Computational Intelligence Laboratory

Lecture 9

Sparse Coding

Thomas Hofmann

ETH Zurich – cil.inf.ethz.ch

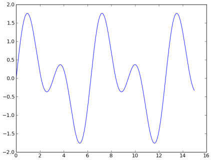
May 12, 2017

Section 1

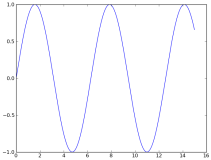
Sparse Coding

Sparse Coding

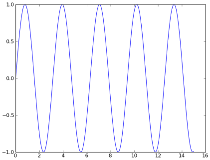
- Signals can be represented in different ways
  - infinite number of possible representations
  - each capturing different characteristics
  - example: Fourier series



$=$



$+$



Sparse Coding

- Natural signals often allow for sparse representation
  - sparsity: many coefficients vanish (≈ 0)
  - due to regularity of signal
  - need to find suitable dictionary $\mathcal{U} = \{\mathbf{u}_1, \dots, \mathbf{u}_L\}$
  - such that accurate signal representation in $\mathrm{span}(\mathcal{U})$

Sparse Coding in Image Compression

- Recall: SVD Compression
  - an instance of sparse coding (Lecture 3)
  - dictionary: rank 1 matrices (left/right singular vector)
- Today: Fourier and wavelet basis
  - fixed orthogonal basis, efficient to compute.

Signal Compression

- Given original signal $\mathbf{x} \in \mathbb{R}^D$ and orthogonal matrix $\mathbf{U}$
- Compute linear transformation = change of basis

$$\mathbf{z} = \underbrace{\mathbf{U}^\top}_{D \times D} \cdot \mathbf{x}$$

- Energy preserving

$$\|\mathbf{U}^\top \mathbf{x}\|^2 = \|\mathbf{x}\|^2$$

  - direct consequence of orthogonality
  - preservation of length

Signal Compression

- Truncate "small" values of $\mathbf{z} \Rightarrow$ estimate $\hat{\mathbf{z}}$
  - encoding only $K \ll D$ non-zero values: compression
  - for instance: employ a threshold ϵ

$$\hat{z}_d = \begin{cases} 0 & \text{if } |z_d| < \epsilon \\ z_d & \text{otherwise} \end{cases}$$

- Reconstruct signal through inverse transform

$$\hat{\mathbf{x}} = \mathbf{U}\hat{\mathbf{z}}, \quad \text{as} \quad \mathbf{U} = (\mathbf{U}^\top)^{-1}$$

  - efficient inversion via transposition
  - key idea: *orthogonality* of $\mathbf{U}$

Decomposition and Reconstruction

- Given $\mathbf{x}$, orthonormal basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_D\}$ (columns of $\mathbf{U}$)

$$\mathbf{x} = \sum_{d=1}^{D} z_d(\mathbf{x}) \cdot \mathbf{u}_d, \quad z_d(\mathbf{x}) := \langle \mathbf{x}, \mathbf{u}_d \rangle$$

- Sparsification $\equiv$ only use K-subset σ of basis functions

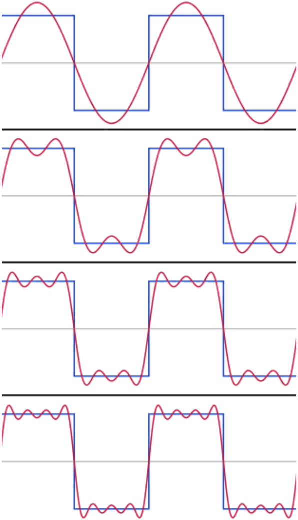
$$\hat{\mathbf{x}} = \sum_{d \in \sigma} z_d(\mathbf{x}) \cdot \mathbf{u}_d$$

- Reconstruction error:

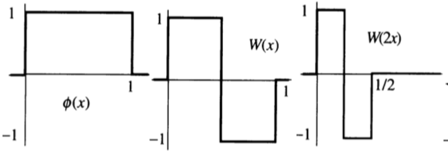
$$\|\mathbf{x} - \hat{\mathbf{x}}\|^2 = \sum_{d \notin \sigma} \|\langle \mathbf{x}, \mathbf{u}_d \rangle \cdot \mathbf{u}_d\|^2 = \sum_{d \notin \sigma} \langle \mathbf{x}, \mathbf{u}_d \rangle^2$$

1-D signal processing

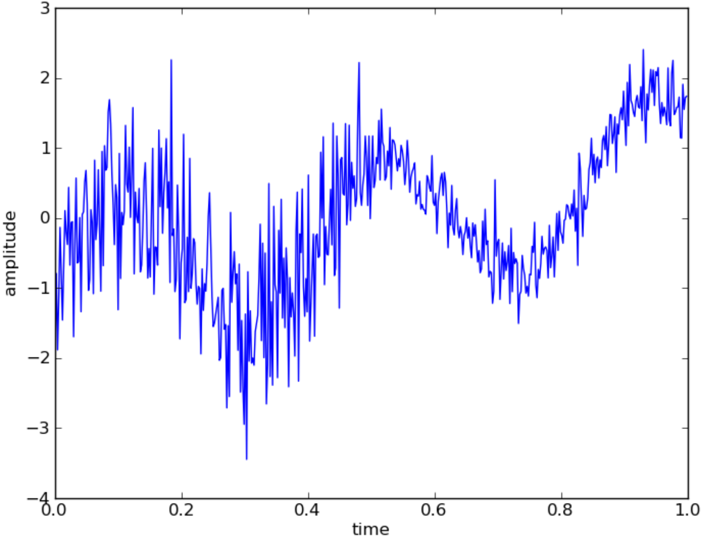
Discrete Fourier Transform



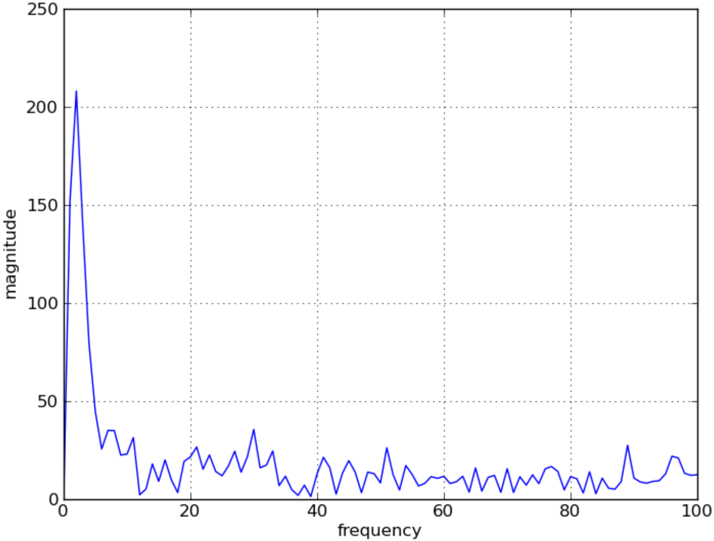
Discrete Wavelet Transform



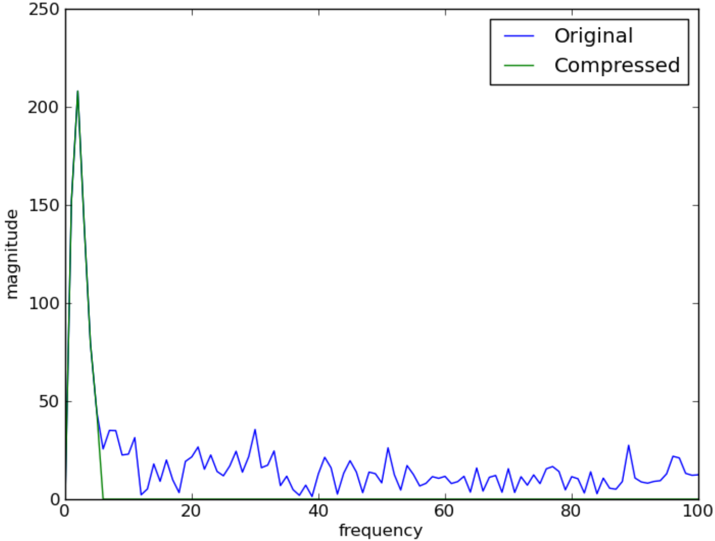
Noisy signal: $\mathbf{x}$



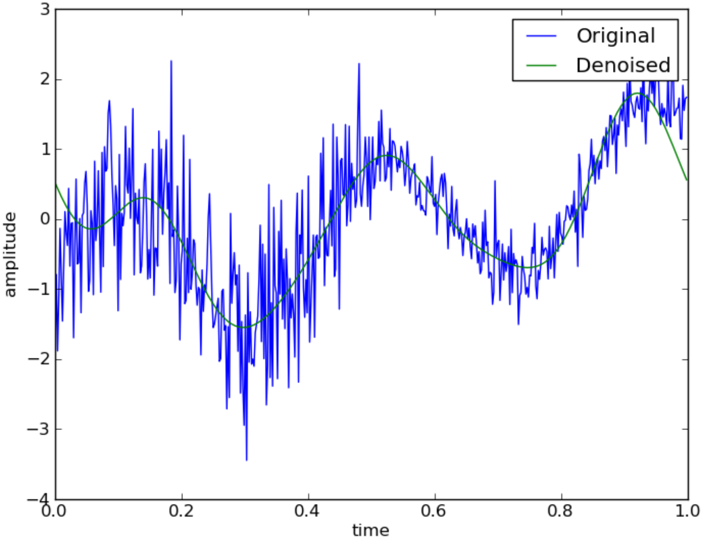
Fourier spectrum: $\mathbf{z} = \mathbf{U}^\top \mathbf{x}$



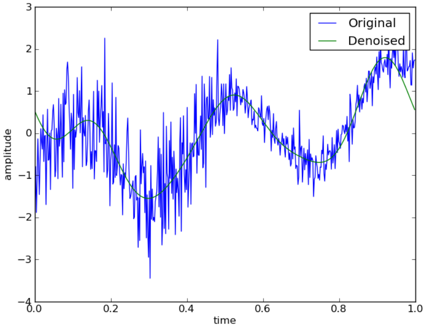
Retain 3% of the coefficients: $\hat{\mathbf{z}}$



Denoised signal: $\hat{\mathbf{x}} = \mathbf{U}\hat{\mathbf{z}}$

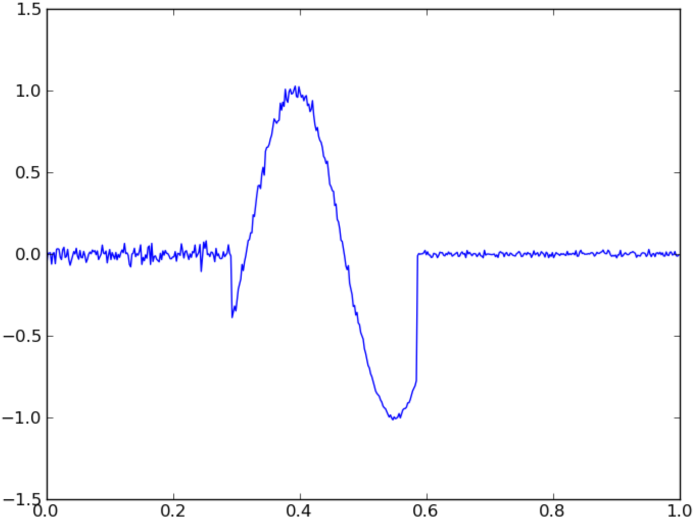


Signal Compression: Observations

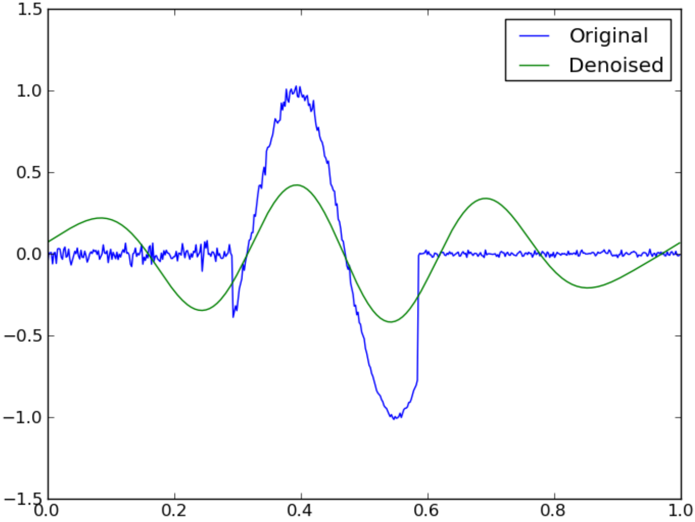


- Signal is compressed by 97%.
- High signal frequencies have small amplitudes in spectrum
- Reconstructed signal is smoother than the original one (low-pass filter)

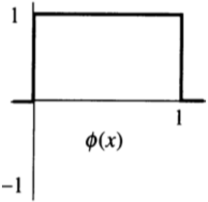
Challenge: Localized signal



Challenge: Poor denoising of localized signal

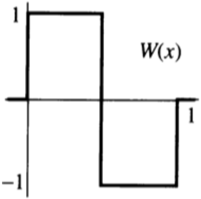


Haar Wavelets



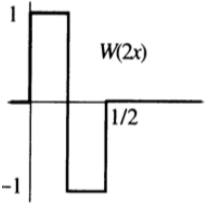
$$\begin{bmatrix}1\\1\\1\\1\end{bmatrix}$$

scaling function



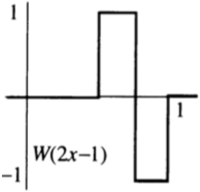
$$\begin{bmatrix}1\\1\\-1\\-1\end{bmatrix}$$

mother wavelet



$$\begin{bmatrix}1\\-1\\0\\0\end{bmatrix}$$

dilated



$$\begin{bmatrix}0\\0\\1\\-1\end{bmatrix}$$

translated

Note that the wavelet basis is *orthogonal*

Haar Wavelets – $D = 4$

- For $D = 4$ we get the following orthogonal matrix

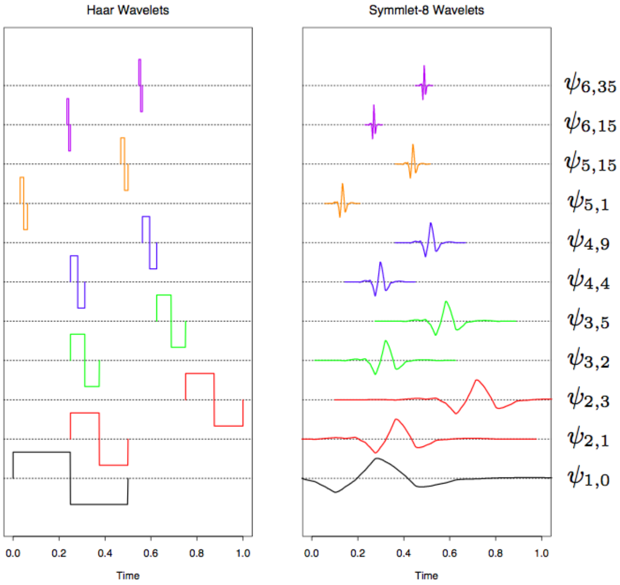
$$\mathbf{U} = \frac{1}{2}\begin{pmatrix} 1 & 1 & \sqrt{2} & 0 \\ 1 & 1 & -\sqrt{2} & 0 \\ 1 & -1 & 0 & \sqrt{2} \\ 1 & -1 & 0 & -\sqrt{2} \end{pmatrix}$$

Haar Wavelets – $D = 8$

- For $D = 8$ we get the following orthogonal matrix

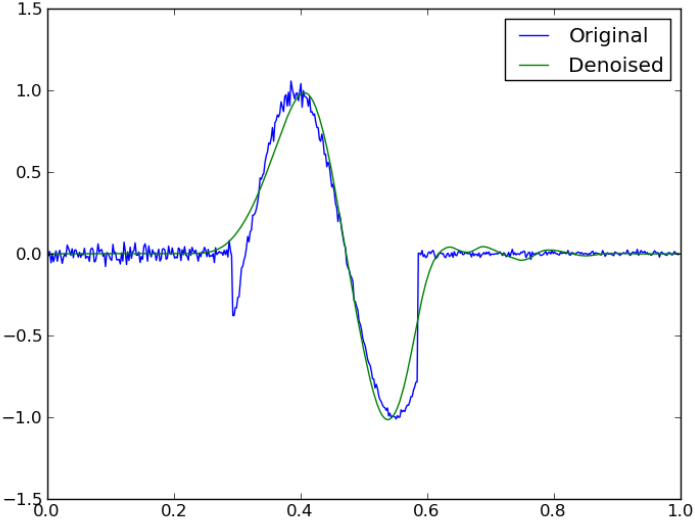
$$\mathbf{U} = \frac{1}{2\sqrt{2}} \begin{pmatrix} 1 & 1 & \sqrt{2} & 0 & 2 & 0 & 0 & 0 \\ 1 & 1 & \sqrt{2} & 0 & -2 & 0 & 0 & 0 \\ 1 & 1 & -\sqrt{2} & 0 & 0 & 2 & 0 & 0 \\ 1 & 1 & -\sqrt{2} & 0 & 0 & -2 & 0 & 0 \\ 1 & -1 & 0 & \sqrt{2} & 0 & 0 & 2 & 0 \\ 1 & -1 & 0 & \sqrt{2} & 0 & 0 & -2 & 0 \\ 1 & -1 & 0 & -\sqrt{2} & 0 & 0 & 0 & 2 \\ 1 & -1 & 0 & -\sqrt{2} & 0 & 0 & 0 & -2 \end{pmatrix}$$

Wavelets

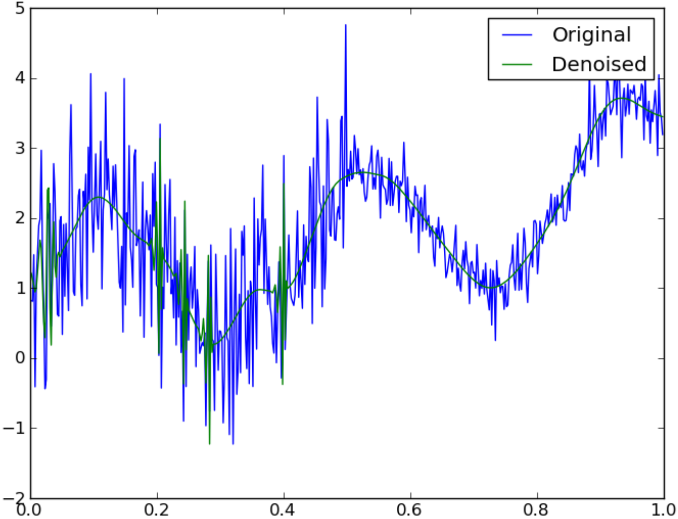


Haar Wavelets
Symmlet-8 Wavelets
0.0
0.2
0.4
0.6
0.8
1.0
Time
0.0
0.2
0.4
0.6
0.8
1.0
Time
$\psi_{6,35}$
$\psi_{6,15}$
$\psi_{5,15}$
$\psi_{5,1}$
$\psi_{4,9}$
$\psi_{4,4}$
$\psi_{3,5}$
$\psi_{3,2}$
$\psi_{2,3}$
$\psi_{2,1}$
$\psi_{1,0}$

Wavelet denoising of localized signal



Wavelet denoising of smooth signal

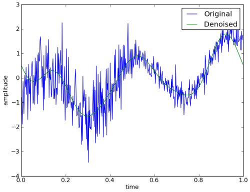


Fourier basis vs Wavelet basis

A priori, there does not exist a choice of a transform that is better than all other choices. It depends on the signal type.

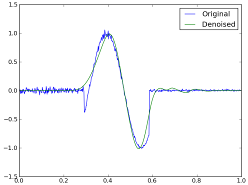
Fourier basis

- Global support
- Good for "sine like" signals
- Poor for localized signal



Wavelet basis

- Local support
- Good for localized signal
- Poor for non-vanishing signals



Principal Component Analysis

- Given $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_N]$ vectors in $\mathbb{R}^D$
- Mean: $\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n$
- Compute centered covariance matrix

$$\mathbf{\Sigma} = \frac{1}{N}(\mathbf{X}-\mathbf{M})(\mathbf{X}-\mathbf{M})^\top, \quad \mathbf{M} := [\underbrace{\bar{\mathbf{x}} \ \dots \ \bar{\mathbf{x}}}_{N \text{ times}}]$$

- Compute eigenvector decomposition

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$$

  - $\mathbf{\Sigma}$: real symmetric matrix, $\mathbf{U}$: orthogonal
  - eigenvalues ordered: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_D$

Principal Component Analysis (cont'd)

- Karhunen-Loeve transform or Hoteling transform
  - "throw away" the $D - K$ directions with smallest variance (dependent on signal set, not individual signal)
  - equivalently: keep K largest eigenvectors

$$\hat{\mathbf{x}} = \mathbf{U}\hat{\mathbf{z}}, \quad \hat{z}_d = \begin{cases} z_d & \text{if } d \leq K \\ 0 & \text{otherwise} \end{cases}$$

  - suffices to define $\mathbf{U}_K$ as

$$\mathbf{U}_K := [\mathbf{u}_1 \cdots \mathbf{u}_K]$$

and to reconstruct via

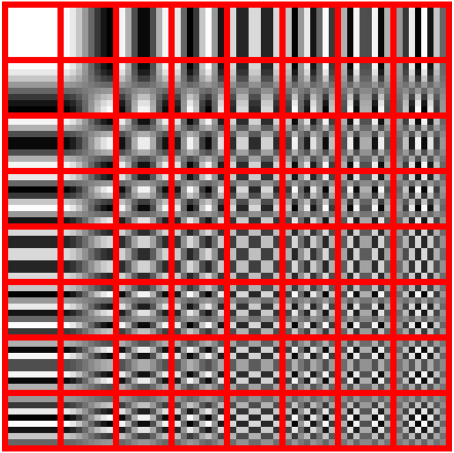
$$\hat{\mathbf{x}} = \mathbf{U}_K \, \mathbf{z}_{[1:K]}$$

Communication Cost

PCA basis
- $\mathbf{U}_K$ is data-dependent, optimal for given $\mathbf{\Sigma}$
- Transmit: eigenvectors $\{\mathbf{u}_d : d \leq K\}$ and $\mathbf{z}_{1:K}$.

Fixed basis
- Sender and receiver agree on basis beforehand, e.g. Haar Wavelets.
- Transmit: non-zero elements of $\hat{\mathbf{z}}$.

2-D Discrete cosine transform



- in JPEG, DCT is applied to 8x8 blocks of an image.
- further optimizations to improve compression.

2-D Discrete cosine transform

- Attention: think of each 8×8 patch as a $D = 64$ vector
- Basis functions are $D = 64$ vectors that can also be displayed as 8×8 patches
- There are 64 basis functions, which can be arranged on a 8×8 grid!
- Each red square is a basis function!

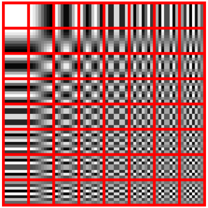
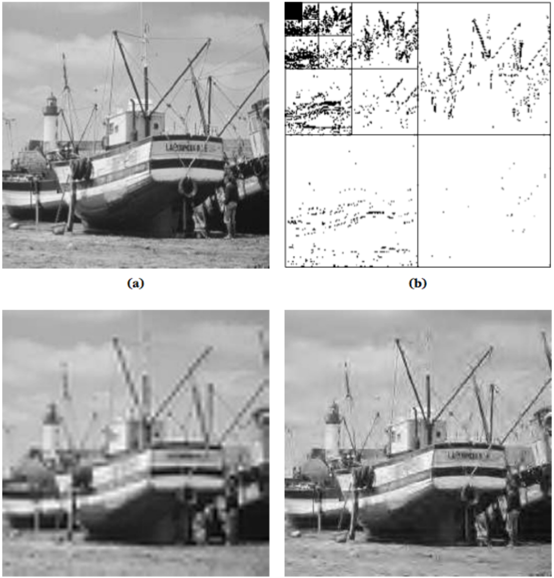


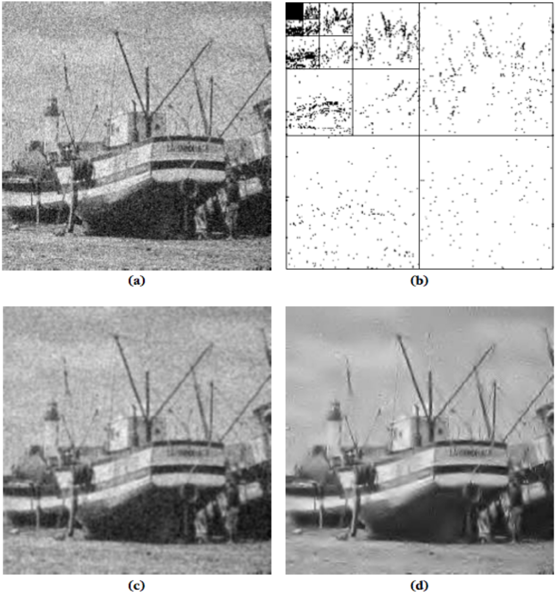
Image compression with wavelets



(a) (b) (c) (d)

(a) Discrete image of 256^2 pixels.
(b) Orthogonal wavelet coefficients at 4 different scales; black points correspond to large coefficients.
(c) Approximation using the three largest scales.
(d) Approximation using the K largest coefficients ($K = \frac{256^2}{16}$).

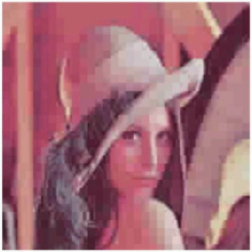
Image denoising with wavelets



(a) Noisy image.
(b) Orthogonal wavelet coefficients at 4 different scales; black points correspond to large coefficients.
(c) Approximation using the three largest scales.
(d) Approximation using the K largest coefficients ($K = \frac{256^2}{16}$).

Image compression

Original Lena Image (256 x 256 Pixels, 24-Bit RGB)



JPEG Compressed (Compression Ratio 43:1)

JPEG2000 Compressed (Compression Ratio 43:1)

Computational Efficiency

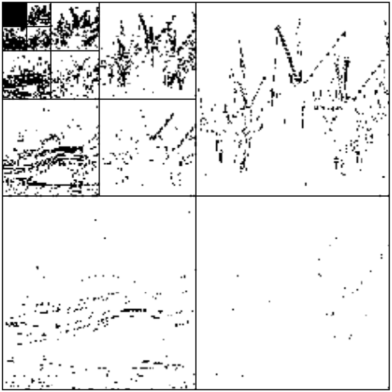
- Basis transform via matrix multiplication $= \mathcal{O}(D^2)$ cost
- In practice: exploit fast transforms
  - Fourier: $\mathcal{O}(D \log D)$
  - Wavelet: $\mathcal{O}(D)$ or $\mathcal{O}(D \log D)$
- Image compression:
  - break-up images into blocks, transform each block
  - avoids quadratic blow-up
  - for example JPEG: DCT on 8x8 blocks

Section 2

Overcomplete Dictionaries

Sparse Representations

Summary: Natural signals have approx. sparse representations in suitable orthogonal bases, e.g. wavelets for natural images.



From *S. Mallat, A Wavelet Tour of Signal Processing – The Sparse Way, Academic Press, 2009*

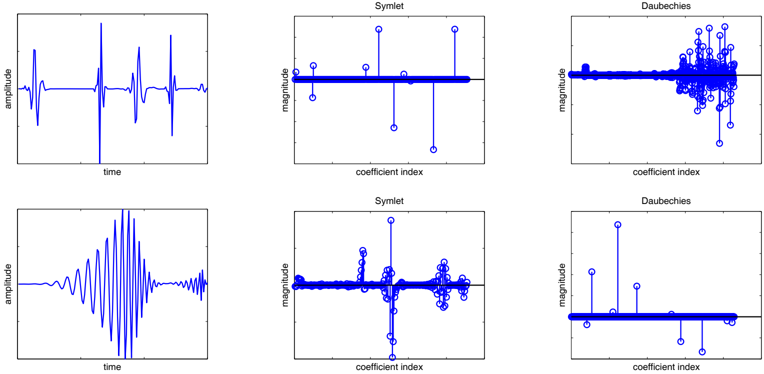
Recall so far...

- Coding via orthogonal transforms
  - given: signal $\mathbf{x}$ and orthonormal matrix $\mathbf{U}$
  - compute linear transformation (change of basis) $\mathbf{z} = \mathbf{U}^\top \mathbf{x}$
  - truncate "small" values, $\mathbf{z} \mapsto \hat{\mathbf{z}}$.
  - compute inverse transform (recall $\mathbf{U}^{-1} = \mathbf{U}^\top$) $\hat{\mathbf{x}} = \mathbf{U}\hat{\mathbf{z}}$.
- Measuring Accuracy
  - reconstruction error $\|\mathbf{x} - \hat{\mathbf{x}}\|$
  - sparsity of the coding vector $\hat{\mathbf{z}}$
- Dictionary choice
  - Fourier dictionary is good for "sine like" signals.
  - wavelet dictionary is good for localized signals.
  - more general dictionaries: overcomplete dictionaries...

Overcomplete Dictionaries

- Beyond a "change of basis"
  - no single basis is optimally sparse for all signal classes
  - overcompleteness ($\mathbf{U} \in \mathbb{R}^{D \times L}$ such that $L > D$): more atoms (dictionary elements) than dimensions
  - union of orthogonal bases and general overcomplete dictionaries: coding algorithm chooses best representation.
  - decoding: involved, no closed form reconstruction formula

Morphology of Signals I



Dictionary selection strategy:

- Manually, by signal inspection
- Try several, choose the one which affords sparsest coding

Morphology of Signals II

From *S. Mallat, A Wavelet Tour of Signal Processing – The Sparse Way, Academic Press, 2009*

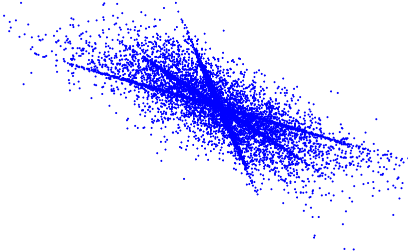
Signal might be a superposition of several characteristics:

- smooth gradients plus oscillating texture
- hence: single orthonormal basis cannot sparsely code both.

Coding idea: Algorithm picks *atoms* (dictionary elements) from a *union of bases*, each one responsible for one characteristic.

General Overcomplete Dictionaries

- Consider data set $\{\mathbf{x}_1, \ldots, \mathbf{x}_{10000}\} \in \mathbb{R}^3$:



- Full coding $(K = 3)$ in spanning basis $\mathbf{U} \in \mathbb{R}^{3\times 3}$
- $K = 2$ coding possible using a four atom dictionary

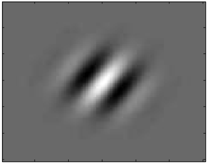
$$\tilde{\mathbf{U}} = [\mathbf{u}_1 \, \mathbf{u}_2 \, \mathbf{u}_3 \, \mathbf{u}_4] \in \mathbb{R}^{3\times 4}$$

  aligned with densely populated subspaces.
- $L > D$ atoms are no longer linearly independent.

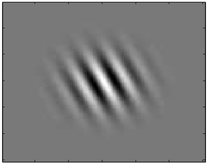
Example: Directional Gabor Wavelets

- Gabor wavelets
  - directional oscillation
  - amplitude modulated by Gaussian window

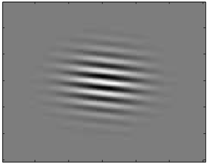
$$g\left(n_1, n_2; \mu_1, \mu_2, f, \theta\right) \propto \exp\left[-\left(n_1-\mu_1\right)^2\right] \exp\left[-\left(n_2-\mu_2\right)^2\right]$$
$$\times \cos\left(f \cdot \left(n_1 \cos\theta + n_2 \sin\theta\right)\right)$$



$(0, 0, 5, 1)$



$(0, 0, 10, 2)$



$(0, 0, 15, 3)$

  - discretizing the parameter range of μ_1, μ_2, f and θ determines the dictionary size, i.e. the overcompleteness factor $\frac{L}{D}$.

Coherence

Increasing the overcompleteness factor $\frac{L}{D}$:

- Increases (potentially) the sparsity of the coding.
- Increases the linear dependency between atoms.

Linear dependency measure for dictionaries: coherence

$$m\left(\mathbf{U}\right) = \max_{i,j:i\neq j} \left|\mathbf{u}_i^\top \mathbf{u}_j\right| .$$

- $m\left(\mathbf{B}\right) = 0$ for an orthogonal basis $\mathbf{B}$.
- $m\left(\left[\mathbf{B}\,\mathbf{u}\right]\right) \geq \frac{1}{\sqrt{D}}$ if atom $\mathbf{u}$ is added to orthogonal $\mathbf{B}$.

Signal Reconstruction (Invertible Dictionary)

$\mathbf{U}$ is orthonormal

- matrix multiplication $\mathbf{x} = \mathbf{U}\mathbf{z}$

$\mathbf{U}$ is spanning basis (D linearly independent atoms)

- $\mathbf{x} = \left(\mathbf{U}^\top\right)^{-1} \mathbf{z}$
- inverting $\mathbf{U}^\top$ can be ill-conditioned

Signal Reconstruction (General Dictionary)

$\mathbf{U} \in \mathbb{R}^{D \times L}$ is overcomplete $(L > D)$:

- *Ill-posed* problem: more unknowns than equations.
- add constraint: find sparsest $\mathbf{z} \in \mathbb{R}^{L}$ such that $\mathbf{x} = \mathbf{U}\mathbf{z}$

Solve mathematical program

$$\begin{aligned} \mathbf{z}^\star \quad &\in \quad \arg\min_{\mathbf{z}} \|\mathbf{z}\|_0 \\ \text{s.t.} \quad & \quad\ \ \mathbf{x} = \mathbf{U}\mathbf{z} \end{aligned}$$

- $\|\mathbf{z}\|_0$ counts the number of non-zero elements in $\mathbf{z}$.

Signal Reconstruction using Convex Optimization

- Sparsest solution, under the equality constraint:

$$\mathbf{z}^\star \in \arg\min_{\mathbf{z}} \ \|\mathbf{z}\|_0, \quad \text{s.t.} \quad \mathbf{x} = \mathbf{U}\mathbf{z}$$

  - NP hard combinatorial problem
  - brute-force: exhaustive search over all atom subsets
  - more efficient approximation: Matching Pursuit *(later)*

- Minimum ℓ_1-norm solution, under the equality constraint:

$$\mathbf{z}^\star \in \arg\min_{\mathbf{z}} \ \|\mathbf{z}\|_1, \quad \text{s.t.} \quad \mathbf{x} = \mathbf{U}\mathbf{z}$$

  - Convex Optimization Problem

Under suitable conditions on $\mathbf{U}$, the solutions of the two problems are equivalent! $\Rightarrow$ can use standard convex optimization methods.

Noisy Observations I

Additive noise:

$$\mathbf{x} = \mathbf{U}\mathbf{z} + \mathbf{n}$$

Assumes each dimension is independently corrupted by zero-mean Gaussian noise with variance σ^2.

$$\mathbf{n} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$$

Finding the code:

$$\begin{aligned} \mathbf{z}^\star \quad &\in \quad \arg\min_{\mathbf{z}} \|\mathbf{z}\|_0 \\ \text{s.t.} \quad & \qquad \|\mathbf{x} - \mathbf{U}\mathbf{z}\|_2^2 < D\sigma^2 \end{aligned}$$

- maximize sparsity of $\mathbf{z}$, ...
- while the squared residual remains below $D\sigma^2$.

Noisy Observations II

Or alternatively solve:

$$\begin{aligned} \mathbf{z}^\star &\in \arg\min_{\mathbf{z}} \|\mathbf{x} - \mathbf{U}\mathbf{z}\|_2 \\ \text{s.t. } \|\mathbf{z}\|_0 &\leq K \end{aligned}$$

- minimize residual, ...
- while selecting K or fewer atoms from the dictionary

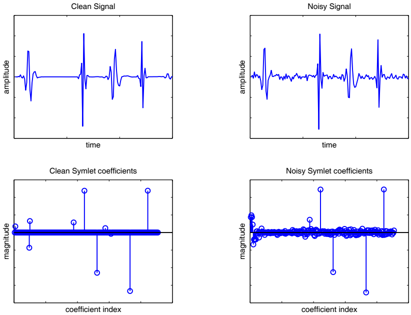
Approximate sparse coding:

Explain signal accurately with few atoms.

Noisy Observations III

- Coding with noise (denote $\mathbf{y}$ such that $\mathbf{n} = \mathbf{U}\mathbf{y}$)

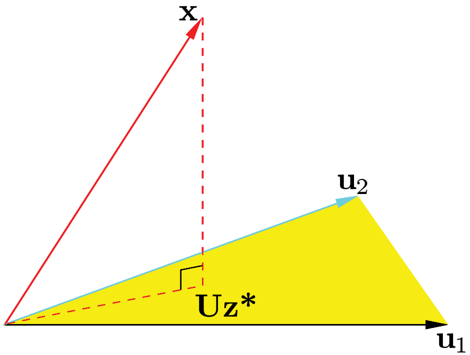
$$\hat{\mathbf{x}} = \mathbf{U}\mathbf{z} + \mathbf{n} = \mathbf{U}\mathbf{z} + \mathbf{U}\mathbf{y} = \mathbf{U}\left(\mathbf{z} + \mathbf{y}\right),$$



- Noise *cannot be sparsely coded* in any dictionary
  - therefore $\mathbf{y}$ has many small coefficients
- Large coefficients are still due to signal $\mathbf{x}$.

Noisy Observations IV

Geometry of sparse coding solution $\mathbf{z}^\star$:
(Will be useful later to understand the matching pursuit algorithm)



Orthogonal projection of $\mathbf{x}$ onto subspace spanned by selected atoms $\{\mathbf{u}_i \mid z_i^\star \neq 0\}$ minimizes $\|\mathbf{x} - \mathbf{Uz}\|_2$.