MMSE Estimation for Sparse Representation Modeling*

Michael Elad
The Computer Science Department
The Technion – Israel Institute of technology
Haifa 32000, Israel

Joint work with
Irad Yavneh & Matan Protter

École Polytechnique

April 6th, 2009
Noise Removal?

In this talk we focus on signal/image denoising …

- Important: (i) Practical application; (ii) A convenient platform for testing basic ideas in signal/image processing.

- Many Considered Directions: Partial differential equations, Statistical estimators, Adaptive filters, Inverse problems & regularization, Wavelets, Example-based techniques, Sparse representations, …

- Main Massage Today: Several sparse representations can be found and used for better denoising performance – we introduce, motivate, discuss, demonstrate, and explain this new idea.
Agenda

1. Background on Denoising with Sparse Representations
2. Using More than One Representation: Intuition
3. Using More than One Representation: Theory
4. A Closer Look At the Unitary Case
5. Summary and Conclusions
Part I

Background on Denoising with Sparse Representations
Many of the proposed signal denoising algorithms are related to the minimization of an energy function of the form:

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

- **\( y \)**: 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 signals** of interest.

---

Thomas Bayes
1702 - 1761
Sparse Representation Modeling

- Every column in \( D \) (dictionary) is a prototype signal (atom).

- The vector \( \alpha \) is generated randomly with few (say \( L \) for now) non-zeros at random locations and with random values.

A fixed Dictionary

\( D \)

\( \alpha \)

A sparse & random vector

\( M \)

\( x \)

\( N \)

\( K \)
Back to Our MAP Energy Function

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

- The vector $\alpha$ is the representation (sparse/redundant).

- Bottom line: Denoising of $y$ is done by minimizing

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

or

$$\min_{\alpha} \|\alpha\|_0 \quad \text{s.t.} \quad \|D\alpha - y\|_2^2 \leq \epsilon^2$$
The Solver We Use: Greed Based

- 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 $\|D \alpha - y\|_2^2$ is below the destination threshold.

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

OMP finds one atom at a time for approximating the solution of

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

Initialization

1. $n = 0, \alpha^0 = 0$
2. $r^0 = y - D\alpha^0 = y$
3. and $S^0 = \{\}$

Main Iteration

1. Compute $E(i) = \min_{z} \|z \cdot d_i - r^{n-1}\|_2$ for $1 \leq i \leq K$
2. Choose $i_0$ s.t. $\forall 1 \leq i \leq K$, $E(i_0) \leq E(i)$
3. Update $S^n : S^n = S^{n-1} \cup \{i_0\}$
4. LS : $\alpha^n = \min_{\alpha} \|D\alpha - y\|_2$ s.t. $\text{supp}\{\alpha\} = S^n$
5. Update Residual : $r^n = y - D\alpha^n$

If $\|r^n\|_2 \leq \varepsilon$ then **Stop**
Part II
Using More than One Representation: Intuition
Consider the denoising problem

$$\min_{\alpha} \|\alpha\|_0^0 \quad \text{s.t.} \quad \|D\alpha - y\|_2^2 \leq \varepsilon^2$$

and suppose that we can find a group of J candidate solutions

$$\{\alpha_j\}_{j=1}^J$$

such that

$$\forall j \left\{ \begin{array}{l} \|\alpha_j\|_0^0 << N \\ \|D\alpha_j - y\|_2^2 \leq \varepsilon^2 \end{array} \right\}$$

Basic Questions:

- **What** could we do with such a set of competing solutions in order to better denoise y?
- **Why** should this help?
- **How** shall we practically find such a set of solutions?

Relevant work: [Larsson & Selen ('07)]
[Larsson & Selen ('07)]
[Schintter et. al. ('08)]
[Elad and Yavneh ('08)]
Motivation - General

Why bother with such a set?

- Because each representation conveys a different story about the desired signal.
- Because pursuit algorithms are often wrong in finding the sparsest representation, and then relying on their solution is too sensitive.
- … Maybe there are “deeper” reasons?
Our Motivation

- An intriguing relationship between this idea and the common-practice in example-based techniques, where several examples are merged.

- Consider the Non-Local-Means [Buades, Coll, & Morel ('05)]. It uses (i) a local dictionary (the neighborhood patches), (ii) it builds several sparse representations (of cardinality 1), and (iii) it merges them.

- Why not take it further, and use general sparse representations?
Generating Many Representations

Our* Answer: Randomizing the OMP

Initialization

\[ n = 0, \quad \alpha^0 = 0 \]
\[ r^0 = y - D\alpha^0 = y \]
and \( S^0 = \{ \} \)

Main Iteration

1. Compute \( E(i) = \min_z \left\| z \cdot d_i - r^{n-1} \right\|_2 \) for \( 1 \leq i \leq K \)
2. Choose \( i_0 \) with probability \( \propto \exp\{-c \cdot E(i)\} \)
3. For now, let's set the parameter \( c \) manually for best performance.
4. Later we shall define a way to set it automatically.
5. \[ \| r^n \|_2 \leq \varepsilon \]

* Larsson and Schnitter propose a more complicated and deterministic tree pruning method.
**Proposed Experiment:**

- Form a random dictionary $D$.
- Multiply by a sparse vector $\alpha_0$ ($\|\alpha_0\|_0 = 10$).
- Add Gaussian iid noise $v$ with $\sigma=1$ and obtain $y$.
- Solve the problem
  \[
  \min_{\alpha} \|\alpha\|_0 \quad \text{s.t.} \quad \|D\alpha - y\|_2^2 \leq 100
  \]
  using OMP, and obtain $\alpha^{\text{OMP}}$.
- Use Random-OMP and obtain $\left\{\alpha_j^{\text{RandOMP}}\right\}_{j=1}^{1000}$.
- Let's look at the obtained representations ...
Some Observations

We see that

• The OMP gives the sparsest solution

• Nevertheless, it is not the most effective for denoising.

• The cardinality of a representation does not reveal its efficiency.
The Surprise (at least for us) ...

Let's propose the average
\[ \hat{\alpha} = \frac{1}{1000} \sum_{j=1}^{1000} \alpha_{j} \]
as our representation.

This representation **IS NOT SPARSE AT ALL** but it gives

\[ \frac{\| D\hat{\alpha} - D\alpha_0 \|_2^2}{\| y - D\alpha_0 \|_2^2} = 0.06 \]
Is It Consistent? ... Yes!

Here are the results of 1000 trials with the same parameters ...

Cases of zero solution
Part III
Using More than One Representation: Theory
Our Signal Model

- **D** is fixed and known.

- The vector \( \alpha \) is built by:
  - Choosing the support \( s \) with probability \( P(s) \) from all the \( 2^K \) possibilities \( \Omega \).
  - **For simplicity, assume that \(|s| = k\) is fixed and known.**
  - Choosing the \( \alpha_s \) coefficients using iid Gaussian entries \( N(0, \sigma_x) \).

- The ideal signal is \( x = D\alpha = D_s\alpha_s \).

The p.d.f. \( P(\alpha) \) and \( P(x) \) are clear and known.
Adding Noise

The noise \( v \) is additive white Gaussian vector with probability \( P_v(v) \).

\[ P(y|x) = C \cdot \exp \left\{ -\frac{\|x - y\|^2}{2\sigma^2} \right\} \]

The conditional p.d.f.'s \( P(y|s) \), \( P(s|y) \), and even also \( P(x|y) \) are all clear and well-defined (although they may appear nasty).
The Key - The Posterior $P(x | y)$

We have access to $P(x | y)$

- MAP
  - $\hat{x}^{\text{MAP}} = \text{ArgMax}_{x} P(x | y)$

- Oracle
  - $\hat{x}^{\text{oracle}}$
  - Oracle known support $s$

- MMSE
  - $\hat{x}^{\text{MMSE}} = E\{x | y\}$

- The estimation of $\alpha$ and multiplication by $D$ is equivalent to the above.
- These two estimators are impossible to compute, as we show next.
MMSE Estimation for Sparse Representation Modeling
By: Michael Elad

\[ P(\alpha | y, s) = P(\alpha_s | y) \]

\[ P(y | \alpha_s) \propto \exp \left\{ - \frac{\| y - D_s \alpha_s \|^2}{2\sigma^2} \right\} \]

\[ P(\alpha_s) \propto \exp \left\{ - \frac{\| \alpha_s \|^2}{2\sigma_x^2} \right\} \]

\[ P(\alpha_s | y) \propto \exp \left\{ - \frac{\| y - D_s \alpha_s \|^2}{2\sigma^2} - \frac{\| \alpha_s \|^2}{2\sigma_x^2} \right\} \]

\[ \hat{\alpha}_s = \left[ \frac{1}{\sigma^2} D_s^T D_s + \frac{1}{\sigma_x^2} I \right]^{-1} \frac{1}{\sigma^2} D_s^T y \]

Comments:

- This estimate is both the MAP and MMSE.
- The oracle estimate of \( x \) is obtained by multiplication by \( D_s^* \).

* When \( s \) is known
The MAP Estimation

$$
\hat{\alpha}_{MAP} = \text{ArgMax}_{\alpha_s \in \Omega, s \in \Omega} P(y | \alpha_s) \cdot P(\alpha_s | y, s)
$$

$$
P(s | y) \propto P(s) \cdot P(y | s) = \ldots
$$

$$
\propto P(s) \cdot \exp \left\{ \frac{h_s^T Q_s^{-1} h_s}{2} + \frac{\log(\det(Q_s^{-1}))}{2} \right\}
$$

$$
\hat{s}_{MAP} = \text{ArgMax}_{s \in \Omega} P(s) \cdot \exp \left\{ \frac{h_s^T Q_s^{-1} h_s}{2} + \frac{\log(\det(Q_s^{-1}))}{2} \right\}
$$
The MAP Estimation

Implications:

\[ \hat{s}^{\text{MAP}} = \arg\max_{s \in \Omega} P(s) \cdot \exp \left\{ \frac{h_s^T Q_s^{-1} h_s}{2} + \frac{\log(\det(Q_s^{-1}))}{2} \right\} \]

- The MAP estimator requires to test all the possible supports for the maximization. In typical problems, this is impossible as there is a combinatorial set of possibilities.

- This is why we rarely use exact MAP, and we typically replace it with approximation algorithms (e.g., OMP).
The MMSE Estimation

\[ \hat{\alpha}_{\text{MMSE}} = \mathbb{E}\{\alpha \mid y\} = \sum_{s \in \Omega} P(s \mid y) \cdot \mathbb{E}\{\alpha \mid y, s\} \]

\[ P(s \mid y) \propto P(s) \cdot P(y \mid s) = ... \]

\[ \propto P(s) \cdot \exp \left\{ \frac{h_s^T Q_s^{-1} h_s}{2} + \frac{\log(\det(Q_s^{-1}))}{2} \right\} \]

\[ \hat{\alpha}_s = Q_s^{-1} h_s \]

\[ \hat{\alpha}_{\text{MMSE}} = \sum_{s \in \Omega} P(s \mid y) \cdot \alpha_s \]
The MMSE Estimation

\[
\hat{\alpha}_{\text{MMSE}} = E\{\alpha \mid y\} = \sum_{s \in \Omega} P(s \mid y) \cdot E\{\alpha \mid y, s\}
\]

Implications:

- The best estimator (in terms of L_2 error) is a weighted average of many sparse representations!!!

- As in the MAP case, in typical problems one cannot compute this expression, as the summation is over a combinatorial set of possibilities. We should propose approximations here as well.
The Case of $|s| = k=1$

$$P(s | y) \propto P(s) \cdot \exp \left\{ \frac{h_s^T Q_s^{-1} h_s}{2} + \frac{\log(\det(Q_s^{-1}))}{2} \right\}$$

- Based on this we can propose a greedy algorithm for both MAP and MMSE:
  - **MAP** – choose the atom with the largest inner product (out of K), and do so one at a time, while freezing the previous ones (almost OMP).
  - **MMSE** – draw at random an atom in a greedy algorithm, based on the above probability set, getting close to $P(s | y)$ in the overall draw.
The MMSE estimation we got requires a sweep through all supports (i.e. combinatorial search) - impractical.

Similarly, an explicit expression for $P(x/y)$ can be derived and maximized - this is the MAP estimation, and it also requires a sweep through all possible supports - impractical too.

The OMP is a (good) approximation for the MAP estimate.

The Random-OMP is a (good) approximation of the Minimum-Mean-Squared-Error (MMSE) estimate. It is close to the Gibbs sampler of the probability $P(s|y)$ from which we should draw the weights.

Back to the beginning: Why Use Several Representations?
Because their average leads to a provable better noise suppression.
Comparative Results

The following results correspond to a small dictionary (20×30), where the combinatorial formulas can be evaluated as well.

Parameters:
- N=20, K=30
- True support=3
- \( \sigma_x=1 \)
- J=10 (RandOMP)
- Averaged over 1000 experiments
Part IV
A Closer Look At the Unitary Case
\[ DD^T = D^T D = I \]
Few Basic Observations

Let us denote \( \beta = D^T y \)

\[
Q_s = \frac{1}{\sigma^2} D_s^T D_s + \frac{1}{\sigma_x^2} I = \frac{\sigma^2 + \sigma_x^2}{\sigma^2 \sigma_x^2} I
\]

\[
h_s = \frac{1}{\sigma^2} D_s^T y = \frac{1}{\sigma^2} \beta_s
\]

\[
\hat{\alpha}_s = Q_s^{-1} h_s = \frac{\sigma^2 \sigma_x^2}{\sigma^2 + \sigma_x^2} \cdot \frac{1}{\sigma^2} \beta_s = c \cdot \beta_s \quad \text{(The Oracle)}
\]
Back to the MAP Estimation

\[ \hat{s}_{\text{MAP}} = \arg\max_{s \in \Omega} \exp \left\{ \frac{h_s^T Q_s^{-1} h_s}{2} - \frac{\log(\det(Q_s^{-1}))}{2} \right\} \]

This means that MAP estimation can be easily evaluated by computing \( \beta \), sorting its entries in descending order, and choosing the \( k \) leading ones.

We assume \(|s|=k\) fixed with equal probabilities.
Closed-Form Estimation

- It is well-known that MAP enjoys a closed form and simple solution in the case of a unitary dictionary $D$.

- This closed-form solution takes the structure of thresholding or shrinkage. The specific structure depends on the fine details of the model assumed.

- It is also known that OMP in this case becomes exact.

What about the MMSE? Could it have a simple closed-form solution too?
The MMSE ... Again

This is the formula we got:

$$\hat{\alpha}_{\text{MMSE}} = c \cdot \sum_{s \in \Omega} P(s | y) \cdot \beta_s$$

We combine linearly many sparse representations (with proper weights)

The result is one effective representation (not sparse anymore)
The MMSE ... Again

This is the formula we got:

\[ \hat{\alpha}_{\text{MMSE}} = c \cdot \sum_{s \in \Omega} P(s | y) \cdot \beta_s \]

- We change the above summation to

\[ \hat{\alpha}_{\text{MMSE}} = \sum_{j=1}^{K} q_j^k \cdot \beta_j \cdot e_j \]

where there are K contributions (one per each atom) to be found and used.

- We have developed a closed-form recursive formula for computing the q coefficients.
Towards a Recursive Formula

We have seen that the governing probability for the weighted averaging is given by

\[
P(s | y) = \ldots \propto \exp \left\{ \frac{c}{2\sigma^2} \cdot \|\beta_s\|^2 \right\}
\]

\[
\hat{\alpha}_{\text{MMSE}} = c \cdot \sum_{s \in \Omega} P(s | y) \cdot \beta_s
\]

\(q_j\) indicating if \(j\) is in \(s\)
The Recursive Formula

\[ q_j^k = \sum_{s \in \Omega} \left( \prod_{i \in s} q_i \right) \cdot l_s(j) = ... = k \cdot \frac{q_j^1(1 - q_j^{k-1})}{1 - \sum_{\ell=1}^{K} q_j^\ell q_j^{\ell-1}} \]

where \( q_j^1 = q_j \)

\[ \left\{ q_j^1 \right\}_{j=1}^{K} \quad \left\{ q_j^2 \right\}_{j=1}^{K} \quad \left\{ q_j^3 \right\}_{j=1}^{K} \quad \left\{ q_j^k \right\}_{j=1}^{K} \]
An Example

This is a synthetic experiment resembling the previous one, but with few important changes:

- \( \mathbf{D} \) is unitary
- The representation’s cardinality is 5 (the higher it is, the weaker the Random-OMP becomes)
- Dimensions are different: \( N=K=64 \)
- \( J=20 \) (RandOMP runs)
Part V

Summary and Conclusions
Today We Have Seen that ... 

Sparsity and Redundancy are used for denoising of signals/images.

How?

By finding the sparsest representation and using it to recover the clean signal.

Today we have shown that averaging several sparse representations for a signal lead to better denoising, as it approximates the MMSE estimator.

Can we do better?

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