Over-Complete & Sparse Representations for Image Decomposition and Inpainting

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

Second International Conference on Computational Harmonic Analysis
The 19th Annual Shanks Lecture
May 24-29th, 2004

Joint work with: Jean-Luc Starck – CEA - Service d'Astrophysique, CEA-Saclay, France
David L. Donoho – Statistics, Stanford.

Collaborators

Jean-Luc Starck

CEA - Service d'Astrophysique
CEA-Saclay
France

David L. Donoho

Statistics Department
Stanford

Background material:

- D. L. Donoho and M. Elad, "Maximal Sparsity Representation via l_1 Minimization", to appear in Proceedings of the Naional Academy of Science.
- J.-L. Starck, M. Elad, and D. L. Donoho, "Image Decomposition: Separation of Texture from Piece-Wise Smooth Content", SPIE annual meeting, 3–8 August 2003, San Diego, California, USA.
- J.-L. Starck, M. Elad, and D.L. Donoho, "Redundant Multiscale Transforms and their Application for Morphological Component Analysis", submitted to the Journal of Advances in Imaging and Electron Physics.
- J.-L. Starck, M. Elad, and D.L. Donoho, "Simultaneous PWS and Texture Image Inpainting using Sparse Representations", to be submitted to the IEEE Trans. On Image Processing.

These papers & slides can be found in: http://www.cs.technion.ac.il/~elad

General

- Sparsity and over-completeness have important roles in analyzing and representing signals.
- Our efforts so far have been concentrated on analysis of the (basis/matching) pursuit algorithms, properties of sparse representations (uniqueness), and applications.
- Today we discuss the image decomposition application (image=cartoon+texture). We present
  - Theoretical analysis serving this application,
  - Practical considerations, and
  - Application – filling holes in images (inpainting)

Agenda

1. Introduction

   Sparsity and Over-completeness!?

2. Theory of Decomposition

   Uniqueness and Equivalence

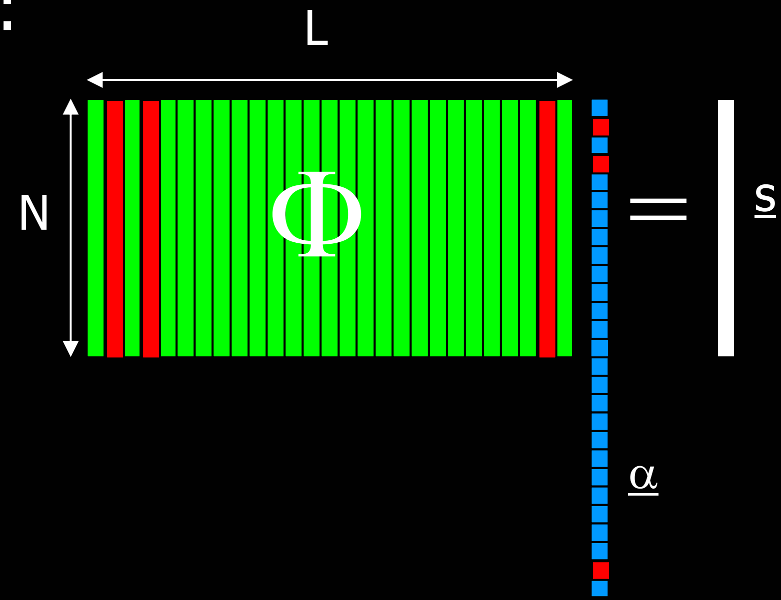
3. Decomposition in Practice

   Practical Considerations, Numerical algorithm

4. Discussion

Atom (De-) Composition

- Given a signal $\underline{s} \in \Re^N$, we are often interested in its representation (transform) as a linear combination of ‘atoms’ from a given dictionary:



- If the dictionary is over-complete (L>N), there are numerous ways to obtain the ‘atom-decomposition’.

- Among those possibilities, we consider the sparsest.

Atom Decomposition?

- Searching for the sparsest representation, we have the following optimization task:

$$P_0: \quad \underset{\underline{\alpha}}{\text{Min}} \; \left\|\underline{\alpha}\right\|_0 \quad \text{s.t.} \quad \underline{s} = \Phi \underline{\alpha}$$

- Hard to solve – complexity grows exponentially with L.

- Replace the l_0 norm by an l_1: Basis Pursuit (BP)

[Chen, Donoho, Saunders. 95']

$$P_1: \quad \underset{\underline{\alpha}}{\text{Min}} \; \left\|\underline{\alpha}\right\|_1 \quad \text{s.t.} \quad \underline{s} = \Phi \underline{\alpha}$$

- Greedy stepwise regression - Matching Pursuit (MP) algorithm [Zhang & Mallat. 93'] or orthonornal version of it (OMP) [Pati, Rezaiifar, & Krishnaprasad. 93'].

Questions about Decomposition

- **Interesting observation**: In many cases the pursuit algorithms successfully find the sparsest representation.
- Why BP/MP/OMP should work well? Are there Conditions to this success?
- Could there be several different sparse representations? What about uniqueness?
- How all this leads to image separation? Inpainting?

Agenda

1. Introduction

   Sparsity and Over-completeness!?

2. Theory of Decomposition

   Uniqueness and Equivalence

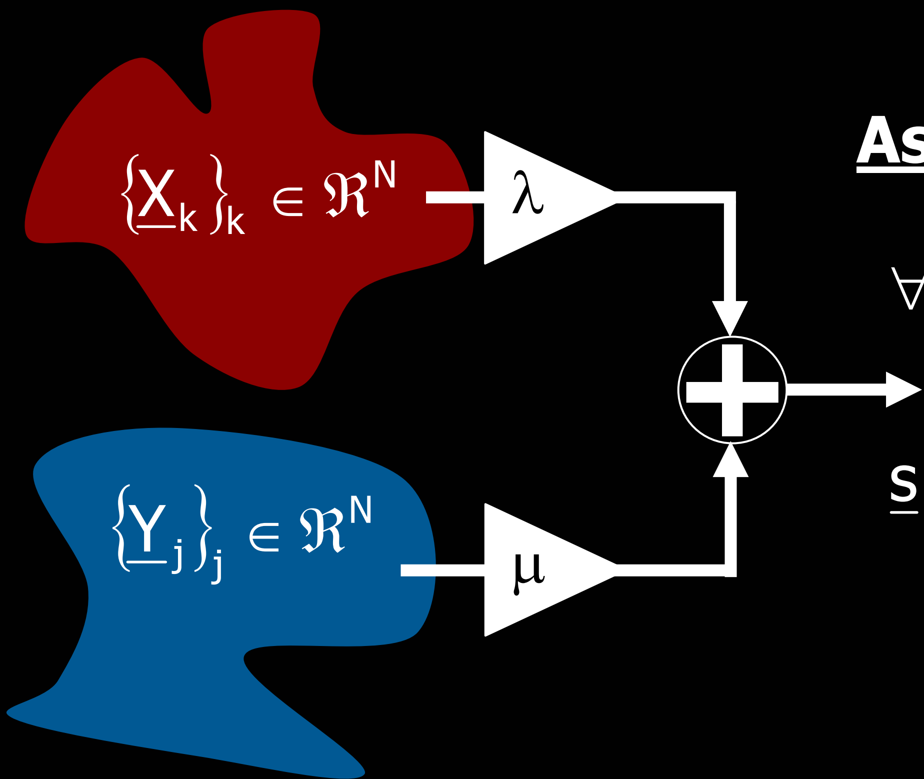
3. Decomposition in Practice

   Practical Considerations, Numerical algorithm

4. Discussion

Decomposition – Definition

Family of Cartoon images



Family of Texture images

<u>Our Assumption</u>

$\forall \underline{s} \quad \exists k, j, \lambda, \mu$

such that

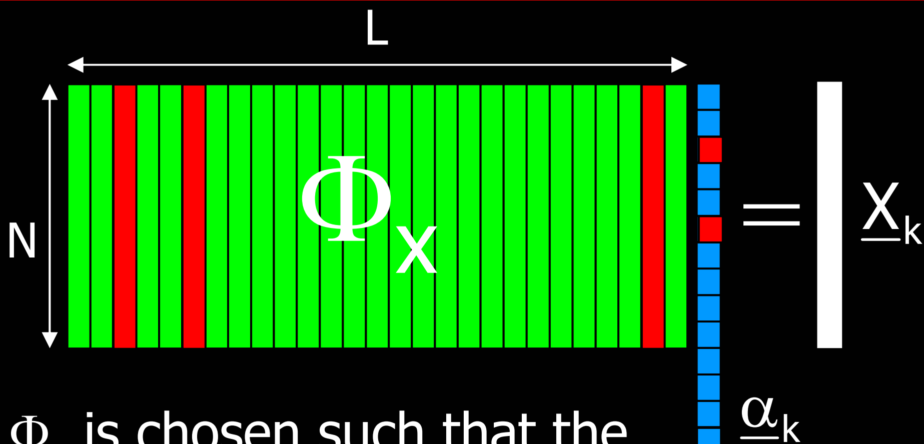
$\underline{s} = \lambda \underline{X}_k + \mu \underline{Y}_j$

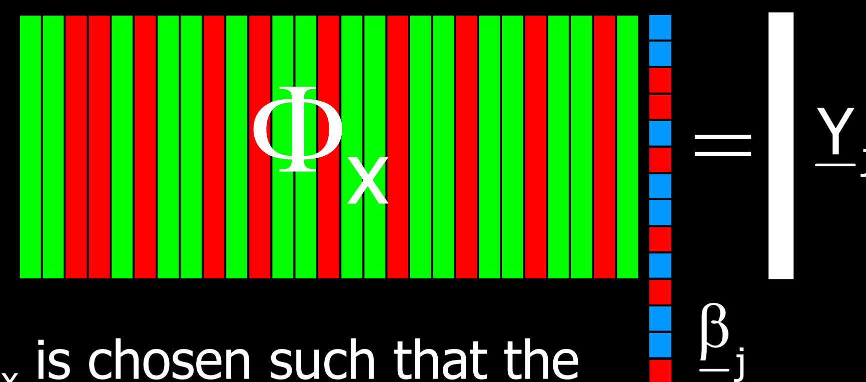
<u>Our Inverse Problem</u>

Given $\underline{s}$, find its building parts and the mixture weights

$\lambda, \mu, \underline{X}_k, \underline{Y}_j$

Use of Sparsity





Φ_x is chosen such that the representation of $\{\underline{X}_k\}_k \in \Re^N$ are sparse:

$$\left\{\underline{\alpha}_k = \underset{\underline{\alpha}}{\text{ArgMin}} \left\|\underline{\alpha}\right\|_0 \ \text{ s.t. } \ \underline{X}_k = \Phi_x \underline{\alpha}\right\}_k$$

$$\Rightarrow \ \forall k \ \left\|\underline{\alpha}_k\right\|_0 << N$$

Φ_x is chosen such that the representation of $\{\underline{Y}_j\}_j \in \Re^N$ are non-sparse:

$$\left\{\underline{\beta}_j = \underset{\underline{\beta}}{\text{ArgMin}} \left\|\underline{\beta}\right\|_0 \ \text{ s.t. } \ \underline{Y}_j = \Phi_x \underline{\beta}\right\}_k$$

$$\Rightarrow \ \forall j \ \left\|\underline{\beta}_j\right\|_0 \rightarrow N$$

We similarly construct Φ_y to sparsify Y’s while being inefficient in representing the X’s.

Choice of Dictionaries

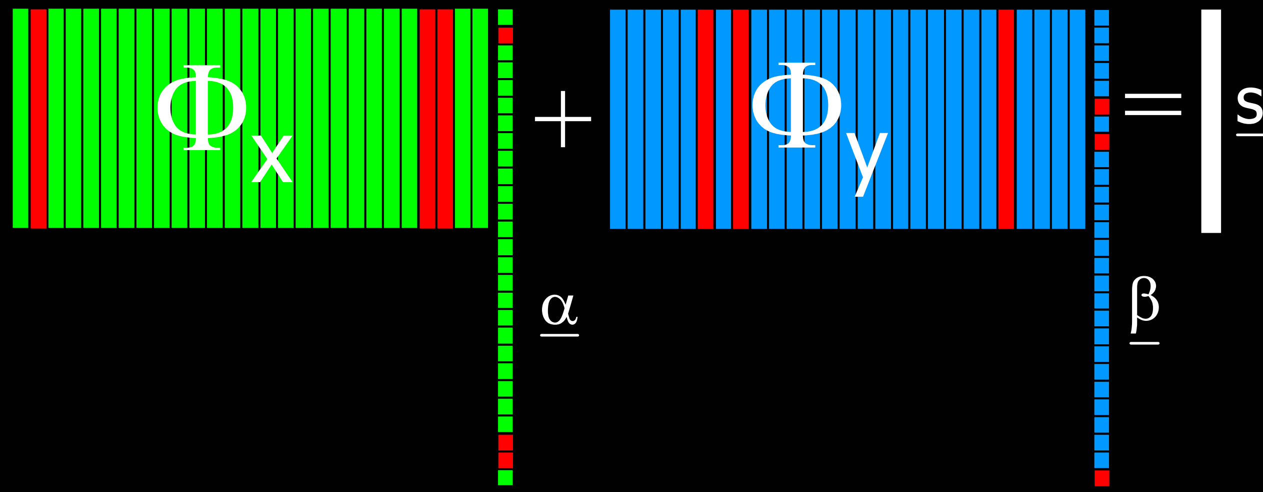
- Training, e.g.

$$\Phi_x = \underset{\Phi}{\text{ArgMin}} \frac{\sum_k \left\|\underline{\alpha}_k\right\|_0}{\sum_j \left\|\underline{\beta}_j\right\|_0} \text{ Subject to}$$

$$\left\{\underline{\alpha}_k = \underset{\underline{\alpha}}{\text{ArgMin}} \left\|\underline{\alpha}\right\|_0 \text{ s.t. } \underline{X}_k = \Phi\underline{\alpha}\right\}_k \quad \& \quad \left\{\underline{\beta}_j = \underset{\underline{\beta}}{\text{ArgMin}} \left\|\underline{\beta}\right\|_0 \text{ s.t. } \underline{Y}_j = \Phi\underline{\beta}\right\}_j$$

- Educated guess: texture could be represented by local overlapped DCT, and cartoon could be built by Curvelets/Ridgelets/Wavelets (depending on the content).
- Note that if we desire to enable partial support and/or different scale, the dictionaries must have multiscale and locality properties in them.

Decomposition via Sparsity



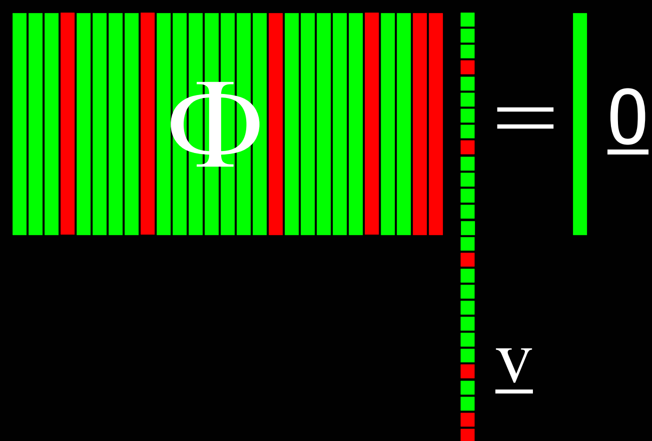
$$\begin{bmatrix}\hat{\underline{\alpha}}\\ \hat{\underline{\beta}}\end{bmatrix} = \underset{\underline{\alpha},\underline{\beta}}{\text{ArgMin}} \|\underline{\alpha}\|_0 + \|\underline{\beta}\|_0 \quad \text{s.t.} \quad \underline{s} = \begin{bmatrix}\Phi_x & \Phi_y\end{bmatrix}\begin{bmatrix}\underline{\alpha}\\ \underline{\beta}\end{bmatrix}$$

Why should this work?

Uniqueness via 'Spark'

- Given a unit norm signal $\underline{s}$, assume we hold two different representations for it using Φ

$$\underline{s} = \Phi\underline{\gamma}_1 = \Phi\underline{\gamma}_2 \quad \Rightarrow \quad \Phi\left(\underline{\gamma}_1 - \underline{\gamma}_2\right) = \underline{0}$$



Definition: Given a matrix Φ, define σ=Spark{Φ} as the smallest number of columns from Φ that are linearly dependent.

Uniqueness Rule

$$\sigma \le \left\|\underline{\gamma}_1\right\|_0 + \left\|\underline{\gamma}_2\right\|_0$$

Any two different representations of the same signal using an arbitrary dictionary cannot be jointly sparse [Donoho & E, 03`].

Theorem 1

If we found a representation that satisfy

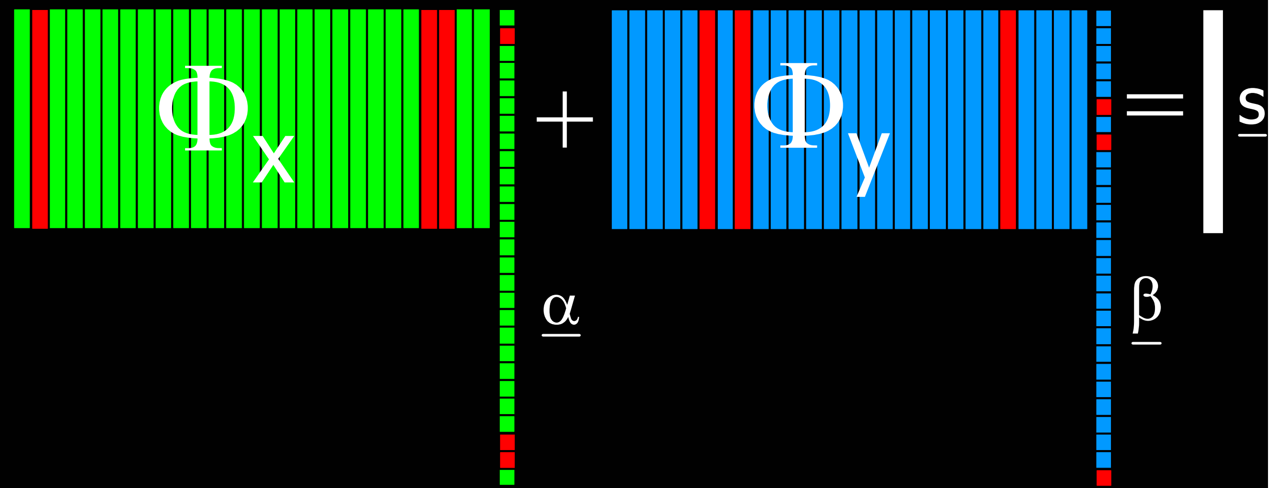
$$\frac{\sigma}{2} > \left\|\underline{\gamma}\right\|_0$$

Then necessarily it is unique (the sparsest).

Uniqueness Rule - Implications

$$\begin{bmatrix}\hat{\underline{\alpha}}\\ \hat{\underline{\beta}}\end{bmatrix} = \underset{\underline{\alpha},\underline{\beta}}{\text{ArgMin}} \left\|\underline{\alpha}\right\|_0 + \left\|\underline{\beta}\right\|_0$$

$$\text{s.t. } \underline{s} = \begin{bmatrix}\Phi_x & \Phi_y\end{bmatrix}\begin{bmatrix}\underline{\alpha}\\ \underline{\beta}\end{bmatrix}$$



- If $\left\|\hat{\underline{\alpha}}\right\|_0 + \left\|\hat{\underline{\beta}}\right\|_0 < 0.5\sigma\left(\begin{bmatrix}\Phi_x & \Phi_y\end{bmatrix}\right)$, it is necessarily the sparsest one possible, and it will be found.
- For dictionaries effective in describing the ‘cartoon’ and ‘texture’ contents, we could say that the decomposition that leads to separation is the sparsest one possible.

Lower bound on the "Spark"

- Define the *Mutual Incoherence* as

$$0 < M = \max_{\substack{1 \le k,j \le L \\ k \ne j}} \left\{ \left| \underline{\phi}_k^H \underline{\phi}_j \right| \right\} \le 1$$

- We can show (based on Gerśgorin disk theorem) that a lower-bound on the spark is obtained by

$$\sigma \ge 1 + \frac{1}{M}.$$

- Since the Gerśgorin theorem is non-tight, this lower bound on the Spark is too pessimistic.

Equivalence – The Result

We also have the following result [Donoho & E 02',Gribonval & Nielsen 03`] :

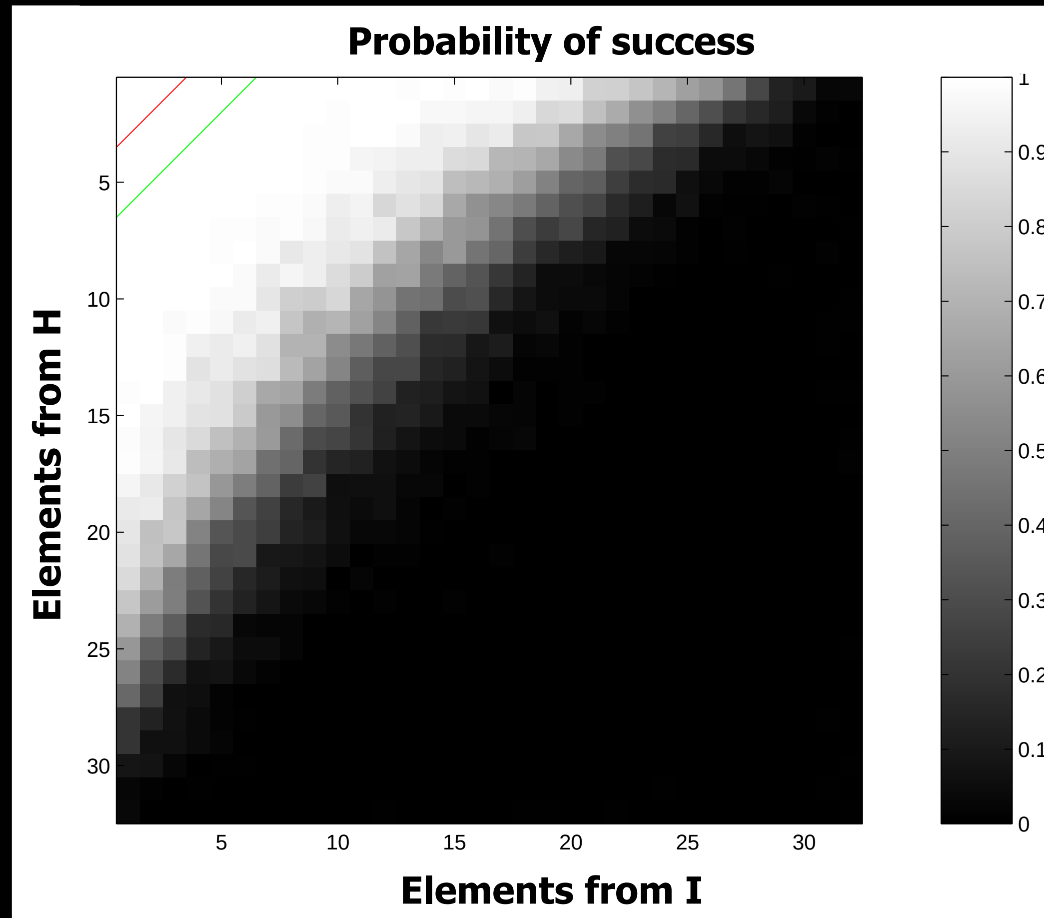
Theorem 2

Given a signal $\underline{s}$ with a representation $\underline{s} = \Phi\underline{\gamma}$,
Assuming that $\|\underline{\gamma}\|_0 < 0.5(1 + 1/M)$, P_1 (BP) is
Guaranteed to find the sparsest solution.

- BP is expected to succeed if sparse solution exists.
- A similar result exists for the greedy algorithms [Tropp 03'].
- In practice, the MP & BP succeed far above the bound.

Equivalence Beyond the Bound

- Dictionary $\Phi=[I,H]$ of size 64×128.
- M=1/8 – Unique. And Equiv. are guaranteed for 4 non-zeros and below.
- Spark=16 – Uniqueness is guaranteed for less than 8 non-zeros.
- As can be seen, the results are successful far above the bounds (empirical test with 100 random experiments per combination).



To Summarize so far …

Over-complete linear transforms – great for sparse representations

Can be used to separate images?

Design/choose proper Dictionaries

Theoretical Justification?

We show a uniqueness result guaranteeing proper separation if sparse enough

Is it practical?

- We show an equivalence result, implying - BP/MP successfully separate the image if sparse enough representation exists.
- We also show encouraging empirical behavior beyond the bounds

Agenda

1. Introduction

   Sparsity and Over-completeness!?

2. Theory of Decomposition

   Uniqueness and Equivalence

3. Decomposition in Practice

   Practical Considerations, Numerical algorithm

4. Discussion

Noise Considerations

$$\begin{bmatrix} \hat{\underline{\alpha}} \\ \hat{\underline{\beta}} \end{bmatrix} = \underset{\underline{\alpha},\underline{\beta}}{\text{ArgMin}} \ \|\underline{\alpha}\|_1 + \|\underline{\beta}\|_1 \quad \text{s.t.} \quad \underline{s} = \begin{bmatrix} \Phi_x & \Phi_y \end{bmatrix} \begin{bmatrix} \underline{\alpha} \\ \underline{\beta} \end{bmatrix}$$

Forcing exact representation is sensitive to additive noise and model mismatch

$$\begin{bmatrix} \hat{\underline{\alpha}} \\ \hat{\underline{\beta}} \end{bmatrix} = \underset{\underline{\alpha},\underline{\beta}}{\text{ArgMin}} \ \|\underline{\alpha}\|_1 + \|\underline{\beta}\|_1 + \lambda \left\|\underline{s} - \Phi_x \underline{\alpha} - \Phi_y \underline{\beta}\right\|_2^2$$

Recent results [Tropp 04', Donoho et.al. 04'] show that the noisy case generally meets similar rules of uniqueness and equivalence

Artifacts Removal

$$\begin{bmatrix} \hat{\underline{\alpha}} \\ \hat{\underline{\beta}} \end{bmatrix} = \underset{\underline{\alpha},\underline{\beta}}{\text{ArgMin}} \ \|\underline{\alpha}\|_1 + \|\underline{\beta}\|_1 + \lambda \left\| \underline{s} - \Phi_x \underline{\alpha} - \Phi_y \underline{\beta} \right\|_2^2$$

We want to add external forces to help the separation succeed, even if the dictionaries are not perfect

$$\begin{bmatrix} \hat{\underline{\alpha}} \\ \hat{\underline{\beta}} \end{bmatrix} = \underset{\underline{\alpha},\underline{\beta}}{\text{ArgMin}} \ \|\underline{\alpha}\|_1 + \|\underline{\beta}\|_1 + \lambda \left\| \underline{s} - \Phi_x \underline{\alpha} - \Phi_y \underline{\beta} \right\|_2^2 + \mu TV\left\{\Phi_x \underline{\alpha}\right\}$$

Complexity

$$\begin{bmatrix} \underline{\hat{\alpha}} \\ \underline{\hat{\beta}} \end{bmatrix} = \underset{\underline{\alpha},\underline{\beta}}{\text{ArgMin}} \left\|\underline{\alpha}\right\|_1 + \left\|\underline{\beta}\right\|_1 + \lambda \left\|\underline{s} - \Phi_x \underline{\alpha} - \Phi_y \underline{\beta}\right\|_2^2 + \mu TV\left\{\Phi_x \underline{\alpha}\right\}$$

Instead of 2N unknowns (the two separated images), we have 2L»2N ones.

Define two image unknowns to be

$$\underline{s}_x = \Phi_x \underline{\alpha} \quad , \quad \underline{s}_y = \Phi_y \underline{\beta}$$

and obtain …

Simplification

$$\begin{bmatrix}\hat{\underline{\alpha}}\\ \hat{\underline{\beta}}\end{bmatrix} = \underset{\underline{\alpha},\underline{\beta}}{\text{ArgMin}} \left\|\underline{\alpha}\right\|_1 + \left\|\underline{\beta}\right\|_1 + \lambda\left\|\underline{s} - \Phi_x\underline{\alpha} - \Phi_y\underline{\beta}\right\|_2^2 + \mu TV\left\{\Phi_x\underline{\alpha}\right\}$$

$$\underline{s}_x = \Phi_x\underline{\alpha} \quad \Rightarrow \quad \begin{matrix}\underline{\alpha} = \Phi_x^+\underline{s}_x + \underline{r}_x \\ \text{where } \Phi_x\underline{r}_x = 0\end{matrix}$$

$$\begin{bmatrix}\hat{\underline{s}}_x\\ \hat{\underline{s}}_y\end{bmatrix} = \underset{\underline{s}_x,\underline{s}_y}{\text{ArgMin}} \left\|\Phi_x^+\underline{s}_x\right\|_1 + \left\|\Phi_y^+\underline{s}_y\right\|_1 + \lambda\left\|\underline{s} - \underline{s}_x - \underline{s}_y\right\|_2^2 + \mu TV\left\{\underline{s}_x\right\}$$

Justifications

Heuristics: (1) Bounding function; (2) Relation to BCR; (3) Relation to MAP.

Theoretic: See recent results by D.L. Donoho.

Algorithm

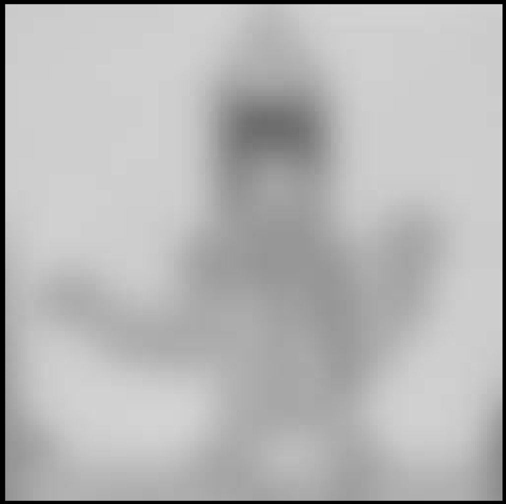
$$\begin{bmatrix} \underline{\hat{s}}_x \\ \underline{\hat{s}}_y \end{bmatrix} = \underset{\underline{s}_x, \underline{s}_y}{\text{ArgMin}} \left\| \Phi_x^+ \underline{s}_x \right\|_1 + \left\| \Phi_y^+ \underline{s}_y \right\|_1 + \lambda \left\| \underline{s} - \underline{s}_x - \underline{s}_y \right\|_2^2 + \mu TV\{\underline{s}_x\}$$

An algorithm was developed to solve the above problem:

- It iterates between an update of $\underline{s}_x$ to update of $\underline{s}_y$.
- Every update (for either $\underline{s}_x$ or $\underline{s}_y$) is done by a forward and backward fast transforms – this is the dominant computational part of the algorithm.
- The update is performed using diminishing soft-thresholding (similar to BCR but sub-optimal due to the non unitary dictionaries).
- The TV part is taken-care-of by simple gradient descent.
- Convergence is obtained after 10-15 iterations.

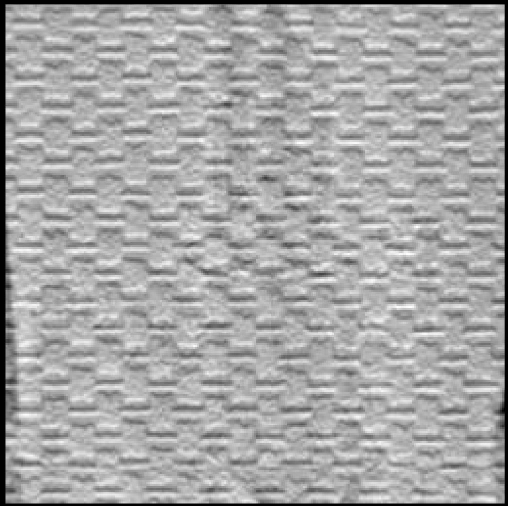
Results 1 – Synthetic Case

Original image composed as a combination of texture and cartoon



The very low freq. content – removed prior to the use of the separation

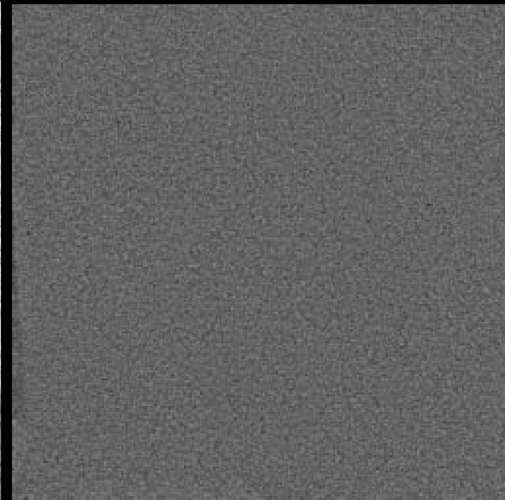
The separated texture (spanned by Global DCT functions)



The separated cartoon (spanned by 5 layer Curvelets functions+LPF)

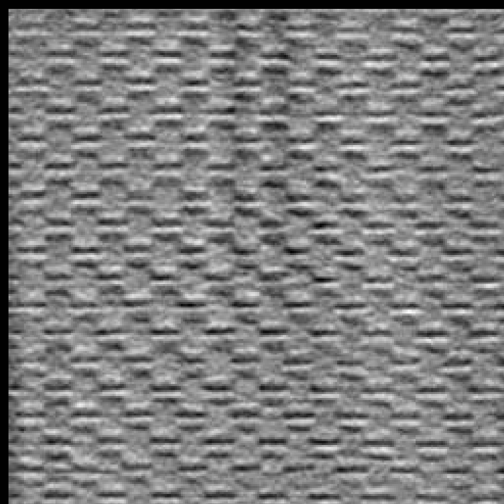
Results 2 – Synthetic + Noise

Original image composed as a combination of texture, cartoon, and additive noise (Gaussian, $\sigma = 10$)



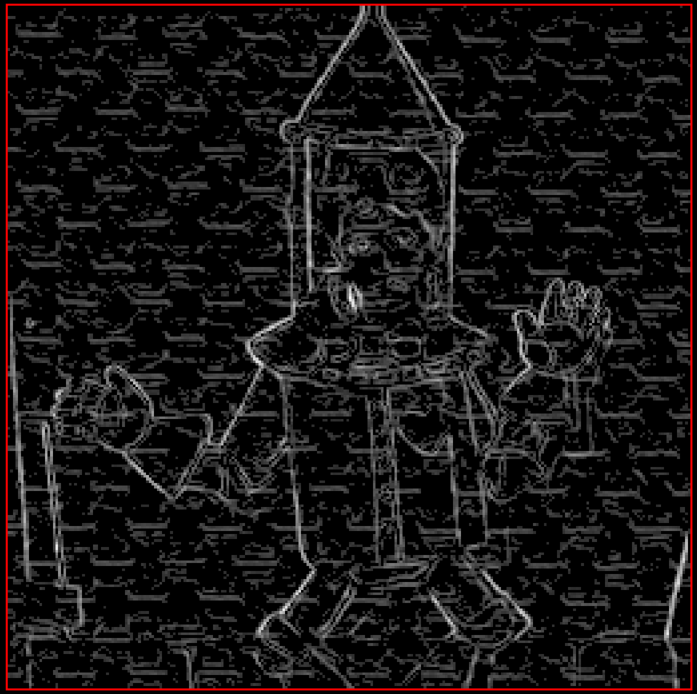
The residual, being the identified noise

The separated texture (spanned by Global DCT functions)



The separated cartoon (spanned by 5 layer Curvelets functions+LPF)

Results 3 – Edge Detection

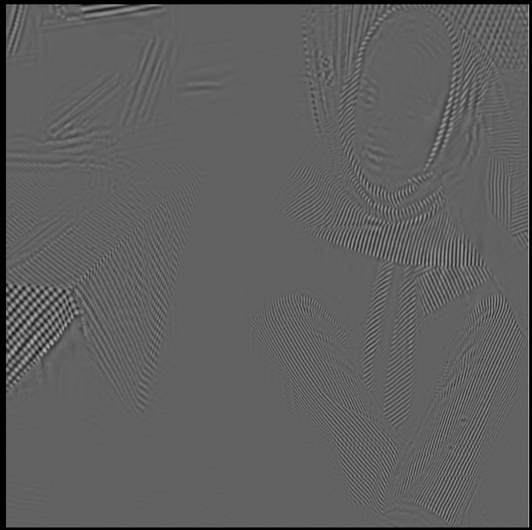


Edge detection on the original image

Edge detection on the cartoon part of the image

Results 4 – Good old 'Barbara'

Original 'Barbara' image

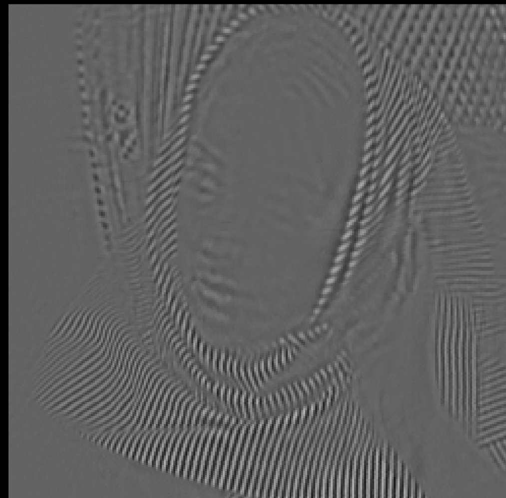


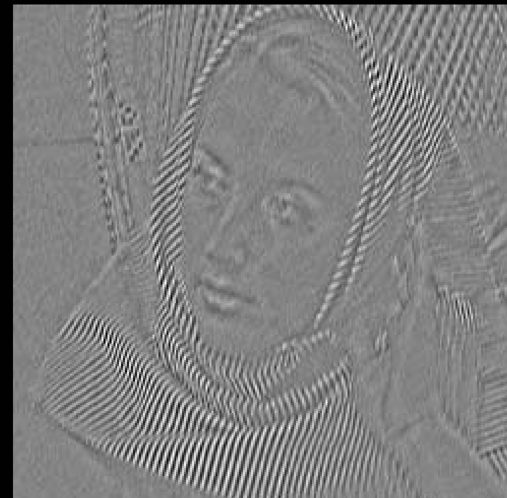
Separated texture using local overlapped DCT (32×32 blocks)

Separated Cartoon using Curvelets (5 resolution layers)

Results 4 – Zoom in

Zoom in on the result shown in the previous slide (the texture part)





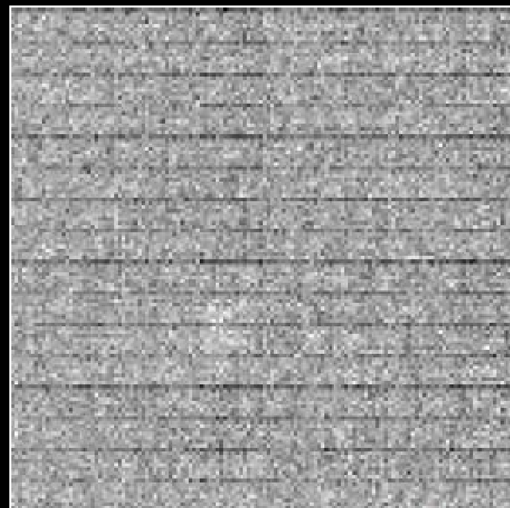
The same part taken from Vese's et. al.

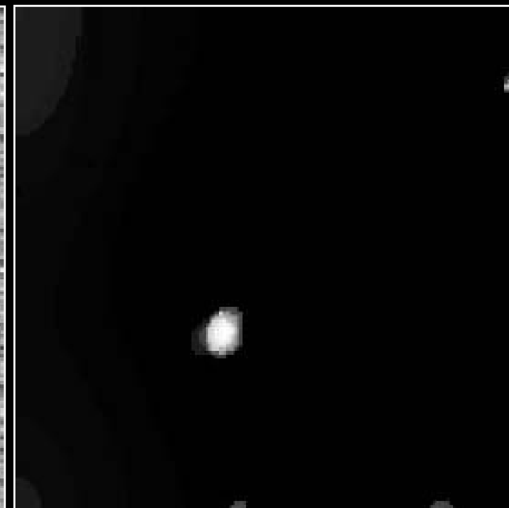
Zoom in on the results shown in the previous slide (the cartoon part)

The same part taken from Vese's et. al.

Results 5 – Gemini

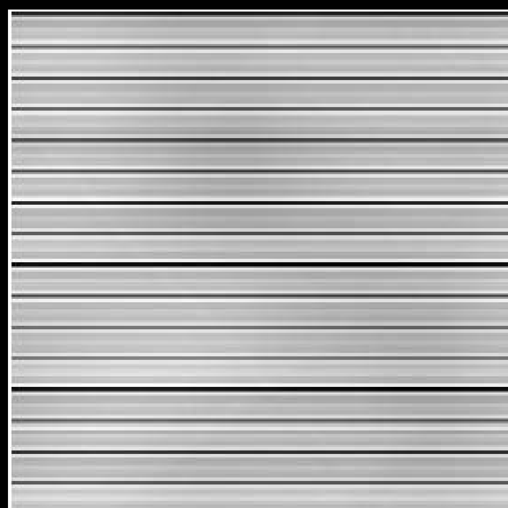
The original image - Galaxy SBS 0335-052 as photographed by Gemini

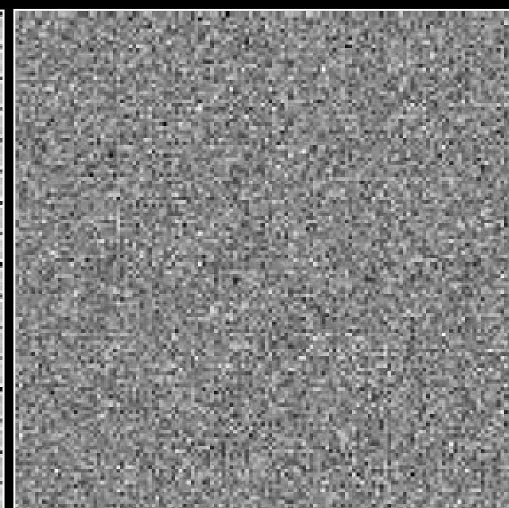




The Cartoon part spanned by wavelets

The texture part spanned by global DCT





The residual being additive noise

Side Story - Inpainting

For separation

$$\begin{bmatrix}\hat{\underline{\alpha}}\\ \hat{\underline{\beta}}\end{bmatrix} = \underset{\underline{\alpha},\underline{\beta}}{\text{ArgMin}} \|\underline{\alpha}\|_1 + \|\underline{\beta}\|_1 + \lambda \|\underline{s} - \Phi_x\underline{\alpha} - \Phi_y\underline{\beta}\|_2^2$$

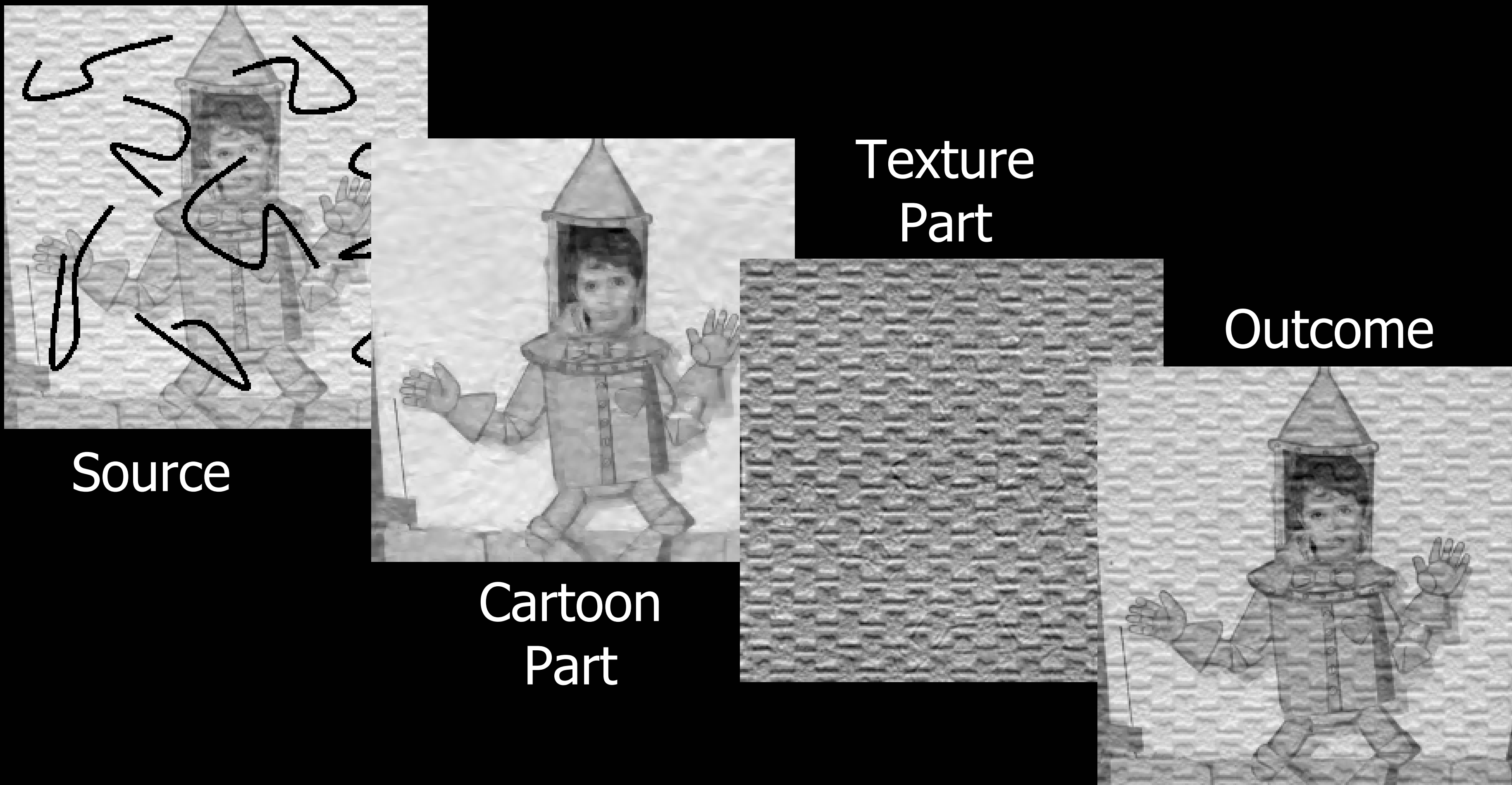
What if some values in $\underline{s}$ are unknown (with known locations!!!)?

$$\begin{bmatrix}\hat{\underline{\alpha}}\\ \hat{\underline{\beta}}\end{bmatrix} = \underset{\underline{\alpha},\underline{\beta}}{\text{ArgMin}} \|\underline{\alpha}\|_1 + \|\underline{\beta}\|_1 + \lambda \|W(\underline{s} - \Phi_x\underline{\alpha} - \Phi_y\underline{\beta})\|_2^2$$

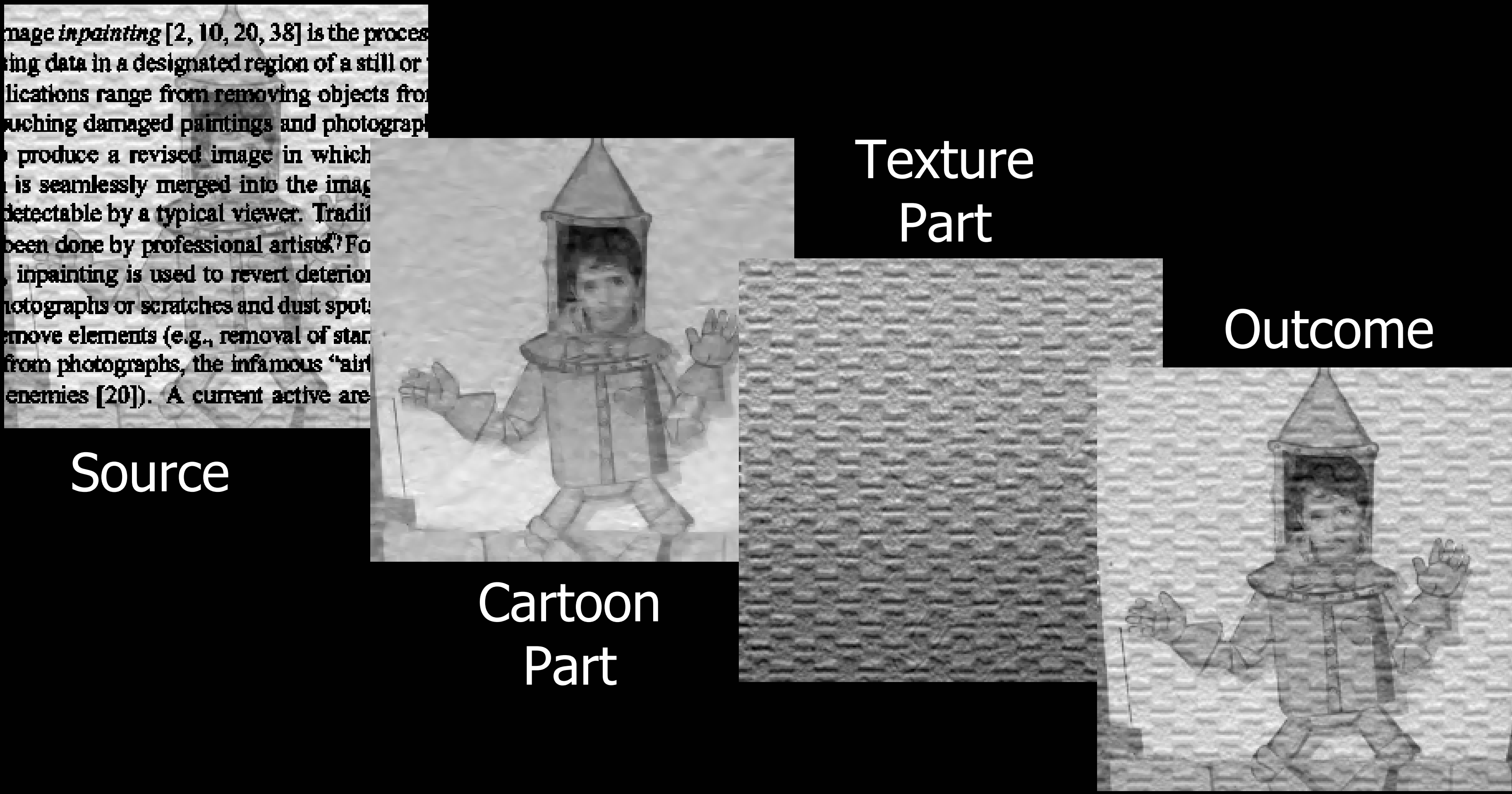
The image $\Phi_x\underline{\alpha} + \Phi_y\underline{\beta}$ will be the inpainted outcome.

Interesting comparison to [Bertalmio et.al. '02]

Results 6 - Inpainting



Results 7 - Inpainting



Source

Cartoon Part

Texture Part

Outcome

Results 8 - Inpainting

Source

Outcome

There are still artifacts –
these are just preliminary results

Agenda

1. Introduction

   Sparsity and Over-completeness!?

2. Theory of Decomposition

   Uniqueness and Equivalence

3. Decomposition in Practice

   Practical Considerations, Numerical algorithm

4. Discussion

Summary

Over-complete and Sparsity are powerful in representations of signals

Application?

Decompose an image to Cartoon+Texture

Theoretical Justification?

We show theoretical results explaining how could this lead to successful separation. Also, we show that pursuit algorithms are expected to succeed

Practical issues?

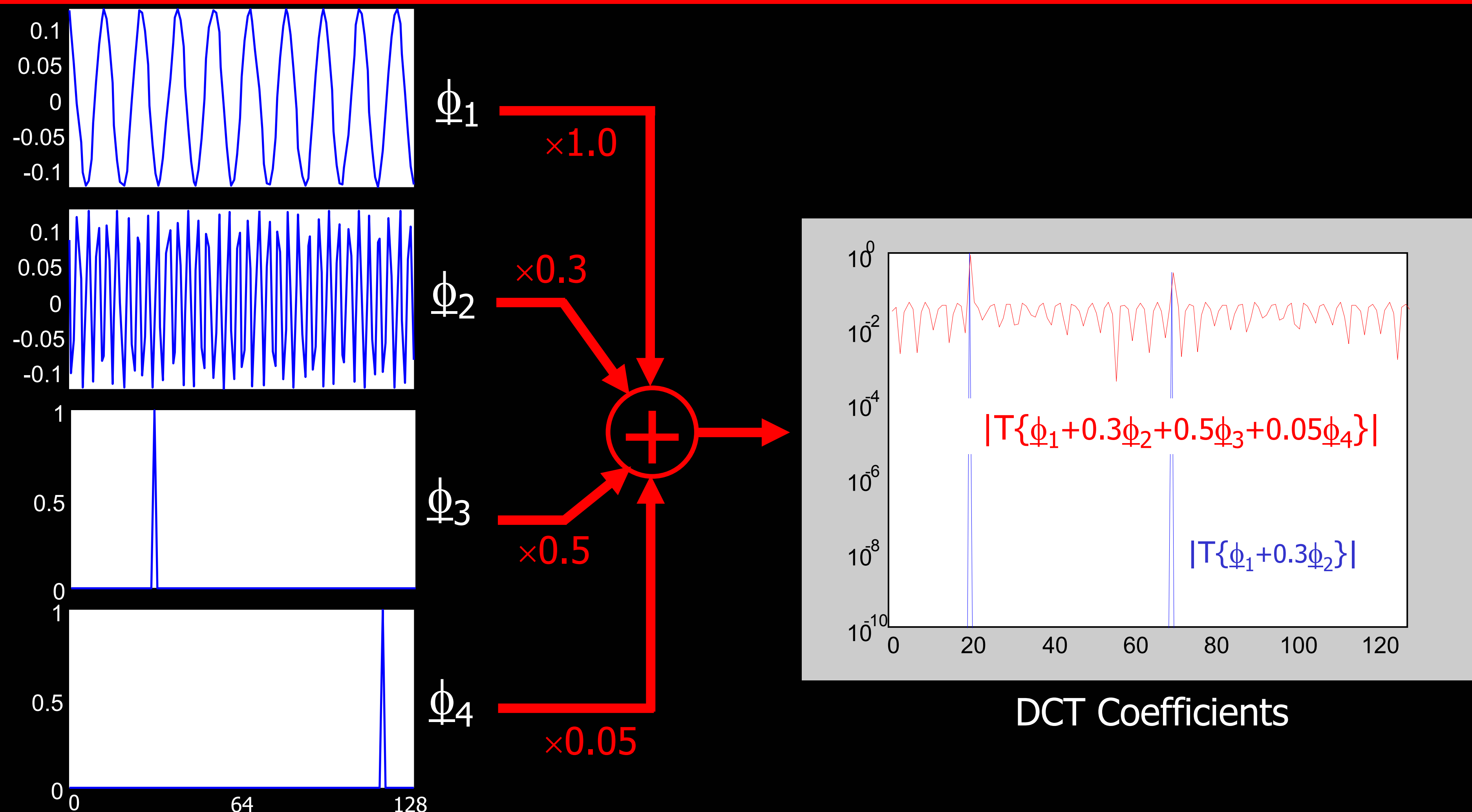
We present ways to robustify the process, and apply it to image inpainting

Where next?

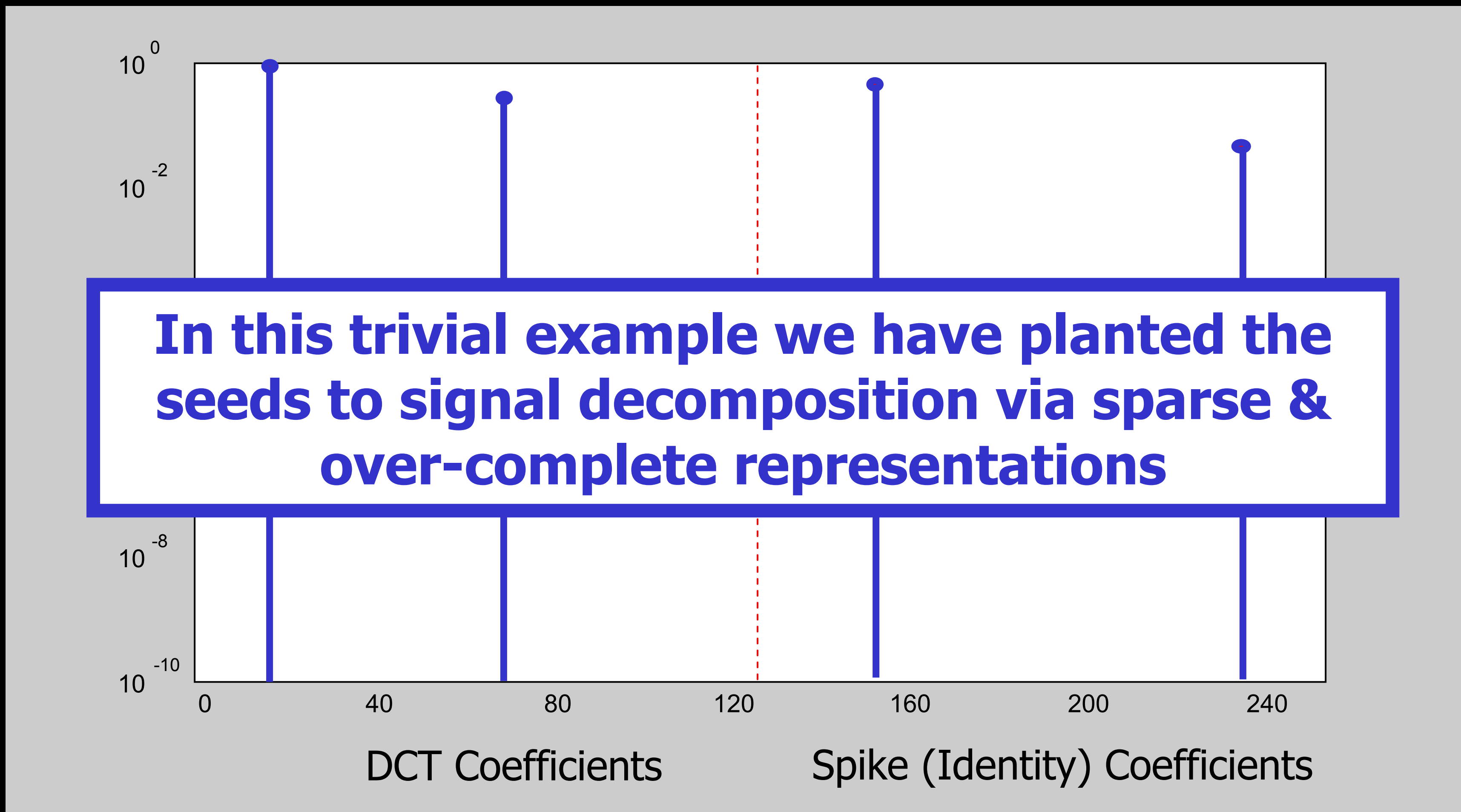
Choice of dictionaries, performance beyond the bounds, Other applications? More ...

These slides and related papers can be found in:
http://www.cs.technion.ac.il/~elad

Why Over-Completeness?

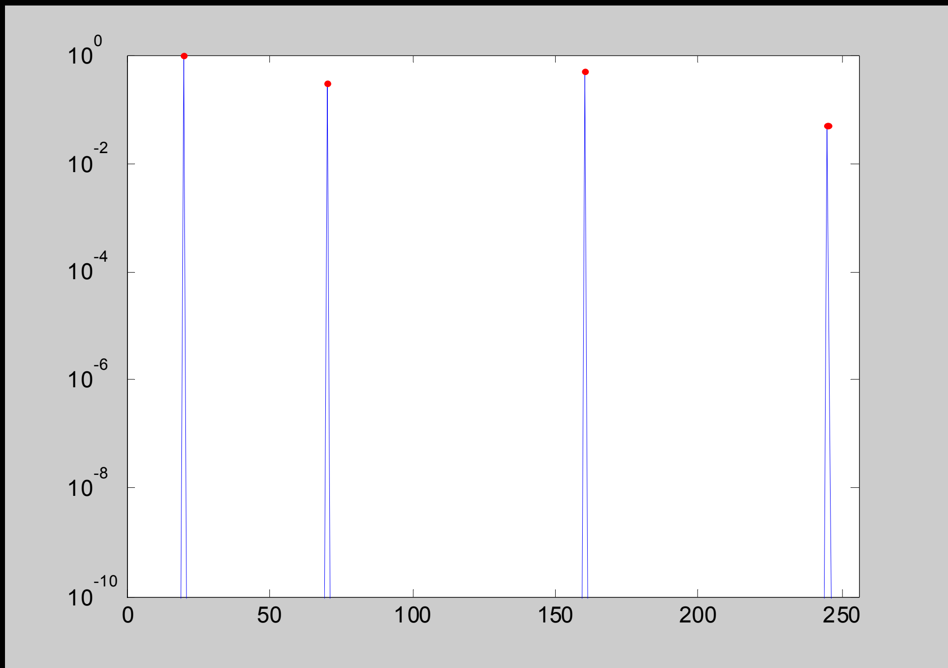


Desired Decomposition



In this trivial example we have planted the seeds to signal decomposition via sparse & over-complete representations

Example – Basis Pursuit



Dictionary Coefficients

- The same problem can be addressed using the (greedy stepwise regression) Matching Pursuit (MP) algorithm [Zhang & Mallat. 93'].
- Why BP/MP should work well? Are there Conditions to this success?
- Could there be a different sparse representation? What about uniqueness?

Appendix A – Relation to Vese's

$$\underset{\underline{s}_x,\underline{s}_y}{\text{Min}} \left\|\Phi_x^+ \underline{s}_x\right\|_1 + \left\|\Phi_y^+ \underline{s}_y\right\|_1 + \lambda \left\|\underline{s} - \underline{s}_x - \underline{s}_y\right\|_2^2$$

If Φ_x^+ is one resolution layer of the non-decimated Haar – we get TV

If Φ_x^+ is the local DCT, then requiring sparsity parallels the requirement for oscilatory behavior

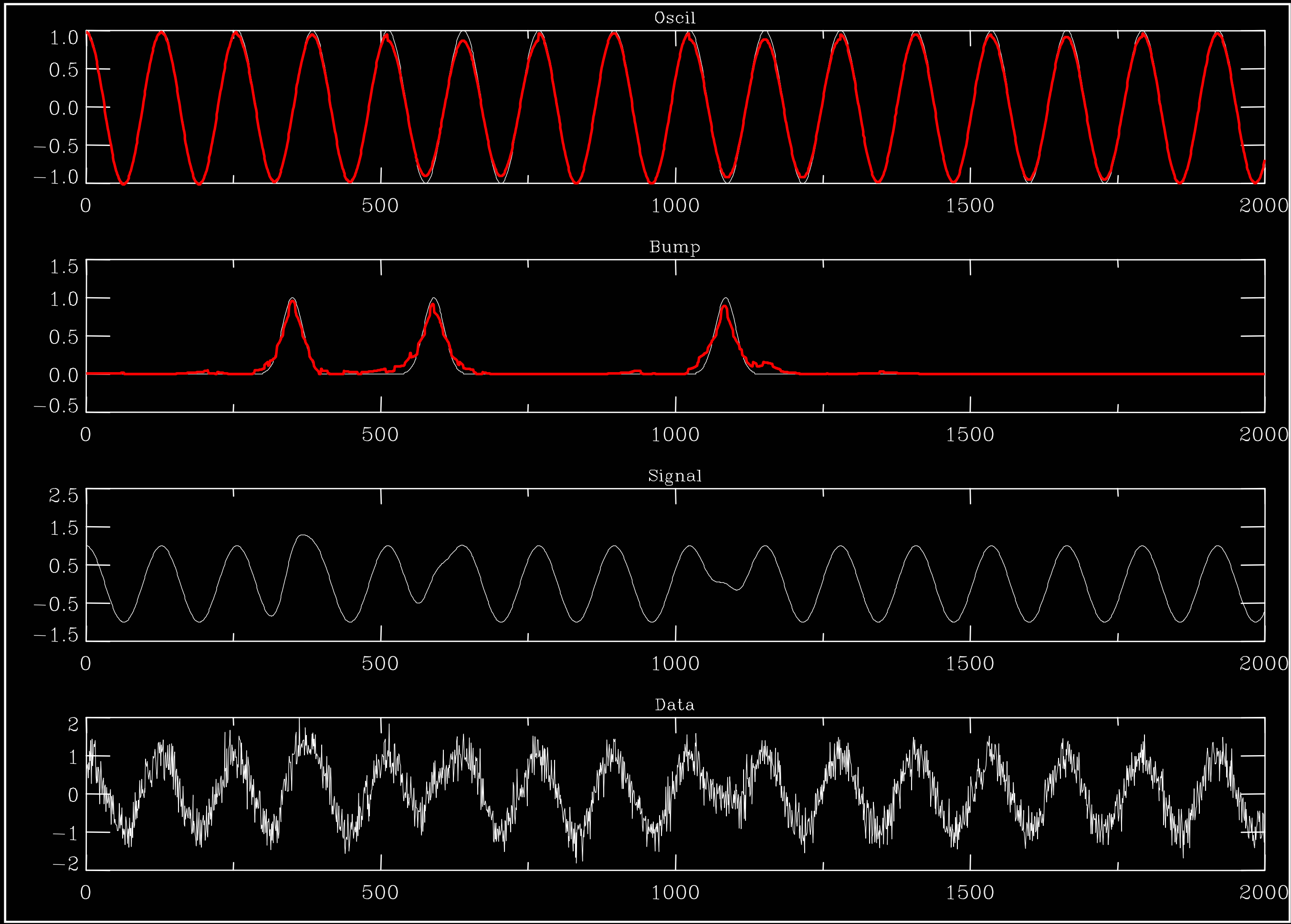
$$\underset{\underline{s}_x,\underline{s}_y}{\text{Min}} \left\|\underline{s}_x\right\|_{BV} + \left\|\underline{s}_y\right\|_{BV*} + \lambda \left\|\underline{s} - \underline{s}_x - \underline{s}_y\right\|_2^2$$

Vese & Osher's Formulation

Results 0 – Zoom in

An oscillating function is added to a function with bumps, and this addition is contaminated with noise.

The separation is done with local-DCT (blocks of 256) and isotropic wavelet.



Why Over-Completeness?

- Many available square linear transforms – sinusoids, wavelets, packets, …
- Definition: Successful transform is one which leads to sparse (sparse=simple) representations.
- Observation: Lack of universality - Different bases good for different purposes.
  - Sound = harmonic music (Fourier) + click noise (Wavelet),
  - Image = lines (Ridgelets) + points (Wavelets).
- Proposed solution: Over-Complete dictionaries, and possibly combination of bases.

To Summarize so far …

Over-complete and Sparse representations

How can we practically use those?

(Basis/Matching) Pursuit algorithms can be used with promising empirical behavior.

Theory? Applications?

In the next part we show how sparsity and over-completeness can drive the image separation application and how we can theoretically guarantee performance.