Dictionary Learning Using Wavelets

Daniel Mckenzie

University of Georgia

25 April 2016

Overview

This talk is about using wavelets for **compression** .

The plan:

- Review wavelets.
- Discuss Dictionary learning.
- Explain how one can combine dictionary learning and wavelets to achieve better compression.

I will only consider discrete, finite signals.

Consider the following signals:

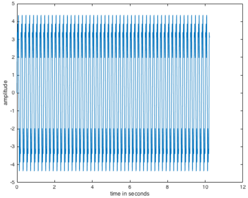


Figure: Signal A

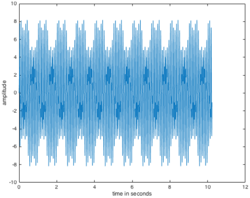


Figure: Signal B

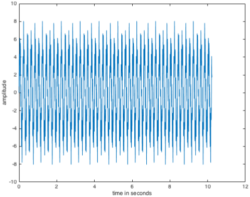


Figure: Signal C

Each is a discrete signal consisting of 1024 double precision floating point values, i.e. 65536 bits per signal.

Now consider their (discrete) Fourier Transforms:

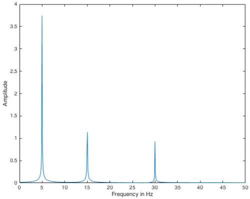


Figure: Signal A

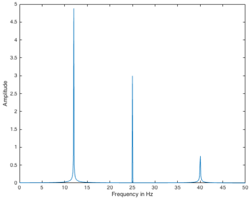


Figure: Signal B

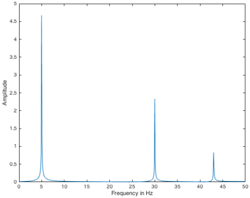


Figure: Signal C

Each signal has three non-zero entries, i.e. now 192 bits per signal.

The Point *Using Prior knowledge of our signals, we switched basis for* $\mathbb{R}^{1024}$ *from* $\{\mathbf{e}_n\}_{n=1}^{1024}$ *to* $\{\sin(2\pi n)\}_{n=1}^{1024}$*. Signals are sparse in new basis and thus can easily be compressed.*

Some remarks on Discrete Fourier Transfom:

- Linear transformation $\mathbb{R}^N \to \mathbb{R}^N$, hence can be represented as a matrix: $\hat{\mathbf{f}} = \mathbf{F}\mathbf{f}$.
- Matrix multiplication takes $\mathcal{O}(N^2)$ operations.
- The Fast Fourier Transform takes $\mathcal{O}(N \log(N))$.
- 'The Fast Fourier Transform is the most important numerical algorithm of our lifetime.' Gilbert Strang. [Str94]

Unable to detect time localized frequency events.

Example (Linear Chirp)

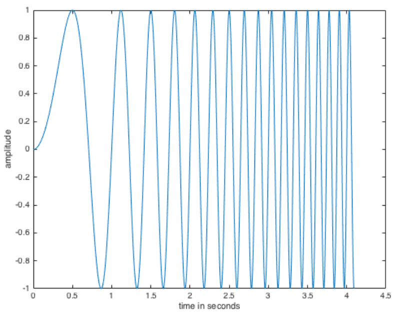


Figure: signal with 4096 entries

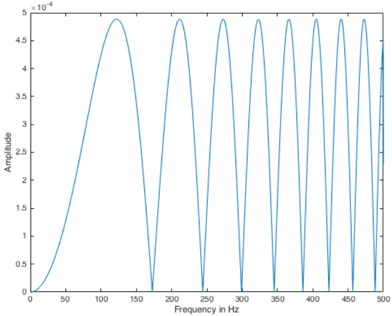


Figure: Fourier Transform of signal

Figure: Graph of a Linear Chirp: $y = \sin(2\pi t^2)$ and its Fourier Transform

One Solution:

Instead of basis of sine waves, use a basis of function with compact support in time.

Example (Haar Wavelets ([JJWW00]))

Consider discrete signals of length 8*, e.g.* $\mathbf{f} = (4, 6, 10, 12, 8, 6, 5, 5)$*, in basis* $\{\mathbf{e}_i\}_{i=1}^{8}$*. Let*

$$\varphi = \frac{1}{\sqrt{2}}(1, 1, 0, \ldots, 0)$$
$$\psi = \frac{1}{\sqrt{2}}(1, -1, 0, \ldots, 0)$$

and define translations:

$$\varphi_k[i] = \varphi[i - 2k]$$
$$\psi_k[i] = \psi[i - 2k]$$

Example (Haar Wavelets cont.)

Then $(\varphi_0, \ldots, \varphi_3 | \psi_0, \ldots, \psi_3)$ *is a basis. In this basis:*

$$\mathbf{f} = \sqrt{2}(5, 11, 7, 5 | -1, -1, 1, 0)$$

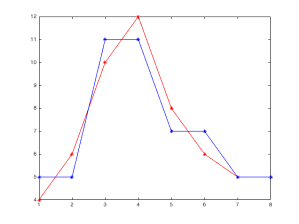


Figure: $\mathbf{f}$ plotted in red and $\sqrt{2}(5, 11, 7, 5 | 0, 0, 0, 0)$ plotted in blue

So $\mathbf{a}_1 = (5, 11, 7, 5)$ *is a (very) good approximation to* $\mathbf{f}$*. Can think of* $\mathbf{a}$ *as low frequency approximation,* $\mathbf{d}_1 = (-1, -1, 1, 0)$ *as high frequency details.*

Example (Haar Wavelets (cont.))

Consider translations and dilations:

$$\varphi_{j,k} = \frac{1}{\sqrt{2^{j+1}}}\varphi\left[floor(\frac{n-k+1}{2^j})\right]$$
$$\psi_{j,k} = \frac{1}{\sqrt{2^{j+1}}}\psi\left[floor(\frac{n-k+1}{2^j})\right]$$

$$\varphi_{1,0} = (1/2, 1/2, 1/2, 1/2, 0, 0, 0, 0)$$

Claim $(\varphi_{2,1}, \varphi_{2,2}|\psi_{2,1}\psi_{2,2}|\psi_{1,1}, \ldots, \psi_{1,4})$ *is a basis. In this basis:*

$$\mathbf{f} = (16, 12| - 6, 2| - \sqrt{2}, -\sqrt{2}, \sqrt{2}, 0)$$

$\mathbf{a}_2 = (16, 12)$ *is still a good approximation to* $\mathbf{f}$. ψ *is the Haar Wavelet and* φ *is its scaling function.*
$\mathbf{f} \rightarrow (\mathbf{a}_1|\mathbf{d}_1)$ *is a 1-level Haar Wavelet transform.*
$\mathbf{f} \rightarrow (\mathbf{a}_2|\mathbf{d}_2|\mathbf{d}_1)$ *is a 2-level Haar Wavelet transform.*

We shall restrict attention to orthogonal, compactly supported wavelets which come from an MRA, specifically Daubechies wavelets. Some properties:

- Changing to wavelet basis is linear tranformation $\mathbb{R}^N \to \mathbb{R}^N$, hence can be represented as a matrix: $\hat{\mathbf{f}} = \mathbf{W}_\psi \mathbf{f}$.
- Has a 'Fast Transform' (Conjugate Mirror Filters).
- Preserves ℓ^2 norm.
- Is an invertible transformation.

A General Compression Scheme

:

1. Given 1-dim signal $\mathbf{f}$, take its wavelet transform:

$$\hat{f} = W_{\psi}\mathbf{f} = (\mathbf{a}_n|\mathbf{d}_n|\ldots|\mathbf{d}_1)$$

2. Either set $\mathbf{d}_\ell = \mathbf{0}$ for $\ell = 1, \ldots, k$ or threshold to get $\hat{\hat{\mathbf{f}}}$, which should be sparse.
3. Encode and store $\hat{\hat{\mathbf{f}}}$.
4. Reconstruct approximation to $\mathbf{f}$ as needed via: $\tilde{f} = \mathbf{W}_{\psi}^{T}\hat{\hat{\mathbf{f}}}$

The Point *Wavelets are good for (lossy) compression.*

Using Wavelets for Image Compression

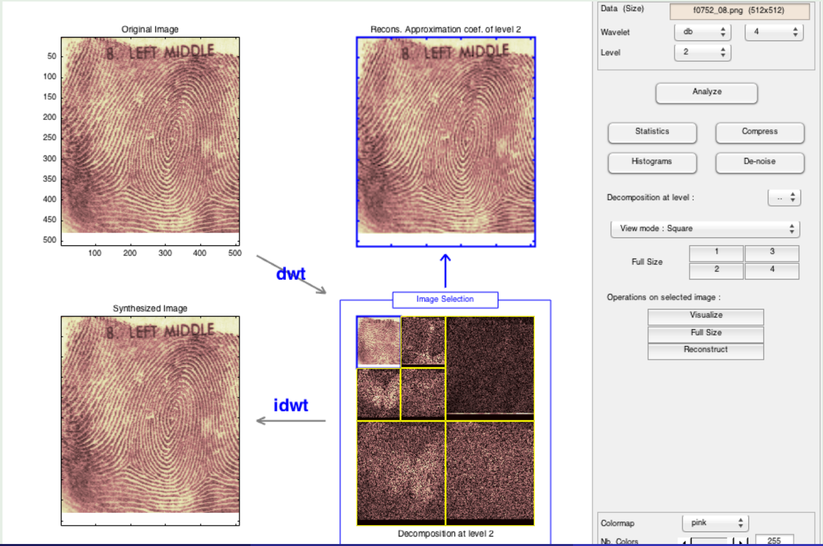
- Given image $\mathbf{I}$, represented as matrix of grayscale values, take an n-level wavelet transform to get:

$$\hat{\mathbf{I}} = (\mathbf{a}_n | \mathbf{h}_n, \mathbf{v}_n, \mathbf{d}_n | \ldots | \mathbf{h}_1, \mathbf{v}_1, \mathbf{d}_1)$$

- Set $\mathbf{v}_\ell = \mathbf{h}_\ell = \mathbf{d}_\ell = \mathbf{0}$ for $\ell = 1, \ldots, k$ or threshold to get $\hat{\hat{\mathbf{I}}}$.
- Encode and store $\hat{\hat{\mathbf{I}}}$

Example

512 $\times$ 512 fingerprint image analysed with db4 wavelet using MATLAB's Wavelet Toolbox.



Any Questions?

Dictionaries and Learning

- Recall that an (orthogonal) wavelet transform is like switching to a new basis.
- Wavelet transforms suffer from the 'Curse of Generality'.
- **Idea:** Given a collection of 'similar' signals $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\} \subset \mathbb{R}^n$ (e.g. all fingerprint images of same size). Determine a basis $\{\mathbf{d}_1, \ldots, \mathbf{d}_n\}$ of $\mathbb{R}^n$ s.t. representations of the $\mathbf{y}_i$ in this basis are as sparse as possible.

Definition (Dictionary)

A collection $\{\mathbf{d}_1, \ldots, \mathbf{d}_K\} \subset \mathbb{R}^n$ *which spans* $\mathbb{R}^n$ *is a* **Dictionary**. *($K \geq n$)*

Frequently shall write dictionary as $n \times K$ matrix $\mathbf{D}$ whose columns are $\mathbf{d}_i$. The $\mathbf{d}_i$ are called 'atoms'.

Definition (Dictionary Learning Problem)

Given $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ *with* $N >> K$ *find a dictionary* $\mathbf{D}^*$ *s.t.* $\mathbf{x}_i^* \approx \mathbf{D}^*\mathbf{y}_i$ *and the* $\mathbf{x}_i$ *are sufficiently sparse. Mathematically:*

$$\begin{aligned}(\mathbf{D}^*, \mathbf{X}^*) &= argmin_{\mathbf{D},\mathbf{X}} \sum_{i=1}^{N} ||\mathbf{y}_i - \mathbf{D}\mathbf{x}_i||_2^2 \text{ subject to } ||\mathbf{x}||_0 \leq T \\ &= argmin_{\mathbf{D},\mathbf{X}} ||\mathbf{Y} - \mathbf{D}\mathbf{X}||_F^2 \text{ subject to } ||\mathbf{x}||_0 \leq T\end{aligned}$$

If $\mathbf{Y}$ $n \times N$ *matrix with column* $\mathbf{y}_i$ *and* $\mathbf{X}$ *a* $K \times N$ *matrix with columns* $\mathbf{x}_i$.

The ‘one-bit’ approach, cf. [AEB06]

- Suppose $T = 1$, and we require that $\mathbf{x}_i$ are binary vectors.
- Problem becomes:

$$(\mathbf{D}^*, \mathbf{X}^*) = \text{argmin}_{\mathbf{D},\mathbf{X}} ||\mathbf{Y} - \mathbf{D}\mathbf{X}||_F^2 \text{ subject to } \mathbf{x}_i = \mathbf{e}_j \in \mathbb{R}^K \quad (1)$$

Efficient Algorithm for solution to (1):

Algorithm 1 k-means

Input $\mathbf{Y}$

Initialize $\mathbf{D}^{(0)} \in \mathbb{R}^{n \times K}$

for $J = 1 : J_{\max}$ **do**

  for $k = 1 : K$ **do**

    $C_k^{(J)} = \{\}$ (an empty list)

  for $i = 1 : N$ **do** ▷ The Sparse Coding Stage

    if $||\mathbf{y}_i - \mathbf{d}_{k^*}^{(J-1)}||_2 = \min_k ||\mathbf{y}_i - \mathbf{d}_k^{(J-1)}||$ **then** Add i to list $C_{k^*}^{(J)}$

  for $k = 1 : K$ **do** ▷ The Dictionary Update Stage

    $\mathbf{d}_k^{(J)} = \frac{1}{|C_k^{(J)}|} \sum_{i \in C_k^{(J)}} \mathbf{y}_i$

Output $\mathbf{D}^* = \mathbf{D}^{(J_{\max})}$ and $\mathbf{x}_i = \mathbf{e}_k$ if $i \in C_k^{J_{\max}}$

The general approach: The K-SVD algorithm [AEB06]

Consider again the general problem:

$$(\mathbf{D}, \mathbf{X}) = \text{argmin}_{\mathbf{D},\mathbf{X}} ||\mathbf{Y} - \mathbf{D}\mathbf{X}||_F^2 \text{ subject to } ||\mathbf{x}||_0 \leq T \qquad (2)$$

Generalize the previous algorithm to the K-SVD algorithm as follows:

- **Input:** sample signals $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$, sparsity parameter T, size of dictionary K.
- Initialize dictionary $\mathbf{D}^{(0)} \in \mathbb{R}^{n \times K}$ with normalized columns.
- Until stopping criteria met, repeat:
  1. (*Sparse coding step*) Find $\mathbf{x}_i$ such that

     $$\mathbf{x}_i = \text{argmin}_{\mathbf{x}} \{||\mathbf{y}_i - \mathbf{D}\mathbf{x}||_2\} \text{ subject to: } ||\mathbf{x}||_0 \leq T \qquad (3)$$

     Using MP,OMP etc.
  2. Let $\mathbf{X} \in \mathbb{R}^{K \times N}$ have columns $\mathbf{x}_i$.

3. (*Dictionary Update*) Update each atom $\mathbf{d}_k$ in turn:
   - Let ω_k be the set of examples that use $\mathbf{d}_k$: $\omega_k = \{j : (\mathbf{x}_j)_k) \neq 0\}$.
   - For each $j \in \omega_k$ compute $\mathbf{e}_{j,k} = \mathbf{y}_j - \sum_{\ell,\ell\neq k}(\mathbf{x}_j)_\ell \mathbf{d}_\ell$, residual without $\mathbf{d}_k$.
   - We are going to update $\mathbf{d}_k$ and the coefficients $(\mathbf{x}_\ell)_k$ (for $\ell \in \omega_k$) so as to minimize this residual error $\mathbf{E}_k$.
   - Let $\mathbf{E}_k \in \mathbb{R}^{n,|\omega_k|}$ have columns $\mathbf{e}_{j,k}$.
   - Solve:

$$(\mathbf{d}_k^*, \xi^*) = \text{argmin}_{\xi,\mathbf{d}} ||\mathbf{E}_k - \mathbf{d}\xi||_F^2 \text{ subject to: } \xi \in \mathbb{R}^{|\omega_k|} \text{ and } ||\mathbf{d}||_2 = 1 \tag{4}$$

   - $\mathbf{d}\xi \in \mathbb{R}^{n,|\omega_k|}$ is rank one. so (4) is solved by choosing $\mathbf{d}\xi$ to be optimal rank one approximation to $\mathbf{E}_k$
   - (by Eckart-Young-Mirsky theorem) if $U\Sigma V^T = \mathbf{E}_k$ is SVD, then $\mathbf{d}_k^* = \mathbf{u}_1$ and $\xi^* = \sigma(1)\mathbf{v}_1$ solves (4).
   - Update $\mathbf{d}_k$ to $\mathbf{d}_k^*$ and $(\mathbf{x}_{j_\ell})_k = \xi_\ell$ where $\omega_k = \{j_1, j_2, \ldots\}$

4. **Output:** Learned dictionary $\mathbf{D}$ and sparse representation matrix $\mathbf{X}$ such that if $\mathbf{X}$ has columns $\mathbf{x}_i$ then $||\mathbf{x}_i||_0 \leq T$ and $\mathbf{D}\mathbf{x}_i \approx \mathbf{y}_i$

Some Remarks

- Updating only coefficients $(\mathbf{x}_j)_k$ for $j \in \omega_k$ ensures that sparsity of the $\mathbf{x}_i$ can only improve.
- If in sparse coding step (3) solution is always found, then representation error $|\mathbf{Y} - \mathbf{DX}||_F^2$ is non-increasing, thus algorithm will converge.
- OMP works well enough for fairly small T to 'practically' ensure convergence

Implementation

- In [AEB06] K-SVD is implemented for images of faces.
- $N = 11\ 000$ and the $\mathbf{y}_i$ for $i = 1, \dots, K$ are 8×8 pixel blocks (i.e. $n = 64$), randomly sampled from a database of 4752×4752 facial images.
- K-SVD run with $K = 441$

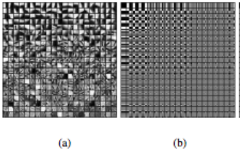


Figure: Learned dictionary on right, Haar dictionary on left (from [AEB06]

- For further details see [AEB06]

Experimental Results

- I randomly sampled image from same database, split into 594 8×8 blocks B_i.
- Fix # bits per coefficient, Q.
- Fix an error goal ϵ
- Encode each B_i to $\tilde{B}_i$ using OMP such that if $e^2 = \frac{||B_i - \tilde{B}_i||^2}{64}$ then $e^2 < \epsilon$.
- Let $PSNR = 10 \log_{10}(\frac{1}{e^2})$ (higher PSNR = better quality)
- Let TNB denote the total number of bits required to encode $\tilde{I}$.

$$TNB = \#\text{blocks} \times a + \#coefs(b + Q)$$

  where a is # bits required to code # coefficients per block, b is # bits required to code index of each atom.
- BPP (Bits Per Pixel) is given by:

$$BPP = \frac{TNB}{\#\text{pixels}}$$

Experimental Results cont.

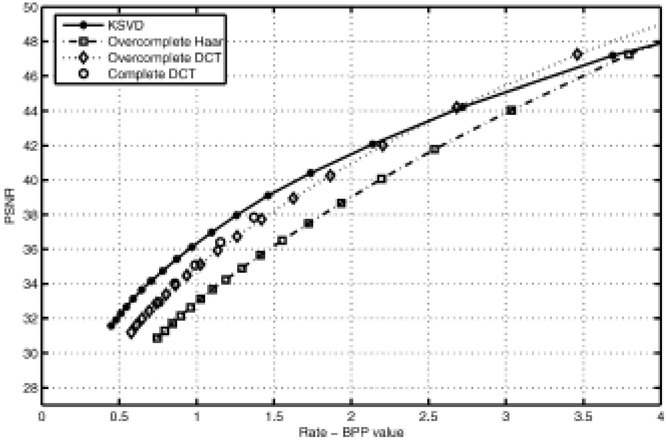


Figure: BPP vs. PSNR for learned dictionary, Haar wavelet transform and DCT. From [AEB06]

Experimental Results cont.

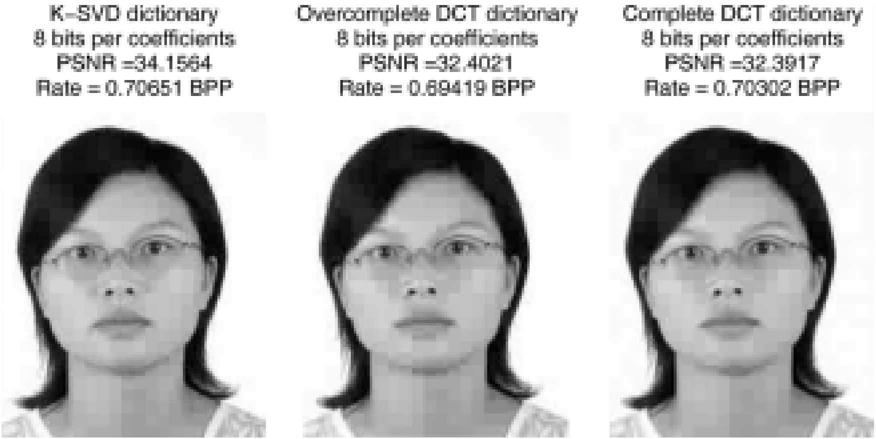


Figure: Reconstruction of compressed images. From [AEB06]

Remarks about Learned Dictionaries

Learned dictionaries appear to offer better compression rates, at least at low PSNR. However:

1. Learning is computationally intensive.
2. Lack of fast transform
3. Not multiscale; loses out on some potential compression.

   Question: *Can we combine strengths of learned and wavelet dictionaries?*

Any questions?

Dictionary Learning Using Wavelets

- [OLE11] attempts to combine a wavelet transform with a learned dictionary.
- **Idea:** First take wavelet transform, then apply a dictionary learning algorithm (K-SVD).
- Formally, solve:

$$(\tilde{\mathbf{D}}^*, \mathbf{X}^*) = \text{argmin}_{\tilde{\mathbf{D}},\mathbf{X}} ||\mathbf{Y} - \mathbf{W}_\psi \tilde{\mathbf{D}}\mathbf{X}||_F^2 = \text{argmin}_{\tilde{\mathbf{D}},\mathbf{X}} ||\mathbf{W}_\psi \mathbf{Y} - \tilde{\mathbf{D}}\mathbf{X}||_F^2$$

  where the columns of $\tilde{\mathbf{D}}$ are constrained to be very sparse.
- Effective dictionary is now $\mathbf{D} = \mathbf{W}_\psi \tilde{\mathbf{D}}$ whose atoms are linear combinations of several wavelet atoms, adapted to training set $\mathbf{Y}$.

Implementation

- In [AEB06] a 3 layer db4 wavelet transform was taken on data base of 20 coastal scenery images.
- 3 levels gives 10 bands: $\mathbf{a}_3, \mathbf{d}_3, \mathbf{v}_3, \mathbf{h}_3, \ldots, \mathbf{h}_1$. Will train a dictionary $\mathbf{D}_b$ for each band (10 in total).
- For each band, take $K = 64$ (# of atoms), and again each atom will be a $8\times$ 8 block
- As before, train $\mathbf{D}_b$ using a training set $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$ of 8×8 blocks randomly drawn from b-th band of images using K-SVD.

Experimental Results

- Given image I, take 3 level wavelet transform $\hat{I}$.
- For each band b, split $\hat{I}_b$ into 8×8 blocks B_i.
- encode each block B_i to $\tilde{B}_i$ using OMP such that *in total* only M atoms are used.
- This is compared to Wavelet transform compression using thresholding to keep only the M largest coefficients, and to regular K-SVD using M coefficients.

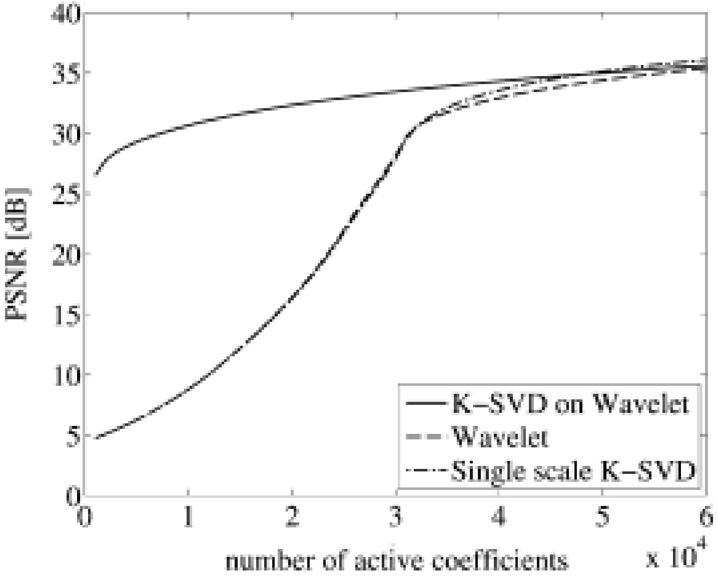


Figure: Comparing PSNR to M for three methods. [OLE11]

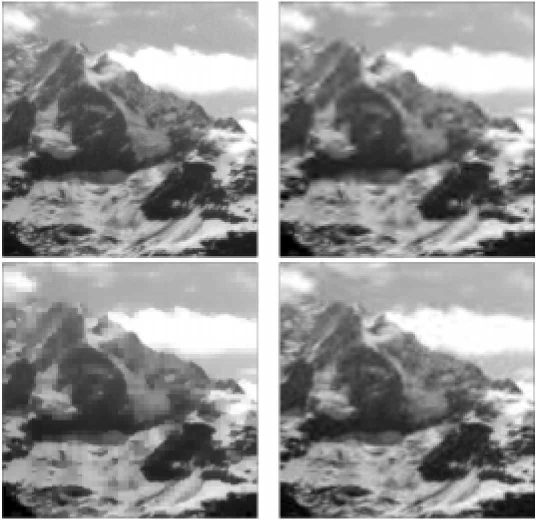


Figure: reconstructions using 32 000 terms. Top left is original, top right is wavelet (db4) reconstruction, bottom left is regular K-SVD, bottom right is K-SVD + Wavelet. [OLE11]

Conclusions

To conclude:

- In order to compress one first needs to choose a dictionary such that signal becomes sparse.
- 'generic dictionaries' (Wavelets etc.) provide fast transforms, but are not adapted to signals at hand.
- 'learned dictionaries' provide good sparsity, but lack of fast transform can be prohibitive.
- learning a dictionary whose atoms are made up of generic atoms may allow one to squeeze out some extra sparsity by adapting to a given class of signals.

Thank You!

Michal Aharon, Michael Elad, and Alfred Bruckstein.
K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation.
IEEE Transactions on Signal Processing, 54(11):4311–4322, 2006.

Sreenivasa Rao Jammalamadaka, Sreenivasa Rao Jammalamadaka, James S. Walker, and James S. Walker.
A Primer on Wavelets and Their Scientific Applications, volume 95.
2000.

B Ophir, M Lustig, and M Elad.
Multi-Scale Dictionary Learning Using Wavelets.
Selected Topics in Signal Processing, IEEE Journal of, 5(5):1014–1024, 2011.

Gilbert Strang.
Wavelets.
American scientist, 82(3):256–266, 1994.