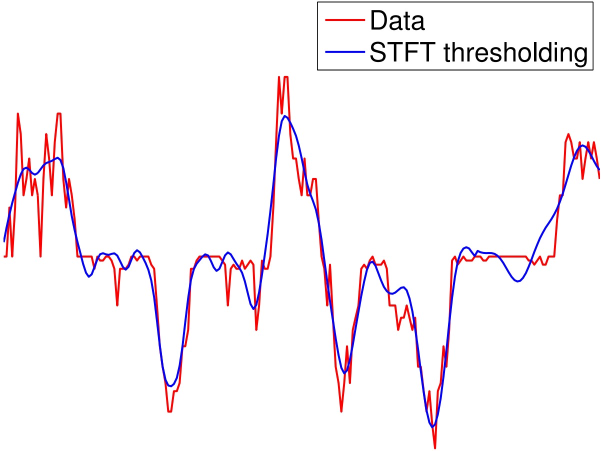
Spectrogram (STFT)



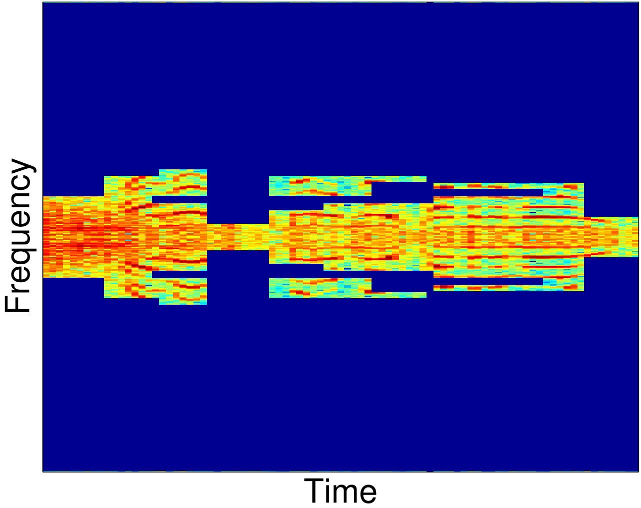
STFT thresholding



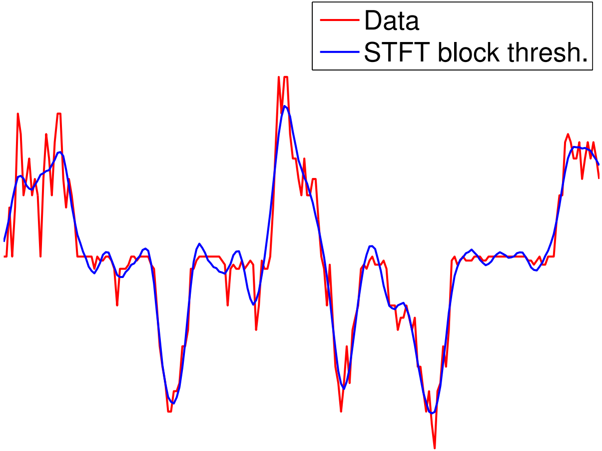
STFT thresholding



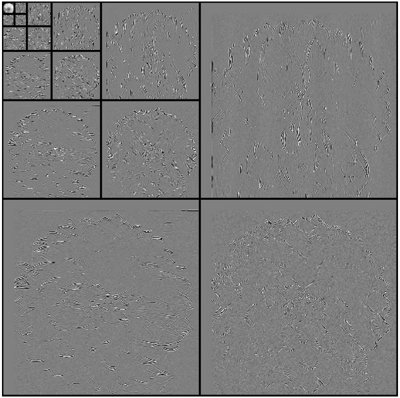
STFT block thresholding



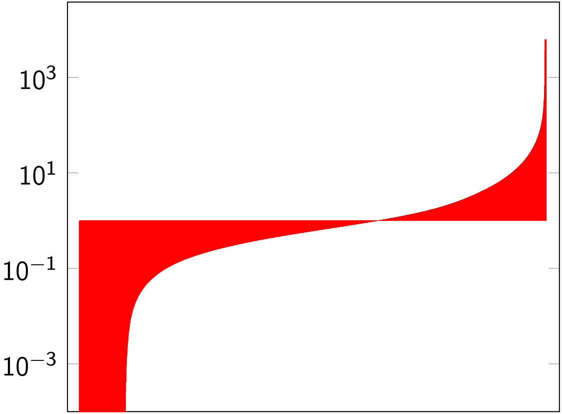
STFT block thresholding



Wavelets



Sorted wavelet coefficients



Denoising via ℓ_1-norm regularized least squares

Synthesis sparse model

$$x = Dc \quad \text{where } c \text{ is sparse}$$

We would like to solve

$$\begin{aligned} &\text{minimize} && ||\tilde{c}||_0 \\ &\text{subject to} && y \approx D\,\tilde{c} \end{aligned}$$

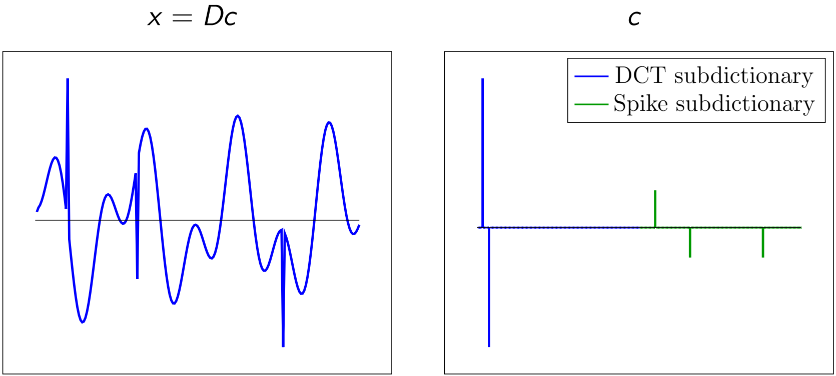
Computationally intractable if D is overcomplete

Basis-pursuit denoising

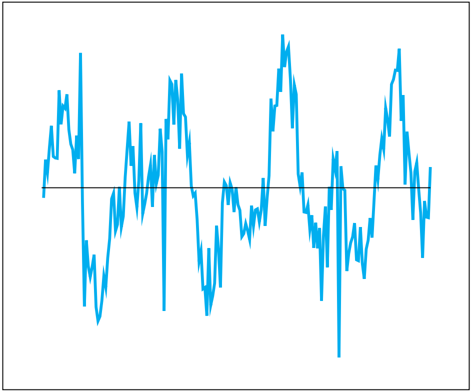
$$\hat{c} = \arg\min_{\tilde{x} \in \mathbb{R}^m} ||y - D\tilde{c}||_2^2 + \lambda\,||\tilde{c}||_1$$

$$\hat{x} = D\hat{c}$$

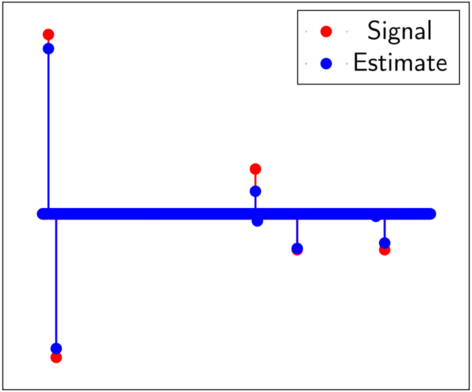
Sines and spikes



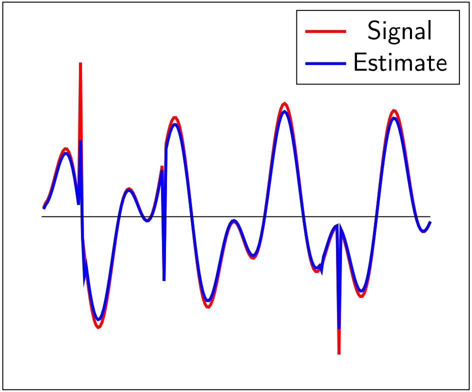
Denoising via ℓ_1-norm regularized least squares



Denoising via ℓ_1-norm regularized least squares



Denoising via ℓ_1-norm regularized least squares



Denoising via ℓ_1-norm regularized least squares

Analysis sparse model

$$D^T x \quad \text{is sparse}$$

We would like to solve

$$\begin{aligned} \text{minimize} \quad & \left|\left|D^T \tilde{x}\right|\right|_0 \\ \text{subject to} \quad & y \approx \tilde{x} \end{aligned}$$

Computationally intractable if D is overcomplete

Instead, we solve

$$\hat{x} = \arg\min_{\tilde{x} \in \mathbb{R}^m} ||y - \tilde{x}||_2^2 + \lambda \left|\left|A^T \tilde{x}\right|\right|_1$$

Significantly more challenging to solve than synthesis formulation

Total variation

Images and some time series tend to be piecewise constant

Equivalently: Sparse gradient (or derivative)

The total variation of an image I is the ℓ_1-norm of the horizontal and vertical components of the gradient

$$\mathrm{TV}(I) := \|\nabla_x I\|_1 + \|\nabla_y I\|_1$$

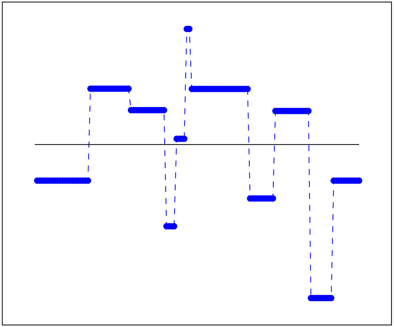
Equivalent to ℓ_1-norm regularization with an overcomplete analysis operator

Denoising via TV regularization

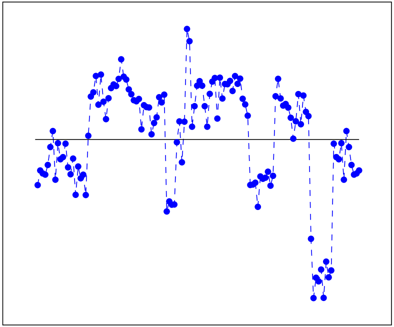
$$\hat{I} = \arg\min_{\tilde{I} \in \mathbb{R}^{n \times n}} \left\| Y - \tilde{I} \right\|_F^2 + \lambda \, \mathrm{TV}\left(\tilde{I}\right)$$

Denoising via TV regularization

Signal

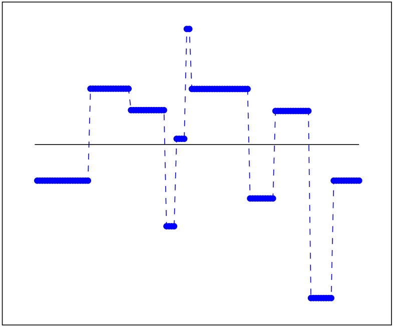


Data

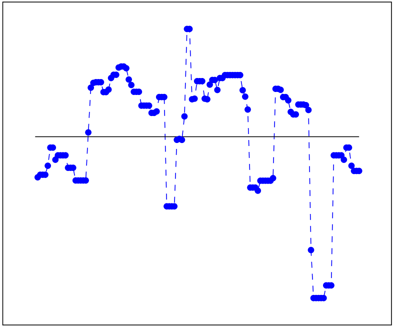


Denoising via TV regularization

Signal

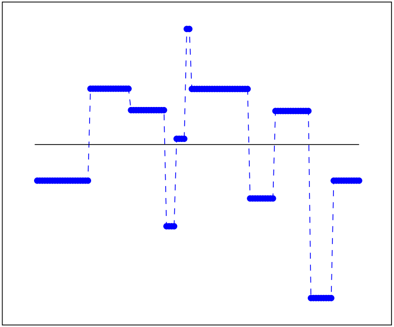


TV reg. (small λ)

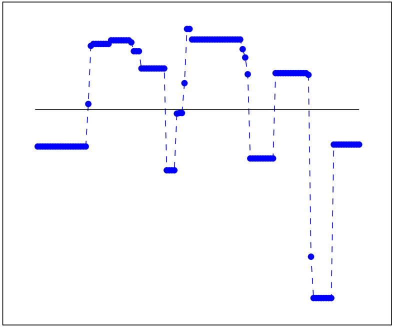


Denoising via TV regularization

Signal

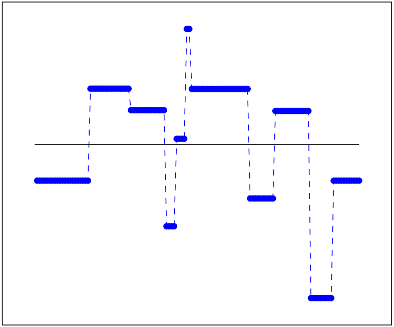


TV reg. (medium λ)

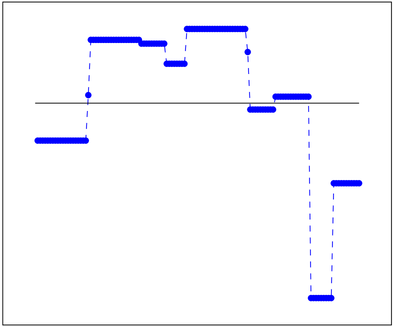


Denoising via TV regularization

Signal



TV reg. (large λ)



Denoising via TV regularization

Denoising via TV regularization

Small λ

Small λ

Medium λ

Medium λ

Large λ

Large λ

Denoising via TV regularization

Original

Noisy

Estimate