

A Weighted Average of Sparse Several Representations is Better than the Sparsest One Alone*

Michael Elad

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



SIAM Conference on
Imaging Science IS`08
Session on Topics in Sparse and
Redundant Representations – Part I
July 8th, 2008 San-Diego

Joint work with Irad Yavneh

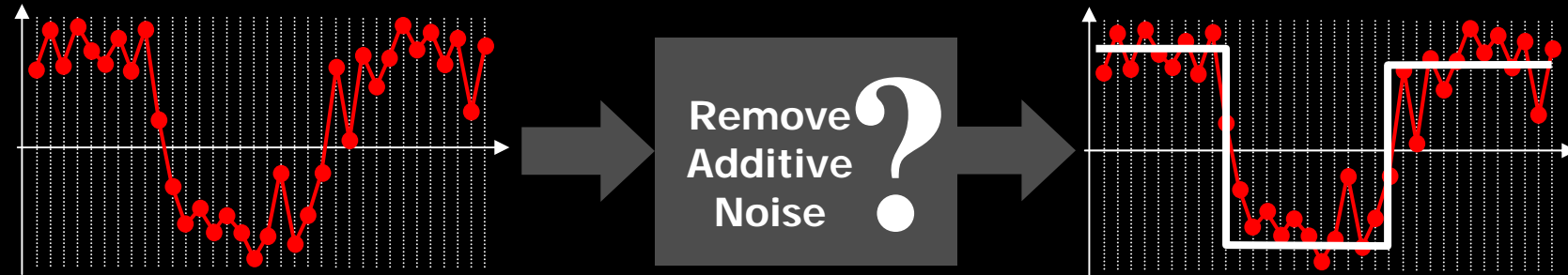


The Computer Science Department The
Technion – Israel Institute of technology
Haifa 32000, Israel



Noise Removal?

Today 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 Message Today:** Several sparse representations can be found and used for better denoising performance – we introduce, demonstrate and explain this new idea.



Part I

The Basics of Denoising by Sparse Representations



Denoising By Energy Minimization

Many of the proposed signal denoising algorithms are related to the minimization of an energy function of the form

$$f(\underline{x}) = \frac{1}{2} \|\underline{x} - \underline{y}\|_2^2 + \text{Pr}(\underline{x})$$

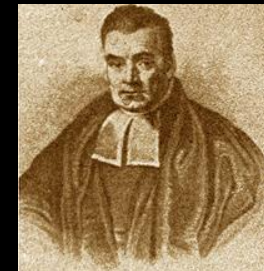
\underline{y} : Given measurements

\underline{x} : Unknown to be recovered

Relation to
measurements

Prior or regularization

- 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



The Evolution Of $\Pr(\underline{x})$

During the past several decades we have made all sort of guesses about the prior $\Pr(\underline{x})$ for signals/images:

$$\Pr(\underline{x}) = \lambda \|\underline{x}\|_2^2$$



Energy

$$\Pr(\underline{x}) = \lambda \|\mathbf{L}\underline{x}\|_2^2$$



Smoothness

$$\Pr(\underline{x}) = \lambda \|\mathbf{L}\underline{x}\|_{\mathbf{W}}^2$$



**Adapt+
Smooth**

$$\Pr(\underline{x}) = \lambda \rho\{\mathbf{L}\underline{x}\}$$



**Robust
Statistics**

$$\Pr(\underline{x}) = \lambda \|\|\nabla \underline{x}\|\|_1$$



**Total-
Variation**

$$\Pr(\underline{x}) = \lambda \|\mathbf{W}\underline{x}\|_1$$



**Wavelet
Sparsity**

$$\Pr(\underline{x}) = \lambda \|\underline{\alpha}\|_0^0$$

for $\underline{x} = \mathbf{D}\underline{\alpha}$

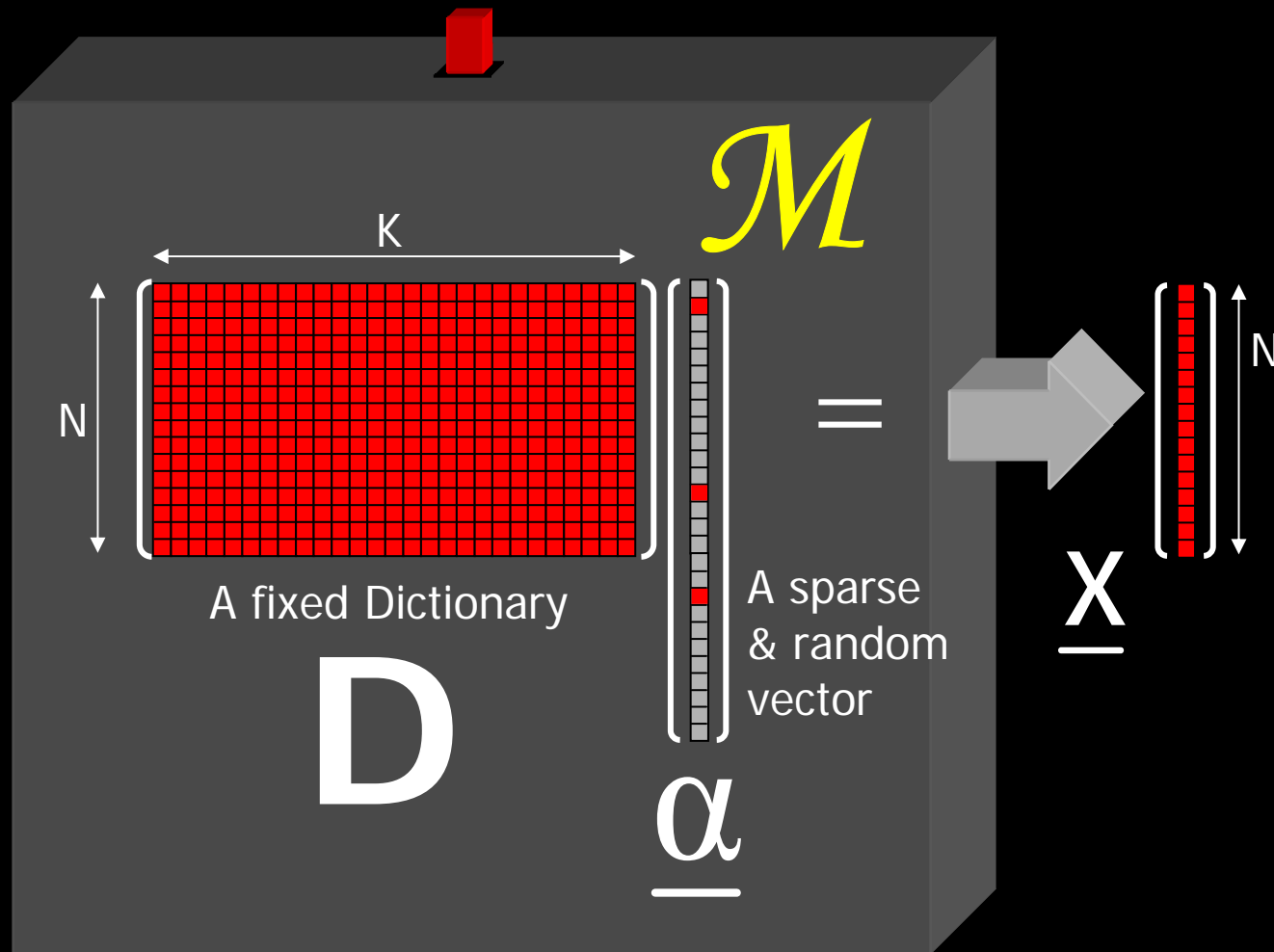


**Sparse &
Redundant**

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



Sparse Modeling of Signals

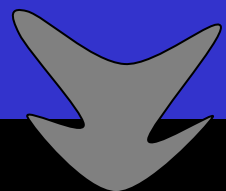


- Every column in D (dictionary) is a prototype signal (atom).
- The vector α is generated randomly with few (say L) non-zeros at random locations and with random values.

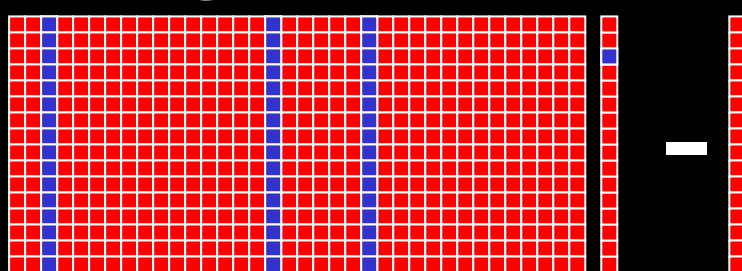


Back to Our MAP Energy Function

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

$$\frac{1}{2} \left\| \begin{array}{c} \underline{x} \\ \underline{y} \end{array} \right\|_2^2$$


- The vector $\underline{\alpha}$ is the representation (**sparse/redundant**).

$$\mathbf{D}\underline{\alpha} - \underline{y} =$$


- Bottom line: Denoising of \underline{y} is done by minimizing

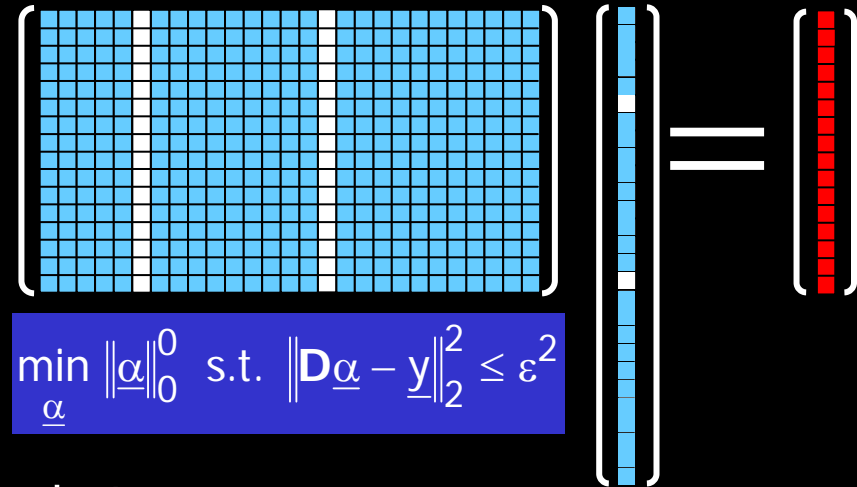
$$\min_{\underline{\alpha}} \left\| \mathbf{D}\underline{\alpha} - \underline{y} \right\|_2^2 \quad \text{s.t.} \quad \left\| \underline{\alpha} \right\|_0 \leq L \quad \text{or} \quad \min_{\underline{\alpha}} \left\| \underline{\alpha} \right\|_0 \quad \text{s.t.} \quad \left\| \mathbf{D}\underline{\alpha} - \underline{y} \right\|_2^2 \leq \varepsilon^2 .$$



The Solver We Use: Greedy 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 $\|\mathbf{D}\underline{\alpha} - \underline{y}\|_2$ is below the destination threshold.
- ❑ The Orthogonal MP (OMP) is an improved version that re-evaluates the coefficients by Least-Squares after each round.



Orthogonal Matching Pursuit

OMP finds one atom at a time for approximating the solution of $\min_{\underline{\alpha}} \|\underline{\alpha}\|_0$ s.t. $\|\mathbf{D}\underline{\alpha} - \underline{y}\|_2^2 \leq \varepsilon^2$

Initialization

$n = 0, \underline{\alpha}^0 = 0$
 $\underline{r}^0 = \underline{y} - \mathbf{D}\underline{\alpha}^0 = \underline{y}$
and $S^0 = \{\}$

$n = n + 1$

Main Iteration

1. Compute $E(i) = \min_z \|z \cdot \underline{d}_i - \underline{r}^{n-1}\|$ 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 : $\underline{\alpha}^n = \min_{\underline{\alpha}} \|\mathbf{D}\underline{\alpha} - \underline{y}\|$ s.t. $\text{supp}\{\underline{\alpha}\} = S^n$
5. Update Residual : $\underline{r}^n = \underline{y} - \mathbf{D}\underline{\alpha}^n$

No

$$\|\underline{r}^n\|_2 \leq \varepsilon$$

Yes

Stop



Part II

Finding & Using More than One Sparse Representation



Back to the Beginning. What If ...

Consider the denoising problem

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

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

$$\{\underline{\alpha}_j\}_{j=1}^J$$

such that

$$\forall j \left\{ \begin{array}{l} \|\underline{\alpha}_j\|_0 \ll N \\ \|\mathbf{D}\underline{\alpha}_j - \underline{y}\|_2^2 \leq \varepsilon^2 \end{array} \right\}$$

Basic Questions:

- ❑ **What** could we do with such a set of competing solutions in order to better denoise \underline{y} ?
- ❑ **Why** should this work?
- ❑ **How** shall we practically find such a set of solutions?



These questions were studied and answered recently

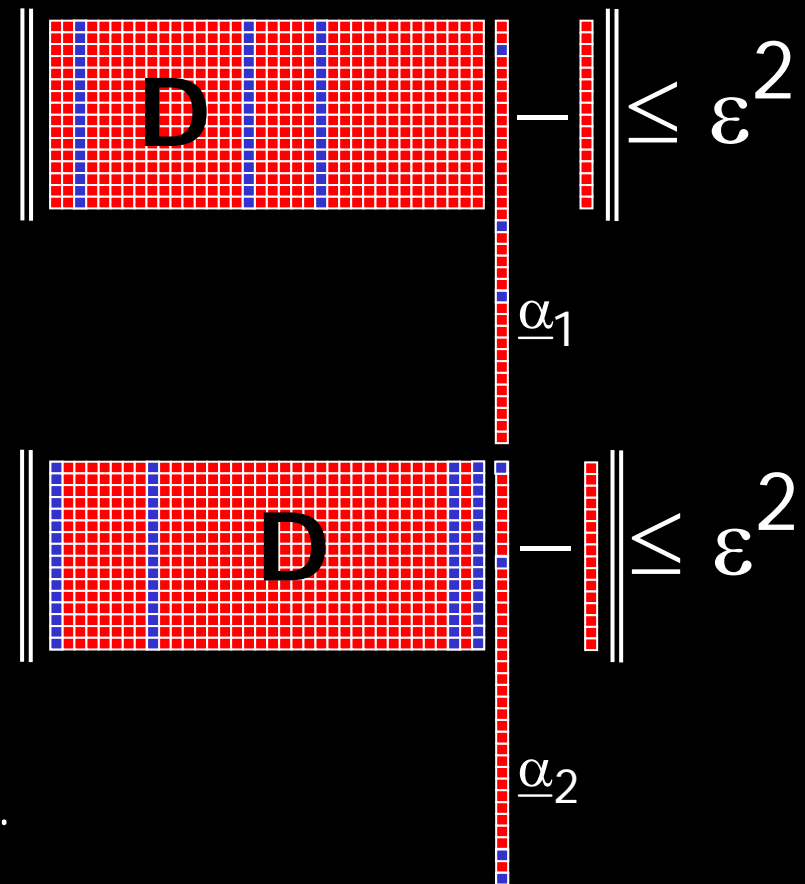
[Elad and Yavneh ('08)]



Motivation

Why bother with such a set?

- Because of the intriguing relation to example-based techniques, where several nearest-neighbors for the signal are used jointly.
- 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 becomes too sensitive.
- ... Maybe there are "deeper" reasons?



Generating Many Representations

Our Answer: Randomizing the OMP

Initialization

$n = 0, \underline{\alpha}^0 = 0$
 $\underline{r}^0 = \underline{y} - \mathbf{D}\underline{\alpha}^0 = \underline{y}$
and $S^0 = \{\}$

$n = n + 1$

Main Iteration

1. Compute $E(i) = \min_z \|z \cdot \underline{d}_i - \underline{r}^{n-1}\|$ for $1 \leq i \leq K$
2. Choose i_0 with probability $\propto \exp\{-c \cdot E(i)\}$
3. Update $S^n : S^n = S^{n-1} \cup \{i_0\}$
4. LS : $\underline{\alpha}^n = \min_{\underline{\alpha}} \|\mathbf{D}\underline{\alpha} - \underline{y}\|$ s.t. $\text{supp}\{\underline{\alpha}\} = S^n$
5. Update Residual : $\underline{r}^n = \underline{y} - \mathbf{D}\underline{\alpha}^n$

No

$$\|\underline{r}^n\|_2 \leq \varepsilon$$

Yes

Stop



Lets Try

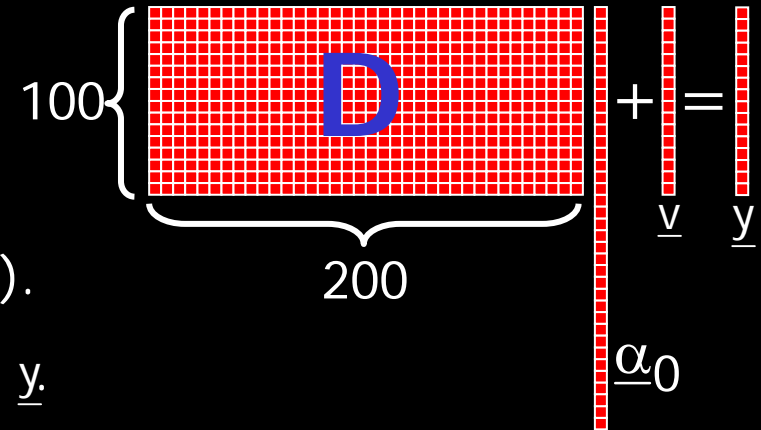
Proposed Experiment :

- ❑ Form a random \mathbf{D} .
- ❑ Multiply by a sparse vector $\underline{\alpha}_0$ ($\|\underline{\alpha}_0\|_0 = 10$).
- ❑ Add Gaussian iid noise ($\sigma=1$) and obtain \underline{y} .
- ❑ Solve the problem

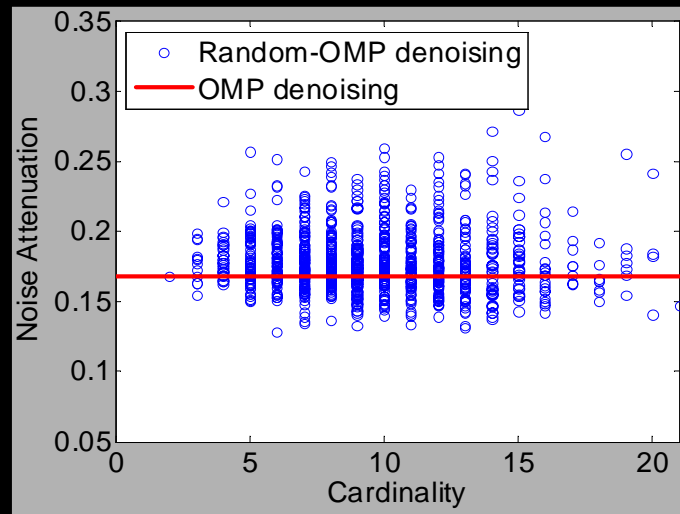
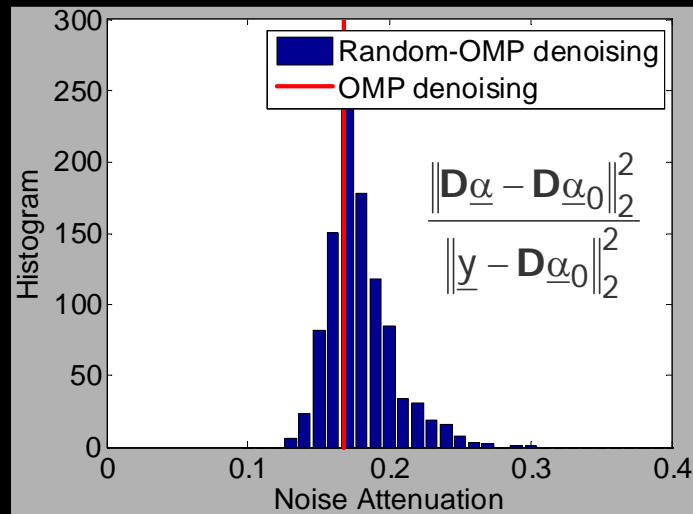
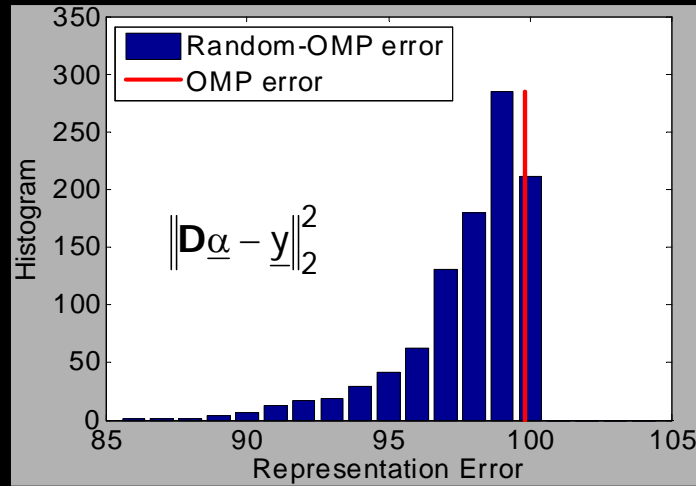
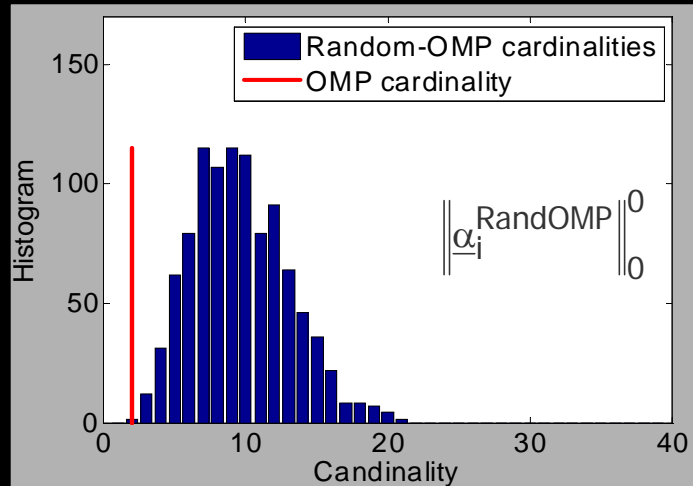
$$\min_{\underline{\alpha}} \|\underline{\alpha}\|_0 \quad \text{s.t.} \quad \|\mathbf{D}\underline{\alpha} - \underline{y}\|_2^2 \leq 100$$

using OMP, and obtain $\underline{\alpha}^{\text{OMP}}$.

- ❑ Use RandOMP and obtain $\{\underline{\alpha}_j^{\text{RandOMP}}\}_{j=1}^{1000}$.
- ❑ Lets 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 ...

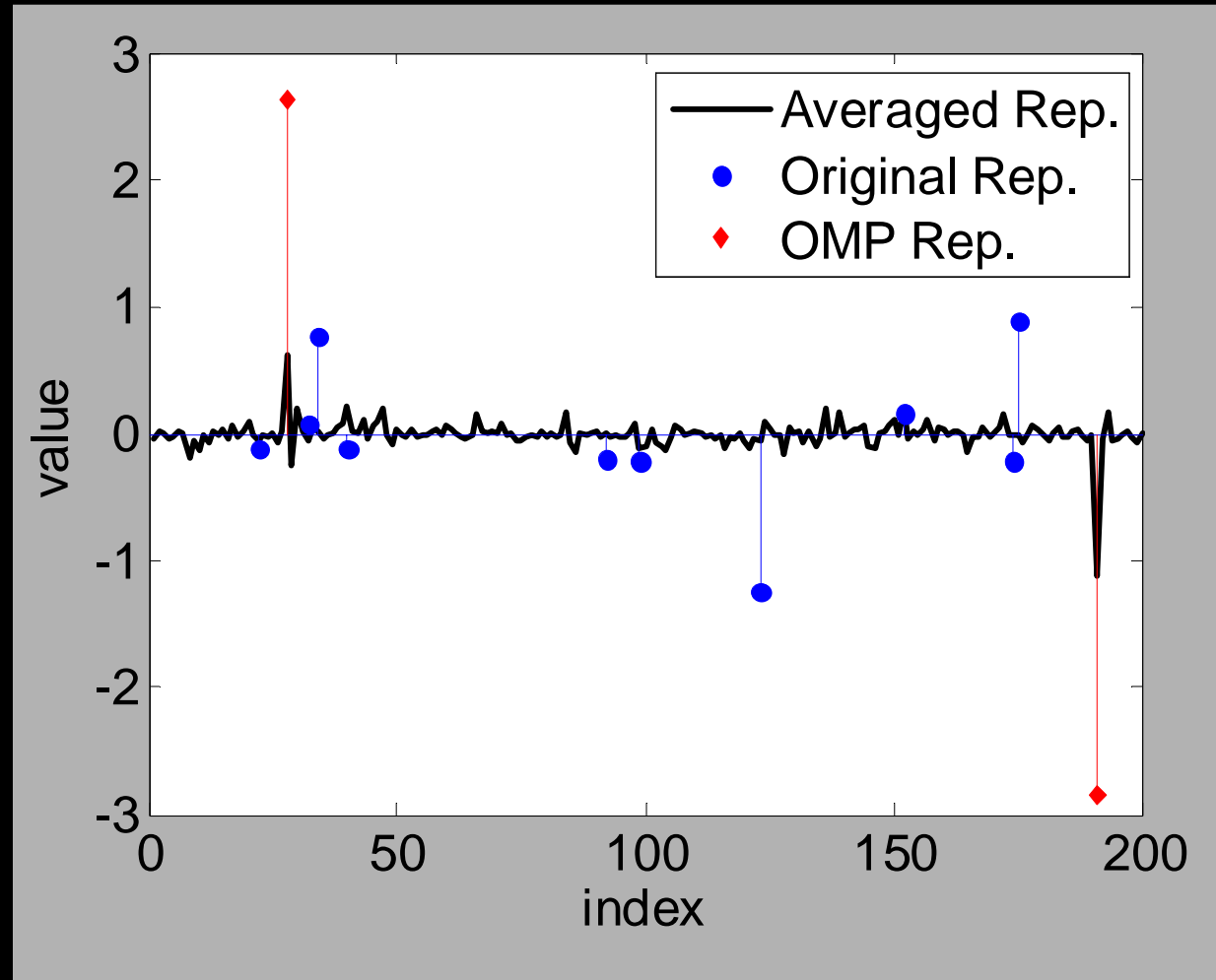
Lets propose the average

$$\hat{\underline{\alpha}} = \frac{1}{1000} \sum_{j=1}^{1000} \underline{\alpha}_j^{\text{RandOMP}}$$

as our representation

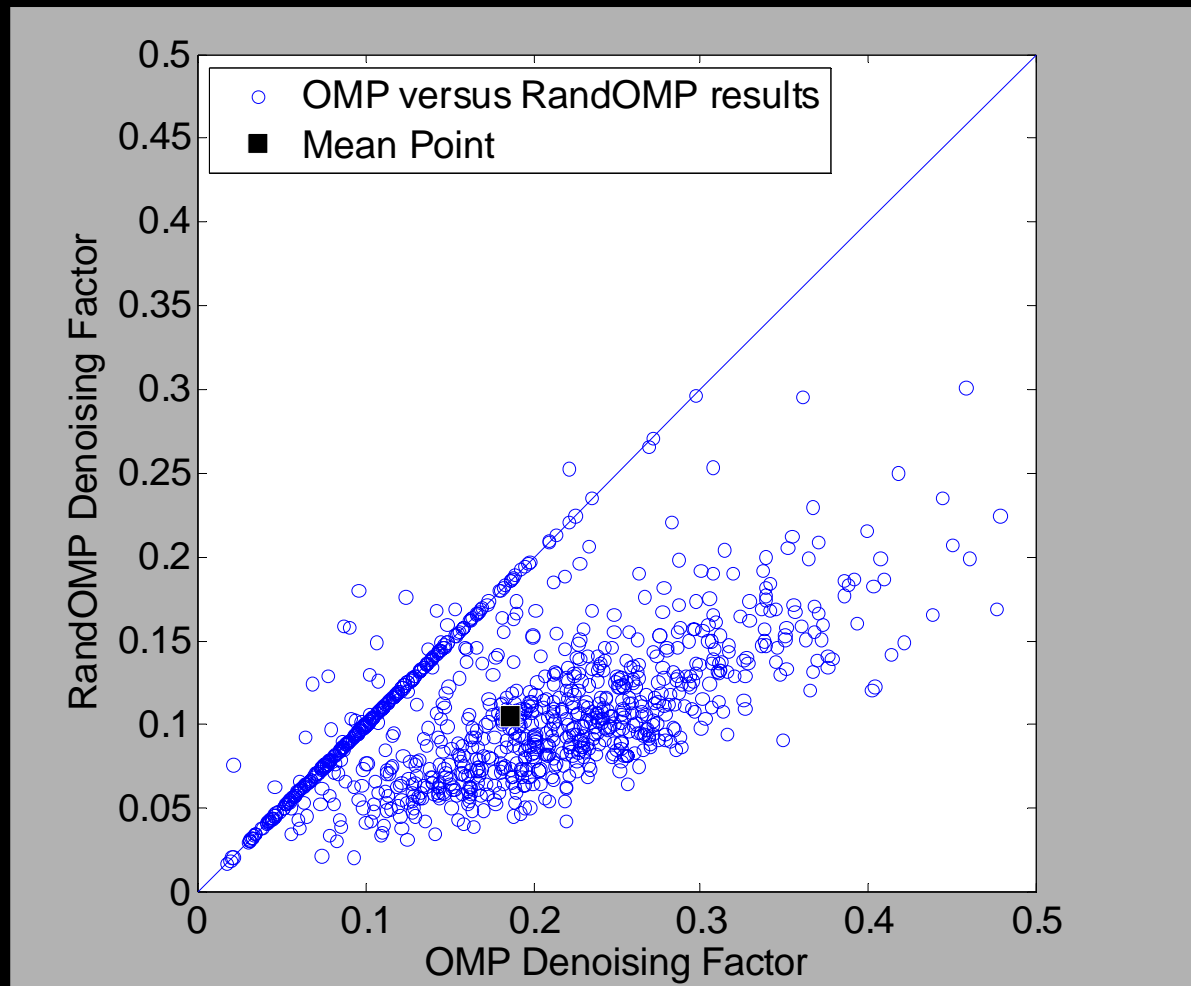


This representation **IS NOT SPARSE AT ALL** but its noise attenuation is: **0.06 (OMP gives 0.16)**

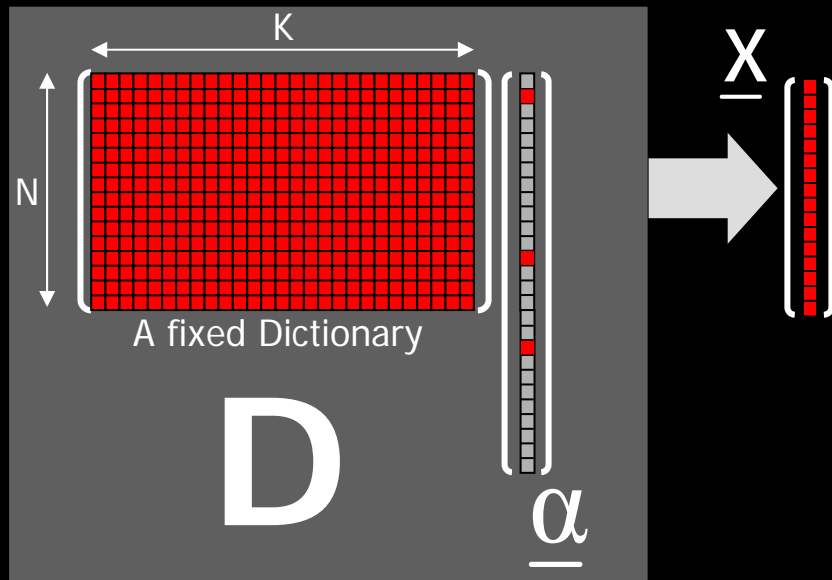


Is It Consistent? Yes!

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

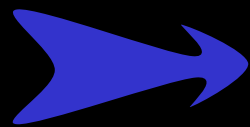


The Explanation – Our Signal Model



Signal Model Assumed

- D is fixed and known
- $\underline{\alpha}$ is built by:
 - Choosing the support S w.p. $P(S)$ of all the 2^K possibilities Ω ,
 - Choosing the coefficients using iid Gaussian entries* $N(0, \sigma_x)$: $P(\underline{x}|S)$.
- The ideal signal is $\underline{x} = D\underline{\alpha}$.

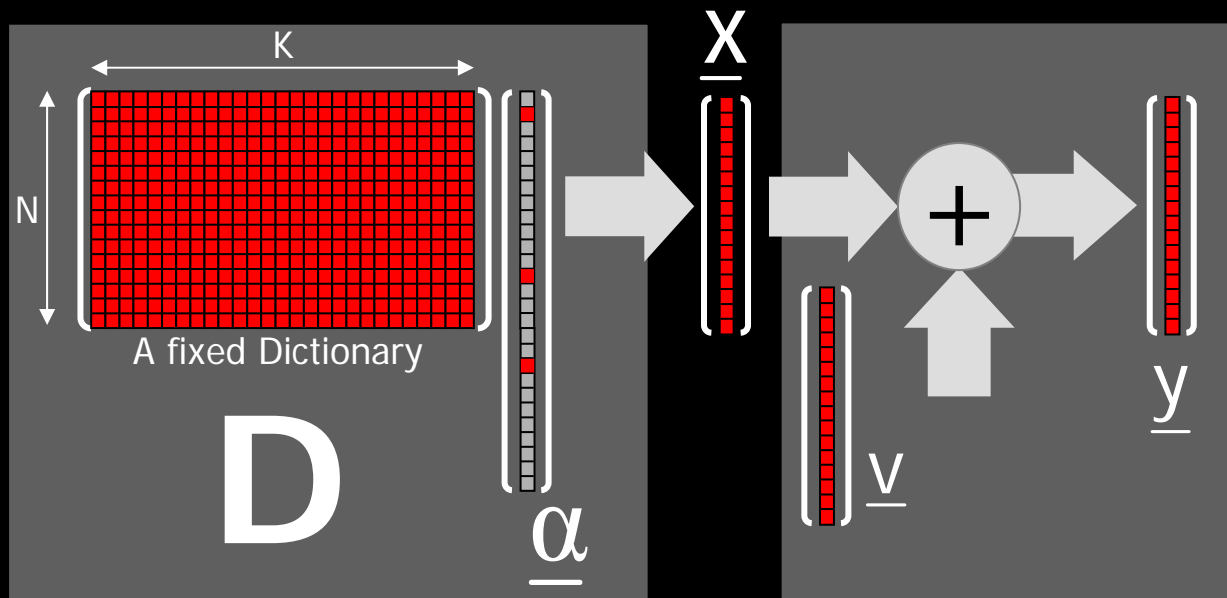


The p.d.f. of the signal $P(\underline{x})$ is:
$$P(\underline{x}) = \sum_{S \in \Omega} P(\underline{x}|S)P(S)$$

* Not exactly, but this does not change our analysis.



The Explanation – Adding Noise



Noise Assumed:

The noise \underline{v} is additive white Gaussian vector with probability $P_v(\underline{v})$

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

The p.d.f. of the noisy signal $P(\underline{y})$, and the conditionals $P(\underline{y}|\underline{S})$ and $P(\underline{S}|\underline{y})$ are clear and well-defined (although nasty).



Some Algebra Leads To

$$\hat{\underline{x}}^{\text{MMSE}} = E\{\underline{x}|\underline{y}\} = \dots = \frac{\sigma_x^2}{\sigma_x^2 + \sigma^2} \sum_{S \in \Omega} P(S|\underline{y}) \cdot \underline{y}_S$$

$$P(S|\underline{y}) \propto \exp\left\{ \frac{\sigma_x^2}{\sigma_x^2 + \sigma^2} \cdot \frac{\|\underline{y}_S\|^2}{2\sigma^2} \right\} P(S)$$

Projection of the signal \underline{y}
onto the support S

$$\underline{y}_S = \mathbf{D} \cdot \left\{ \underset{\underline{\alpha}}{\text{ArgMin}} \|\underline{y} - \mathbf{D}\underline{\alpha}\| \text{ s.t. } \text{supp}\{\underline{\alpha}\} = S \right\}$$

Implications:

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



As It Turns Out ...

- ❑ The MMSE estimation we got requires a sweep through all 2^K supports (i.e. combinatorial search) – impractical.
- ❑ Similarly, an explicit expression for $P(\underline{x}/\underline{y})$ 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 RandOMP is a (good) approximation of the Minimum-Mean-Squared-Error (MMSE) estimate. It is close to the Gibbs sampler of the probability $P(S|\underline{y})$ from which we should draw the weights.

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

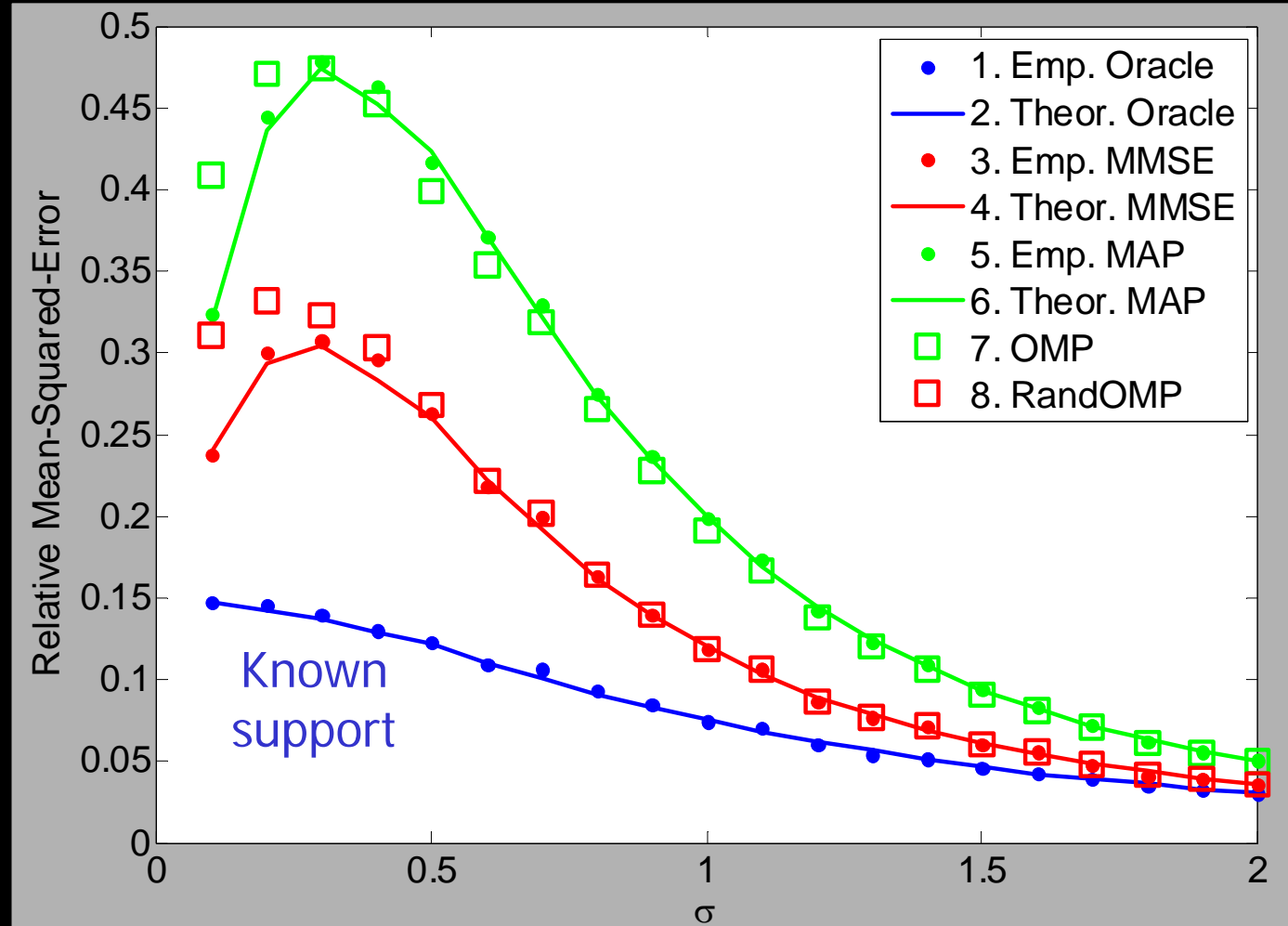


Comparative Results

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
- $\sigma_x=1$
- Averaged over 1000 experiments



Part III

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

What do we do?

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

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

and today

More on these (including the slides, the papers, and a Matlab toolbox) in <http://www.cs.technion.ac.il/~elad>

