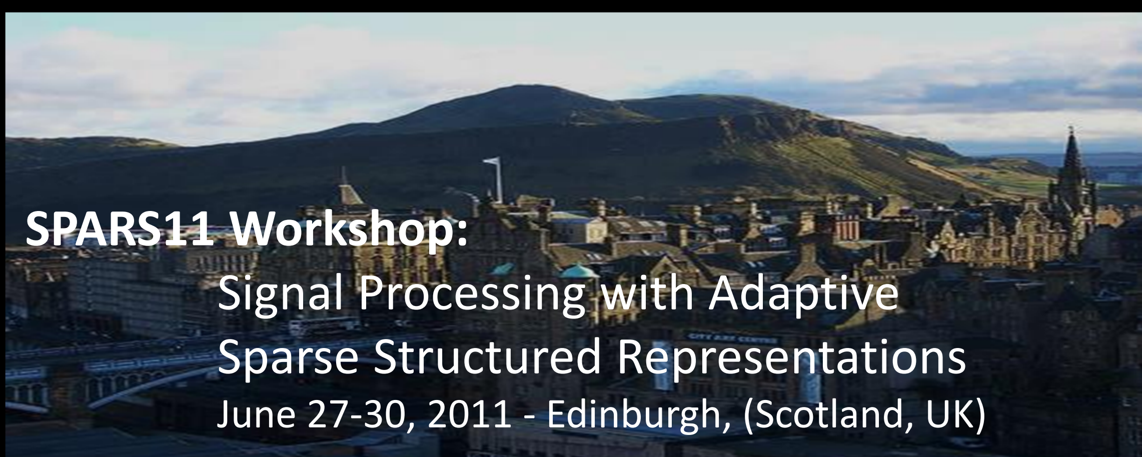
K-SVD Dictionary-Learning for Analysis Sparse Models *

Michael Elad

The Computer Science Department

The Technion – Israel Institute of technology

Haifa 32000, Israel



Joint work with

Ron Rubinstein

and

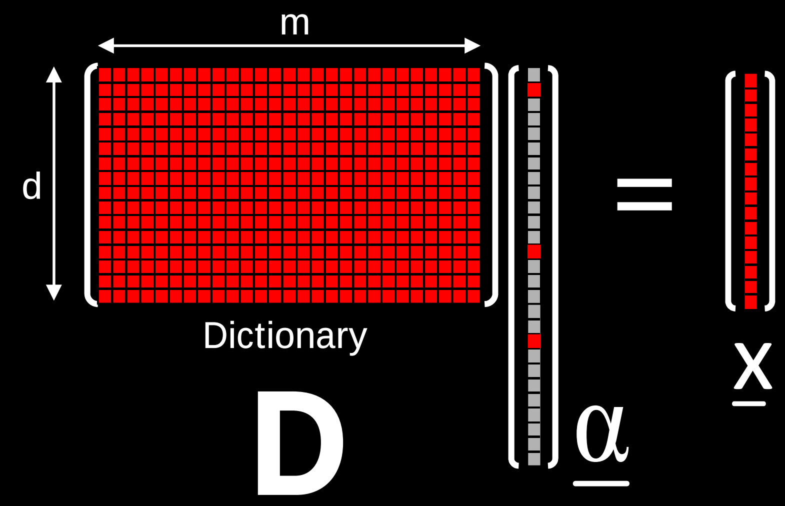
Remi Gribonval, Mark Plumbley, Mike Davies, Sangnam Nam, Boaz Ophir, Nancy Bertin

Part I - Background

Recalling the Synthesis Model and the K-SVD

The Synthesis Model – Basics

- The synthesis representation is expected to be sparse: $\|\underline{\alpha}\|_0 = k << d$



- Adopting a Bayesian point of view:
  - Draw the support at random
  - Choose the non-zero coefficients randomly (e.g. iid Gaussians)
  - Multiply by **D** to get the synthesis signal
- Such synthesis signals belong to a Union-of-Subspaces (UoS):

$$\underline{x} \in \bigcap_{|T|=k} \text{span}\{\mathbf{D}_T\} \quad \text{where} \quad \mathbf{D}_T \underline{\alpha}_T = \underline{x}$$

- This union contains $\binom{m}{k}$ subspaces, each of dimension k.

The Synthesis Model – Pursuit

- Fundamental problem: Given the noisy measurements,

$$\underline{y} = \underline{x} + \underline{v} = \mathbf{D}\underline{\alpha} + \underline{v}, \quad \underline{v} \sim \mathbf{N}\left\{\underline{0}, \sigma^2 \mathbf{I}\right\}$$

  recover the clean signal $\underline{x}$ – This is a denoising task.

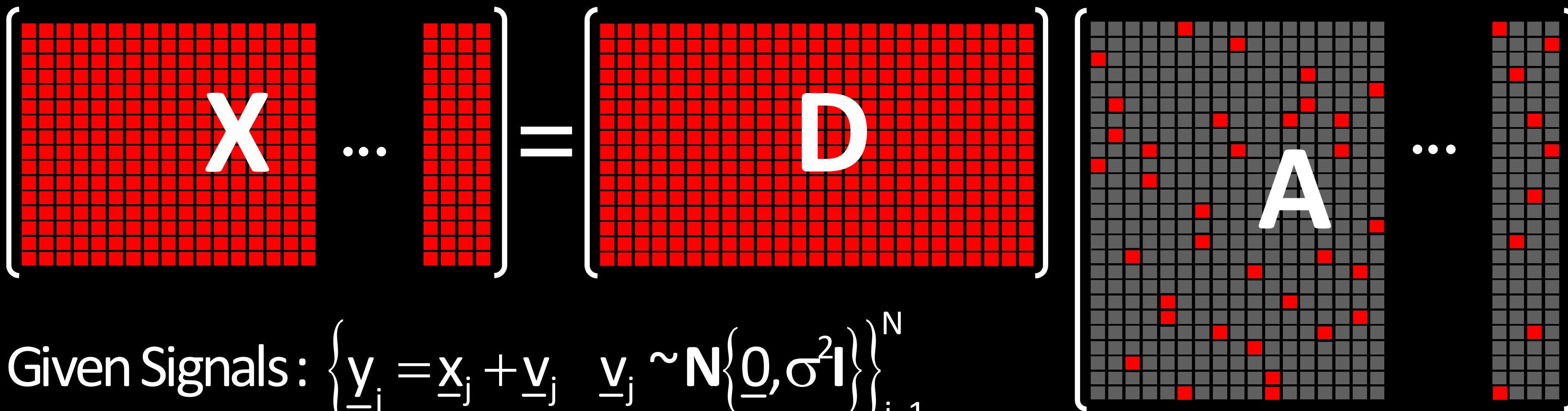
- This can be posed as: $\hat{\underline{\alpha}} = \underset{\underline{\alpha}}{\text{ArgMin}} \left\| \underline{y} - \mathbf{D}\underline{\alpha} \right\|_2^2 \text{ s.t. } \left\| \underline{\alpha} \right\|_0 = k \Rightarrow \hat{\underline{x}} = \mathbf{D}\hat{\underline{\alpha}}$

- While this is a (NP-) hard problem, its approximated solution can be obtained by
  - Use L_1 instead of L_0 (Basis-Pursuit)
  - Greedy methods (MP, OMP, LS-OMP)
  - Hybrid methods (IHT, SP, CoSaMP)

  } Pursuit Algorithms

- Theoretical studies provide various guarantees for the success of these techniques, typically depending on k and properties of **D**.

The Synthesis Model – Dictionary Learning



Given Signals: $\left\{\underline{y}_j = \underline{x}_j + \underline{v}_j \quad \underline{v}_j \sim \mathbf{N}\{\underline{0}, \sigma^2 \mathbf{I}\}\right\}_{j=1}^{N}$

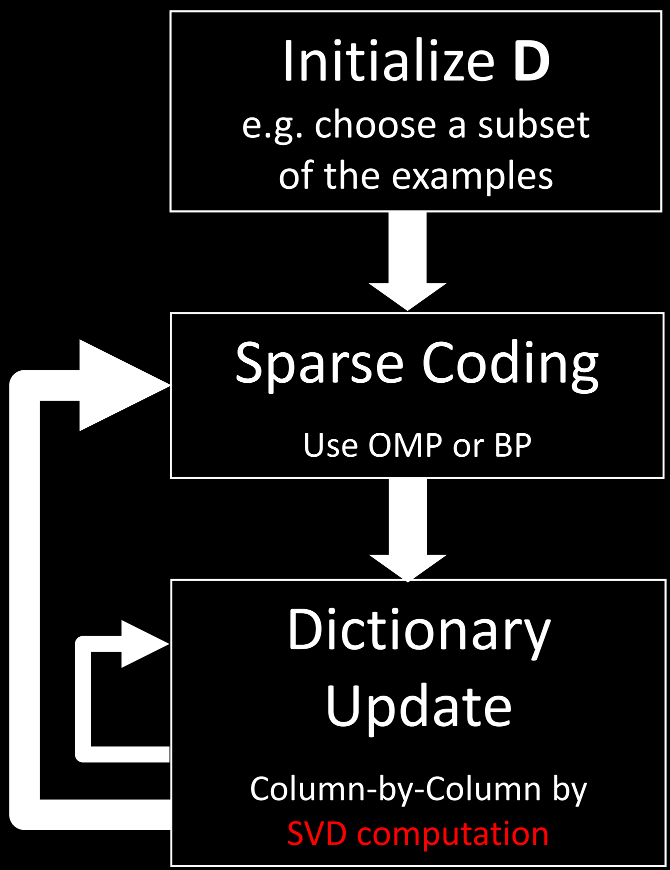
$$\underset{\mathbf{D},\mathbf{A}}{\text{Min}} \sum_{j=1}^{N} \left\| \mathbf{D}\underline{\alpha}_j - \underline{y}_j \right\|_2^2 \quad \text{s.t.} \quad \forall j = 1,2,\ldots,N \quad \left\| \underline{\alpha}_j \right\|_0 \le k$$

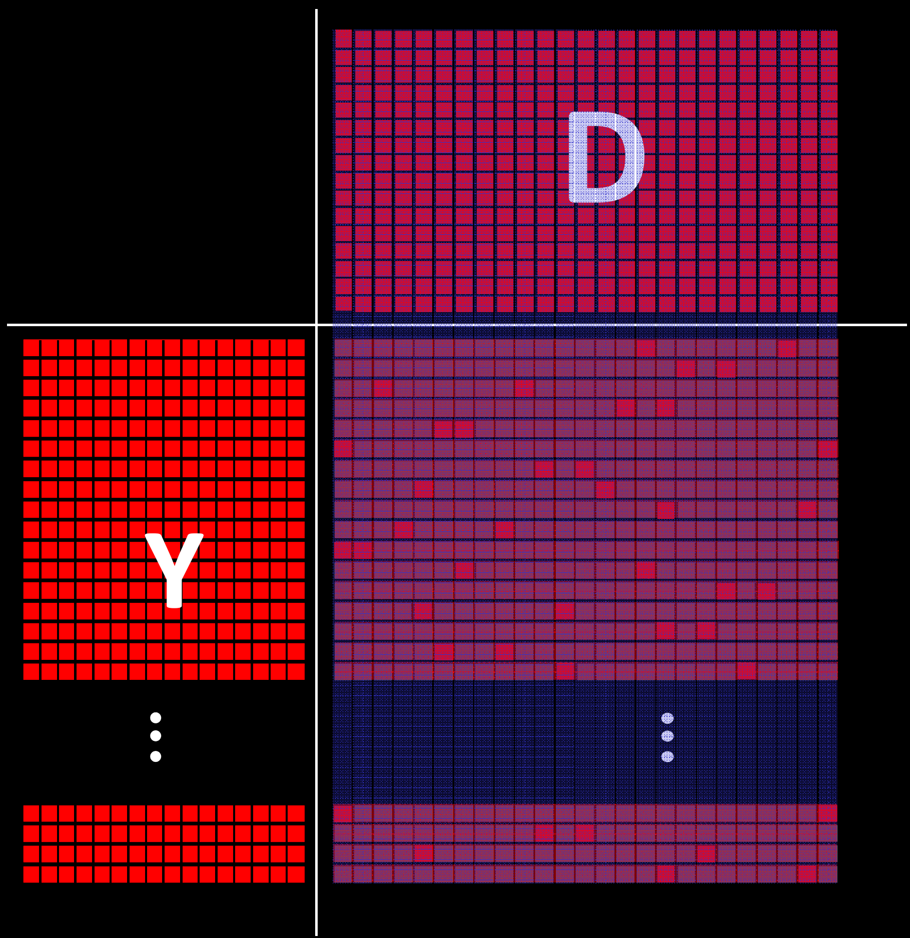
Each example is a linear combination of atoms from **D**

Each example has a sparse representation with no more than k atoms

Field & Olshausen (96')
Engan et. al. (99')
…
Gribonval et. al. (04')
Aharon et. al. (04')
…

The Synthesis Model – K-SVD Aharon, E., & Bruckstein (`04)



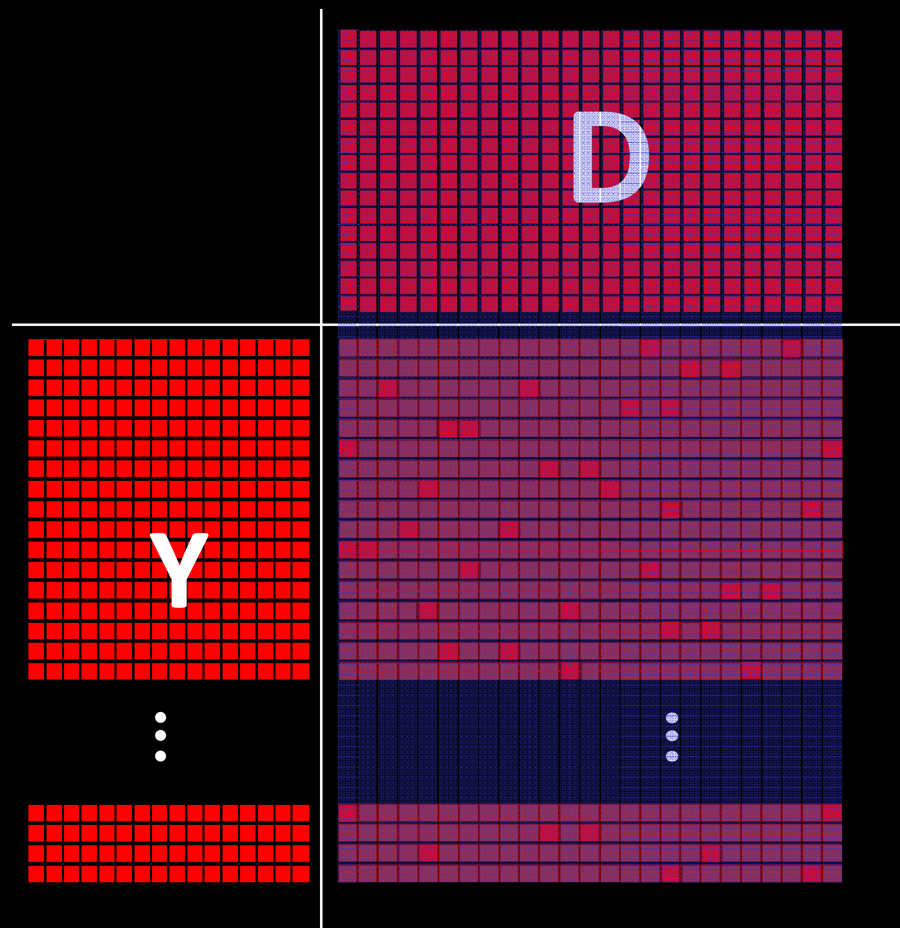


The Synthesis Model – K-SVD Aharon, E., & Bruckstein (`04)

Initialize D

e.g. choose a subset of the examples

Recall: the dictionary update stage in the K-SVD is done one atom at a time, updating it using ONLY those examples who use it, while fixing the non-zero supports.



Part II - Analysis

The Basics of the Analysis Model

1. S. Nam, M.E. Davies, M. Elad, and R. Gribonval, "Co-sparse Analysis Modeling - Uniqueness and Algorithms" , ICASSP, May, 2011.
2. S. Nam, M.E. Davies, M. Elad, and R. Gribonval, "The Co-sparse Analysis Model and Algorithms" , Submitted to ACHA, June 2011.

The Analysis Model – Basics

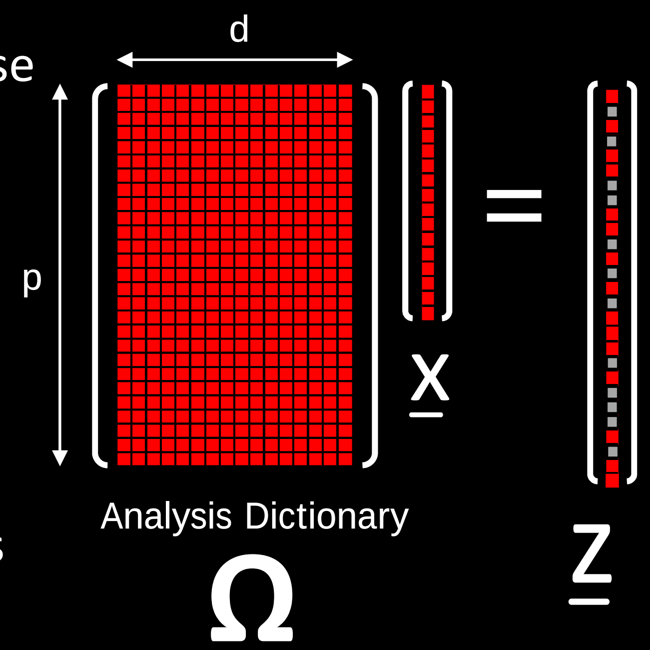
- The analysis representation $\underline{z}$ is expected to be sparse

$$\|\mathbf{\Omega}\underline{x}\|_0 = \|\underline{z}\|_0 = p - \ell$$

- Co-sparsity: ℓ - the number of zeros in $\underline{z}$.
- Co-Support: Λ - the rows that are orthogonal to $\underline{x}$

$$\mathbf{\Omega}_\Lambda \underline{x} = \underline{0}$$

- If Ω is in general position*, then $0 \le \ell < d$ and thus we cannot expect to get a truly sparse analysis representation – Is this a problem? No!
- Notice that in this model we put an emphasis on the zeros in the analysis representation, $\underline{z}$, rather then the non-zeros. In particular, the values of the non-zeroes in $\underline{z}$ are not important to characterize the signal.

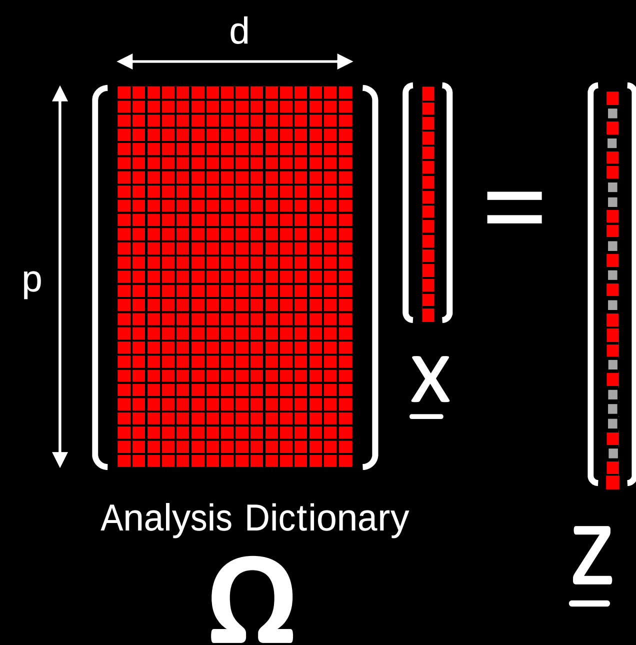


* $\text{spark}\{\mathbf{\Omega}^T\} = d + 1$

The Analysis Model – Bayesian View

- Analysis signals, just like synthesis ones, can be generated in a systematic way:

	Synthesis Signals	**Analysis Signals**
Support:	Choose the support T (|T|=k) at random	Choose the co-support Λ (|Λ|=ℓ) at random
Coef. :	Choose $\underline{\alpha}_T$ at random	Choose a random vector $\underline{v}$
Generate:	Synthesize by: $\mathbf{D}_T\underline{\alpha}_T=\underline{x}$	Orhto $\underline{v}$ w.r.t. $\mathbf{\Omega}_\Lambda$: $\underline{x}=\left[\mathbf{I}-\mathbf{\Omega}_\Lambda^\dagger\mathbf{\Omega}_\Lambda\right]\underline{v}$

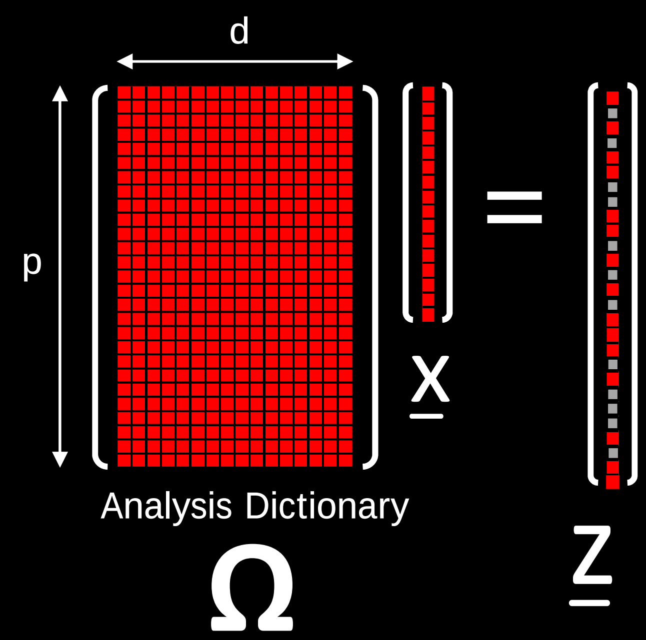


- Bottom line: an analysis signal $\underline{x}$ satisfies: $\exists\left|\Lambda\right|=\ell$ s.t. $\mathbf{\Omega}_\Lambda\underline{x}=\underline{0}$

The Analysis Model – UoS

- Analysis signals, just like synthesis ones, belong to a union of subspaces:

	Synthesis Signals	Analysis Signals
What is the Subspace Dimension:	k	d-ℓ
How Many Subspaces:	$\binom{m}{k}$	$\binom{p}{\ell}$
Who are those Subspaces:	$\text{span}\{\mathbf{D}_T\}$	$\text{span}^{\perp}\{\mathbf{\Omega}_{\Lambda}\}$



- Example: p=m=2d:
  - Synthesis: k=1 (one atom) – there are 2d subspaces of dimensionality 1
  - Analysis: ℓ=d-1 leads to $\binom{2d}{d-1}$>>O(2^d) subspaces of dimensionality 1

The Analysis Model – Pursuit

- Fundamental problem: Given the noisy measurements,

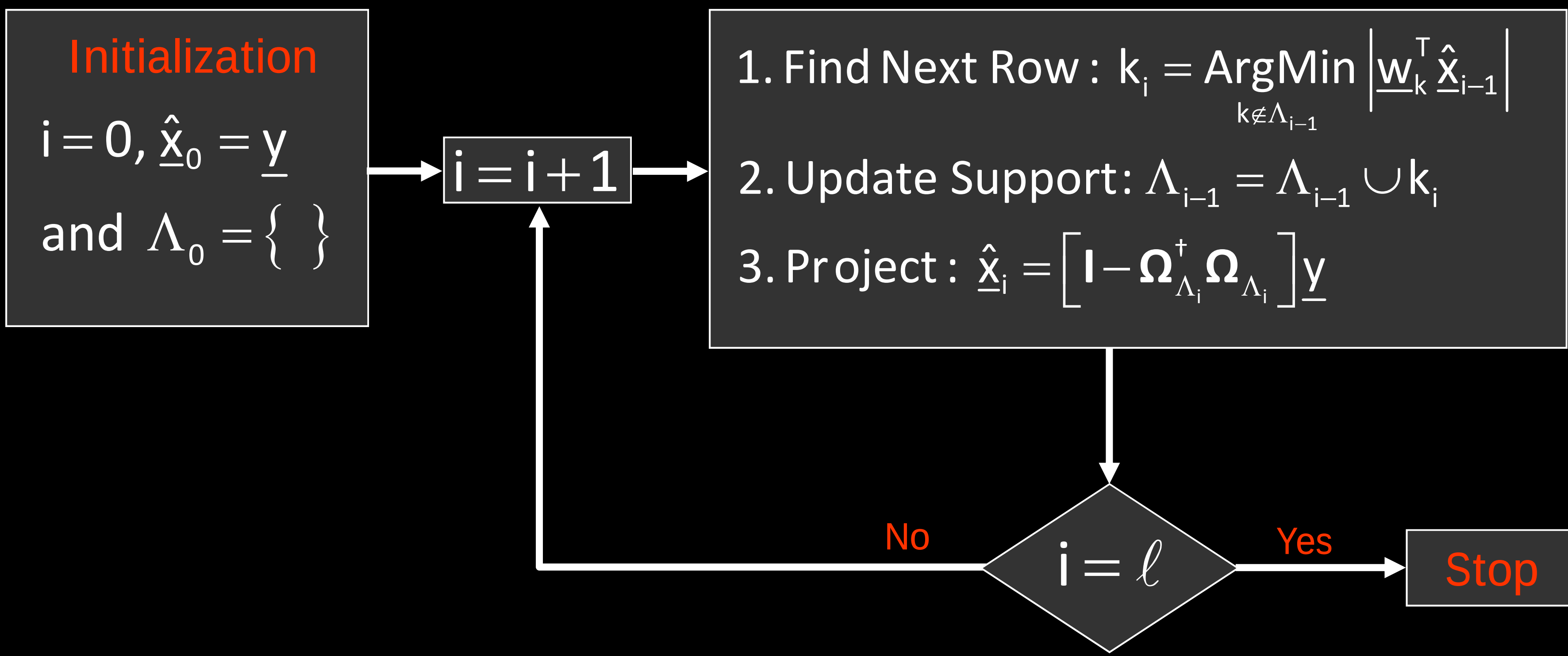
$$\underline{y} = \underline{x} + \underline{v}, \quad \exists |\Lambda| = \ell \text{ s.t. } \mathbf{\Omega}_\Lambda \underline{x} = \underline{0}, \quad \underline{v} \sim \mathbf{N}\{\underline{0}, \sigma^2 \mathbf{I}\}$$

  recover the clean signal $\underline{x}$ – This is a denoising task.

- This goal can be posed as: $\hat{\underline{x}} = \underset{\underline{\alpha}}{\text{ArgMin}} \|\underline{y} - \underline{x}\|_2^2 \text{ s.t. } \|\mathbf{\Omega}\underline{x}\|_0 = p - \ell$
- This is a (NP-) hard problem, just as in the synthesis case (and even harder!!!)
- We can approximate its solution by
  - L_1 replacing L_0 (BP-analysis)
  - Greedy methods (OMP, …), and
  - Hybrid methods (IHT, SP, CoSaMP, …).
- Theoretical studies should provide guarantees for the success of these techniques, typically depending on the co-sparsity and properties of Ω.

The Analysis Model – Backward Greedy

BG finds one row at a time from Λ for approximating the solution of $\hat{\underline{x}} = \underset{\underline{\alpha}}{\text{ArgMin}} \left\| \underline{y} - \underline{x} \right\|_2^2 \ \text{s.t.} \ \left\| \mathbf{\Omega} \underline{x} \right\|_0 = p - \ell$



The Analysis Model – Backward Greedy

BG finds one row at a time from Λ for approximating the solution of

$$\hat{\underline{x}} = \underset{\underline{\alpha}}{\text{ArgMin}} \left\| \underline{y} - \underline{x} \right\|_2^2 \quad \text{s.t.} \quad \left\| \mathbf{\Omega} \underline{x} \right\|_0 = p - \ell$$

Initialization

$i = 0,\ \hat{\underline{x}}_0 = \underline{y}$ and $\Lambda_0 = \{\ \}$

Variations and Improvements:

- ❑ Gram-Schmidt applied to the accumulated rows speeds-up the algorithm.
- ❑ An exhaustive alternative, xBG, can be used, where per each candidate row we test the decay in the projection energy and choose the smallest of them as the next row.
- ❑ One could think of a forward alternative that detects the non-zero rows (GAP) – talk with Sangnam.

$\ldots\text{gMin}_{\notin \Lambda_{i-1}} \left| \underline{w}_k^T \hat{\underline{x}}_{i-1} \right|$

$\ldots = \Lambda_{i-1} \cup k_i$

$\ldots_{\Lambda_i} \Big] \underline{y}$

Yes

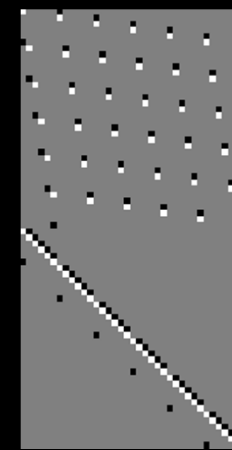
Stop

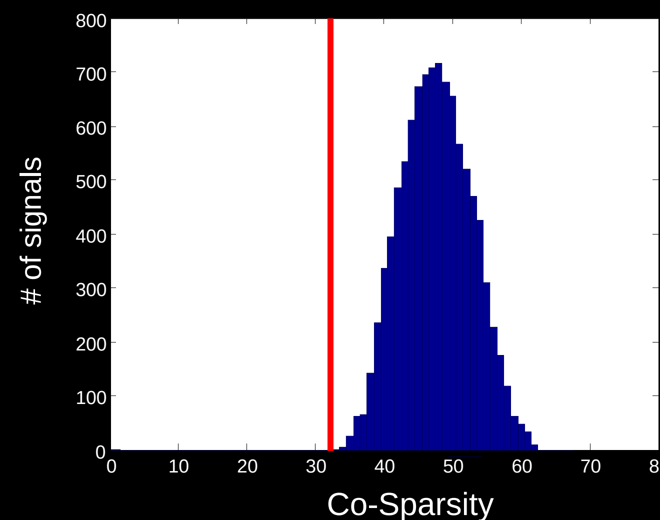
The Analysis Model – Low-Spark Ω

- What if spark(Ω^T)<<d ?
- For example: a TV-like operator for image-patches of size 6×6 pixels (Ω size is 72×36)
- Here are analysis-signals generated for co-sparsity (ℓ) of 32:

- Their true co-sparsity is higher – see graph:
- In such a case we may consider $\ell > \text{d}$
- More info: S. Nam, M.E. Davies, M. Elad, and R. Gribonval

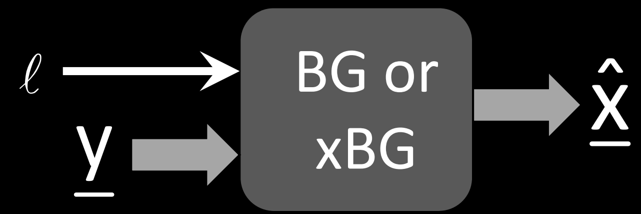
$\Omega =$ [Horizontal Derivative / - - - - - - / Vertical Derivative] =

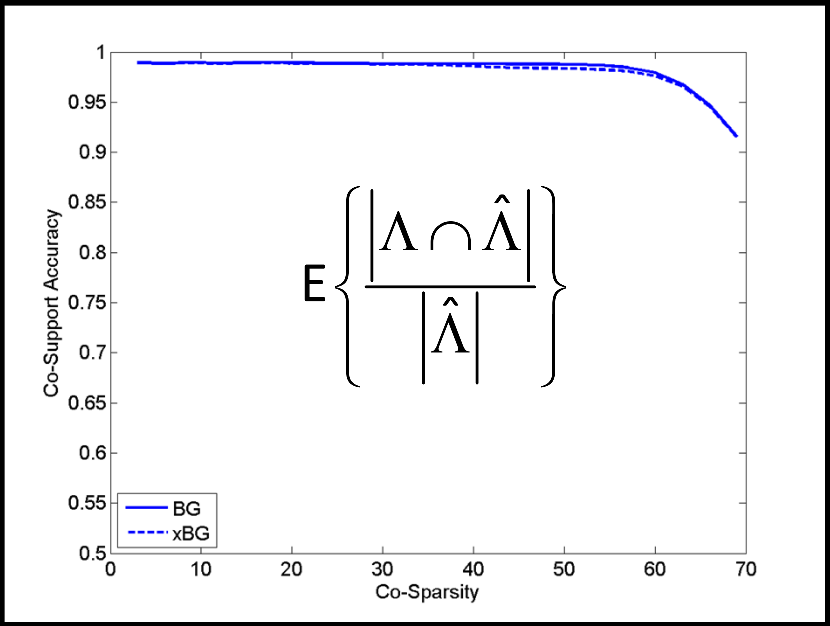


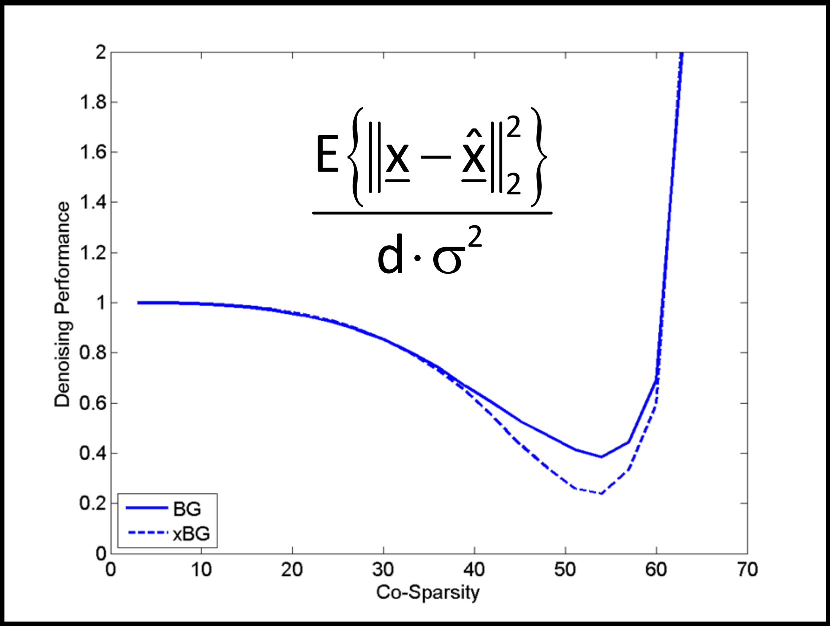


The Analysis Model – Low-Spark Ω – Pursuit

- An example – performance of BG (and xBG) for these TV-like signals:
- 1000 signal examples, SNR=25
  - Accuracy of the co-support recovered
  - Denoising performance

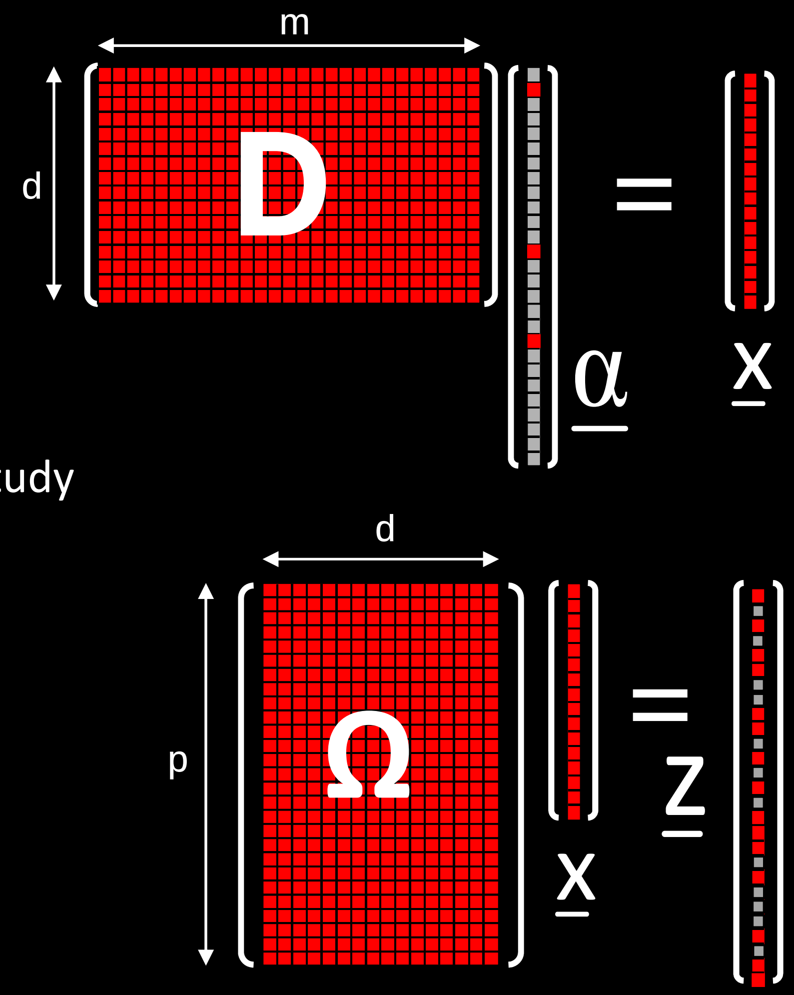






The Analysis Model – Summary

- The analysis and the synthesis models are similar, and yet very different
- The two align for p=m=d : non-redundant
- Just as the synthesis, we should work on:
  - Pursuit algorithms (of all kinds) – Design
  - Pursuit algorithms (of all kinds) – Theoretical study
  - Dictionary learning from example-signals
  - Applications …
- Our experience on the analysis model:
  - Theoretical study is harder
  - Different applications should be considered

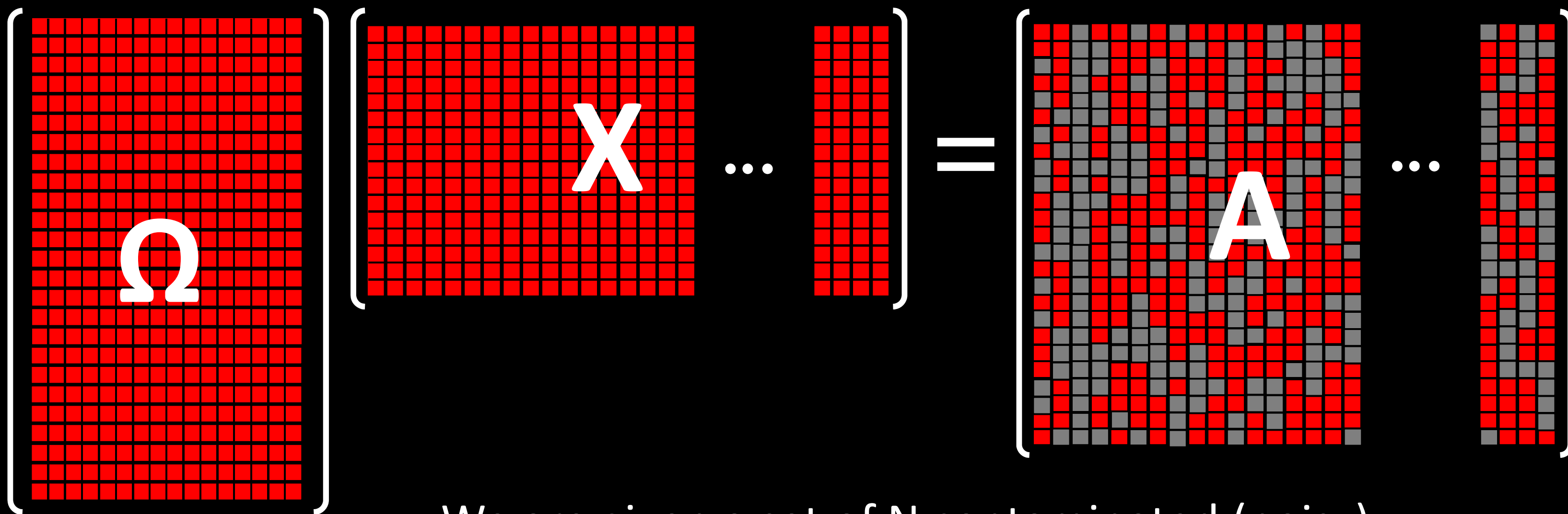


Part III – Dictionaries

Analysis Dictionary-Learning by K-SVD-Like Algorithm

1. B. Ophir, M. Elad, N. Bertin and M.D. Plumbley, "Sequential Minimal Eigenvalues - An Approach to Analysis Dictionary Learning", EUSIPCO, August 2011.
2. R. Rubinstein and M. Elad, "The Co-sparse Analysis Model and Algorithms" , will be submitted (very) soon to IEEE-TSP ....

Analysis Dictionary Learning – The Signals



We are given a set of N contaminated (noisy) analysis signals, and our goal is to recover their analysis dictionary, $\boldsymbol{\Omega}$

$$\left\{ \underline{y}_j = \underline{x}_j + \underline{v}_j, \quad \exists \left|\Lambda_j\right| = \ell \ \text{ s.t. } \ \boldsymbol{\Omega}_{\Lambda_j} \underline{x}_j = \underline{0}, \quad \underline{v} \sim \mathbf{N}\left\{\underline{0}, \sigma^2 \mathbf{I}\right\} \right\}_{j=1}^{N}$$

Analysis Dictionary Learning – Goal

Synthesis

$$\underset{\mathbf{D},\mathbf{A}}{\text{Min}} \sum_{j=1}^{N} \left\| \mathbf{D}\underline{\alpha}_j - \underline{y}_j \right\|_2^2 \quad \text{s.t. } \forall j = 1,2,\ldots,N \quad \left\| \underline{\alpha}_j \right\|_0 \le k$$

Analysis

$$\underset{\mathbf{\Omega},\underline{\mathbf{X}}}{\text{Min}} \sum_{j=1}^{N} \left\| \underline{x}_j - \underline{y}_j \right\|_2^2 \quad \text{s.t. } \forall j = 1,2,\ldots,N \quad \left\| \mathbf{\Omega}\underline{x}_j \right\|_0 \le p - \ell$$

We shall adopt a similar approach to the K-SVD for approximating the minimization of the analysis goal

Analysis Dictionary – Sparse-Coding

$$\underset{\boldsymbol{\Omega},\underline{\mathbf{X}}}{\text{Min}} \sum_{j=1}^{N} \left\| \underline{x}_j - \underline{y}_j \right\|_2^2 \quad \text{s.t. } \forall j=1,2,\ldots,N \;\; \left\| \boldsymbol{\Omega}\underline{x}_j \right\|_0 \le p-\ell$$

Assuming that Ω is fixed, we aim at updating $\underline{X}$

$$\left\{ \hat{\underline{x}}_j = \underset{\underline{\mathbf{x}}}{\text{ArgMin}} \left\| \underline{x} - \underline{y}_j \right\|_2^2 \quad \text{s.t. } \left\| \boldsymbol{\Omega}\underline{x} \right\|_0 \le p-\ell \right\}_{j=1}^{N}$$

These are N separate analysis-pursuit problems. We suggest to use the BG or the xBG algorithms.

Analysis Dictionary – Dic. Update (1)

$$\min_{\Omega,\underline{X}} \sum_{j=1}^{N} \left\| \underline{x}_j - \underline{y}_j \right\|_2^2 \quad \text{s.t. } \forall j=1,2,\ldots,N \quad \left\| \Omega \underline{x}_j \right\|_0 \le p-\ell$$

Assuming that $\underline{X}$ has been updated (and thus Λ_j are known), we now aim at updating a row (e.g. $\underline{w}_k^T$) from Ω

$$\min_{\underline{w}_k, \mathbf{X}_k} \left\| \mathbf{X}_k - \mathbf{Y}_k \right\|_2^2 \quad \text{s.t.} \quad \begin{Bmatrix} \forall j \in S_k \quad \Omega_j \underline{x}_j = \underline{0} \\ \underline{w}_k^T \mathbf{X}_k = \underline{0} \\ \left\| \underline{w}_k \right\|_2 = 1 \end{Bmatrix}$$

We use only the signals S_k that are found orthogonal to $\underline{w}_k$

Each example should keep its co-support $\Lambda_j \backslash k$

Each of the chosen examples should be orthogonal to the new row $\underline{w}_k$

Avoid trivial solution

Analysis Dictionary – Dic. Update (2)

$$\underset{\underline{w}_k, \mathbf{X}_k}{\text{Min}} \ \left\| \mathbf{X}_k - \mathbf{Y}_k \right\|_2^2 \ \text{ s.t. } \begin{Bmatrix} \forall j \in S_k \quad \mathbf{\Omega}_j \underline{x}_j = \underline{0} \\ \underline{w}_k^T \mathbf{X}_k = \underline{0} \\ \left\| \underline{w}_k \right\|_2 = 1 \end{Bmatrix}$$

This problem we have defined is too hard to handle

Intuitively, and in the spirit of the K-SVD, we could suggest the following alternative

$$\underset{\underline{w}_k, \mathbf{X}_k}{\text{Min}} \ \left\| \mathbf{X}_k - \left[\mathbf{I} - \mathbf{\Omega}_j^{\dagger} \mathbf{\Omega}_j \right] \mathbf{Y}_k \right\|_2^2 \ \text{ s.t. } \begin{Bmatrix} \underline{w}_k^T \mathbf{X}_k = \underline{0} \\ \left\| \underline{w}_k \right\|_2 = 1 \end{Bmatrix}$$

Analysis Dictionary – Dic. Update (2)

$$\underset{\underline{w}_k,\underline{\mathbf{X}}_k}{\text{Min}} \left\| \mathbf{X}_k - \mathbf{Y}_k \right\|_2^2 \quad \text{s.t.} \quad \left\{ \begin{array}{c} \forall j \in S_k \quad \boldsymbol{\Omega}_j \underline{x}_j = \underline{0} \\ \underline{w}_k^T \mathbf{X}_k = \underline{0} \\ \left\| \underline{w}_k \right\|_2 = 1 \end{array} \right\}$$

This problem we have defined is too hard to handle

Intuitively, and in the spirit of the K-SVD, we could suggest the following alternative

WRONG!

$$\underset{\underline{w}_k,\underline{\mathbf{X}}_k}{\text{Min}} \left\| \ldots \boldsymbol{\Omega}_j \right] \mathbf{Y}_k \right\|_2^2 \quad \text{s.t.} \quad \left\{ \begin{array}{c} \underline{w}_k^T \mathbf{X}_k = \underline{0} \\ \left\| \underline{w}_k \right\|_2 = 1 \end{array} \right\}$$

Analysis Dictionary – Dic. Update (3)

$$\underset{\underline{w}_k, \mathbf{X}_k}{\text{Min}} \left\| \mathbf{X}_k - \left[\mathbf{I} - \mathbf{\Omega}_j^{\dagger} \mathbf{\Omega}_j \right] \mathbf{Y}_k \right\|_2^2 \quad \text{s.t.} \quad \left\{ \begin{array}{l} \underline{w}_k^T \mathbf{X}_k = \underline{0} \\ \left\| \underline{w}_k \right\|_2 = 1 \end{array} \right\}$$

This lacks in one of the forces on $\underline{w}_k$ that the original problem had

A better approximation for our original problem is

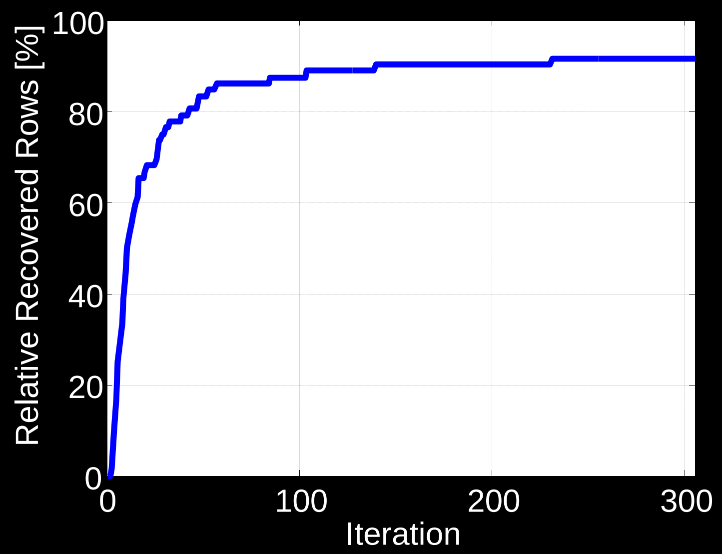
$$\underset{\underline{w}_k, \mathbf{X}_k}{\text{Min}} \left\| \mathbf{X}_k - \mathbf{Y}_k \right\|_2^2 \quad \text{s.t.} \quad \left\{ \begin{array}{l} \underline{w}_k^T \mathbf{X}_k = \underline{0} \\ \left\| \underline{w}_k \right\|_2 = 1 \end{array} \right\} \Rightarrow \underset{\underline{w}_k,}{\text{Min}} \left\| \underline{w}_k^T \mathbf{Y}_k \right\|_2^2 \quad \text{s.t.} \quad \left\| \underline{w}_k \right\|_2 = 1$$

The obtained problem is a simple Rank-1 approximation problem, easily given by SVD

Analysis Dictionary Learning – Results (1)

Synthetic experiment #1: TV-Like Ω

- We generate 30,000 TV-like signals of the same kind described before (Ω: 72×36, ℓ=32)
- We apply 300 iterations of the Analysis K-SVD with BG (fixed ℓ), and then 5 more using the xBG
- Initialization by orthogonal vectors to randomly chosen sets of 35 examples
- Additive noise: SNR=25. atom detected if: $1-\left|\underline{w}^T\underline{\hat{w}}\right|<0.01$

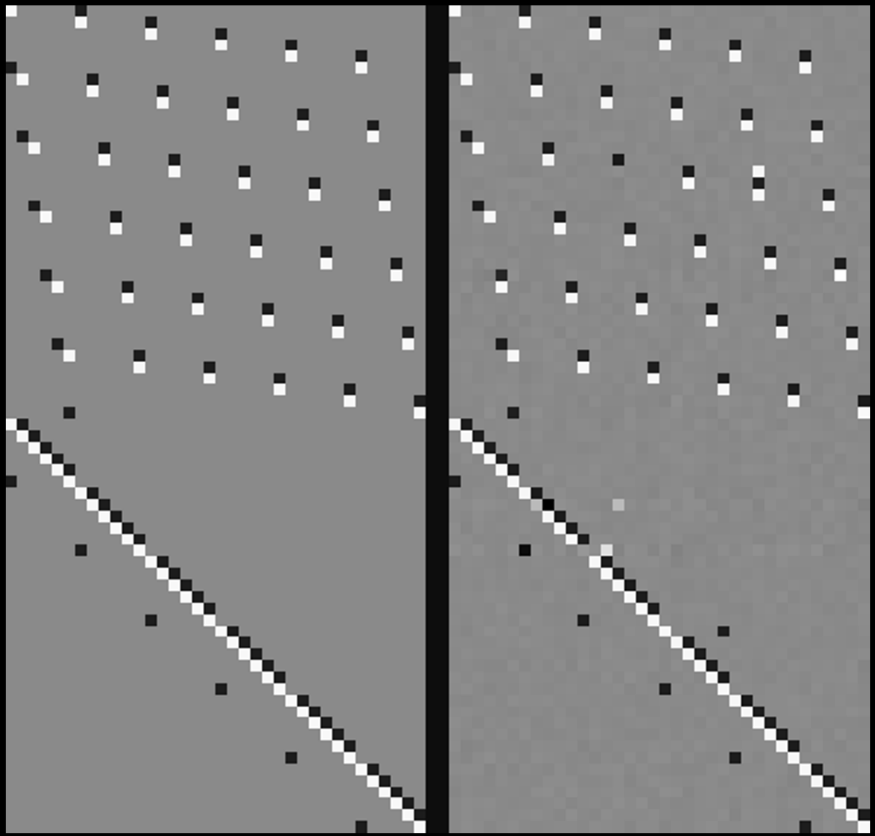


Even though we have not identified Ω completely (~92% this time), we got an alternative feasible analysis dictionary with the same number of zeros per example, and a residual error within the noise level.

Analysis Dictionary Learning – Results (1)

Synthetic experiment #1: TV-Like Ω

Original Analysis Dictionary

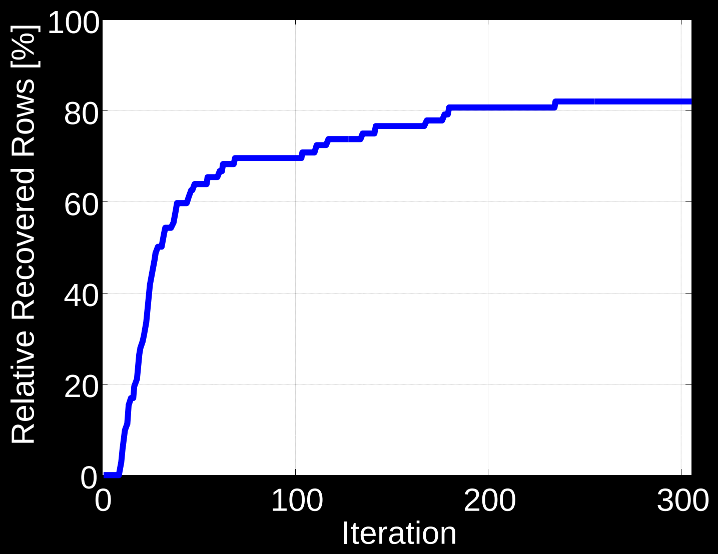


Learned Analysis Dictionary

Analysis Dictionary Learning – Results (2)

Synthetic experiment #2: Random Ω

- Very similar to the above, but with a random (full-spark) analysis dictionary Ω: 72×36
- Experiment setup and parameters: the very same as above
- In both algorithms: replacing BG by xBG (in both experiments) leads to a consistent descent in the relative error, and better recovery results. However, the run-time is ~50 times longer

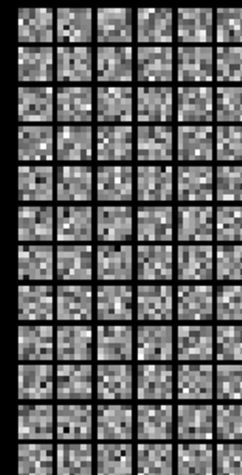


As in the previous example, even though we have not identified Ω completely (~80% this time), we got an alternative feasible analysis dictionary with the same number of zeros per example, and a residual error within the noise level.

Analysis Dictionary Learning – Results (3)

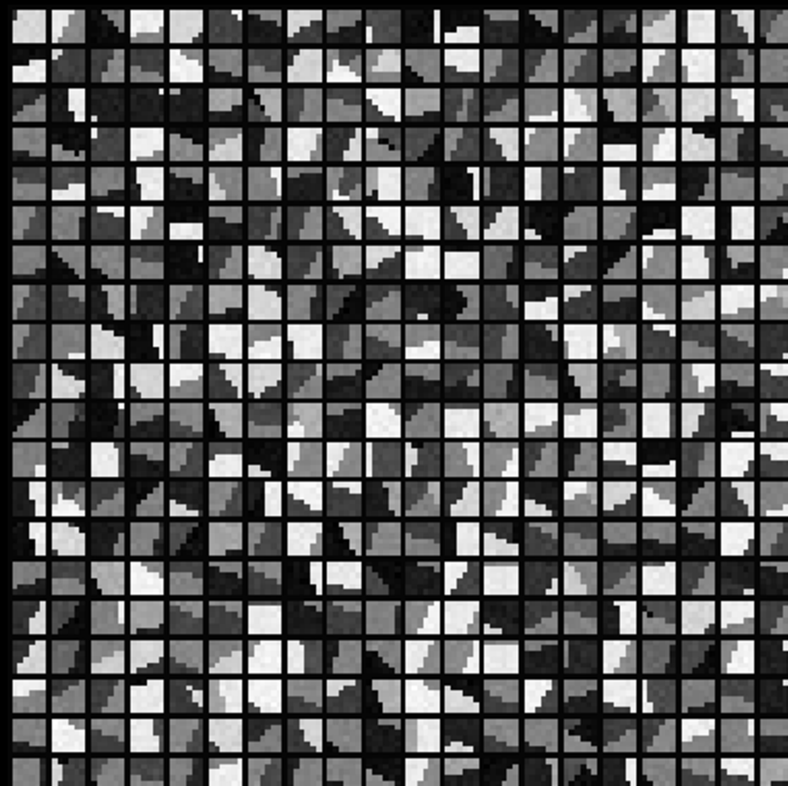
Experiment #3: Piece-Wise Constant Image

- We take 10,000 patches (+noise $\sigma=5$) to train on
- Here is what we got:

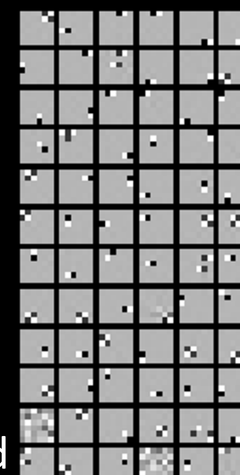


Initial Ω

Original Image



Patches used for training



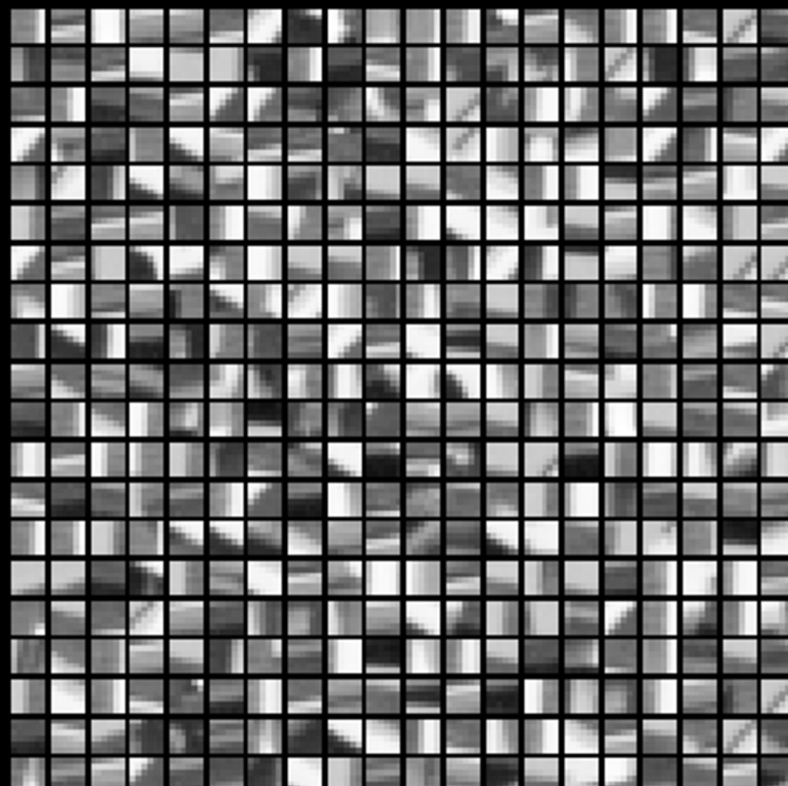
Trained (100 iterations) Ω

Analysis Dictionary Learning – Results (4)

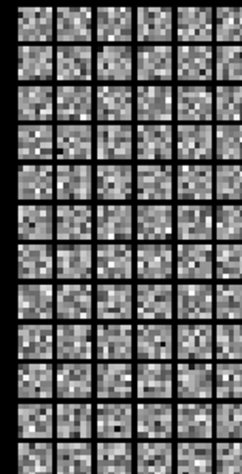
Experiment #3: The Image "House"

- We take 10,000 patches (+noise $\sigma=10$) to train on
- Here is what we got:

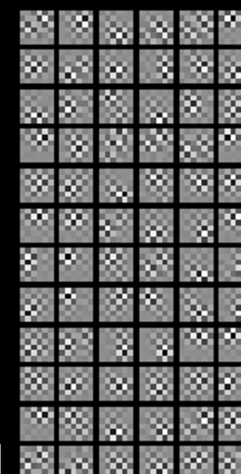
Original Image



Patches used for training



Initial Ω



Trained (100 iterations) Ω

Part IV – We Are Done

Summary and Conclusions

Today We Have Seen that …

Sparsity and Redundancy are practiced mostly in the context of the synthesis model

Is there any other way?

Yes, the analysis model is a very appealing (and different) alternative, worth looking at

So, what to do?

In the past few years there is a growing interest in better defining this model, suggesting pursuit methods, analyzing them, etc.

What about Dictionary learning?

We propose new algorithms (e.g. K-SVD like) for this task. The next step is applications that will benefit from this

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