The Dichotomy Between Global Processing & Local Modeling *

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

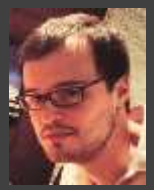
December 7-11, 2015
Technical University of Berlin

2. International
Matheon Conference on
Compressed Sensing and its Applications

* Joint work with


Dima Batenkov


Jere. Sulam


Vardan Papyan


Yaniv Romano

The research leading to these results has received funding from the European Research Council under European Union's Seventh Framework Program, ERC Grant agreement no. 320649, and by the Intel Collaborative Research Institute for Computational Intelligence

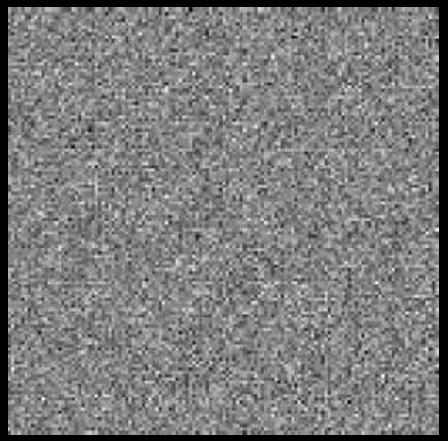
Part I

Motivating this Discussion

Our Starting Point: Image Denoising

Original Image

White Gaussian Noise



Noisy Image

Many of the proposed image denoising algorithms are cast as the minimization of an energy function of the form

$$f(\underline{X}) = \frac{1}{2}\|\underline{X} - \underline{Y}\|_2^2 + G(\underline{X})$$

Relation to measurements

Prior or regularization

$\underline{Y}$: Given measurements

$\underline{X}$: Unknown to be recovered

Leading Image Denoising Methods

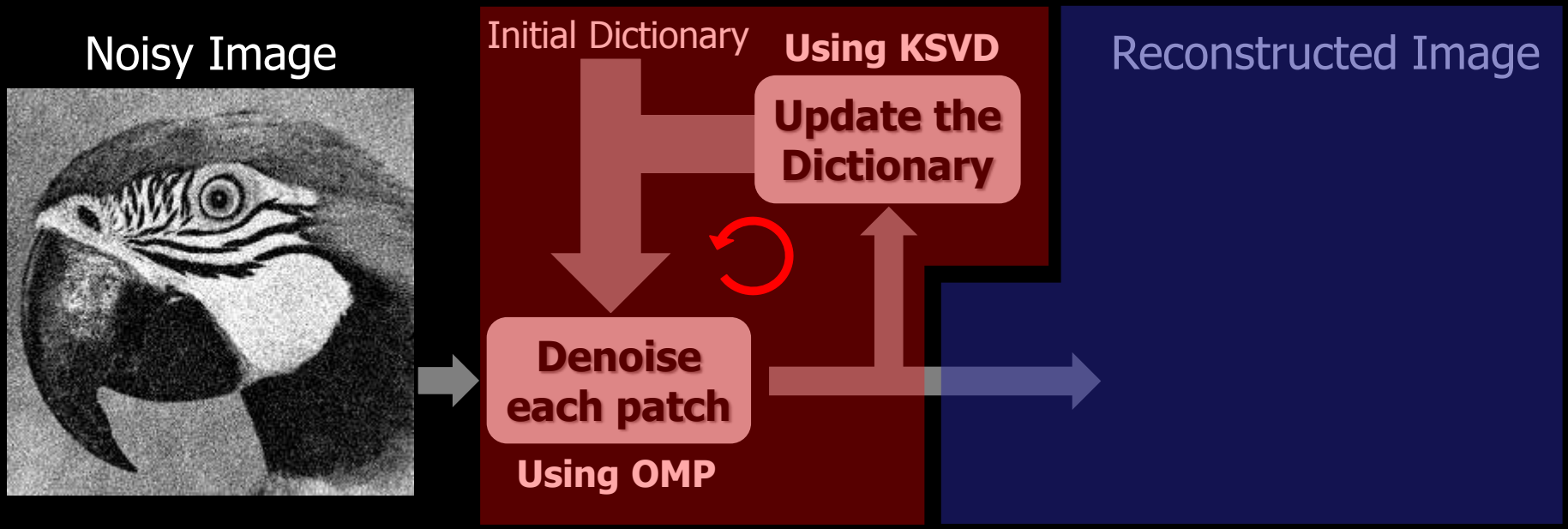
are built upon powerful patch-based (local) image models:

- K-SVD: sparse representation modeling of image patches [Elad & Aharon, '06]
- BM3D: combines a sparsity and self-similarity [Dabov, Foi, Katkovnik & Egiazarian '07]
- EPLL: uses GMM of the image patches [Yu, Sapiro & Mallat '10] [Zoran & Weiss '11]
- CSR: clustering and sparsity on patches [Dong, Li, Lei & Shi '11]
- …

IN THIS TALK

- ❑ We aim to dive into this strange choice of modeling signals/images locally for regularizing global inverse problems. What is the rationale behind this? Is this enough? Can it be improved?
- ❑ We shall start with a practical view and gradually move to a theory.
- ❑ We will have more open question than answers …

Consider this Algorithm [Elad & Aharon, '06]

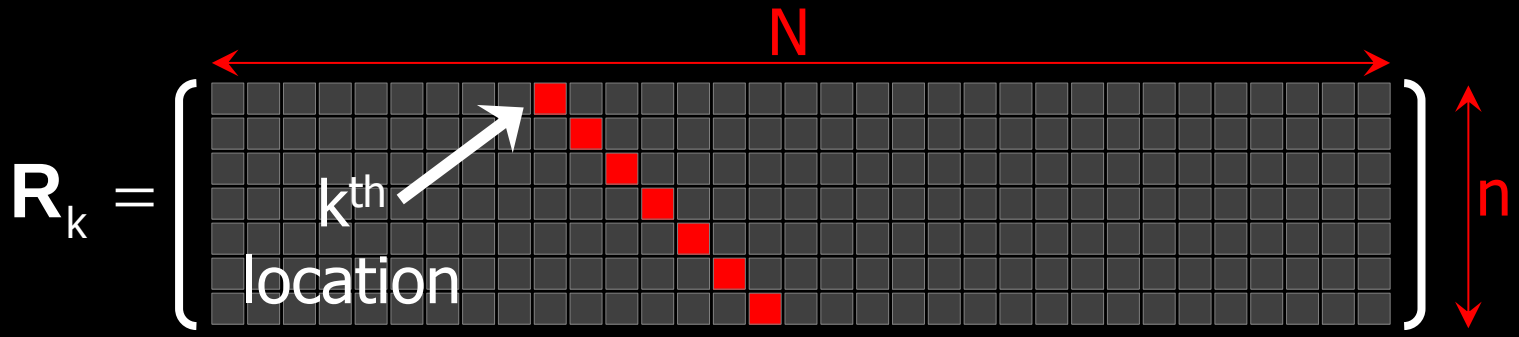


- This method is very effective for image denoising. Many variants of it were developed over the years in order to extend/improve it.
- The above is only one in the large family **of patch-based algorithms** that denoise an image by decomposing it into patches, processing them separately and then merging them back by **plain averaging** to form the final outcome.

This Algorithm's Origin

$$\hat{\underline{X}} = \underset{\underline{X},\{\underline{\alpha}_k\}_k,\mathbf{D}}{\text{ArgMin}} \frac{1}{2}\|\underline{X}-\underline{Y}\|_2^2 + \mu\sum_k \|\mathbf{R}_k\underline{X}-\mathbf{D}\underline{\alpha}_k\|_2^2 \text{ s.t. } \|\underline{\alpha}_k\|_0 \le L$$

- The expression $\mathbf{R}_k\underline{X}$ extracts a patch of size n from $\underline{X}\in\mathbb{R}^N$:



- $\mathbf{R}_k^T\underline{z}$ puts the patch $\underline{z}$ in the k^{th} location in the N-dim. vector.
- Considering cyclic patch-extraction, we have: $\frac{1}{n}\sum_k \mathbf{R}_k^T\mathbf{R}_k = \mathbf{I}$

This Algorithm's Origin

$$\hat{\underline{X}} = \underset{\underline{X},\{\underline{\alpha}_k\}_k,\mathbf{D}}{\text{ArgMin}} \frac{1}{2}\|\underline{X}-\underline{Y}\|_2^2 + \mu\sum_k \|\mathbf{R}_k\underline{X}-\mathbf{D}\underline{\alpha}_k\|_2^2 \quad \text{s.t.} \quad \|\underline{\alpha}_k\|_0 \le L$$

Log-Likelihood

Prior or regularization

The Proposed Regularization / Model:

Every patch in the unknown image is expected to have a sparse representation w.r.t. the dictionary **D**

- This algorithm seeks the "most appropriate" dictionary to fulfil this expectation, along with the sparse representations.
- The denoising itself is obtained by projecting each of the patches to this model (via OMP).
- When optimizing over $\underline{X}$, this amounts to plain patch-averaging.

So, What is Missing?

- Over the past several years, many researchers kept revisiting this algorithm and the line of thinking behind it, with a clear feeling that the final word has not been said, and that key features are still lacking.
- What is missing? Here is what we thought of …
  - A multi-scale treatment [Ophir, Lustig, & Elad '11] [Sulam, Ophir & Elad '14] [Papyan & Elad '15]
  - Exploiting self-similarities [Ram & Elad '13] [Romano, Protter & Elad, '14]
  - Pushing to better agreement on the overlaps [Romano & Elad '13] [Romano & Elad '15][2]
  - Enforcing the local model on the final patches (EPLL) [Sulam & Elad '15]
- Beyond all these, a key part that is missing is **A Theoretical Backbone** for the local model as a way to characterize the unknown image.

Theoretical Backbone?

The core model Assumption on $\underline{X}$:

$$\forall k \quad \mathbf{R}_k \underline{X} = \mathbf{D}\underline{\alpha}_k \quad \text{where} \quad \left\|\underline{\alpha}_k\right\|_0 \leq L$$

Every patch in the unknown signal is expected to have a sparse representation w.r.t. the dictionary **D**

Questions to consider:

- Who are those signals belonging to this model?
- Under which conditions on **D** would this model be feasible?
- How does one sample from this model?
- How should we perform pursuit properly (& locally) under this model?
- How should we learn **D** if this is indeed the model?
- …

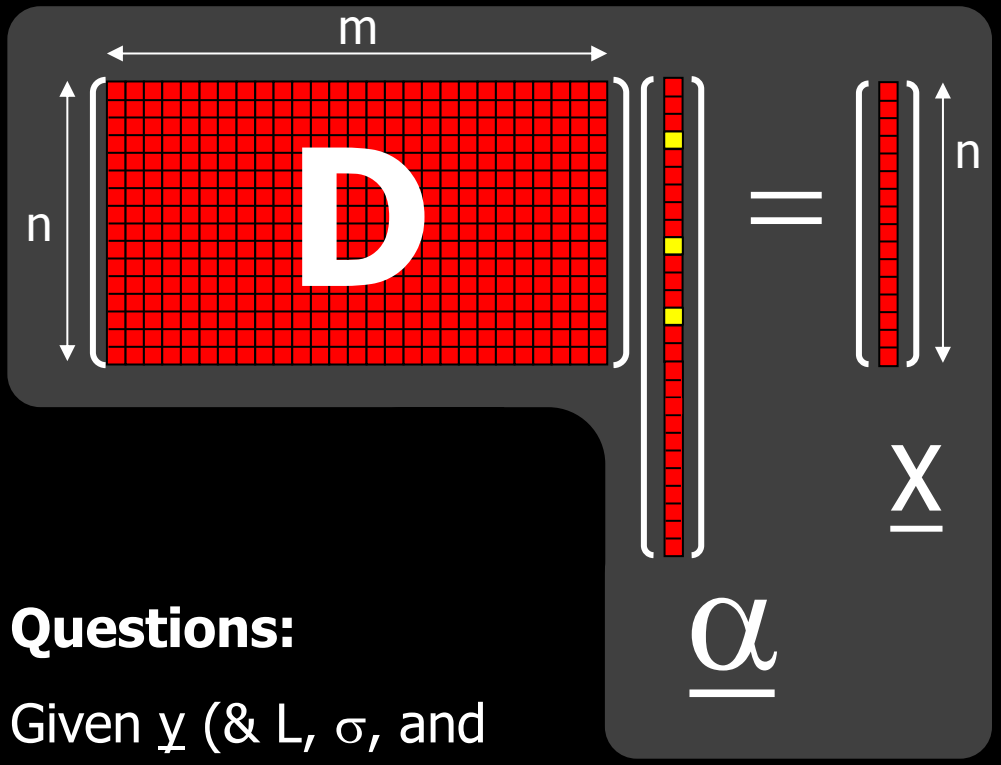
Why Sampling ?

Surely, you are familiar with this line of work …

Generate a random vector $\underline{\alpha}$ of length m and with L<<m non-zeros

Multiply $\underline{\alpha}$ by the dictionary **D** of size n-by-m, and obtain $\underline{x}=\mathbf{D}\underline{\alpha}$

Add noise (WAGN) with STD=σ to the signal and get $\underline{y}=\underline{x}+\underline{v}$



Questions:

Given $\underline{y}$ (& L, σ, and some properties of **D**),

(i) Can we recover the true support of $\underline{\alpha}$?
(ii) How efficiently can we denoise $\underline{y}$?

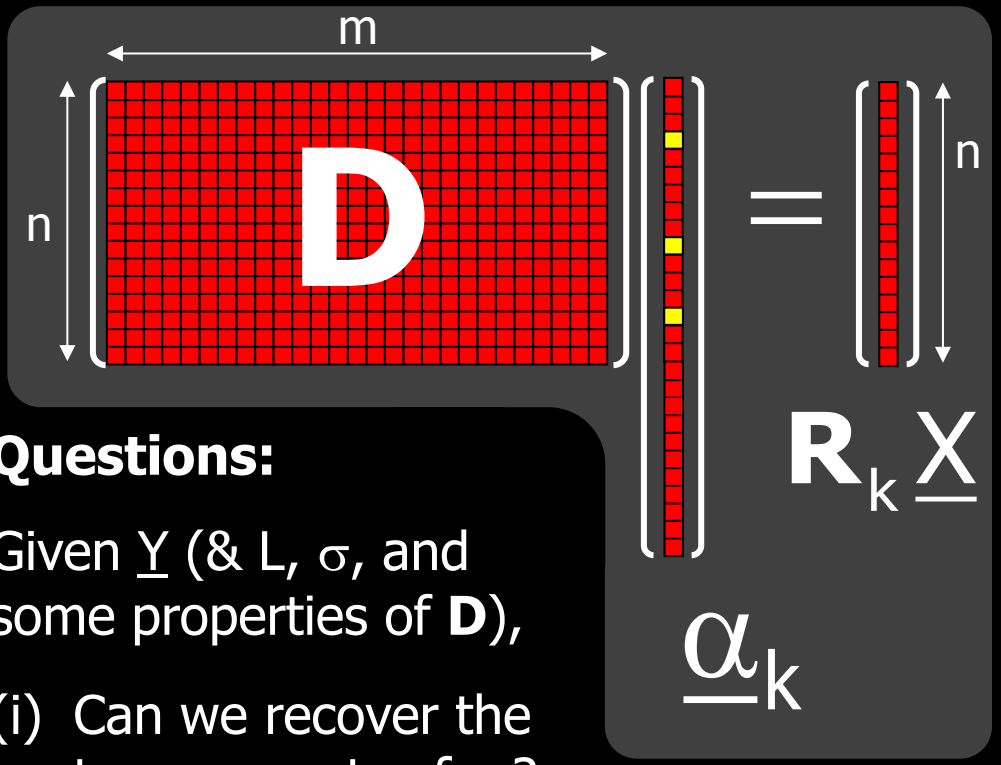
Why Sampling ?

Generate a signal $\underline{X}$ such that each of its patches has a sparse representation

$$\forall k \quad \mathbf{R}_k \underline{X} = \mathbf{D}\underline{\alpha}_k$$

$$\text{where } \|\underline{\alpha}_k\|_0 \le L$$

Add noise (WAGN) with STD=σ to the signal and get $\underline{Y}=\underline{X}+\underline{V}$



Questions:

Given $\underline{Y}$ (& L, σ, and some properties of **D**),

(i) Can we recover the true supports of $\underline{\alpha}_k$?

(ii) How efficiently can we denoise $\underline{Y}$? What can the oracle do?

(iii) Can we do all this by local processing?

This Talk

Part I – Motivating this Discussion

Part II – A Toy Problem: The Gaussian Case

Part III – Back to Sparse Representations: Who are these Signals ?

Part IV – Pursuit and Denoising for these Signals

Part V – A Closer Look at the case of Conv. Sparsity

Part VI – Concluding Remarks

Take Home Message

- We propose an interesting extension of the classic Sparse Approximation theory to local models that are imposed on global signals.
- For such signals, we address fundamental questions that are of great relevance to image processing.

Part II

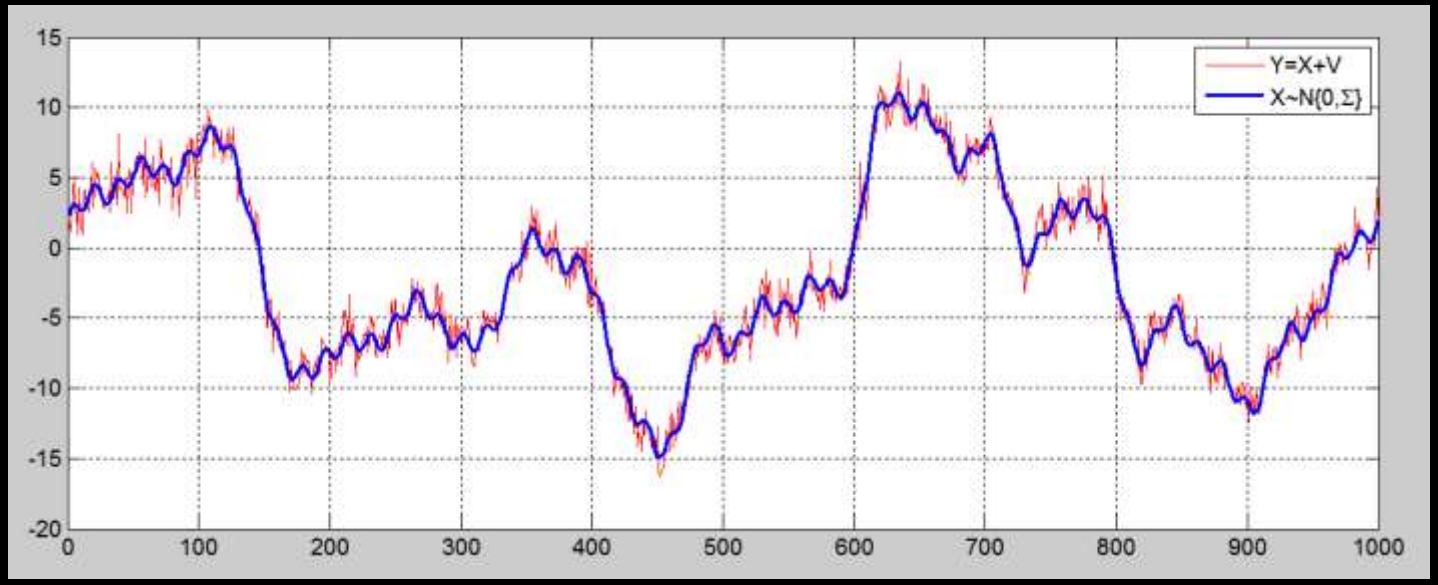
Toy Problem: The Gaussian Case

Problem Definition

- We are given a Gaussian signal $\underline{X} \in \mathbb{R}^N$ of known statistics, $\underline{X} \sim N\{\underline{0}, \Sigma\}$ contaminated by WAGN:

$$\underline{Y} = \underline{X} + \underline{V}, \quad \underline{V} \sim N\{\underline{0}, I\}$$

- Our goal – denoise $\underline{Y}$ to get as close as possible to $\underline{X}$.



* For simplicity we assume zero mean signal X and a unit variance noise.

Sounds Easy – Wiener Filtering !

- Use the Wiener Filter for the denoising, as it gives the Minimum Mean Square Error (MMSE) result:

$$\min_{\underline{X}} \frac{1}{2}\left\|\underline{Y}-\underline{X}\right\|_2^2+\frac{1}{2}\underline{X}^T\Sigma^{-1}\underline{X}$$

This is –log P($\underline{X}$|$\underline{Y}$)

$$\underline{\hat{X}}_{\text{Global}}=\Sigma\left(\mathbf{I}+\Sigma\right)^{-1}\underline{Y}$$

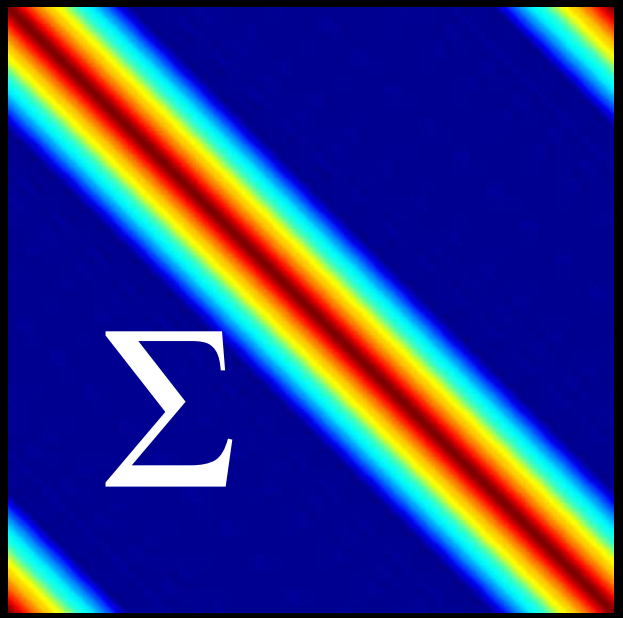
- This is the best we could perform! However,
  - It calls for knowing Σ, and
  - Inverting a matrix of size N-by-N.

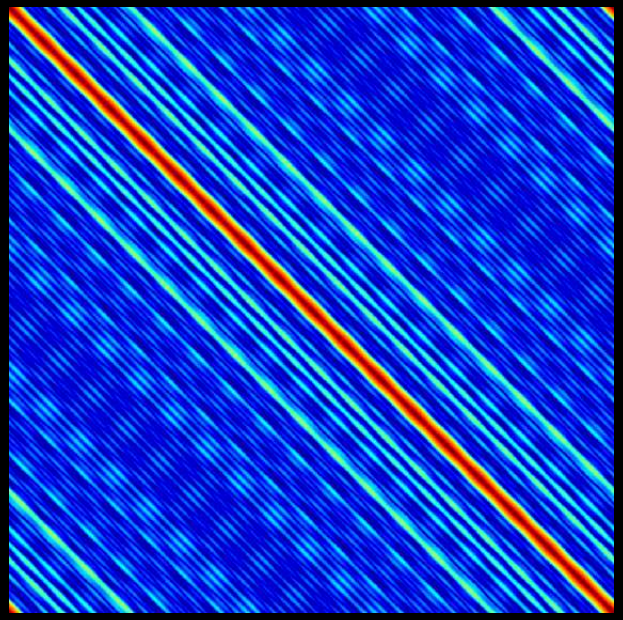
Sounds Easy – Wiener Filtering !

- For example, using the following Σ (N=1000), the obtained linear filter is this:

$$\hat{\underline{X}}_{\text{Global}} = \Sigma\left(\mathbf{I}+\Sigma\right)^{-1}\underline{Y}$$

We are choosing Σ to be a circulant matrix for reasons that will be clear shortly





Logarithmic Scale

- Are there simpler alternatives?

Local Processing by Patch-Averaging

- A patch of length n<<N extracted from $\underline{X}$ by $\mathbf{R}_k\underline{X}$ is also Gaussian-distributed, obtained as the marginal distribution:

$$\underline{x}_k = \mathbf{R}_k\underline{X} \sim N\left\{\underline{0}, \mathbf{R}_k\Sigma\mathbf{R}_k^T\right\}$$

- We could apply the Wiener filter to each of these patches (overlapped) and then seek a global solution that best fits these local results:

$$\min_{\underline{X}} \sum_k \left\|\mathbf{R}_k\underline{X} - \hat{\underline{x}}_k\right\|_2^2$$

Local Processing by Patch-Averaging

- This process amounts to a simple aggregation by averaging:

$$\underline{\hat{x}}_k = \mathbf{R}_k \Sigma \mathbf{R}_k^T \left(\mathbf{R}_k \Sigma \mathbf{R}_k^T + \mathbf{I} \right)^{-1} \mathbf{R}_k \underline{Y}$$

$$\underline{\hat{X}}_{LPA} = \frac{1}{n} \sum_k \mathbf{R}_k^T \cdot \underline{\hat{x}}_k$$

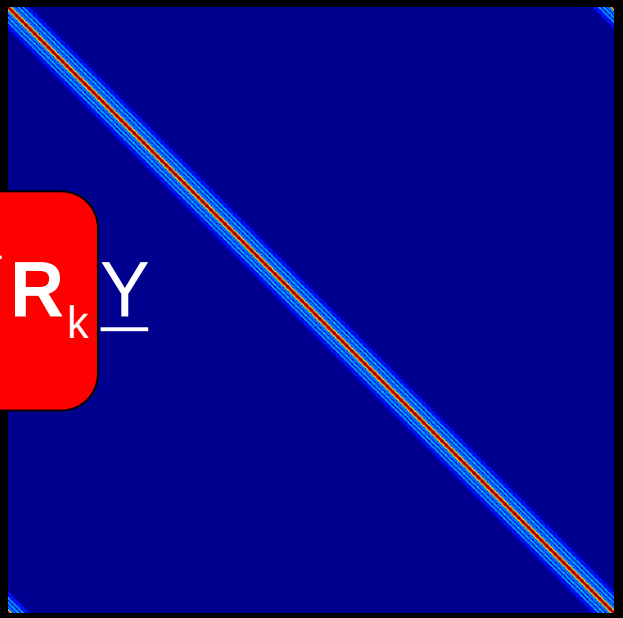
$$= \frac{1}{n} \sum_k \mathbf{R}_k^T \cdot \mathbf{R}_k \Sigma \mathbf{R}_k^T \left(\mathbf{R}_k \Sigma \mathbf{R}_k^T + \mathbf{I} \right)^{-1} \mathbf{R}_k \underline{Y}$$

Local Patch-Averaging (LPA)

- By assuming a circulant Σ, all local models are the same, which is the typical case practiced in image processing of assuming the same model for all patches.

- The obtained linear filter now is this:

$$\underline{\hat{X}}_{LPA} = \frac{1}{n}\sum_k \mathbf{R}_k^T\mathbf{R}_k\Sigma\mathbf{R}_k^T\left(\mathbf{R}_k\Sigma\mathbf{R}_k^T+\mathbf{I}\right)^{-1}\mathbf{R}_k\underline{Y}$$



- Naturally, this matrix is circulant and banded, with width 2n-1 (n=40).

Could We Do Better ? Yes – EPLL !

- We could refer to the global signal as the unknown, but impose the local Gaussian models. This is the rational of the EPLL (Expected Patch Log Likelihood) [Zoran & Weiss '10].

- The MMSE in this case reads

$$\min_{\underline{X}} \frac{1}{2}\left\|\underline{Y}-\underline{X}\right\|_2^2 + \frac{c}{2n}\sum_k \underline{X}^T\mathbf{R}_k^T\left(\mathbf{R}_k\Sigma\mathbf{R}_k^T\right)^{-1}\mathbf{R}_k\underline{X}$$

$$\underline{\hat{X}}_{EPLL} = \left[\mathrm{I}+\frac{c}{n}\sum_k \mathbf{R}_k^T\left(\mathbf{R}_k\Sigma\mathbf{R}_k^T\right)^{-1}\mathbf{R}_k\right]^{-1}\underline{Y}$$

EPLL via Local Processing (ADMM)

- Wait! This looks no simpler than the global Wiener, so where is the benefit in thinking locally?

$$\underline{\hat{X}}_{\text{EPLL}} = \left[I + \frac{c}{n} \sum_k \mathbf{R}_k^T \left(\mathbf{R}_k \Sigma \mathbf{R}_k^T \right)^{-1} \mathbf{R}_k \right]^{-1} \underline{Y}$$

- Answer: ADMM

$$\left\{ \underline{\hat{z}}_k = \left(\mathbf{R}_k \Sigma \mathbf{R}_k^T \right) \left(\mathbf{R}_k \Sigma \mathbf{R}_k^T + \frac{c}{n\lambda} I \right)^{-1} \left(\mathbf{R}_k \underline{\hat{X}} - \underline{\hat{u}}_k \right) \right\}_{k=1,2,\ldots} \quad \text{: A Wiener filter on each patch}$$

$$\underline{\hat{X}} = \left(I + \lambda \sum_k \mathbf{R}_k^T \mathbf{R}_k \right)^{-1} \left(\underline{Y} + \lambda \sum_k \mathbf{R}_k^T \left(\underline{\hat{z}}_k + \underline{\hat{u}}_k \right) \right) \quad \text{: Patch averaging}$$

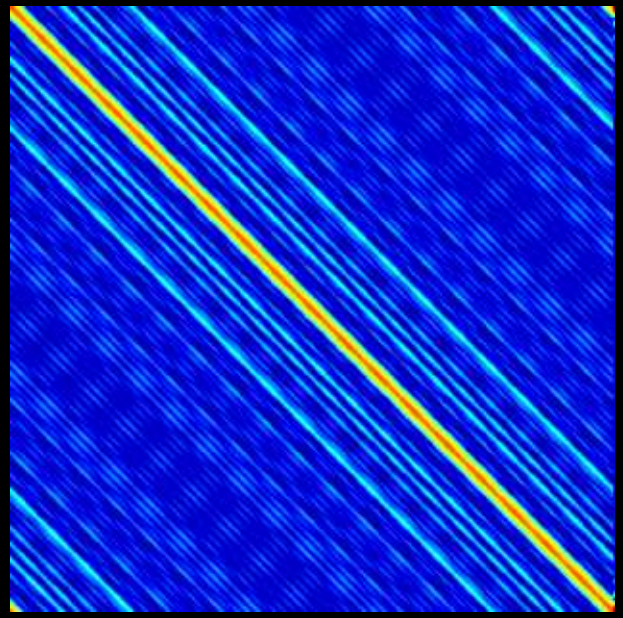
$$\underline{\hat{u}}_k = \underline{\hat{u}}_k - \mathbf{R}_k \underline{\hat{X}} + \underline{\hat{z}}_k$$

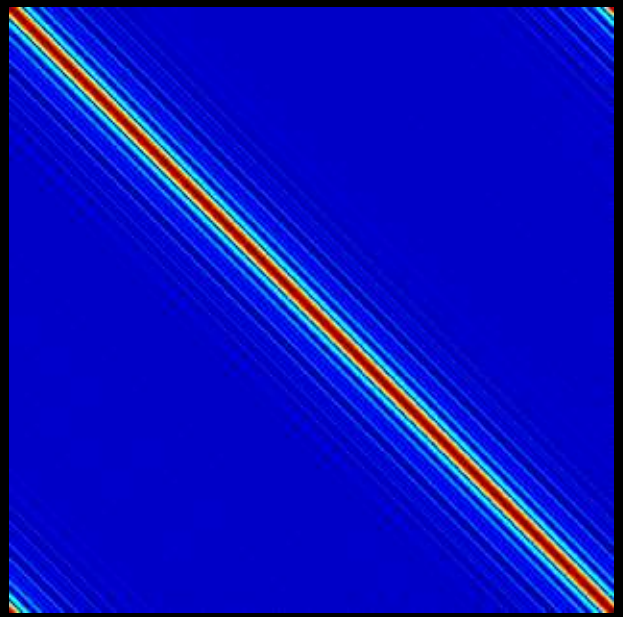
- Notice that the first iteration (with $\underline{u}=\underline{0}$) coincides with LPA.

EPLL – Back to the Example

C=0.25

$$\hat{\underline{X}}_{EPLL} = \left[I + \frac{c}{n} \sum_k \mathbf{R}_k^T \left(\mathbf{R}_k \Sigma \mathbf{R}_k^T \right)^{-1} \mathbf{R}_k \right]^{-1} \underline{Y}$$





$$\hat{\underline{X}}_{Global} = \Sigma \left(\mathbf{I} + \Sigma \right)^{-1} \underline{Y}$$

Going Multi-Scale …

- One could work with fixed-sized patches, related to different scales of the signal. This may enable better proximity to the global filter. *Scale invariance* may prove valuable …

- For example …

$$\min_{\underline{X}} \frac{1}{2}\left\|\underline{Y}-\underline{X}\right\|_2^2 + \frac{c_1}{2n}\sum_k \underline{X}^T\mathbf{R}_k^T\left(\mathbf{R}_k\Sigma\mathbf{R}_k^T\right)^{-1}\mathbf{R}_k\underline{X}$$

- In our notations,

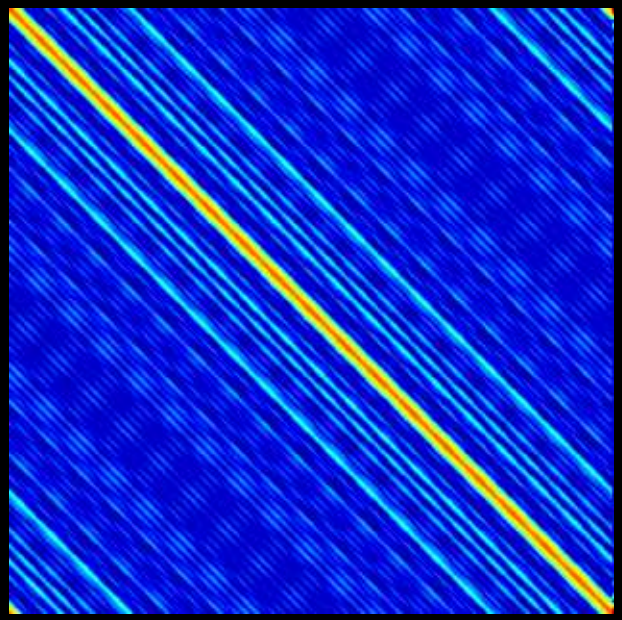
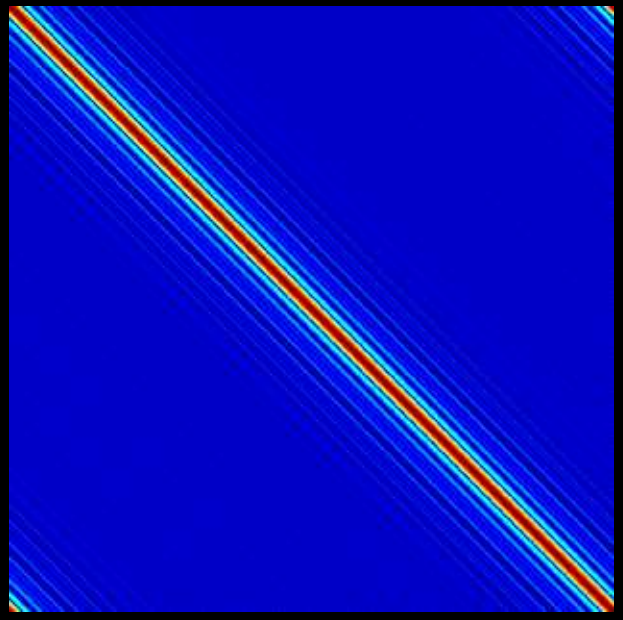
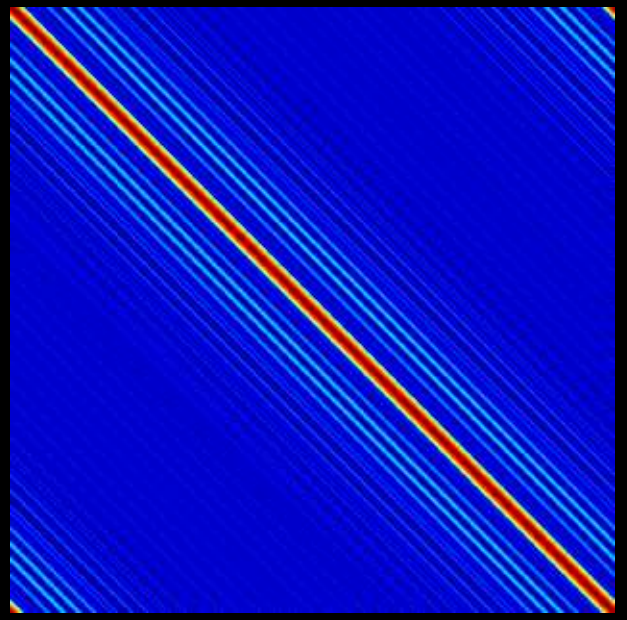
$$+\frac{c_2}{2n}\sum_k \underline{X}^T\mathbf{Q}_k^T\left(\mathbf{Q}_k\Sigma\mathbf{Q}_k^T\right)^{-1}\mathbf{Q}_k\underline{X}$$

  - $\mathbf{R}_k$ extracts a patch of length n from the signal
  - $\mathbf{Q}_k$ extracts a patch of size 2n from the signal and then reduce it to length n by filtering followed by 2:1 decimation

- Naturally, this can be repeated in several scales …

Going Multi-Scale …

$$\hat{\underline{X}}_{\text{Global}} = \Sigma\left(\mathbf{I}+\Sigma\right)^{-1}\underline{Y} \qquad \hat{\underline{X}}_{\text{MS}} = \bullet\ \bullet\ \bullet \ \underline{Y}$$

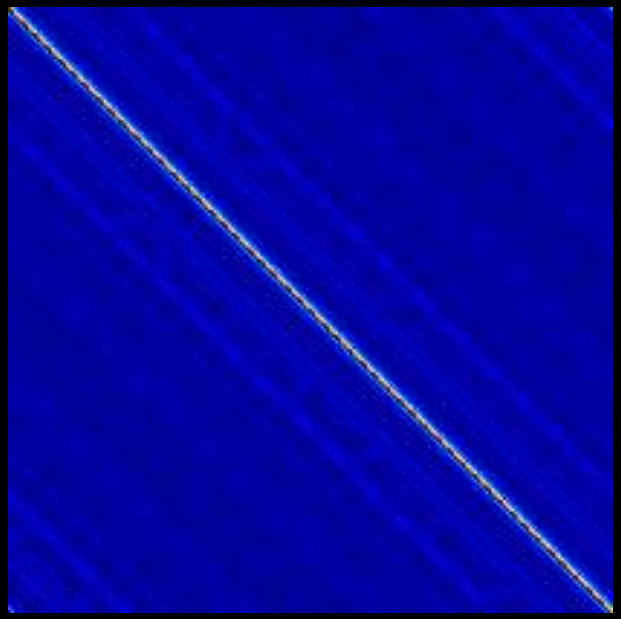




$$\hat{\underline{X}}_{\text{EPLL}} = \left[\mathrm{I}+\frac{\mathrm{c}}{\mathrm{n}}\sum_{\mathrm{k}}\mathbf{R}_{\mathrm{k}}^{\mathrm{T}}\left(\mathbf{R}_{\mathrm{k}}\Sigma\mathbf{R}_{\mathrm{k}}^{\mathrm{T}}\right)^{-1}\mathbf{R}_{\mathrm{k}}\right]^{-1}\underline{Y}$$

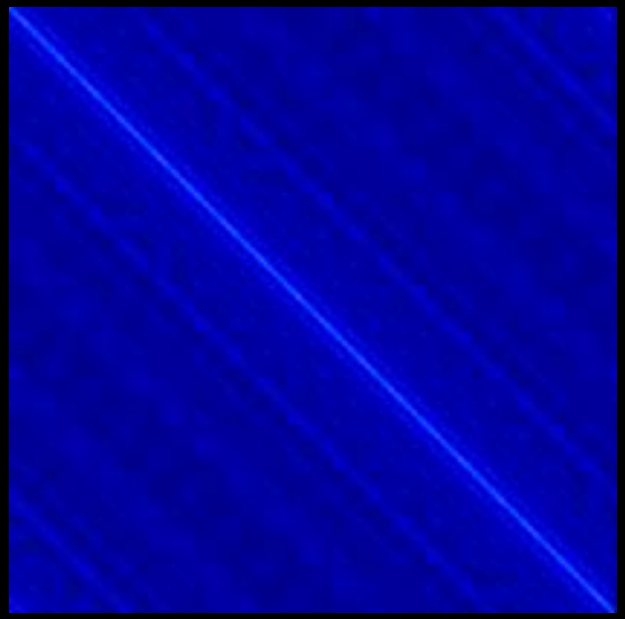
Going Multi-Scale …

These are the absolute error matrices of the filters versus the global Wiener

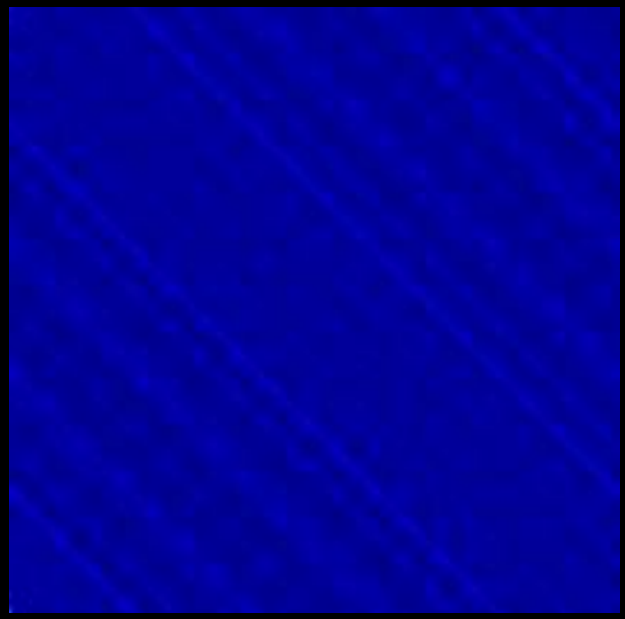
True Scale



$|\mathbf{W}_{Global} - \mathbf{W}_{LPA}|$

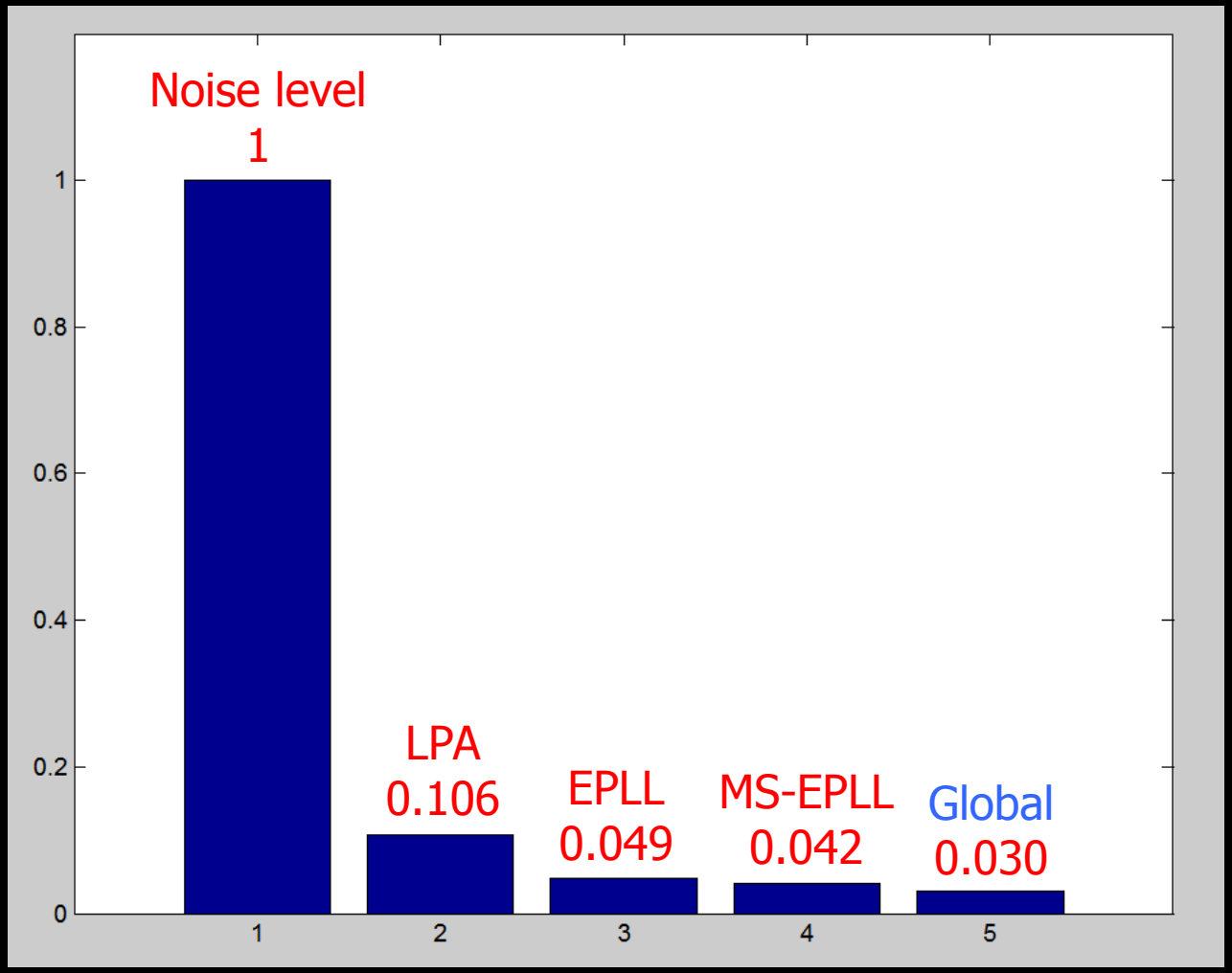


$|\mathbf{W}_{Global} - \mathbf{W}_{EPLL}|$



$|\mathbf{W}_{Global} - \mathbf{W}_{MS}|$

Average Denoising Performance



The Gaussian Case: A Summary

What can be learned from this toy problem?

- The signal $\underline{X}$ has a clear global statistical model, which induces the ultimate denoising approach – the Wiener filtering.
- Nevertheless, we may handle the denoising task by operating on overlapped patches, using only local marginals, and by operating locally we can get to near-ideal performance.
- Under some conditions on Σ, the local and the global methods may lead to the same performance.
- In image processing, identifying/formulating a global model is simply impossible, while learning a local model is within reach.
- We are using local marginals to "characterize" (but not to reconstruct) the global distribution.

Part III

Back to Sparse Representations: Who Are These Signals?

The Model to Explore

The model Assumption on $\underline{X}$:

$$\forall k \quad \mathbf{R}_k \underline{X} = \mathbf{D}\underline{\alpha}_k \quad \text{where} \quad \left\|\underline{\alpha}_k\right\|_0 \leq L$$

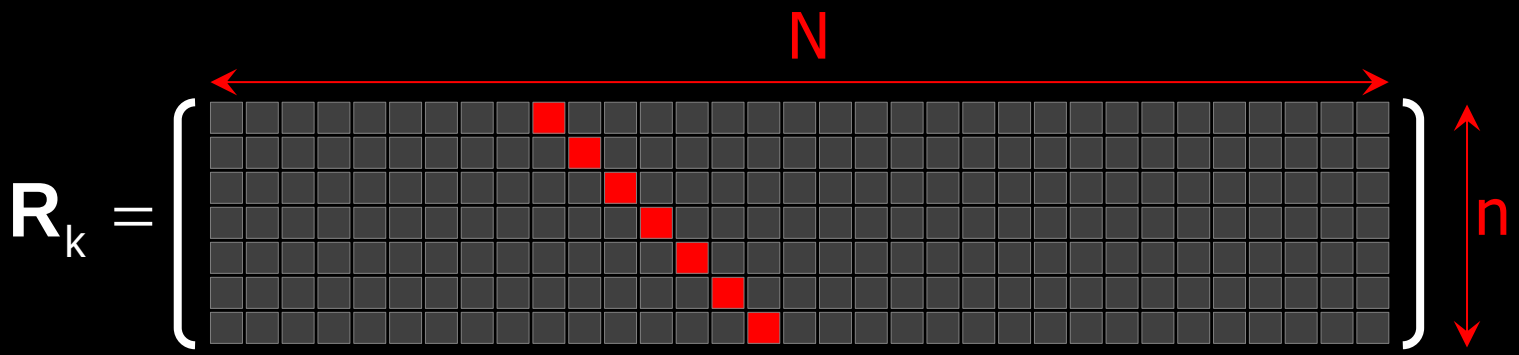
Every patch in the unknown signal is expected to have a sparse representation w.r.t. the dictionary **D**

Questions to consider:

- Who are those signals belonging to this model?
- Under which conditions on **D** would this model be feasible?
- How does one sample from this model?
- How should we perform pursuit properly (& locally) under this model?
- How should we learn **D** if this is indeed the model?
- …

Globalizing the Model (1)

- We start with the model: $\forall k \quad \mathbf{R}_k \underline{X} = \mathbf{D}\underline{\alpha}_k, \quad \|\underline{\alpha}_k\|_0 \le L$
- $\mathbf{R}_k$ extracts a patch of size n from $\underline{X} \in \mathbb{R}^N$:



- Exploiting $\frac{1}{n}\sum_k \mathbf{R}_k^T \mathbf{R}_k = \mathbf{I} \Rightarrow \underline{X} = \frac{1}{n}\sum_k \mathbf{R}_k^T \mathbf{R}_k \underline{X}$

i.e., the global signal is built of averaged local pieces.

Globalizing the Model (2)

$$\underline{X} = \frac{1}{n}\sum_k \mathbf{R}_k^T\mathbf{R}_k\underline{X} = \sum_k \frac{1}{n}\mathbf{R}_k^T\mathbf{D}\underline{\alpha}_k$$

- Defining $\underline{\Gamma} = \begin{bmatrix} \underline{\alpha}_1 \\ \underline{\alpha}_2 \\ \vdots \\ \underline{\alpha}_N \end{bmatrix}$ leads to

$$\begin{bmatrix} | & | & & | \\ \frac{1}{n}\mathbf{R}_1^T\mathbf{D} & \frac{1}{n}\mathbf{R}_2^T\mathbf{D} & \cdots & \frac{1}{n}\mathbf{R}_N^T\mathbf{D} \\ | & | & & | \end{bmatrix}\underline{\Gamma} = \underline{X} = \mathbf{D}_G\underline{\Gamma}$$

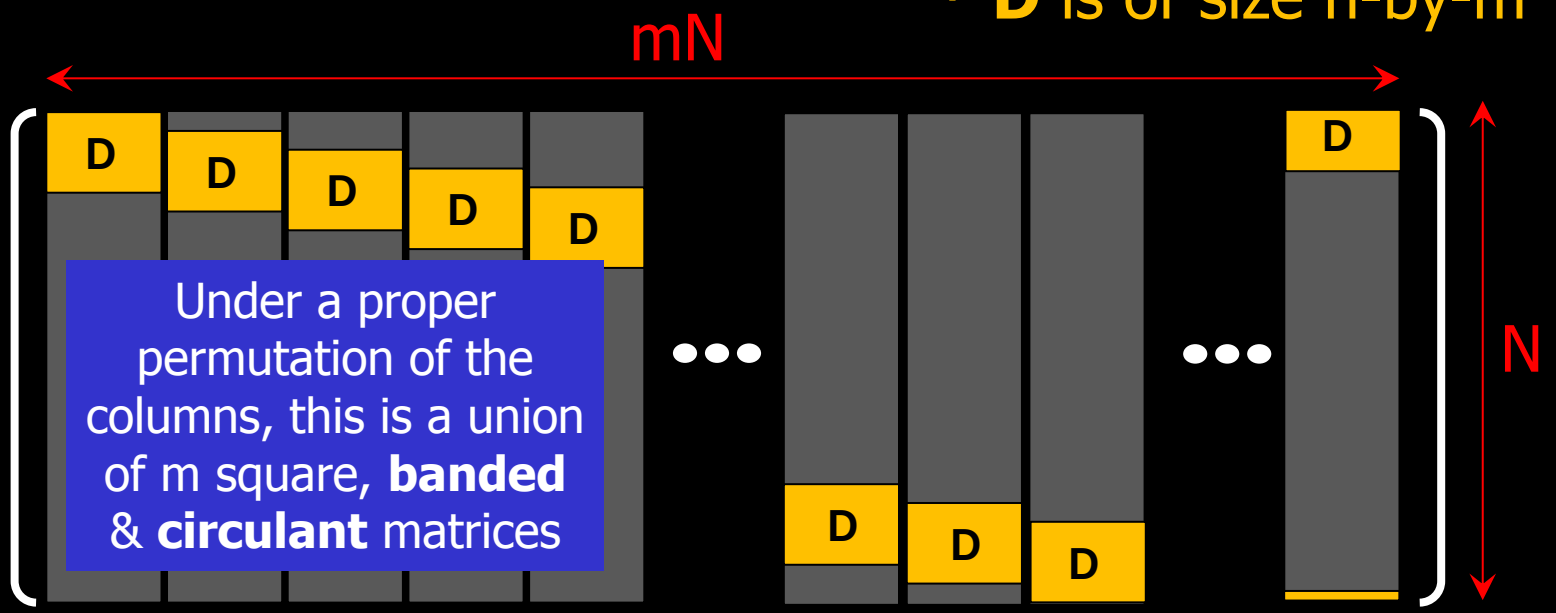
- This suggests the existence of a global sparsity-based model of the kind we are familiar with …

Globalizing the Model (3)

$$\underline{X} = \left[\begin{array}{ccc} | & | & | \\ \frac{1}{n}\mathbf{R}_1^T\mathbf{D} & \frac{1}{n}\mathbf{R}_2^T\mathbf{D} & \cdots \\ | & | & | \end{array} \right. \left. \begin{array}{c} | \\ \frac{1}{n}\mathbf{R}_N^T\mathbf{D} \\ | \end{array} \right] \underline{\Gamma} = \mathbf{D}_G \underline{\Gamma}$$

* **D** is of size n-by-m

$$n\mathbf{D}_G =$$



Globalizing the Model (4)

- However, the vector $\underline{\Gamma}$ is structured …

$$\underline{X} = \sum_k \frac{1}{n} \mathbf{R}_k^T \mathbf{D} \underline{\alpha}_k = \mathbf{D}_G \underline{\Gamma}$$

$$\underline{\Gamma} = \begin{bmatrix} \underline{\alpha}_1 \\ \underline{\alpha}_2 \\ \vdots \\ \underline{\alpha}_N \end{bmatrix}$$

$$\forall j \quad \mathbf{R}_j \underline{X} = \sum_k \frac{1}{n} \mathbf{R}_j \mathbf{R}_k^T \mathbf{D} \underline{\alpha}_k = \mathbf{R}_j \mathbf{D}_G \underline{\Gamma} = \mathbf{D} \underline{\alpha}_j$$

$$= \begin{bmatrix} 0 & \cdots & \mathbf{D} & \cdots & 0 \end{bmatrix} \underline{\Gamma}$$

$$n\mathbf{D}_G = \begin{bmatrix} \mathbf{D} & \mathbf{D} & \mathbf{D} & \mathbf{D} & \mathbf{D} & \bullet\bullet\bullet & & \bullet\bullet\bullet & \mathbf{D} \end{bmatrix}$$

Globalizing the Model (5)

$$\forall j \quad \mathbf{R}_j \mathbf{D}_G \underline{\Gamma} = \begin{bmatrix} 0 & \cdots & \mathbf{D} & \cdots & 0 \end{bmatrix} \underline{\Gamma}$$

$$\underline{\Gamma} = \begin{bmatrix} \underline{\alpha}_1 \\ \underline{\alpha}_2 \\ \vdots \\ \underline{\alpha}_N \end{bmatrix}$$

- This defines N sets of n equations each, which can be expressed as …

$$\mathbf{M}\underline{\Gamma} = \underline{0}$$

where **M** is of size nN×mN

- **M** is some sort of "Laplacian" in nature, capturing the fact that the averaged patches over the overlaps should have a complete agreement on the content.

Globalizing the Model (6)

- Another limitation on $\underline{\Gamma}$ is the fact that $\forall k \left\| \underline{\alpha}_k \right\|_0 \leq L$, which we will denote as

$$\underline{\Gamma} = \begin{bmatrix} \underline{\alpha}_1 \\ \underline{\alpha}_2 \\ \vdots \\ \underline{\alpha}_N \end{bmatrix}$$

$$\left\| \underline{\Gamma} \right\|_{0,\infty} = L$$

- A word about our notation: this definition works on blocks of m elements. Later on we will see a similar definition with a sliding window …

$$\left\| \underline{\Gamma} \right\|_{0,\infty}^{m} = L \iff \left\| \underline{\Gamma} \right\|_{0,\infty}^{m-\text{Sliding}} = L$$

Globalizing the Model – Summary

We started with a local model

$$\left\{ \begin{array}{c} \mathbf{R}_k \underline{X} = \mathbf{D}\underline{\alpha}_k \\ \left\| \underline{\alpha}_k \right\|_0 \le L \end{array} \right\}_k$$

$\mathbf{D}, L, n$ → ..... → $\mathbf{D}_G, \mathbf{M}$

We ended up with an equivalent globalized model

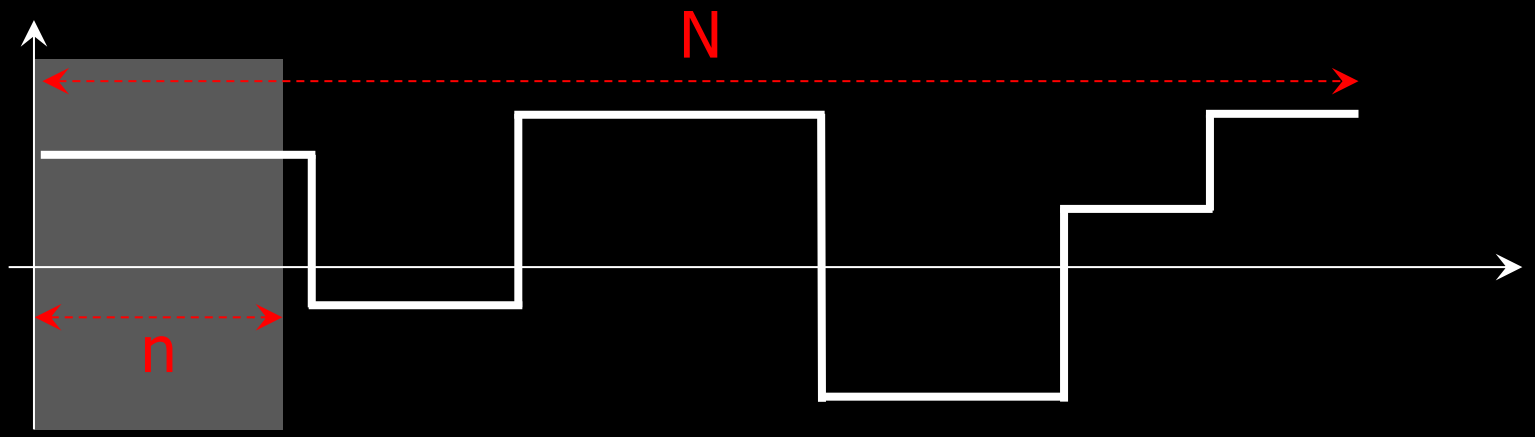
$$\left\{ \begin{array}{c} \underline{X} = \mathbf{D}_G \underline{\Gamma} \\ \mathbf{M}\underline{\Gamma} = \underline{0}, \left\| \underline{\Gamma} \right\|_{0,\infty} = L \end{array} \right\}$$

- However, for many choices of $\mathbf{D}$, $\left\{ \underline{\Gamma} \,\middle|\, \mathbf{M}\underline{\Gamma} = \underline{0},\ \left\| \underline{\Gamma} \right\|_{0,\infty} = L \right\}$ is empty for any $0<L<n$

  → implying a useless model !

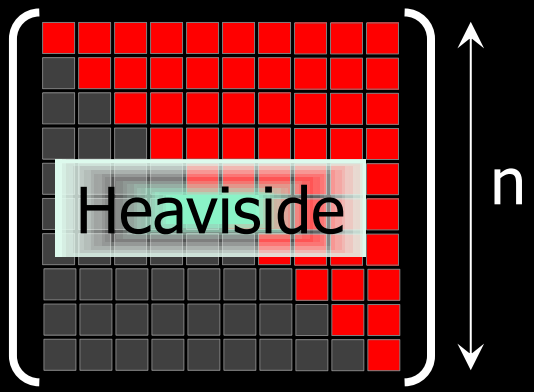
- Even if this model is feasible, how can we sample from it?
- All this might seem hopeless, but … do not despair …

Special Case 1: PWC Signals



- Piece-Wise-Constant signals obey the local model: If every patch of length n contains (up to) L-1 steps, then these patches can be described as

$$\forall k \quad \mathbf{R}_k \underline{X} = \mathbf{D}\underline{\alpha}_k, \quad \left\|\underline{\alpha}_k\right\|_0 \le L, \qquad \mathbf{D} =$$



Special Case 1: PWC Signals

$$\underline{X} = \mathbf{D}_G \underline{\Gamma}, \quad \mathbf{M}\underline{\Gamma} = \underline{0}, \quad \left\|\underline{\Gamma}\right\|_{0,\infty} = L$$

- In the PWC case, **M** has a null space of dimension N. We denote this null-space by **Z** (of size nN×N),

$$\mathbf{Z}\underline{\Phi} = \underline{\Gamma}$$

- A proper choice for describing **Z** leads to the regular global sparse model with the

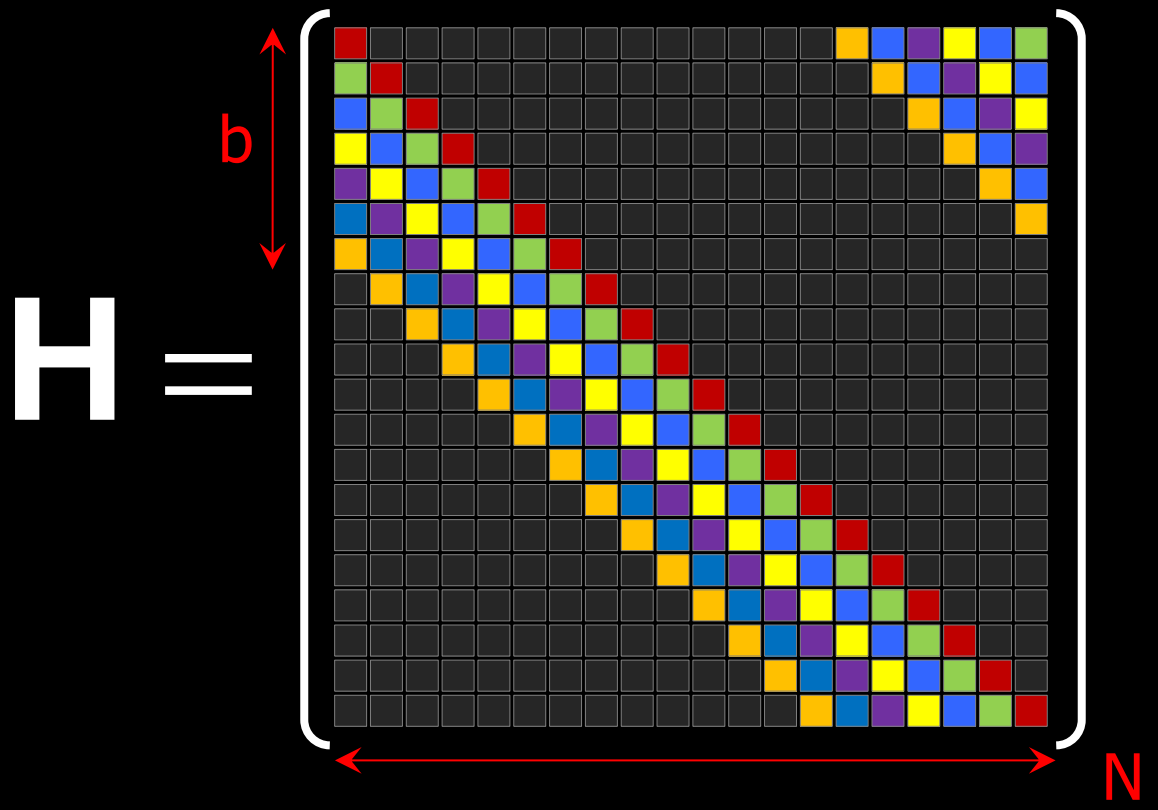
$$\underline{X} = \mathbf{D}_G \underline{\Gamma} = \mathbf{D}_G \mathbf{Z}\, \underline{\Phi}$$

Global Heaviside dictionary

- The representation $\underline{\Phi}$ satisfies $\left\|\underline{\Phi}\right\|_{0,\infty}^{n-\text{Sliding}} \le L - 1$

Special Case 2: Convolutional Sparsity

- Consider a global sparsity-based model for signals $\underline{X}=\mathbf{H}\underline{\Phi}$, in which the dictionary is **circulant** and **banded** in the following way :



[Lewicki & Sejnowski '99]

[Hashimoto & Kurata, '00]

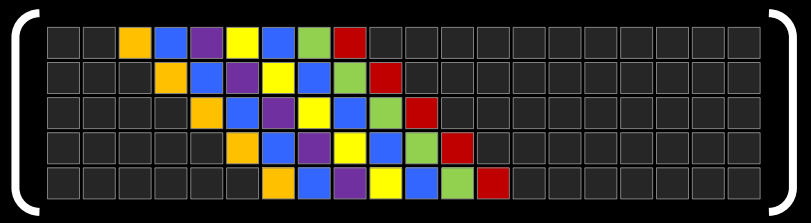
...

[Wohlberg '15]

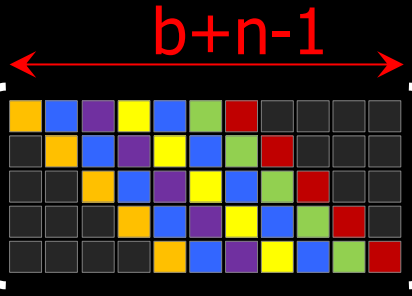
Special Case 2: Convolutional Sparsity

- In this case, every patch of length $n \leq b$ will have a sparse representation w.r.t. the same local dictionary:

$$\underline{x}_k = \mathbf{R}_k \mathbf{H} \underline{\Phi} = \left(\quad \right) \underline{\Phi}$$



$$= \mathbf{R}_k \mathbf{H} \mathbf{S}_k^T \mathbf{S}_k \underline{\Phi} = \left(\quad \right) \underline{\alpha}_k$$



These small representation vectors are overlapping, as they are extracted from the global rep. vector $\underline{\Phi}$

- This suggests a family of local dictionaries that enable our local-global model.

Special Case 2: Convolutional Sparsity

$$\underline{X} = \mathbf{D}_G \underline{\Gamma}, \quad \mathbf{M}\underline{\Gamma} = \underline{0}, \quad \left\| \underline{\Gamma} \right\|_{0,\infty} = L$$

- In terms of the above, we get that **M** has a null space of dimension N, and the situation is similar to the one encountered in the PWC case.
- The same holds true for a global signal emerging from a sparsity-based model with a union of circulant and banded matrices

$$\underline{X} = \sum_j \mathbf{H}_j \underline{\Phi}_j$$

we will come back to this later on ...

Our Global-Local Model: A Summary

Local Model: $\forall k \quad \mathbf{R}_k \underline{X} = \mathbf{D}\underline{\alpha}_k, \quad \|\underline{\alpha}_k\|_0 \le L$

Globalized Model: $\underline{X} = \mathbf{D}_G \underline{\Gamma}, \quad \mathbf{M}\underline{\Gamma} = \underline{0}, \quad \|\underline{\Gamma}\|_{0,\infty}^{n} = L$

We have identified special cases of local dictionaries **D** that enable this model.
Furthermore, in these cases we can easily synthesize signals.

Our Global-Local Model: A Summary

Clarification:

The migration from the local to the globalized model is not constructive. So why bother? Because this will serve us later on when we define the pursuit algorithms.

Open Questions :

- ❑ Are there other local dictionaries **D** that could enable the local-global model? We do know of other cases …
- ❑ Is it guaranteed that any such permissible **D** will have a global model?

Part IV

Pursuit & Denoising for These Signals

The Model to Explore

The model Assumption on $\underline{X}$:

$$\forall k \quad \mathbf{R}_k \underline{X} = \mathbf{D}\underline{\alpha}_k \quad \text{where} \quad \left\|\underline{\alpha}_k\right\|_0 \leq L$$

Every patch in the unknown signal is expected to have a sparse representation w.r.t. the dictionary **D**

Questions to consider:

- Who are those signals belonging to this model?
- Under which conditions on **D** would this model be feasible?
- How does one sample from this model?
- How should we perform pursuit properly (& locally) under this model?
- How should we learn **D** if this is indeed the model?
- …

This Brings us Back to Denoising

- The signal $\underline{X}$ obeys our local model

$$\forall k \quad \mathbf{R}_k \underline{X} = \mathbf{D}\underline{\alpha}_k \quad \text{where} \quad \left\|\underline{\alpha}_k\right\|_0 \le L$$

or the globalized one,

$$\underline{X} = \mathbf{D}_G \underline{\Gamma}, \quad \mathbf{M}\underline{\Gamma} = \underline{0}, \quad \left\|\underline{\Gamma}\right\|_{0,\infty} = L$$

- Given $\underline{Y}=\underline{X}+\underline{V}$, ($\sim N\left\{\underline{0}, \sigma^2 I\right\}$), our goal is to recover $\underline{X}$.
- The specific questions we aim to address are:
  - How would the oracle perform in this case?
  - How can we perform this denoising using local pursuit and how well will this work?

The Pursuit Goal

The signal is believed to belong to the following model

$$\underline{X} = \mathbf{D}_G \underline{\Gamma}, \quad \mathbf{M}\underline{\Gamma} = \underline{0}, \quad \left\|\underline{\Gamma}\right\|_{0,\infty} = L$$

Pursuit is simply a projection onto the model

$$\min_{\underline{\Gamma}} \left\|\mathbf{D}_G \underline{\Gamma} - \underline{Y}\right\|_2^2 \quad \text{s.t.} \quad \mathbf{M}\underline{\Gamma} = \underline{0}, \quad \left\|\underline{\Gamma}\right\|_{0,\infty} = L$$

- This is very similar to the regular pursuit we are accustomed to, and it is just as complex (NP-Hard).
- Remember: while we write the pursuit in terms of $\mathbf{D}_G$ and $\mathbf{M}$, our true goal is to break this into local pursuit steps while being equivalent to this formulation.

The Oracle (1)

- The oracle knows the locations of the $|S| \leq LN$ non-zeros in $\underline{\Gamma}$. The operator extracting these non-zeros is $\mathbf{P}$ (size $|S| \times mN$). Thus,

$$\mathbf{P}^T\mathbf{P}\underline{\Gamma} = \mathbf{P}^T\underline{\Gamma}_s = \underline{\Gamma}$$

- Our pursuit task simplifies to the search of the optimal $\underline{\Gamma}_s$:

$$\min_{\underline{\Gamma}_s} \left\| \mathbf{D}_G\mathbf{P}^T\underline{\Gamma}_s - \underline{Y} \right\|_2^2 \quad \text{s.t.} \quad \mathbf{M}\mathbf{P}^T\underline{\Gamma}_s = \underline{0}$$

- Denoting by $\mathbf{B}$ (size $nN \times d$ ($d \leq |S|$)), the null-space of $\mathbf{M}\mathbf{P}^T$, we have $\underline{\Gamma}_s = \mathbf{B}\underline{\theta}$, and the pursuit becomes

$$\min_{\underline{\theta}} \left\| \mathbf{D}_G\mathbf{P}^T\mathbf{B}\underline{\theta} - \underline{Y} \right\|_2^2$$

The Oracle (2)

$$\min_{\underline{\theta}} \left\| \mathbf{D}_G\mathbf{P}^T\mathbf{B}\,\underline{\theta} - \underline{Y} \right\|_2^2$$

- This matrix is of size N×d, and it is full-rank.
- … We started with a noise power of $N\sigma^2$ in the signal $\underline{Y}$, and the oracle ends with noise energy of $d\sigma^2$.
- Special cases of interest:
  - In the PWC case, d=number of flat regions, while $|S| \approx d \cdot n$.
  - In the convolutional sparsity case (with one circulant matrix),

$$d = \left\| \underline{\Phi} \right\|_0 , \quad |S| = d \cdot (2n-1)$$

Oracle via LPA

- Under the regime of an oracle (the supports of $\underline{\alpha}_k$ are known), lets apply the Local Patch Averaging (LPA):

$$\min_{\underline{X}} \sum_k \left\| \mathbf{R}_k \underline{X} - \underbrace{\mathbf{D}_{S_k} \mathbf{D}_{S_k}^{\dagger} \mathbf{R}_k \underline{Y}}_{\hat{\underline{x}}_k} \right\|_2^2 =$$

$$\Rightarrow \quad \underline{X} = \frac{1}{n} \sum_k \mathbf{R}_k^T \hat{\underline{x}}_k = \frac{1}{n} \sum_k \mathbf{R}_k^T \mathbf{D}_{S_k} \mathbf{D}_{S_k}^{\dagger} \mathbf{R}_k \underline{Y}$$

Theorem: Applying the oracle's LPA algorithm iteratively leads to the optimal solution.

Oracle via EPLL

- Under the regime of an oracle (the supports of $\underline{\alpha}_k$ are known), lets apply the EPLL:

$$\min_{\underline{X}} \frac{1}{2}\|\underline{Y}-\underline{X}\|_2^2 + \frac{\lambda}{2}\sum_k \left\| \underbrace{\left(\mathbf{I}-\mathbf{D}_{S_k}\mathbf{D}_{S_k}^{\dagger}\right)}_{\mathbf{P}_k} \mathbf{R}_k \underline{X} \right\|_2^2$$

$$\Rightarrow \quad \underline{\hat{X}}_{\text{EPLL}} = \left[\mathrm{I}+\lambda\sum_k \mathbf{R}_k^T\mathbf{P}_k^T\mathbf{P}_k\mathbf{R}_k\right]^{-1}\underline{Y}$$

Theorem: Applying the oracle's EPLL **iteratively** with arbitrary $\lambda>0$ leads to the optimal solution.

The Oracle: A Summary

- As expected, the "complexity" of the signal (the value d) governs the denoising performance.
- The above two results regarding LPA and EPLL are misleading, as they may suggest that it is enough to apply local pursuit (projection onto an appropriate subspace) and average in order to lead to the ideal solution.
- While this is indeed the case if the supports are known, as we depart from the oracle regime, we will see that the local pursuits must "communicate" in order to lead to a more successful global recovery.

Patch-Based Pursuit

Recall our pursuit talks:

$$\min_{\underline{\Gamma}} \left\| \mathbf{D}_G \underline{\Gamma} - \underline{Y} \right\|_2^2 \quad \text{s.t.} \quad \mathbf{M}\underline{\Gamma} = 0, \ \left\| \underline{\Gamma} \right\|_{0,\infty} = L$$

Theorem: The above g
task is equiva

$$\min_{\{\underline{\alpha}_k\}_k} \sum_k \left\| \underline{\alpha}_k \right\|_0 \quad \text{s.t.} \quad \underline{\alpha}_k = \underline{\alpha}_{k+1}, \ \left\| \underline{\alpha}_k \right\|_0 \le L \Big\}_k$$

Can be solved by local OMP via the use of ADMM !

This is the missing communication force

$\mathbf{S}_B$ – Extract the bottom n-1 elements
$\mathbf{S}_T$ – Extract the top n-1 elements

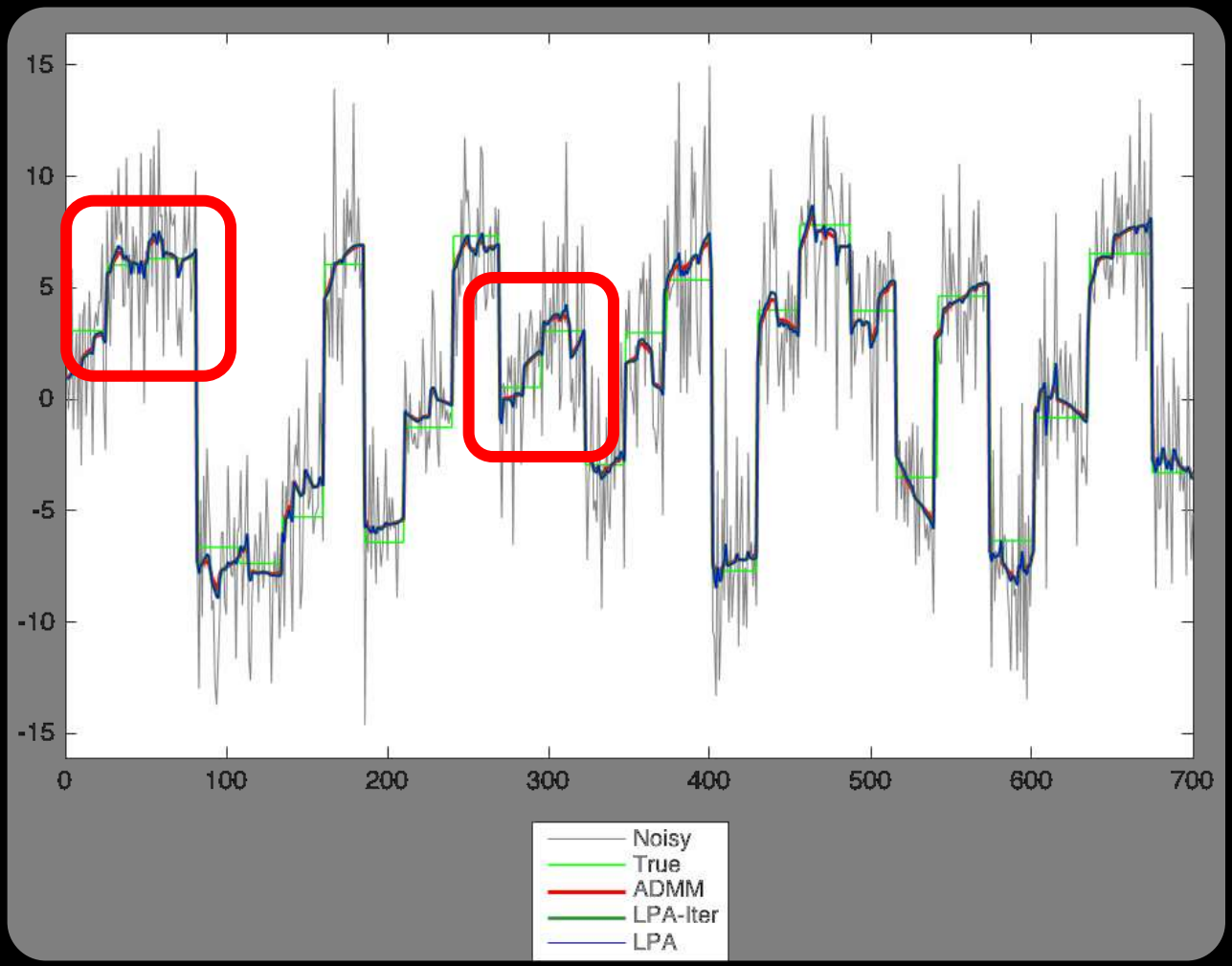
Patch-Based Pursuit - Demo

Details:

- Signal length: N = 700
- Patch size: n = 25
- Dictionary = PWC
- Global sparsity: k = 25
- Noise with $\sigma = 3$

Results:

- Noise (per sample): σ^2
- ADMM – $0.078\sigma^2$
- LPA – $0.106\sigma^2$
- Iter-LPA - $0.113\sigma^2$



Part V

A Closer Look at the Convolutional Sparsity (The Noiseless Case)

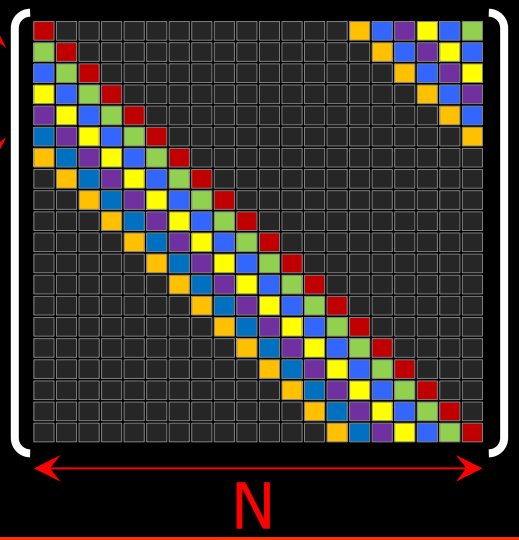
The Convolutional Sparsity Case

- Let us return to the special case of convolutional sparsity.
- We will describe this not via the ~~local~~ model or its ~~globalized~~ form, but rather through the global model.
- Consider a global sparsity-based model of the form

$$\underline{X} = \sum_{j=1}^{p} \mathbf{H}_j \underline{\Phi}_j$$

where $\mathbf{H}_j$ are banded and circulant.

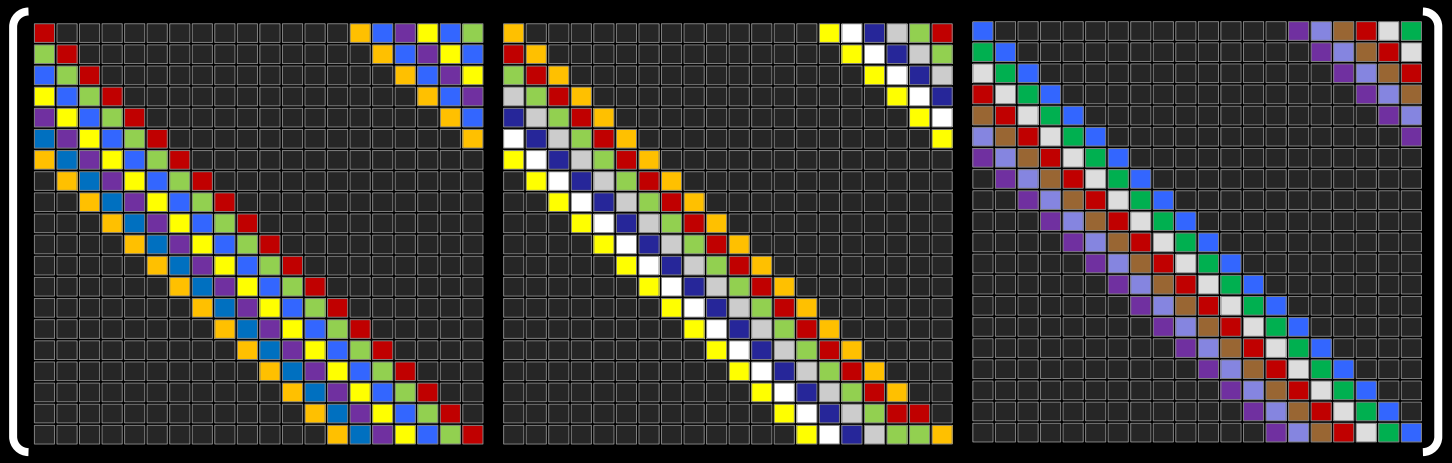
$\mathbf{H}_j =$



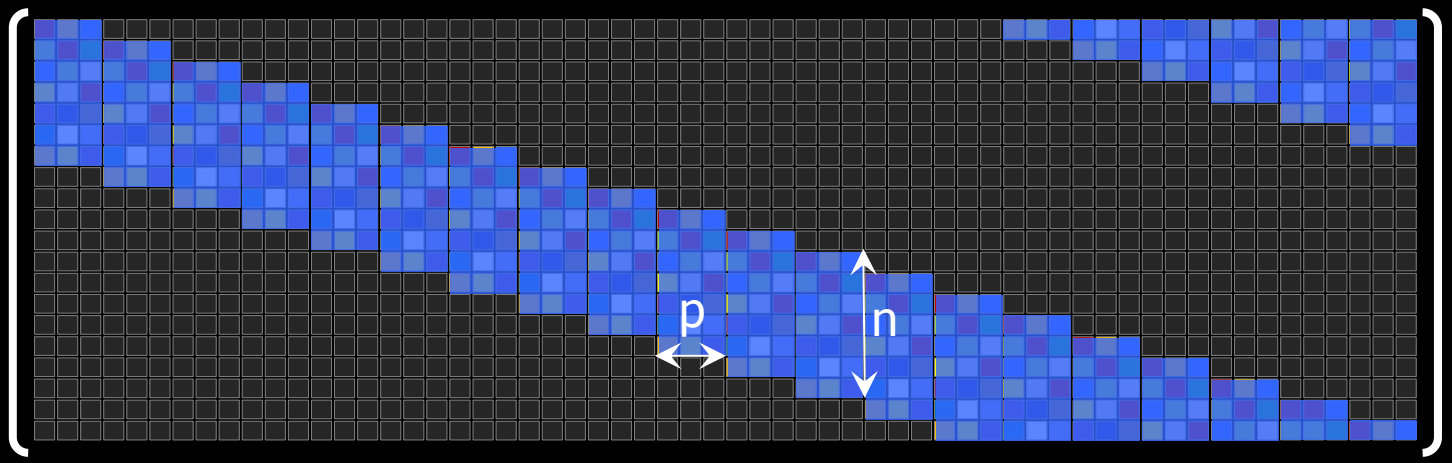
- For simplicity we assume that b=n.

The Global Dictionary

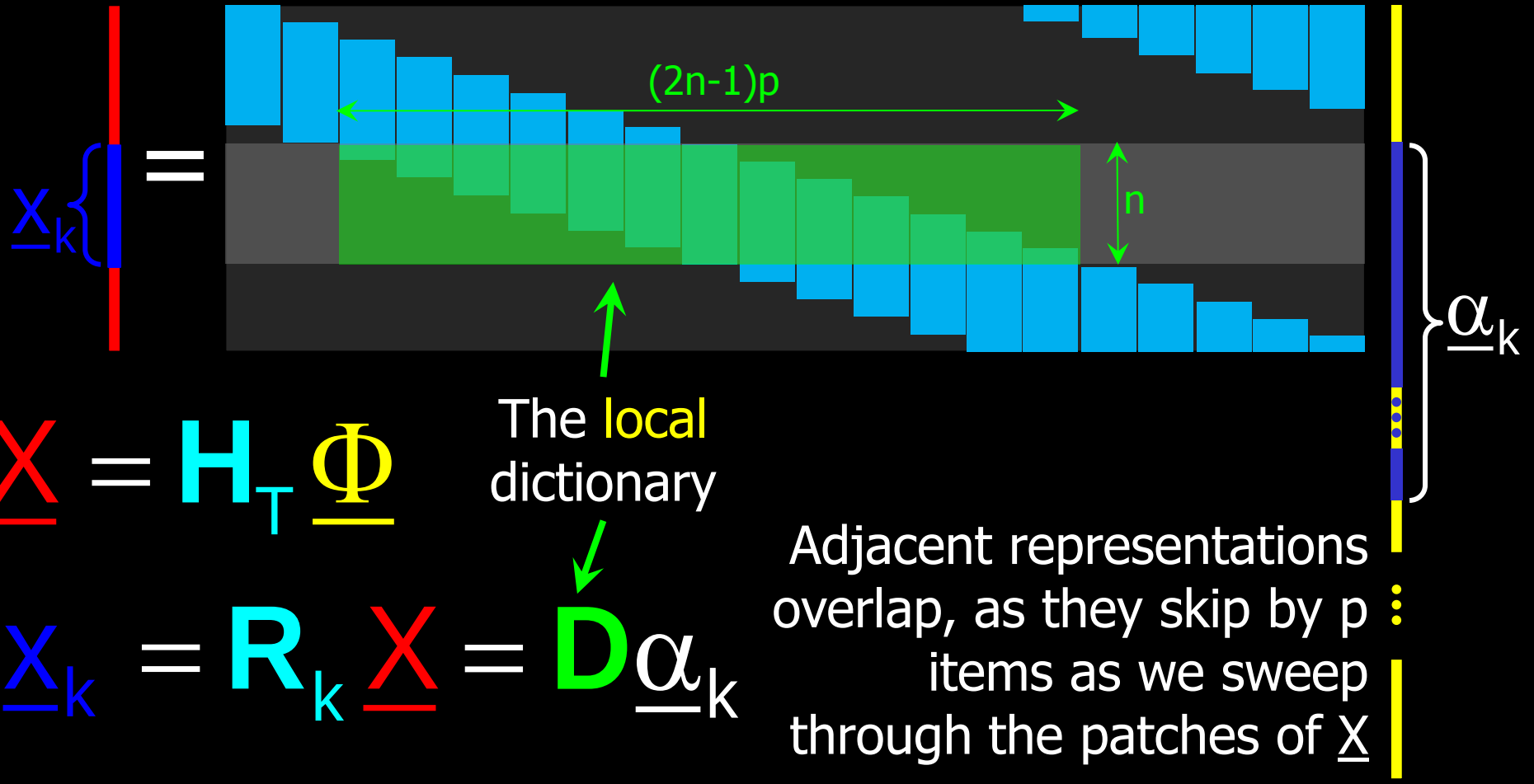
$$[H_1\ H_2\ H_3] =$$



$$H_T =$$



The Local Representations $\underline{\alpha}_k$



$$\underline{X} = \mathbf{H}_T \underline{\Phi}$$

$$\underline{x}_k = \mathbf{R}_k \underline{X} = \mathbf{D} \underline{\alpha}_k$$

Adjacent representations overlap, as they skip by p items as we sweep through the patches of $\underline{X}$

The Representation Problem

$$\left(\mathbf{P}_{0,\infty}\right) \quad \min_{\underline{\Phi}} \|\underline{\Phi}\|_{0,\infty} \quad \text{s.t.} \quad \underline{X} = \mathbf{H}_T \underline{\Phi}$$

where we have defined $\|\underline{\Phi}\|_{0,\infty} = \max_k \|\underline{\alpha}_k\|_0$

The Main Questions We Aim to Address:

- Uniqueness of the solution to this problem ?
- Guaranteed Recovery of the solution via global OMP/BP ?
- The same recovery done via local operations ?

Stripe Spark and Uniqueness

We should be excited about this result and later results because they pose a local constraint for a global guarantee, and as such, they are far more optimistic compared to the comparable global guarantees

Theorem: If a solution is found for ($\mathbf{P}_{0,\infty}$) such that

$$\|\underline{\Phi}\|_{0,\infty} < \frac{1}{2}\sigma_S(\mathbf{H}_T)$$

Then this is necessarily the globally optimal solution to this problem.

Stripe Spark vs. Mutual Coherence

$$\mu(\mathbf{H}_T) \triangleq \max_{i \neq j} \left| \mathbf{H}_T{}^T \mathbf{H}_T \right|$$

This is the classic coherence, defined over the global dictionary, assuming normalized columns

Theorem: The relation between the stripe spark and the global coherence is

$$\sigma_S(\mathbf{H}_T) \geq 1 + \frac{1}{\mu(\mathbf{H}_T)}$$

Thus uniqueness of the solution is guaranteed if

$$\|\underline{\Phi}\|_{0,\infty} < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{H}_T)}\right)$$

Recovery Guarantees

Lets solve this problem via OMP or BP, applied globally

$$\left(\mathbf{P}_{0,\infty}\right) \quad \min_{\underline{\Phi}} \left\|\underline{\Phi}\right\|_{0,\infty} \quad \text{s.t.} \quad \underline{X} = \mathbf{H}_T \underline{\Phi}$$

Theorem: If a solution of $(\mathbf{P}_{0,\infty})$ satisfies

$$\left\|\underline{\Phi}\right\|_{0,\infty} < \frac{1}{2}\left(1 + \frac{1}{\mu\left(\mathbf{H}_T\right)}\right)$$

Then global OMP and BP are guaranteed to find it.

Local Recovery Guarantee via ADMM

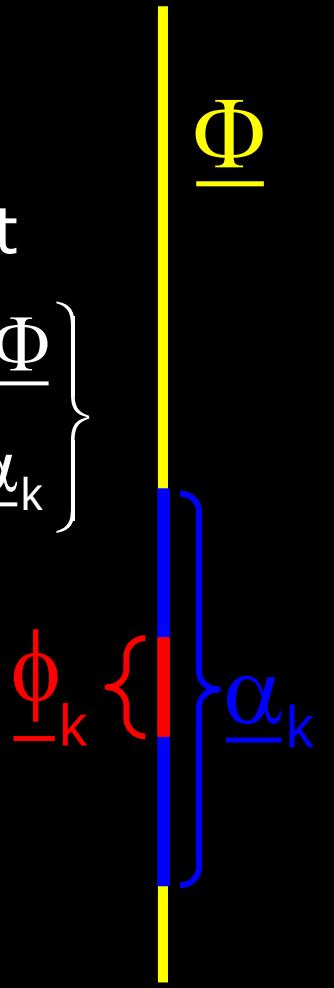
$$\left(\mathbf{P}_1^{\varepsilon}\right) \quad \min_{\underline{\Phi}} \lambda \left\|\underline{\Phi}\right\|_1 + \frac{1}{2}\left\|\underline{X} - \mathbf{H}_T \underline{\Phi}\right\|_2^2$$

.... can be converted to the following equivalent format

$$\min_{\underline{\Phi},\{\underline{\phi}_k\},\{\underline{\alpha}_k\}} \sum_k \lambda \left\|\underline{\phi}_k\right\|_1 + \frac{1}{2}\left\|\mathbf{R}_k \underline{X} - \mathbf{D}\underline{\alpha}_k\right\|_2^2 \quad \text{s.t.} \quad \left\{\begin{array}{l} \underline{\alpha}_k = \mathbf{Q}_k \underline{\Phi} \\ \underline{\phi}_k = \mathbf{S}_k \underline{\alpha}_k \end{array}\right\}$$

which can be solved iteratively by

- Updating $\underline{\phi}_k$ via simple soft shrinkage.
- Updating $\underline{\alpha}_k$ via simple multiplication by a matrix, and
- Updating $\underline{\Phi}$ via patch-averaging.



Local Recovery Guarantee via ADMM

$$\left(\mathbf{P}_1^{\varepsilon}\right) \quad \min_{\underline{\Phi}} \lambda \left\|\underline{\Phi}\right\|_1 + \frac{1}{2}\left\|\underline{X} - \mathbf{H}_T \underline{\Phi}\right\|_2^2$$

Theorem: If the solution of $(\mathbf{P}_{0,\infty})$ satisfies

$$\left\|\underline{\Phi}\right\|_{0,\infty} < \frac{1}{2}\left(1 + \frac{1}{\mu\left(\mathbf{H}_T\right)}\right)$$

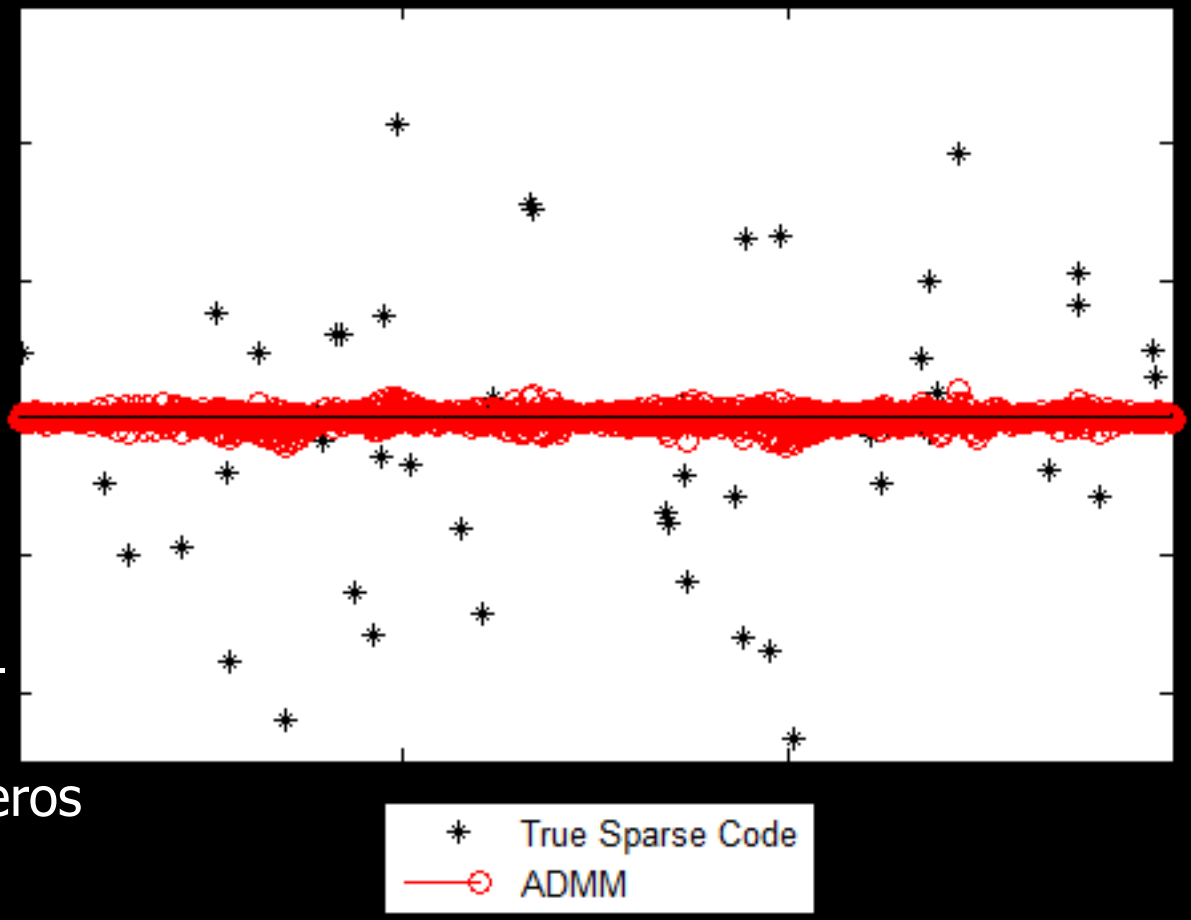
Then the ADMM solver of the global BP, with $\lambda \to 0$, is guaranteed to find it.

Local Pursuit in Action

Details:

- Signal length: N = 300
- Patch size: n = 25
- Dictionary = Convolutional
- Unique atoms: p = 5
- Global sparsity: k = 50
- Number Iterations: 1000
- $\lambda = 1/50$

The graph shows $\underline{\Phi}$ - a vector of length 1500, with 50 non-zeros



Part VI

Concluding Remarks

Local Model for a Global Signal

- This work suggests an interesting extension of sparse approximation theory to these new breed of models.
- Key questions to address:
  - Who are those signals obeying this model ?
  - Who are the appropriate local dictionaries ?
  - How should these signals be processed via local operations ?
  - Can we derive performance bounds for such algorithms ?
  - How should the dictionary be learned ?
- As we have seen today, some of these questions were answered. Much work remains in order to fully map this field …
- Stay tuned – we are working on this.

Thank You for Your Time & Patience

… and …

Thanks to The Organizers:

Holger Boche - Technical University Munich

Giuseppe Caire - Technical University Berlin

Robert Calderbank - Duke University

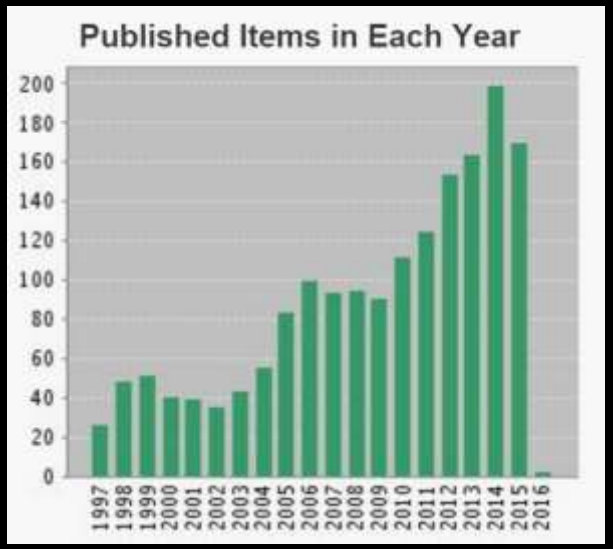
Gitta Kutyniok - Technical University Berlin

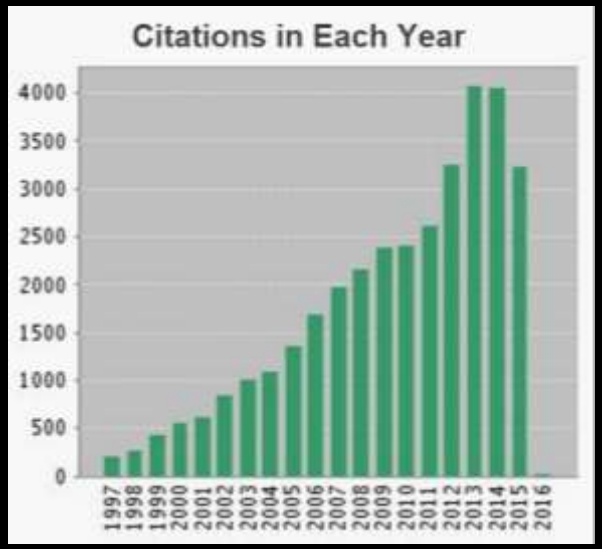
Rudolf Mathar - RWTH Aachen University

Questions?

Image Denoising is a Popular Problem

Probably the most studied problem in image processing …





- There are ~22,000 journal papers on image denoising.
- Searching "`image and Gaussian and noise and (denois* or remov* or filter* or clean)`" in ISI WoS leads to ~1800 papers

Why is it so popular? Here are few possible explanations:

(i) It does come up in many applications
(ii) It is the simplest inverse problem, platform for new ideas
(iii) Many other problems can be recast as an iterated denoising, and …
(iv) It is misleadingly simple

Restricted Isometry Property …

$$\min_{\underline{\Gamma}} \left\|\mathbf{D}_G\underline{\Gamma}-\underline{Y}\right\|_2^2 \quad \text{s.t.} \quad \mathbf{M}\underline{\Gamma}=\underline{0},\ \left\|\underline{\Gamma}\right\|_{0,\infty}=L$$

Definition: The globalized model, as characterized by $\mathbf{D}_G$, $\mathbf{M}$ and k, is said to have the generalized RIP with δ_k if:

$$(1-\delta_k)\left\|\underline{V}\right\|_2^2 \le \left\|\mathbf{D}_G\underline{V}\right\|_2^2 \le (1+\delta_k)\left\|\underline{V}\right\|_2^2$$

for any vector obeying $\mathbf{M}\underline{V}=\underline{0},\ \left\|\underline{V}\right\|_{0,\infty}=k$

Armed with this definition, we can derive a stability result for the above pursuit problem, show its near-oracle performance, derive a similar result for L_1 replacing the $L_{0,\infty}$, and then via ADMM, show that this can be achieved by local operations