Sparse & Redundant Representation Modeling of Images: Theory and Applications

Michael Elad
The Computer Science Department
The Technion
Haifa 32000, Israel

School of ICASSP Wednesday 22nd of April 2015

The research leading to these results has been received funding from the European union's Seventh Framework Program (FP/2007-2013) ERC grant Agreement ERC-SPARSE- 320649
This Talk Gives an **Overview** On ... 

15 years of tremendous progress in the field of Sparse and Redundant Representations

- Theory
- Numerical Problems
- Applications
Sparsity and Redundancy are valuable and well-founded tools for modeling data.

When used in image processing, they lead to state-of-the-art results.
Part I
Denoising by Sparse & Redundant Representations
Noise Removal?

Our story begins with image denoising ...

- **Important**: (i) Practical application; (ii) A convenient platform (the simplest) for testing basic ideas in image processing; (iii) Given a good denoising algorithm, one could solve many other problems.

- **Many Considered Directions**: Partial differential equations, Statistical estimators, Adaptive filters, Inverse problems & regularization, Wavelets, Example-based techniques, **Sparse representations**, ...
Many of the proposed image denoising algorithms are related to the minimization of an energy function of the form

\[
f(x) = \frac{1}{2} \|x - y\|_2^2 + G(x)
\]

- \(y\): Given measurements
- \(x\): Unknown to be recovered

- This is in-fact a Bayesian point of view, adopting the Maximum-A-posteriori Probability (MAP) estimation.
- Clearly, the wisdom in such an approach is within the choice of the prior – **modeling the images** of interest.
The Evolution of $G(x)$

During the past several decades we have made all sort of guesses about the prior $G(x)$ for images:

- Hidden Markov Models,
- Compression algorithms as priors,
- ... 

$$G(x) = \lambda \|x\|_2^2$$  \hspace{1cm}  Energy

$$G(x) = \lambda \|Lx\|_2^2$$  \hspace{1cm}  Smoothness

$$G(x) = \lambda \|Lx\|_w^2$$  \hspace{1cm}  Adapt+\hspace{1cm} Smooth

$$G(x) = \lambda \rho \{Lx\}$$  \hspace{1cm}  Robust\hspace{1cm} Statistics

$$G(x) = \lambda \|\nabla x\|_1$$  \hspace{1cm}  Total-\hspace{1cm} Variation

$$G(x) = \lambda \|WX\|_1$$  \hspace{1cm}  Wavelet\hspace{1cm} Sparsity

$$G(x) = \lambda \|\alpha\|_0^0$$  \hspace{1cm}  Sparse &\hspace{1cm} Redundant

for $x = D\alpha$

- Hidden Markov Models,
- Compression algorithms as priors,
- ...
Sparse Modeling of Signals

- Every column in $D$ (dictionary) is a prototype signal (atom).
- The vector $\alpha$ is generated randomly with few (say L) non-zeros at random locations and with random values.
- We shall refer to this model as Sparseland.
Interesting Model:

- **Simple:** Every generated signal is built as a linear combination of few atoms from our dictionary $D$.
- **Rich:** A general model: the obtained signals are a union of many low-dimensional Gaussians.
- **Familiar:** We have been using this model in other context for a while now (wavelet, JPEG, ...).
Sparse & Redundant Rep. Modeling?

As $p \to 0$, we get a count of the non-zeros in the vector. Our signal model is thus:

$$x = D\alpha$$ where $\alpha$ is sparse

$$\|\alpha\|_0 \leq L$$

Sparse and Redundant Representation Modeling of Signals – Theory and Applications
By: Michael Elad
Back to Our MAP Energy Function

- We $L_0$ norm is effectively counting the number of non-zeros in $\alpha$.

- The vector $\alpha$ is the representation (sparse/redundant) of the desired signal $x$.

- The core idea: while few ($L$ out of $K$) atoms can be merged to form the true signal, the noise cannot be fitted well. Thus, we obtain an effective projection of the noise onto a very low-dimensional space, thus getting denoising effect.

$$\frac{1}{2} \| x - y \|_2^2$$
Wait! There are Some Issues

- **Numerical Problems:** How should we solve or approximate the solution of the problem

\[
\min_{\alpha} \|D\alpha - y\|_2^2 \quad \text{s.t.} \quad \|\alpha\|_0 \leq L
\]

or

\[
\min_{\alpha} \|\alpha\|_0 \quad \text{s.t.} \quad \|D\alpha - y\|_2^2 \leq \varepsilon^2
\]

or

\[
\min_{\alpha} \lambda \|\alpha\|_0 + \|D\alpha - y\|_2^2
\]

- **Theoretical Problems:** Is there a unique sparse representation? If we are to approximate the solution somehow, how close will we get?

- **Practical Problems:** What dictionary \(D\) should we use, such that all this leads to effective denoising? Will all this work in applications?
To Summarize So Far …

Image denoising (and many other problems in image processing) requires a model for the desired image.

What do we do?

We proposed a model for signals/images based on sparse and redundant representations.

Great! No?

There are some issues:
1. Theoretical
2. How to approximate?
3. What about $D$?
Part II
Theoretical & Numerical Foundations
Let's start with the noiseless problem.

Suppose we build a signal by the relation \( D\alpha = x \).

We aim to find the signal's representation:

\[
\hat{\alpha} = \text{ArgMin}_{\alpha} ||\alpha||_0^0 \quad \text{s.t.} \quad x = D\alpha
\]

Why should we necessarily get \( \hat{\alpha} = \alpha \)?

It might happen that eventually \( ||\hat{\alpha}||_0^0 < ||\alpha||_0^0 \).
Matrix “Spark”

**Definition:** Given a matrix $\mathbf{D}$, $\sigma = \text{Spark}\{\mathbf{D}\}$ is the smallest number of columns that are linearly dependent.

**Example:**

\[
\begin{bmatrix}
1 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
\end{bmatrix}
\]

- **Rank** = 4
- **Spark** = 3

*In tensor decomposition, Kruskal defined something similar already in 1989.*
Uniqueness Rule

Suppose this problem has been solved somehow

\[ \hat{\alpha} = \text{ArgMin}_{\alpha} \|\alpha\|_0^0 \text{ s.t. } x = D\alpha \]

Uniqueness

If we found a representation that satisfy

\[ \|\hat{\alpha}\|_0 < \frac{\sigma}{2} \]

Then necessarily it is unique (the sparsest).

This result implies that if \( M \) generates signals using “sparse enough” \( \alpha \), the solution of the above will find it exactly.

Donoho & E. (‘02)
Our Goal

This is a combinatorial problem, proven to be NP-Hard!

Here is a recipe for solving this problem:

1. Set $L=1$
2. Gather all the supports $\{S_i\}_i$ of cardinality $L$
3. Solve the LS problem
   \[ \min_{\alpha} \|D\alpha - y\|_2^2 \text{ s.t. } \text{supp}(\alpha) = S_i \]
   for each support
4. Set $L=L+1$
5. There are $\binom{K}{L}$ such supports
6. Assume: $K=1000$, $L=10$ (known!), 1 nano-sec per each LS
7. We shall need $\sim8e+6$ years to solve this problem !!!!!

Done
Lets Approximate

\[
\min_{\alpha} \|\alpha\|_0 \quad \text{s.t.} \quad \|D\alpha - y\|_2^2 \leq \varepsilon^2
\]

Relaxation methods
Smooth the $L_0$ and use continuous optimization techniques

Greedy methods
Build the solution one non-zero element at a time
Relaxation – The Basis Pursuit (BP)

Instead of solving

\[
\min_{\alpha} \| \alpha \|_0 \quad \text{s.t.} \quad \| D\alpha - y \|_2 \leq \varepsilon
\]

Solve Instead

\[
\min_{\alpha} \| \alpha \|_1 \quad \text{s.t.} \quad \| D\alpha - y \|_2 \leq \varepsilon
\]

- This is known as the Basis-Pursuit (BP) [Chen, Donoho & Saunders (‘95)].
- The newly defined problem is convex (quad. programming).
- Very efficient solvers can be deployed:
  - Interior point methods [Chen, Donoho, & Saunders (‘95)] [Kim, Koh, Lustig, Boyd, & D. Gorinevsky (‘07)].
  - Sequential shrinkage for union of ortho-bases [Bruce et.al. (‘98)].
  - Iterative shrinkage [Figuerido & Nowak (‘03)] [Daubechies, Defrise, & De-Mole (‘04)] [E. (‘05)] [E., Matalon, & Zibulevsky (‘06)] [Beck & Teboulle (‘09)] ...
Go Greedy: Matching Pursuit (MP)

- The MP is one of the greedy algorithms that finds one atom at a time [Mallat & Zhang ('93)].

- Step 1: find the one atom that **best matches** the signal.

- Next steps: given the previously found atoms, find the next **one** to **best fit** the residual.

- The algorithm stops when the error $\|D\alpha - y\|_2$ is below the destination threshold.

- The Orthogonal MP (OMP) is an improved version that re-evaluates the coefficients by Least-Squares after each round.
Pursuit Algorithms

There are various algorithms designed for approximating the solution of this problem:

- **Greedy Algorithms:**
  - Relaxation Algorithms: Basis Pursuit (a.k.a. LASSO), Dantzig Selector & numerical ways to handle them [1995-today].

Why should they work?
The Mutual Coherence

- Compute

\[ D^T D \]

Assume normalized columns

- The **Mutual Coherence** \( \mu \) is the largest off-diagonal entry in absolute value.

- The Mutual Coherence is a property of the dictionary (just like the “Spark”). In fact, the following relation can be shown:

\[
\sigma \geq 1 + \frac{1}{\mu}
\]
BP and MP Equivalence (No Noise)

Given a signal $x$ with a representation $\hat{x} = D\alpha$, assuming that $\|\alpha\|_0 < 0.5(1 + 1/\mu)$, BP and MP are guaranteed to find the sparsest solution.

$$\hat{\alpha} = \text{ArgMin}_{\alpha} \|\alpha\|_0 \text{ s.t. } x = D\alpha$$

- MP and BP are different in general (hard to say which is better).
- The above result corresponds to the worst-case, and as such, it is too pessimistic.
- Average performance results are available too, showing much better bounds [Donoho (‘04)] [Candes et.al. (‘04)] [Tanner et.al. (‘05)] [E. (‘06)] [Tropp et.al. (‘06)] ... [Candes et. al. (‘09)].
BP Stability for the Noisy Case

Given a signal \( y = D\alpha + v \) with a representation satisfying \( \|\alpha\|_0^0 < 1 / 3\mu \) and a white Gaussian noise \( v \sim N(0, \sigma^2 I) \), BP will show* stability, i.e.,

\[
\|\hat{\alpha}_{BP} - \alpha\|_2^2 < \text{Const}(\lambda) \cdot \log K \cdot \|\alpha\|_0^0 \cdot \sigma^2
\]

Ben-Haim, Eldar & E. (‘09)

- For \( \sigma = 0 \) we get a weaker version of the previous result.
- This result is the oracle's error, multiplied by \( C \cdot \log K \).
- Similar results exist for other pursuit algorithms (Dantzig Selector, Orthogonal Matching Pursuit, CoSaMP, Subspace Pursuit, …)

\[
\min_{\alpha} \lambda \|\alpha\|_1 + \|D\alpha - y\|_2^2
\]

* With very high probability
To Summarize So Far …

Image denoising (and many other problems in image processing) requires a model for the desired image.

What do we do?

We proposed a model for signals/images based on sparse and redundant representations.

Problems?

The Dictionary $\textbf{D}$ should be found somehow !!!

What next?

We have seen that there are approximation methods to find the sparsest solution, and there are theoretical results that guarantee their success.
Part III
Dictionary Learning: The K-SVD Algorithm
What Should $\mathbf{D}$ Be?

$$\hat{\alpha} = \arg \min_{\alpha} \| \alpha \|_0^0 \quad \text{s.t.} \quad \frac{1}{2} \| \mathbf{D} \alpha - y \|_2^2 \leq \epsilon^2 \quad \Rightarrow \quad \hat{x} = \mathbf{D} \hat{\alpha}$$

Our Assumption: Good-behaved Images have a sparse representation

$\mathbf{D}$ should be chosen such that it sparsifies the representations

One approach to choose $\mathbf{D}$ is from a known set of transforms (Steerable wavelet, Curvelet, Contourlets, Bandlets, Shearlets ...)

The approach we will take for building $\mathbf{D}$ is training it, based on Learning from Image Examples

Sparse and Redundant Representation Modeling of Signals – Theory and Applications
By: Michael Elad
Given these $P$ examples and a fixed size $[N \times K]$ dictionary $\mathbf{D}$:

1. Is $\mathbf{D}$ unique?
2. How would we find $\mathbf{D}$?
Each example is a linear combination of atoms from $D$

$$\text{Min}_{D,A} \sum_{j=1}^{P} \|D\alpha_j - x_j\|_2^2$$

s.t. $\forall j, \|\alpha_j\|_0 \leq L$

Each example has a sparse representation with no more than $L$ atoms

[Field & Olshausen ('96)]
[Engan et. al. ('99)]
[Lewicki & Sejnowski ('00)]
[Cotter et. al. ('03)]
[Gribonval et. al. ('04)]
[Aharon, E. & Bruckstein ('04)]
[Aharon, E. & Bruckstein ('05)]
K–Means For Clustering

Clustering: An extreme sparse representation

1. Initialize $D$
2. Sparse Coding
   - Nearest Neighbor
3. Dictionary Update
   - Column-by-Column by
   - Mean computation over the relevant examples

By: Michael Elad
The K–SVD Algorithm – General

[Aharon, E. & Bruckstein (‘04,’05)]

1. Initialize $D$
2. Sparse Coding
   - Use Matching Pursuit
3. Dictionary Update
   - Column-by-Column by SVD computation over the relevant examples

Sparse and Redundant Representation Modeling of Signals – Theory and Applications
By: Michael Elad
K–SVD: Sparse Coding Stage

\[ \min_A \sum_{j=1}^{P} \| D\alpha_j - x_j \|_2^2 \quad \text{s.t.} \quad \forall j, \| \alpha_j \|_p \leq L \]

(D is known! For the j^{th} item we solve)

\[ \min_{\alpha} \| D\alpha - x_j \|_2^2 \quad \text{s.t.} \quad \| \alpha \|_p \leq L \]

Solved by A Pursuit Algorithm
K–SVD: Dictionary Update Stage

We refer only to the examples that use the column $d_k$.

Fixing all $A$ and $D$ apart from the $k^{th}$ column, and seek both $d_k$ and the $k^{th}$ column in $A$ to better fit the residual!

We should solve:

$$\min_{d_k, \alpha_k} \| E + A_k^T \alpha_k d_k \|_F^2$$
A Synthetic Experiment

Create a 20×30 random dictionary with normalized columns

Generate 2000 signal examples with 3 atoms per each and add noise

Train a dictionary using the KSVD and MOD and compare

Results

Results graph:
- MOD performance
- K-SVD performance
To Summarize So Far …

Image denoising (and many other problems in image processing) requires a model for the desired image.

What do we do?

We proposed a model for signals/images based on sparse and redundant representations.

Problems?

We have seen approximation methods that find the sparsest solution, and theoretical results that guarantee their success. We also saw a way to learn $D$.

What next?

Will it all work in applications?
Part IV
Back to Denoising … and Beyond – Combining it All
From Local to Global Treatment

- The K-SVD algorithm is reasonable for low-dimension signals (N in the range 10-400). As N grows, the complexity and the memory requirements of the K-SVD become prohibitive.

- So, how should large images be handled?

- The solution: Force shift-invariant sparsity - on each patch of size N-by-N (N=8) in the image, including overlaps.

$$\hat{x} = \text{ArgMin}_{x, \{\alpha_{ij}\}_{ij}} \frac{1}{2} \|x - y\|^2_2 + \mu \sum_{ij} \|R_{ij}x - D\alpha_{ij}\|^2_2$$

s.t. $$\|\alpha_{ij}\|_0^0 \leq L$$

Extracts a patch in the ij location

Our prior
What Data to Train On?

**Option 1:**
- Use a database of images,
- We tried that, and it works fine (~0.5-1dB below the state-of-the-art).

**Option 2:**
- Use the corrupted image itself !!
- Simply sweep through all patches of size N-by-N (overlapping blocks),
- Image of size $1000^2$ pixels $\rightarrow \sim 10^6$ examples to use – more than enough.
- This works much better!
K-SVD Image Denoising

\[ \hat{x} = \operatorname{ArgMin}_{x, \{\alpha_{ij}\}_{ij}} \frac{1}{2} \|x - y\|_2^2 + \mu \sum_{ij} \| R_{ij} x - D \alpha_{ij} \|_2^2 \quad \text{s.t.} \quad \|\alpha_{ij}\|_0 \leq L \]

\( x = y \) and \( D \) known

\( x \) and \( \alpha_{ij} \) known

\( D \) and \( \alpha_{ij} \) known

Compute \( \alpha_{ij} \) per patch

\[ \alpha_{ij} = \operatorname{Min}_{\alpha} \| R_{ij} x - D \alpha \|_2^2 \quad \text{s.t.} \quad \|\alpha\|_0 \leq L \]

using the matching pursuit

Compute \( D \) to minimize

\[ \operatorname{Min}_{\alpha} \sum_{ij} \| R_{ij} x - D \alpha \|_2^2 \]

using SVD, updating one column at a time

Compute \( x \) by

\[ x = \left[ I + \mu \sum_{ij} R_{ij}^T R_{ij} \right]^{-1} \left[ y + \mu \sum_{ij} R_{ij}^T D \alpha_{ij} \right] \]

which is a simple averaging of shifted patches

Sparse and Redundant Representation Modeling
of Signals – Theory and Applications
By: Michael Elad
The results of this algorithm compete favorably with the state-of-the-art.

This algorithm can be extended by using joint sparse representation on the patches, introducing a non-local force in the denoising, thus leading to improved results [Mairal, Bach, Ponce, Sapiro & Zisserman ('09)].

What about EPLL? ...

\[ \sigma = 20 \]
When turning to handle color images, the main difficulty is in defining the relation between the color layers – R, G, and B.

The solution with the above algorithm is simple – consider 3D patches or 8-by-8 with the 3 color layers, and the dictionary will detect the proper relations.
Denoising (Color) [Mairal, E. & Sapiro ('08)]

Our experiments lead to state-of-the-art denoising results, giving ~1dB better results compared to [Mcauley et. al. ('06)] which implements a learned MRF model (Field-of-Experts)

Original  |  Noisy (12.77dB)  |  Result (29.87dB)
Video Denoising [Protter & E. ('09)]

Our experiments lead to state-of-the-art video denoising results, giving ~0.5dB better results on average compared to [Boades, Coll & Morel ('05)] and comparable to [Rusanovskyy, Dabov, & Egiazarian ('06)].
In Computer-Tomography (CT) reconstruction, an image is recovered from a set of its projections.

In medicine, CT projections are obtained by X-ray, and it typically requires a high dosage of radiation in order to obtain a good quality reconstruction.

A lower-dosage projection implies a stronger noise (Poisson distributed) in data to work with.

Armed with sparse and redundant representation modeling, we can denoise the data and the final reconstruction ... enabling CT with lower dosage.
Image Inpainting – The Basics

- Assume: the signal $x$ has been created by $x = D\alpha_0$ with very sparse $\alpha_0$.
- Missing values in $x$ imply missing rows in this linear system.
- By removing these rows, we get $\tilde{D}\alpha = \tilde{x}$.
- Now solve

$$\min_{\alpha} \|\alpha\|_0 \quad \text{s.t.} \quad \tilde{x} = \tilde{D}\alpha$$

- If $\alpha_0$ was sparse enough, it will be the solution of the above problem! Thus, computing $D\alpha_0$ recovers $x$ perfectly.
**Side Note: Compressed-Sensing**

- **Compressed Sensing** is leaning on the very same principal, leading to alternative sampling theorems.
- Assume: the signal $x$ has been created by $x = D\alpha_0$ with very sparse $\alpha_0$.
- Multiply this set of equations by the matrix $Q$ which reduces the number of rows.
- The new, smaller, system of equations is $QD\alpha = Qx \Rightarrow \tilde{D}\alpha = \tilde{x} \times$
- If $\alpha_0$ was sparse enough, it will be the sparsest solution of the new system, thus, computing $D\alpha_0$ recovers $x$ perfectly.
- Compressed sensing focuses on conditions for this to happen, guaranteeing such recovery.
Inpainting [Mairal, E. & Sapiro (08)]

Our experiments lead to state-of-the-art inpainting results.

Original 80% missing  Result
Inpainting [Mairal, E. & Sapiro ('08)]

The same can be done for video, very much like the denoising treatment: (i) 3D patches, (ii) no need to compute the dictionary from scratch for each frame, and (iii) no need for explicit motion estimation.

Original  80% missing  Result
Our experiments lead to state-of-the-art demosaicing results, giving \( \sim 0.2 \) dB better results on average, compared to [Chang & Chan (’06)].

Today’s cameras are sensing only one color per pixel, leaving the rest for interpolated.

Generalizing the inpainting scheme to handle demosaicing is tricky because of the possibility to learn the mosaic pattern within the dictionary.

In order to avoid “over-fitting”, we handle the demosaicing problem while forcing strong sparsity and applying only few iterations.
The problem: Compressing photo-ID images.

General purpose methods (JPEG, JPEG2000) do not take into account the specific family.

By adapting to the image-content (PCA/K-SVD), better results could be obtained.

For these techniques to operate well, train dictionaries locally (per patch) using a training set of images is required.

In PCA, only the (quantized) coefficients are stored, whereas the K-SVD requires storage of the indices as well.

Geometric alignment of the image is very helpful and should be done [Goldenberg, Kimmel, & E. ('05)].
Image Compression

- Detect main features and warp the images to a common reference (20 parameters)
- Divide the image into disjoint 15-by-15 patches. For each compute mean and dictionary
- Per each patch find the operating parameters (number of atoms L, quantization Q)
- Warp, remove the mean from each patch, sparse code using L atoms, apply Q, and dewarp

Training set (2500 images)

On the training set

On the test image

Sparse and Redundant Representation Modeling of Signals – Theory and Applications
By: Michael Elad
Image Compression Results

Results for 820 Bytes per each file
Image Compression Results

Results for 550 Bytes per each file

Original  JPEG  JPEG-2000  Local-PCA  K-SVD

15.81  13.89  10.66  6.60
14.67  12.41  9.44  5.49
15.30  12.57  10.27  6.36

Sparse and Redundant Representation Modeling of Signals – Theory and Applications
By: Michael Elad
Image Compression Results

Results for **400** Bytes per each file
Deblock the Results [Bryt and E. (09)]

550 bytes
K-SVD results with and without deblocking

Deblock (6.24)  Deblock (5.27)  Deblock (6.03)  Deblock (11.32)
Poisson Denoising

\[ Y = X + V \]
\[ V \sim N(0, \sigma^2 I) \]

\[ P(y | x) = \frac{x^y}{y!} e^{-x} \]

\[ \text{peak} \overset{\Delta}{=} \max_{i,j} \{ x_{i,j} \} \]
Poisson Denoising  [Salmon et. al., 2011] [Giryes et. al., 2013]

- Anscombe transform converts Poisson distributed noise into an approximately Gaussian one, with variance 1 using the following formula [Anscombe, 1948]:

\[
    f_{\text{Anscombe}}(y) = 2 \sqrt{y + \frac{3}{8}}
\]

- However, this is of reasonable accuracy only if peak > 4.

- For lower peaks (poor illumination), we use the patch-based approach with dictionary learning, BUT ... in the exponent domain:

\[
\begin{align*}
    x &= D\alpha \\
    \text{where} & \quad \|\alpha\|_0 \leq L
\end{align*}
\]

\[
\begin{align*}
    x &= \exp \{ D\alpha \} \\
    \text{where} & \quad \|\alpha\|_0 \leq L
\end{align*}
\]
Poisson Denoising – Results (1)

Original                   Noisy (peak=1)                   Result (PSNR=22.59dB)

Dictionary learned atoms:
Poisson Denoising – Results (2)
Other Applications?

- Poisson Inpainting
- Super-Resolution
- Blind deblurring
- Audio inpainting
- Dynamic MRI reconstruction
- Clutter reduction in Ultrasound
- Single image interpolation
- Anomaly detection
- ...

Sparse and Redundant Representation Modeling of Signals – Theory and Applications
By: Michael Elad
To Summarize So Far ...

Image denoising (and many other problems in image processing) requires a model for the desired image.

What do we do?

We proposed a model for signals/images based on sparse and redundant representations.

Well, does this work?

Yes! We have seen a group of applications where this model is showing very good results: denoising of bw/color stills/video, CT improvement, inpainting, super-resolution, and compression.

So, what next?

Well, many more things ...
Part V
Summary and Conclusion
Today We Have Seen that ...

**Sparsity, Redundancy, and the use of examples** are important ideas that can be used in designing better tools in signal/image processing.

**In our work on we cover theoretical, numerical, and applicative issues related to this model and its use in practice.**

**What do we do?**

**We keep working on:**
- Improving the model
- Improving the dictionaries
- Demonstrating on other applications
- ...

**What next?**

Sparse and Redundant Representation Modeling of Signals – Theory and Applications
By: Michael Elad
Thank You

All this Work is Made Possible Due to

my teachers and mentors

A.M. Bruckstein  D.L. Donoho

colleagues & friends collaborating with me

G. Sapiro  J.L. Starck  I. Yavneh  M. Zibulevsky

and my students

If you are Interested ...

More on this topic (including the slides, the papers, and Matlab toolboxes) can be found in my webpage:
http://www.cs.technion.ac.il/~elad

A book on these topics was published in August 2010.
Thank You all!

Questions?

More on these (including the slides and the relevant papers) can be found in http://www.cs.technion.ac.il/~elad
Dictionary Learning: Uniqueness?

Uniqueness

If \[ \{ x_j \}_{j=1}^p \] is rich enough* and if

\[ L < \frac{\text{Spark}\{D\}}{2} \]

then \( D \) is unique.

Comments:

- “Rich Enough”: The signals from \( M \) could be clustered to \( \binom{K}{L} \) groups that share the same support. At least \( L+1 \) examples per each are needed. More recent results (see Schnass and Wright’s work) improve this dramatically.
- This result is proved constructively, but the number of examples needed to pull this off is huge – we will show a far better method next.
- A parallel result that takes into account noise is yet to be constructed.

Aharon, E., & Bruckstein (’05)
Improved Dictionary Learning

\[
\min_{D, A} \sum_{j=1}^{P} \|D\alpha_j - x_j\|_2^2 \quad \text{s.t. } \forall j, \|\alpha_j\|_0^0 \leq L
\]

**MOD Algorithm**
- Fix \( D \) and update \( A \)
- Fix \( A \) and update \( D \)

**K-SVD Algorithm**
- Fix \( D \) and update \( A \)
  
  for \( j=1:1:K \)
  - Fix \( A \) & \( D \) apart from the \( j \)-th atom its coefficients
  - Update \( d_j \) and its coef. in \( A \)
  end
**Improved Dictionary Learning**

**Improved Algorithm**

- **Fix** $D$ and **update** $A$
- **Fix** the supports in $A$ and **update** both $D$ and the non-zeros

**MOD and K-SVD can be considered as crude approximation of this method**

This can be done in two ways:

1. **Apply** several rounds of the atoms’ update in the K-SVD, or
2. **Extend** the MOD to update the non-zero elements in $A$

**Min** $D,A$ $\sum_{j=1}^{P} \left\| D\alpha_j - x_j \right\|_2^2$ s.t. $\forall j$, $\|\alpha_j\|_0^0 < L$

$D, A \odot M = 0$
The algorithm we proposed updates $x$ only once at the end.

Why not repeat the whole process several times?

The rationale: The sparse representation model should be imposed on the patches of the FINAL image. After averaging, this is ruined.
EPLL Improvement \[\text{[Sulam and E. ('15)]}\]

- Expected Patch Log Likelihood (EPLL) is an algorithm that came to fix this problem \[\text{[Zoran and Weiss, ('11)]}\] in the context of a GMM prior.

- An extension of EPLL to Spars-Land is proposed in \[\text{[Sulam and E. ('15)]}\]. The core idea is:
  - After the image has been computed, we proceed the iterative process, and apply several such overall rounds of updates.
  - Sparse coding must be done with a new threshold, based on the remaining noise in the image. This is done by evaluating the noise level based on the linear projections (disregarding the support detection by the OMP).
  - This algorithm leads to state-of-the-art results, with 0.5-1dB improvement over the regular K-SVD algorithm shown before.
EPLL Improvement [Sulam and E. (’15)]

Noisy image has $\sigma=25$

KSVD PSNR 31.42 dB

EPLL PSNR 31.83 dB
Inpainting Formulation

\[ \hat{x} = \text{ArgMin}_{x, \{\alpha_{ij}\}, D} \frac{1}{2} \| Mx - y \|^2_2 + \mu \sum_{ij} \| R_{ij} x - D_{ij} \alpha_{ij} \|^2_2 \quad \text{s.t.} \quad \| \alpha_{ij} \|^0_0 \leq L \]

The matrix \( M \) is a mask matrix, obtained by the identity matrix with some of its rows omitted, corresponding to the missing samples.
Inpainting Formulation [Mairal, E. & Sapiro ('08)]

\[
\hat{x} = \text{ArgMin}_{x, \{\alpha_{ij}\}_{ij}, D} \frac{1}{2} \left\| Mx - y \right\|_2^2 + \mu \sum_{ij} \left\| R_{ij} x - D\alpha_{ij} \right\|_2^2 \quad \text{s.t.} \quad \left\| \alpha_{ij} \right\|_0^0 \leq L
\]

\(x = y\) and \(D\) known

Compute \(\alpha_{ij}\) per patch

\[
\alpha_{ij} = \text{Min}_\alpha \left\| M_{ij} \left( R_{ij} x - D\alpha \right) \right\|_2^2
\]

\[
\text{s.t.} \; \left\| \alpha \right\|_0^0 \leq L
\]

using the matching pursuit

\(x\) and \(\alpha_{ij}\) known

Compute \(D\) to minimize

\[
\text{Min}_\alpha \sum_{ij} \left\| M_{ij} \left( R_{ij} x - D\alpha \right) \right\|_2^2
\]

using SVD, updating one column at a time

\[\sim K\text{-SVD}\]

\(D\) and \(\alpha_{ij}\) known

Compute \(x\) by

\[
x = \left[ M^T M + \mu \sum_{ij} R_{ij}^T R_{ij} \right]^{-1} \left[ M^T y + \mu \sum_{ij} R_{ij}^T D\alpha_{ij} \right]
\]

which is again a simple averaging of patches
Inpainting [Mairal, E. & Sapiro ('08)]

For the Peppers image

<table>
<thead>
<tr>
<th>Alg.</th>
<th>RMSE for 25% missing</th>
<th>RMSE for 50% missing</th>
<th>RMSE for 75% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>No-overlap</td>
<td>14.55</td>
<td>19.61</td>
<td>29.70</td>
</tr>
<tr>
<td>Overlap</td>
<td>9.00</td>
<td>11.55</td>
<td>18.18</td>
</tr>
<tr>
<td>K-SVD</td>
<td>8.1</td>
<td>10.05</td>
<td>17.74</td>
</tr>
</tbody>
</table>

This is a more challenging case, where the DCT is not a suitable dictionary.

- For Redundant DCT we get RMSE=16.13, and
- For K-SVD (15 iterations) we get RMSE=12.74
Super-Resolution [Zeyde, Protter, & E. (‘11)]

- Given a low-resolution image, we desire to enlarge it while producing a sharp looking result. This problem is referred to as “Single-Image Super-Resolution”.

- Image scale-up using bicubic interpolation is far from being satisfactory for this task.

- Recently, a sparse and redundant representation technique was proposed [Yang, Wright, Huang, and Ma (‘08)] for solving this problem, by training a coupled-dictionaries for the low- and high res. images.

- We extended and improved their algorithms and results.
This book is about convex optimization, a special class of mathematical optimization problems, which includes least-squares and linear programming problems. It is well known that least-squares and linear programming problems have a fairly complete theory, arise in a variety of applications, and can be solved numerically very efficiently. The basic point of this book is that the same can be said for the larger class of convex optimization problems.

While the mathematics of convex optimization has been studied for about a century, several related recent developments have stimulated new interest in this topic. The first is the recognition that interior-point methods, developed in the 1980s to solve linear programming problems, can be used to solve convex optimization problems as well. These new methods allow us to solve certain new classes of convex optimization problems, such as semidefinite programs and second-order cone programs, almost as easily as linear programs.

The second development is the discovery that convex optimization problems (beyond least-squares and linear programs) are more prevalent in practice than was previously thought. Since 1990 many applications have been discovered in areas such as automatic control systems, estimation and signal processing, communications and networks, electronic circuit design, data analysis and modeling, and statistics, and finance. Convex optimization has also found wide application in combinatorial optimization and global optimization, where it is used to find bounds on the optimal value, as well as approximate solutions. We believe that many other applications of convex optimization are still waiting to be discovered.

There are great advantages to recognizing or formulating a problem as a convex optimization problem. The most basic advantage is that the problem can then be solved, very reliably and efficiently, using interior-point methods or other specialized methods for convex optimization. These solution methods are reliable enough to be embedded in a computer-aided design or analysis tool, or even a real-time reactive or automatic control system. There are also theoretical or conceptual advantages of formulating a problem as a convex optimization problem. The associated dual

An amazing variety of practical problems (design, analysis, and operation) can be formulated as optimization problem, or some variation such. Indeed, mathematical optimization has been widely used in engineering, in electrical systems, and optimal design problems in aerospace engineering. Optimization design and operation, finance, supply chain, and other areas. The list of applications is still growing.

For most of these applications, mathematicians refer to a human decision maker, system designer, or process, checks the results, and modifies the system as necessary. This human decision making process is motivated by the optimization problem, e.g., buying a portfolio.
Super-Resolution – Results (2)

Given image

Scaled-Up (factor 2:1) using the proposed algorithm, PSNR=29.32dB (3.32dB improvement over bicubic)
Super-Resolution – Results (2)

The Original  |  Bicubic Interpolation  |  SR result
Super-Resolution – Results (2)

The Original

Bicubic Interpolation

SR result