Practical compressive sampling with frames

Deanna Needell

CLAREMONT McKENNA COLLEGE

Jan. 2015

2015 Joint Math Meetings

Outline

- Part I : Compressed sensing with frames
  - Mathematical Formulation & Methods
  - Practical CS
    - Other notions of sparsity
    - The need for frames
    - Algorithmic results and challenges
    - New theoretical results for recovery with frames
- Part II (G. Chen) : Dictionary Learning
  - Background and description
  - Applications
  - Existing approaches (K-SVD, GMRA)
  - Summary

The mathematical problem

1. Signal of interest $f \in \mathbb{C}^d (= \mathbb{C}^{N \times N})$

2. Measurement operator $\mathscr{A} : \mathbb{C}^d \to \mathbb{C}^m$ $(m \ll d)$

3. Measurements $y = \mathscr{A} f + \xi$

$$\begin{bmatrix} \\ y \\ \\ \end{bmatrix} = \begin{bmatrix} & & \\ & \mathscr{A} & \\ & & \end{bmatrix} \begin{bmatrix} f \\ \\ \\ \\ \end{bmatrix} + \begin{bmatrix} \\ \xi \\ \\ \end{bmatrix}$$

4. *Problem:* Reconstruct signal f from measurements y

Sparsity

Measurements $y = \mathscr{A} f + \xi$.

$$\begin{bmatrix} \\ y \\ \\ \end{bmatrix} = \begin{bmatrix} & & \\ & \mathscr{A} & \\ & & \end{bmatrix} \begin{bmatrix} \\ f \\ \\ \\ \\ \end{bmatrix} + \begin{bmatrix} \\ \xi \\ \\ \end{bmatrix}$$

Assume f is *sparse*:

- ✧ In the coordinate basis: $\|f\|_0 \stackrel{\text{def}}{=} |\mathrm{supp}(f)| \leq s \ll d$
- ✧ In orthonormal basis: $f = Bx$ where $\|x\|_0 \leq s \ll d$

In practice, we encounter *compressible* signals.

- ✦ f_s is the best s-sparse approximation to f

Many applications...

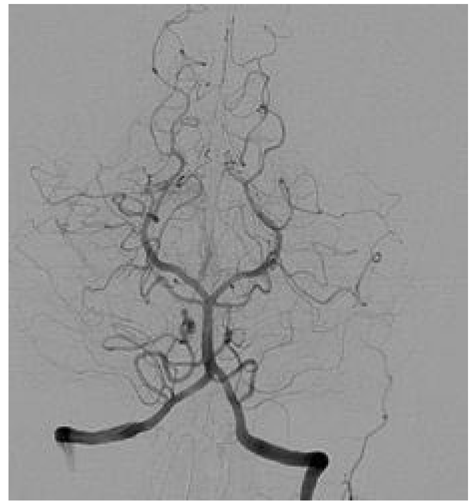
- Radar, Error Correction
- Computational Biology, Geophysical Data Analysis
- Data Mining, classification
- Neuroscience
- Imaging
- Sparse channel estimation, sparse initial state estimation
- Topology identification of interconnected systems
- ...

Notation

- ℓ_p-norms: $\|z\|_p \stackrel{\text{def}}{=} \left(\sum_i |z_i|^p\right)^{1/p}$
- Usual (Euclidean ℓ_2) distance: $\|z\|_2 \stackrel{\text{def}}{=} \left(\sum_i |z_i|^2\right)^{1/2}$
- ℓ_1 (Taxicab) distance: $\|z\|_1 \stackrel{\text{def}}{=} \left(\sum_i |z_i|\right)$
- The ℓ_2-ball: $\{z : \|z\|_2 \leq 1\}$ (circle/sphere)
- The ℓ_1-ball: $\{z : \|z\|_1 \leq 1\}$ (diamond/octahedron)
- For signal f, f_s (f_s^B) is its best s-sparse representation (in basis B)
- $\hat{f}$ will denote the reconstruction of f
- $h = \operatorname{argmin}_z g(z)$ is the *arg*ument z which *min*imizes $g(z)$

Sparsity...

Sparsity in coordinate basis: f=x



How should we reconstruct f?

✦ Easy Theorem:

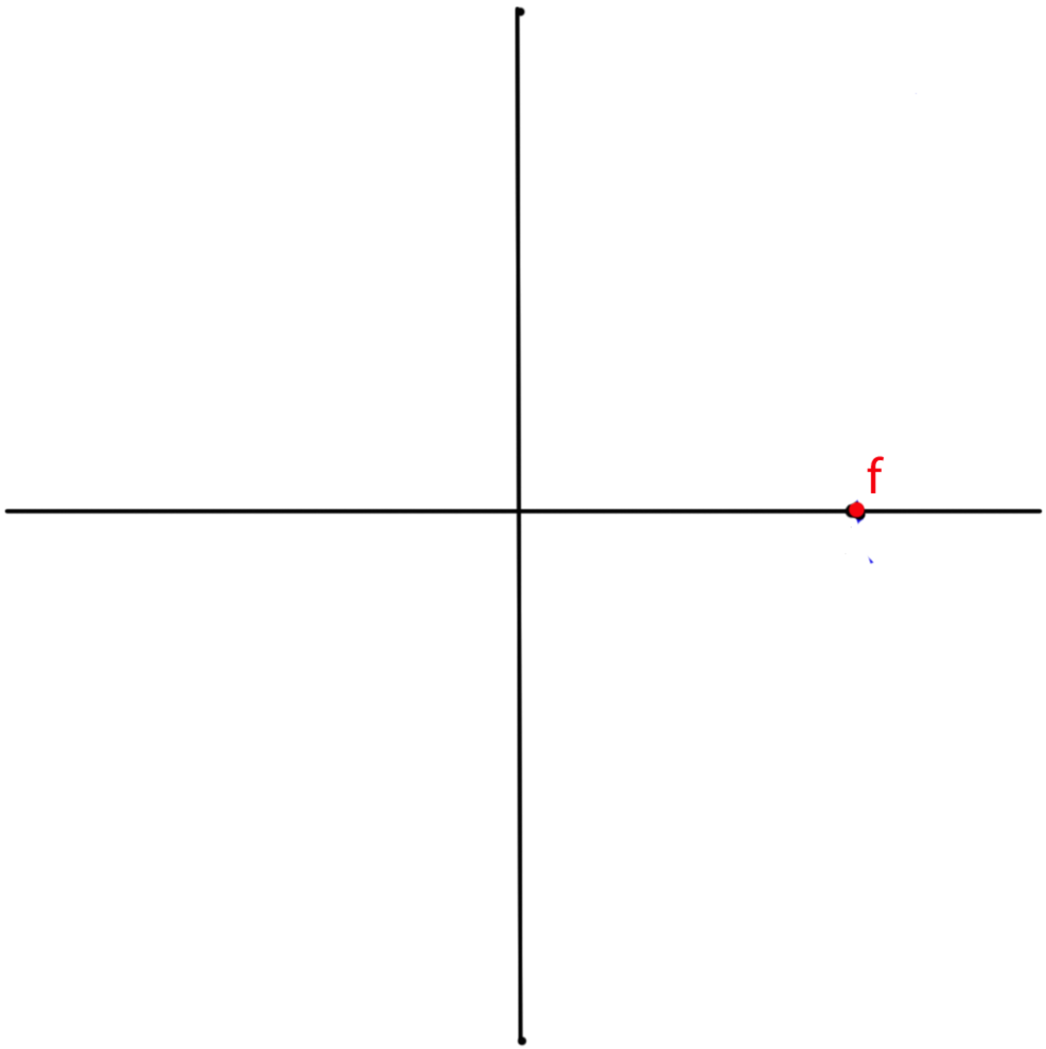
Assume A is one-to-one on all s-sparse signals. Assume there is no noise. Reconstruct an s-sparse signal f by:

$$\hat{f} = \underset{z}{\operatorname{argmin}} \|z\|_0 \quad \text{such that} \quad Az = y.$$

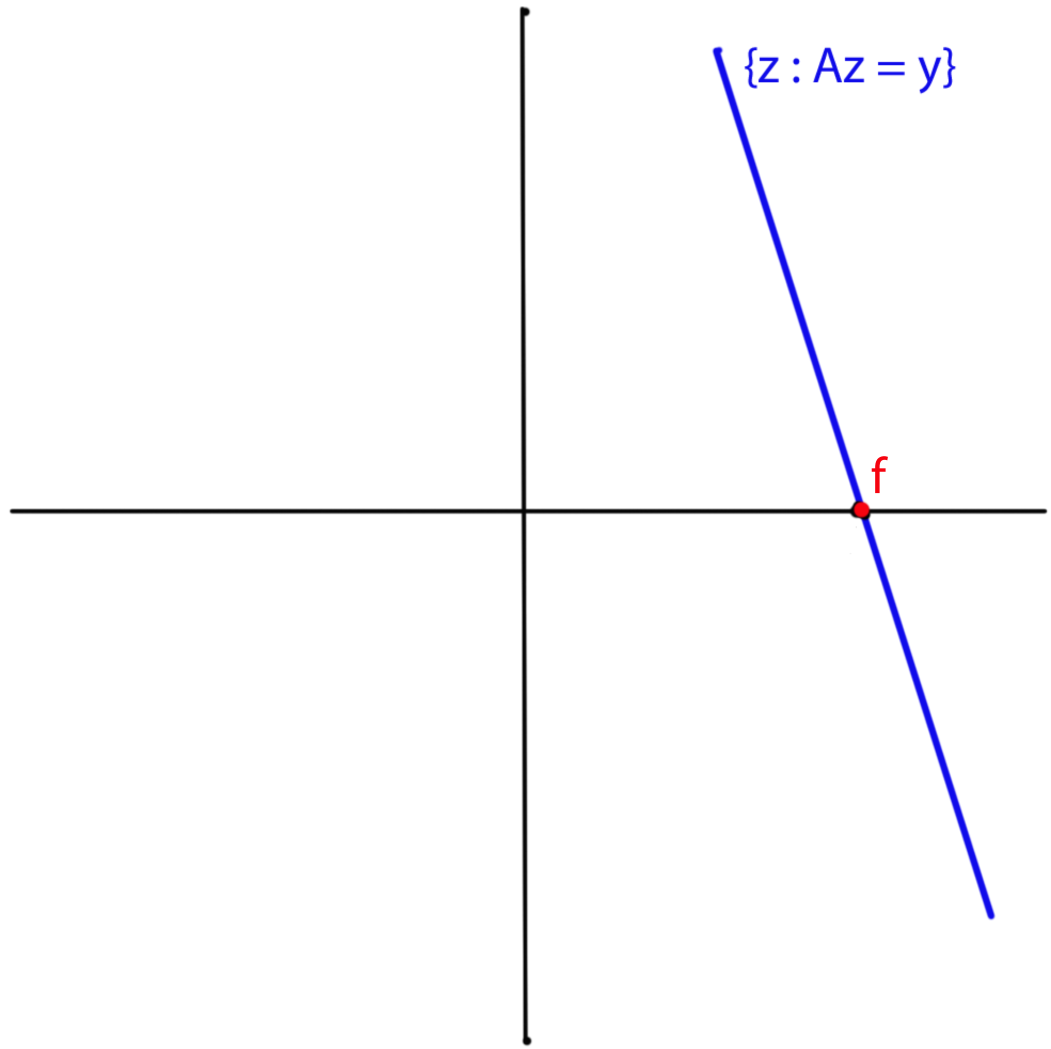
Then we reconstruct f perfectly: $\hat{f} = f$.

✦ Unfortunately, this problem is NP-Hard!

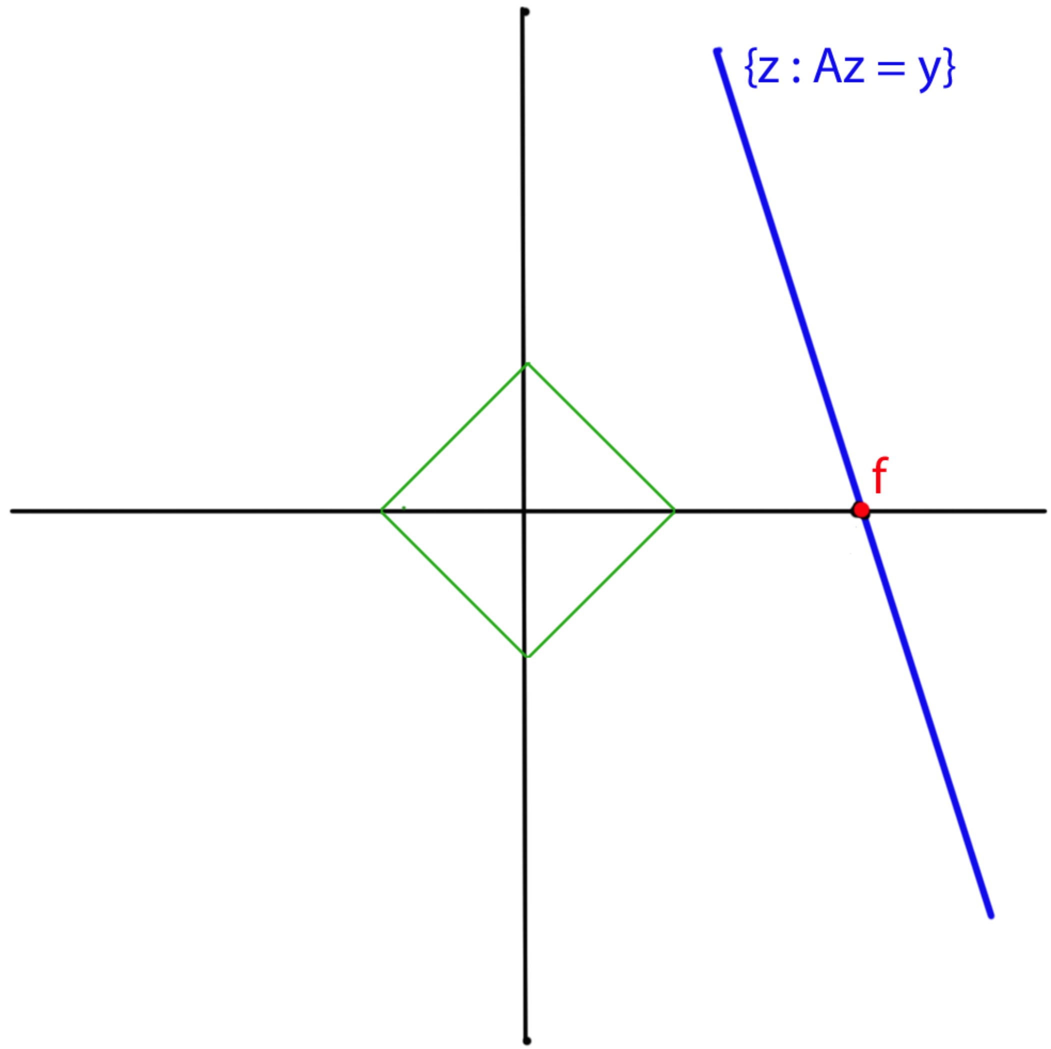
How should we reconstruct f?



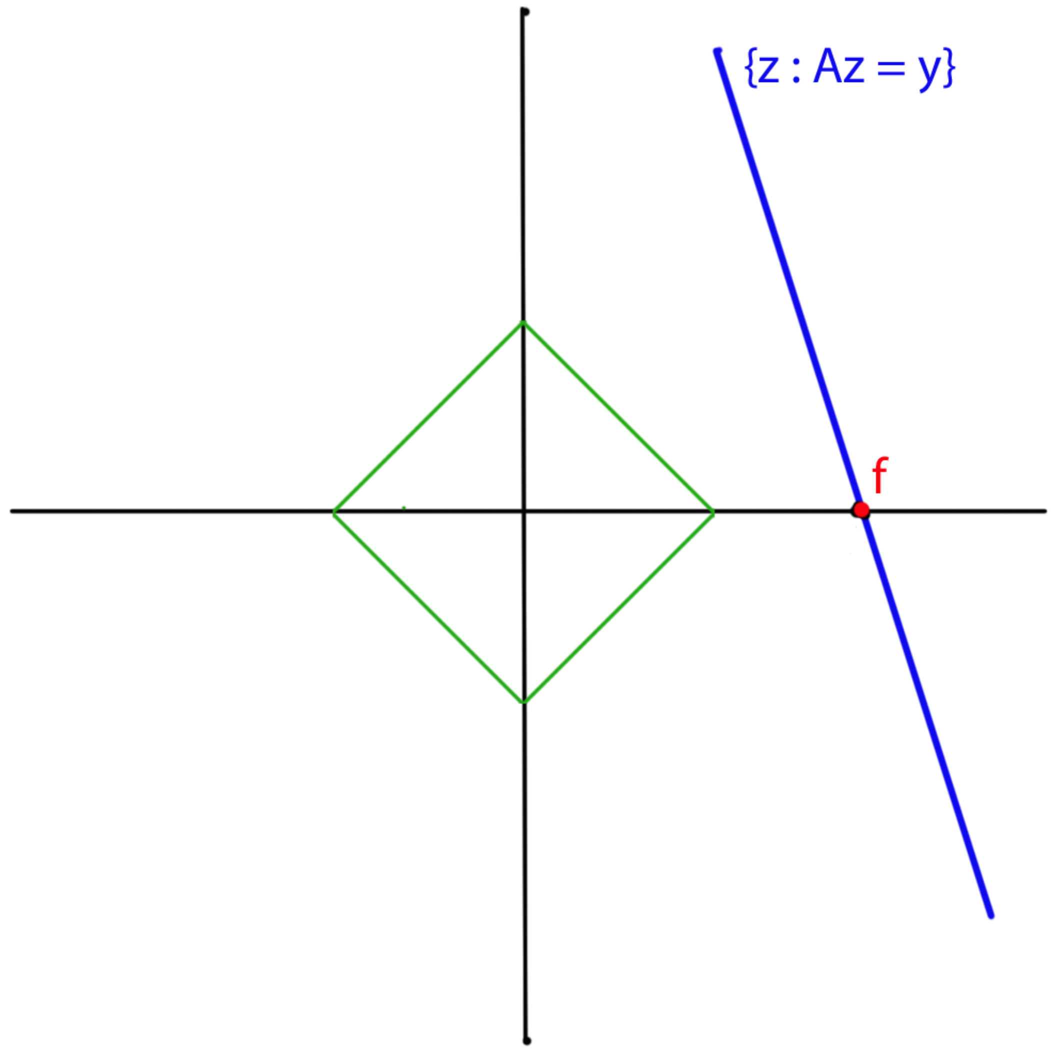
How should we reconstruct f?



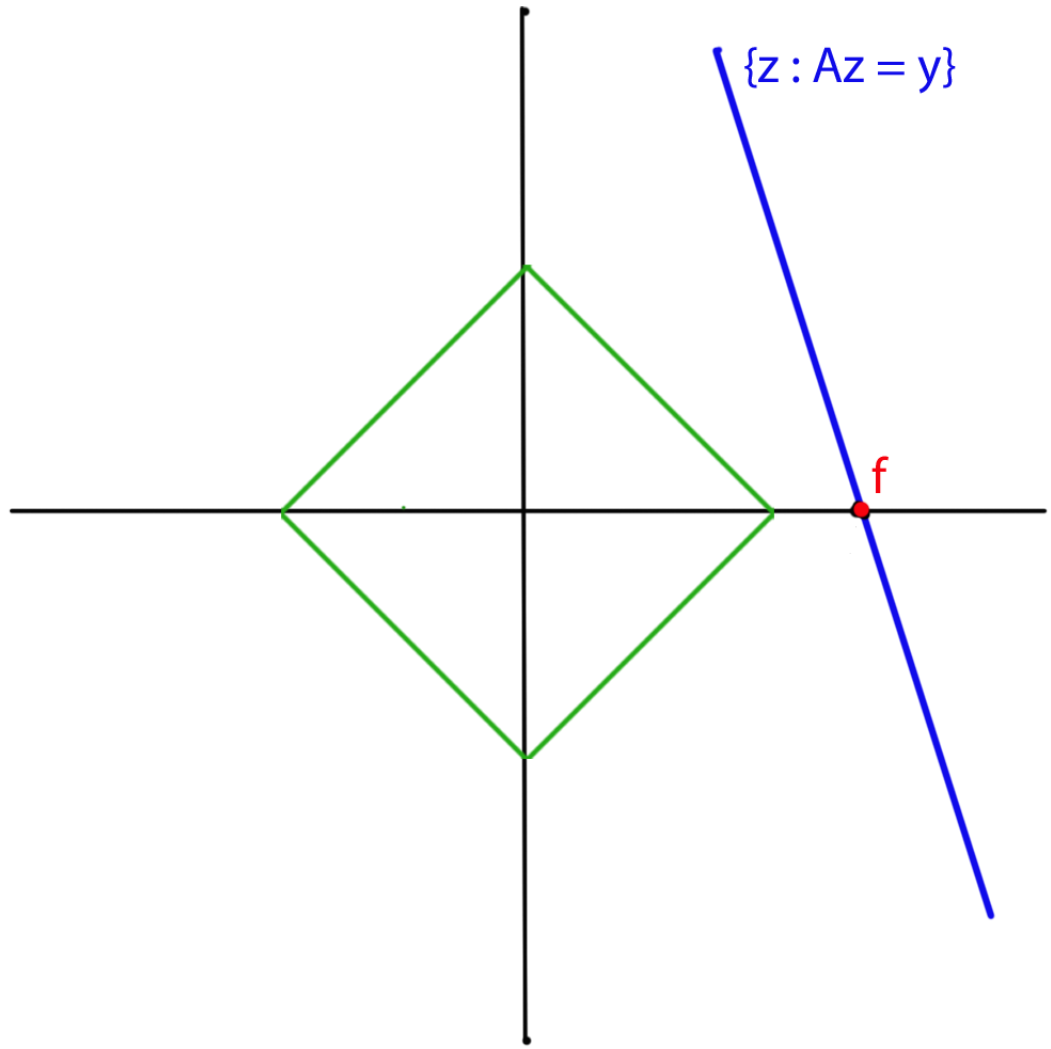
How should we reconstruct f?



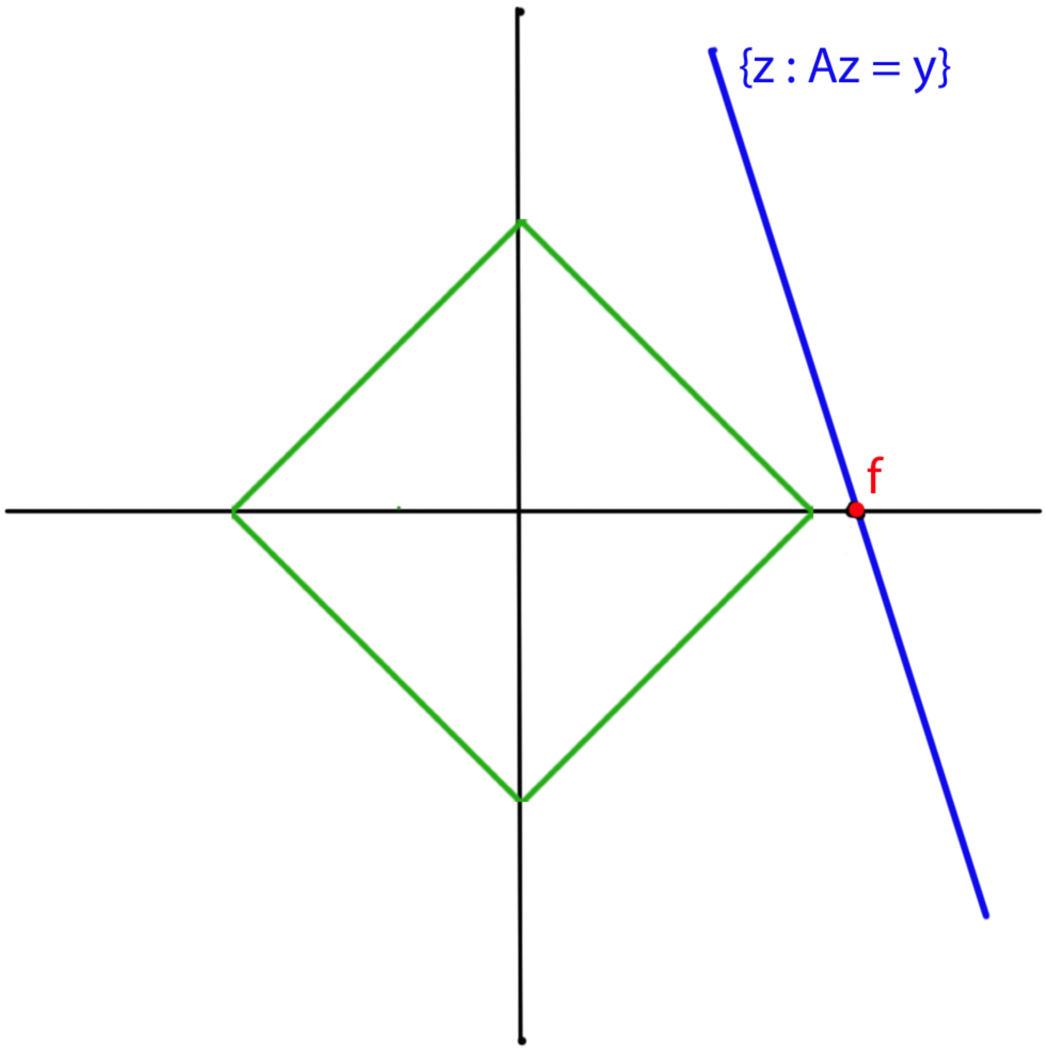
How should we reconstruct f?



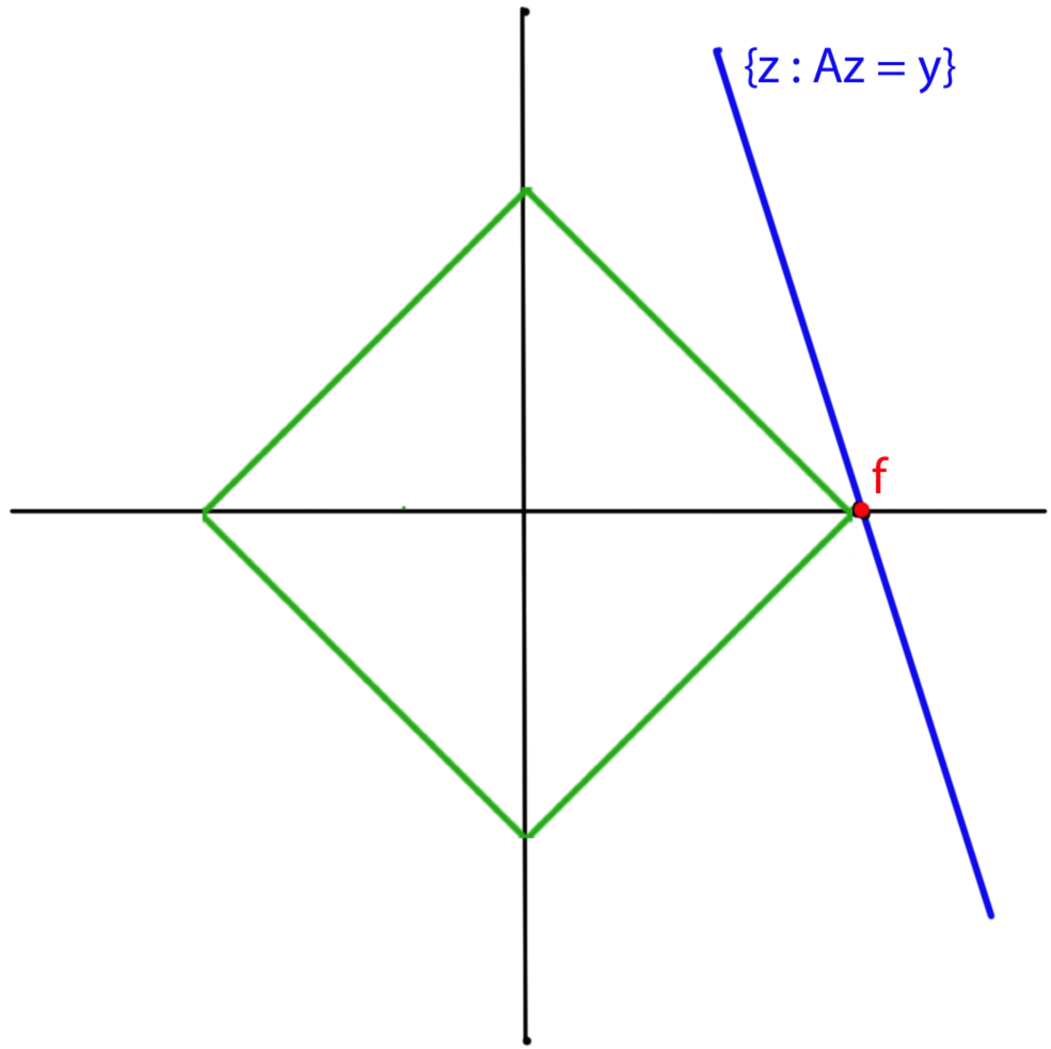
How should we reconstruct f?



How should we reconstruct f?



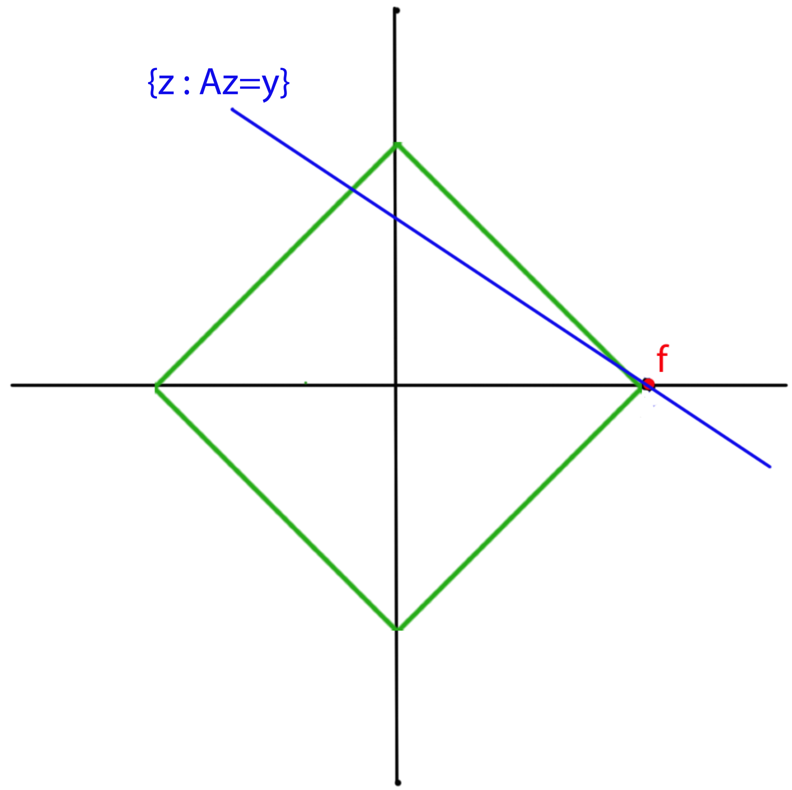
How should we reconstruct f?



Was that contrived?

Will the picture always look this way?

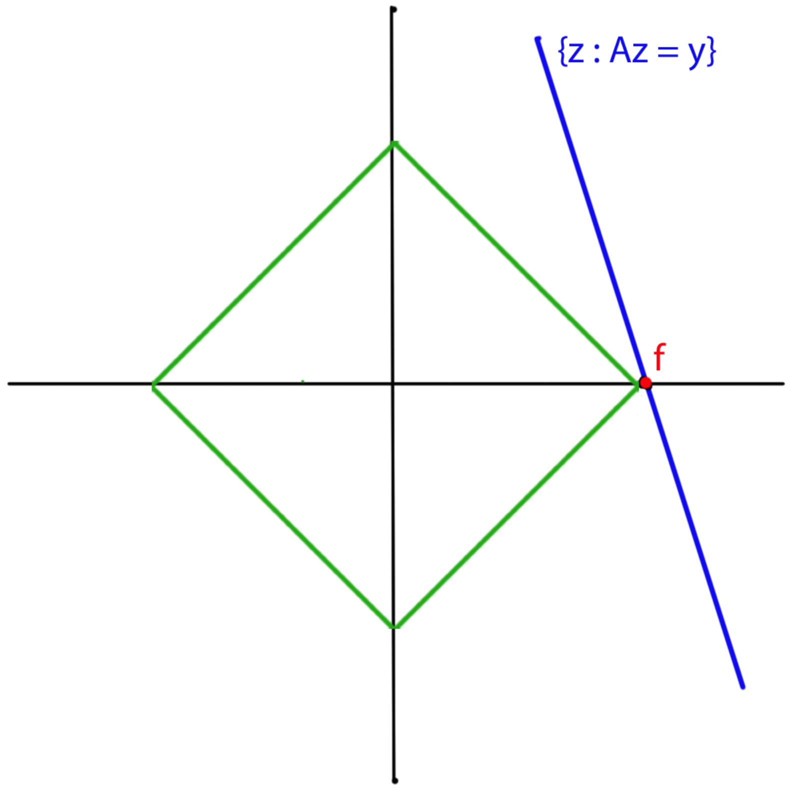
Was that contrived?



But in higher dimensions, for “sufficiently random” operators A, this picture happens with extremely low probability!

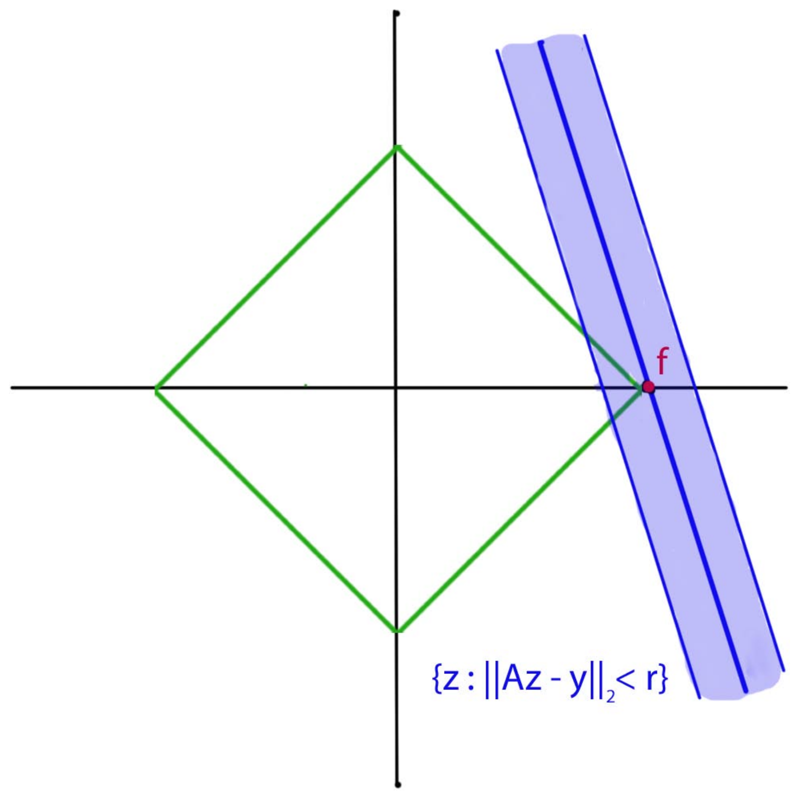
What about noise?

Recall $y = Af + \xi$.



What about noise?

Recall $y = Af + \xi$.



Reconstructing the signal f from measurements y

- ℓ_1-minimization [Candès-Romberg-Tao]

Let A satisfy the *Restricted Isometry Property* and set:

$$\hat{f} = \underset{g}{\operatorname{argmin}} \|g\|_1 \quad \text{such that} \quad \|\mathscr{A} f - y\|_2 \leq \varepsilon,$$

where $\|\xi\|_2 \leq \varepsilon$. Then we can stably recover the signal f:

$$\|f - \hat{f}\|_2 \lesssim \varepsilon + \frac{\|x - x_s\|_1}{\sqrt{s}}.$$

This error bound is optimal.

Restricted Isometry Property

- $\mathscr{A}$ satisfies the Restricted Isometry Property (RIP) when there is $\delta < c$ such that

$$(1-\delta)\|f\|_2 \le \|\mathscr{A} f\|_2 \le (1+\delta)\|f\|_2 \quad \text{whenever } \|f\|_0 \le s.$$

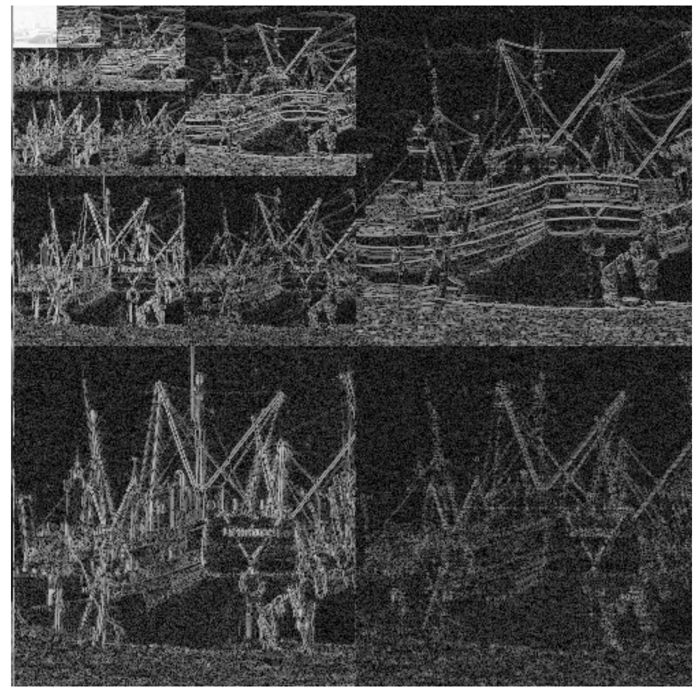
- $m \times d$ Gaussian or Bernoulli measurement matrices satisfy the RIP with high probability when

$$m \gtrsim s \log d.$$

- Random Fourier and others with fast multiply have similar property: $m \gtrsim s \log^4 d$.

Sparsity...

In orthonormal basis: $f = Bx$



Natural Images

Images are compressible in *Wavelet bases*.

$$f = \sum_{j,k=1}^{N} x_{j,k} H_{j,k}, \qquad x_{j,k} = \langle f, H_{j,k} \rangle, \qquad \|f\|_2 = \|x\|_2,$$

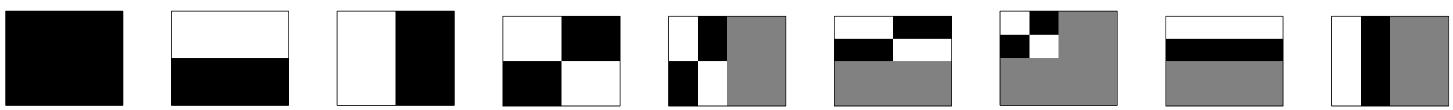


Figure 1: Haar basis functions

Wavelet transform is *orthonormal* and multi-scale. Sparsity level of image is higher on detail coefficients.

Sparsity in orthonormal basis B

✦ L1-minimization Method

For orthonormal basis B, $f = Bx$ with x sparse, one may solve the ℓ_1-minimization program:

$$\hat{f} = \underset{\tilde{f} \in \mathbb{C}^n}{\operatorname{argmin}} \|B^{-1}\tilde{f}\|_1 \quad \text{subject to} \quad \|\mathscr{A}\tilde{f} - y\|_2 \leq \varepsilon.$$

Same results hold.

Iterative methods too

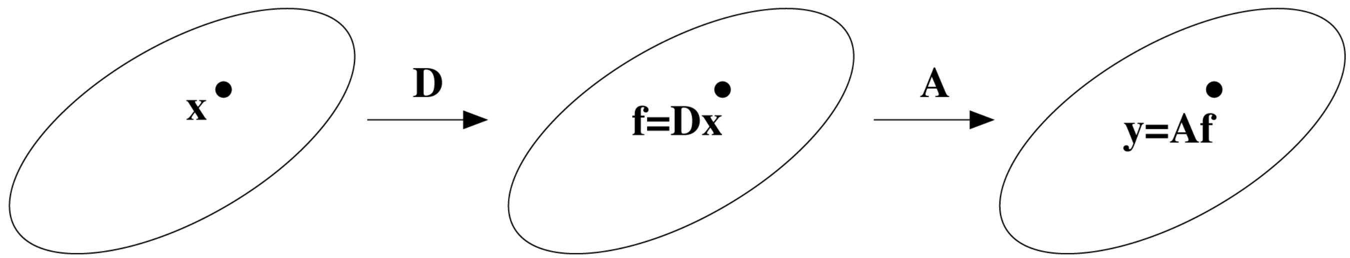
CoSaMP (N-Tropp)

input: Sampling operator A, measurements y, sparsity level s
initialize: Set $x^0 = 0$, $i = 0$.
repeat
signal proxy: Set $p = A^*(y - Ax^i)$, $\Omega = \text{supp}(p_{2s})$, $T = \Omega \cup \text{supp}(x^i)$.
signal estimation: Using least-squares, set $b|_T = A_T^\dagger y$ and $b|_{T^c} = 0$.
prune and update: Increment i and to obtain the next approximation, set $x^i = b_s$.
output: s-sparse reconstructed vector $\widehat{x} = x^i$

Sparsity...

In arbitrary dictionary: $f = Dx$

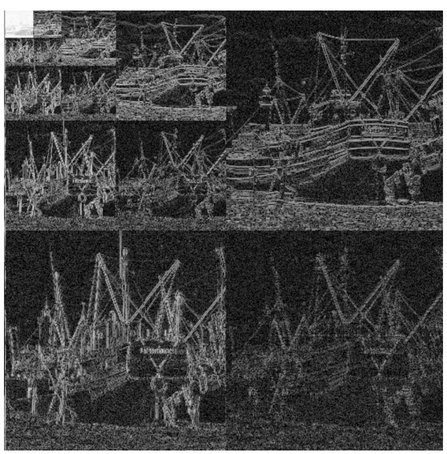
The CS Process

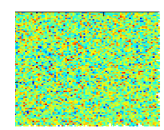


Coefficient Domain

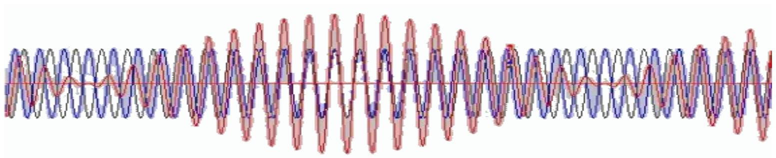
Signal Domain

Compressed Sensing Domain



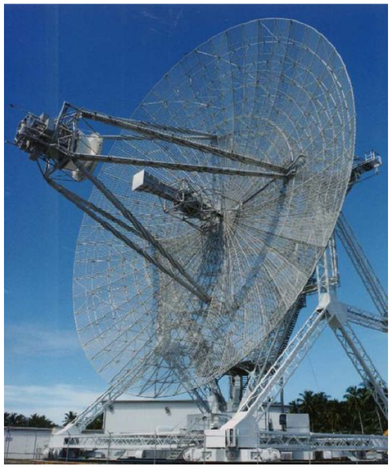


Example: Oversampled DFT



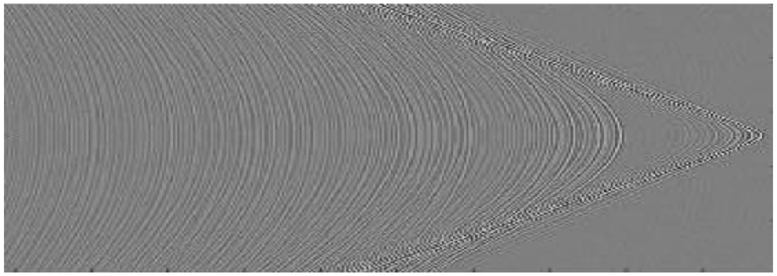
- $n \times n$ DFT: $d_k(t) = \frac{1}{\sqrt{n}} e^{-2\pi i k t/n}$
- Sparse in the DFT → superpositions of sinusoids with frequencies in the lattice.
- Instead, use the *oversampled DFT*:
- Then D is an overcomplete frame with highly coherent columns → *conventional CS does not apply*.

Example: Gabor frames



- Gabor frame: $G_k(t) = g(t - k_2 a)e^{2\pi i k_1 b t}$
- Radar, sonar, and imaging system applications use Gabor frames and wish to recover signals in this basis.
- Then D is an overcomplete frame with possibly highly coherent columns $\rightarrow$ *conventional CS does not apply*.

Example: Curvelet frames



- A Curvelet frame has some properties of an ONB but is overcomplete.

- Curvelets approximate well the curved singularities in images and are thus used widely in image processing.

- Again, this means D is an overcomplete dictionary $\rightarrow$ *conventional CS does not apply*.

Example: UWT

- The undecimated wavelet transform has a translation invariance property that is missing in the DWT.
- The UWT is overcomplete and this redundancy has been found to be helpful in image processing.
- Again, this means D is a redundant dictionary → *conventional CS does not apply*.

ℓ_1-Synthesis Method

✦ For arbitrary tight frame D, one may solve the ℓ_1-synthesis program:

$$\hat{f} = D\left(\underset{\tilde{x}\in\mathbb{C}^n}{\operatorname{argmin}} \|\tilde{x}\|_1 \quad \text{subject to} \quad \|\mathscr{A} D\tilde{x} - y\|_2 \le \varepsilon\right).$$

Some work on this method [Candès et.al., Rauhut et.al., Elad et.al.,...]

✦ *Open:* Understand the ℓ_1-synthesis problem, necessary assumptions, recovery guarantees.

ℓ_1-Analysis Method

- For arbitrary tight frame D, one may solve the ℓ_1-analysis program:

$$\hat{f} = \underset{\tilde{f} \in \mathbb{C}^n}{\operatorname{argmin}} \|D^* \tilde{f}\|_1 \quad \text{subject to} \quad \|\mathscr{A} \tilde{f} - y\|_2 \leq \varepsilon.$$

Condition on A?

✦ D-RIP

We say that the measurement matrix $\mathscr{A}$ obeys the *restricted isometry property adapted to D* (D-RIP) if there is $\delta < c$ such that

$$(1-\delta)\|Dx\|_2^2 \leq \|\mathscr{A}Dx\|_2^2 \leq (1+\delta)\|Dx\|_2^2$$

holds for all s-sparse x.

✦ Similarly to the RIP, many classes of $m \times d$ random matrices satisfy the D-RIP with $m \approx s\log(d/s)$. In fact, any matrix that satisfies RIP will satisfy D-RIP after applying random signs to the columns [Krahmer-Ward '11]

CS with tight frame dictionaries

✦ Theorem [Candès-Eldar-N-Randall]

Let D be an arbitrary tight frame and let $\mathscr{A}$ be a measurement matrix satisfying D-RIP. Then the solution $\hat{f}$ to ℓ_1-analysis satisfies

$$\|\hat{f} - f\|_2 \lesssim \varepsilon + \frac{\|D^* f - (D^* f)_s\|_1}{\sqrt{s}}.$$

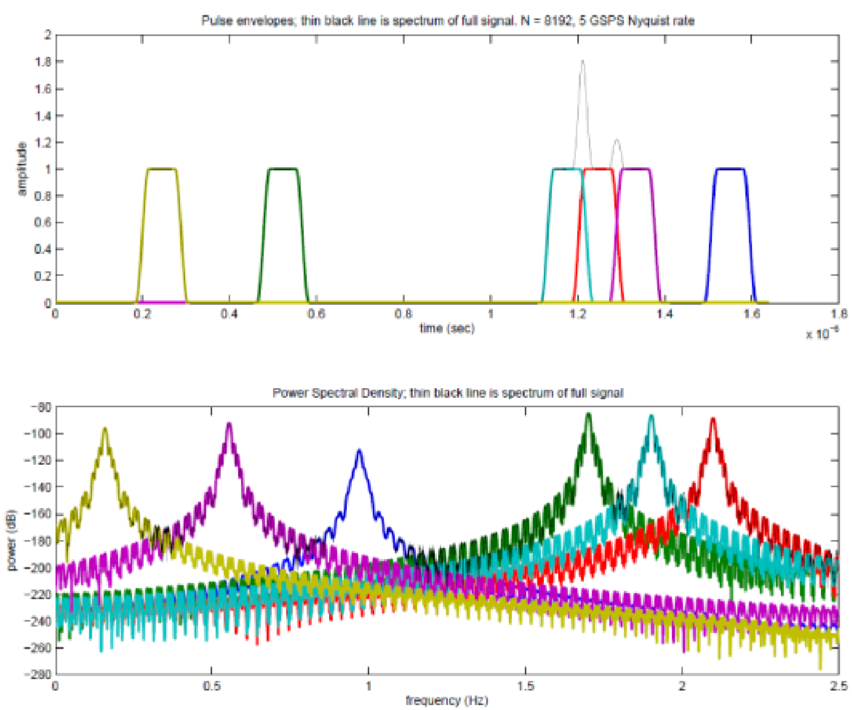
✦ In other words, This result says that ℓ_1-analysis is very accurate when $D^* f$ has rapidly decaying coefficients and D is a tight frame. ✦ For a dictionary $D \in \mathbb{C}^{n\times N}$ and a sparsity level s, we define the *unrecoverable energy* as

$$\varepsilon^\star_{s,D} = \varepsilon^\star \stackrel{\text{def}}{=} \sup_{\|Dz\|_2=1, \|z\|_0 \le s} \frac{\|D^* D z - (D^* D z)_s\|_1}{\sqrt{s}}.$$

ℓ_1-analysis: Experimental Setup

$n = 8192, m = 400, d = 491,520$

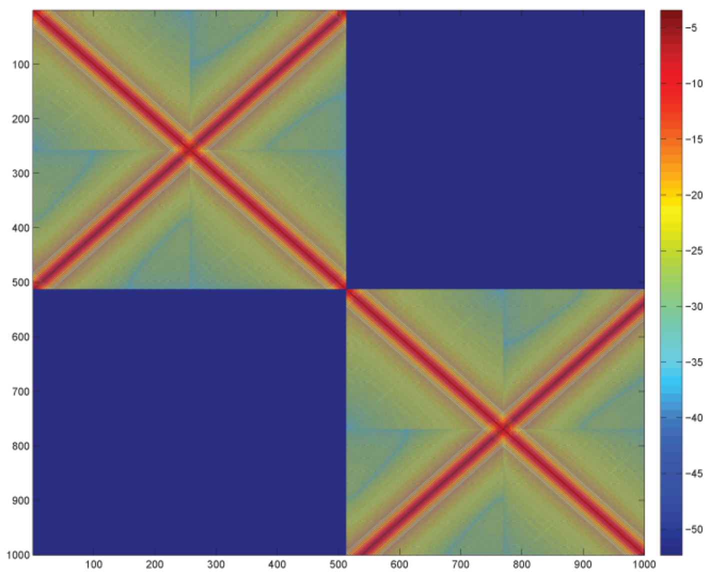
A: $m \times n$ Gaussian, D: $n \times d$ Gabor

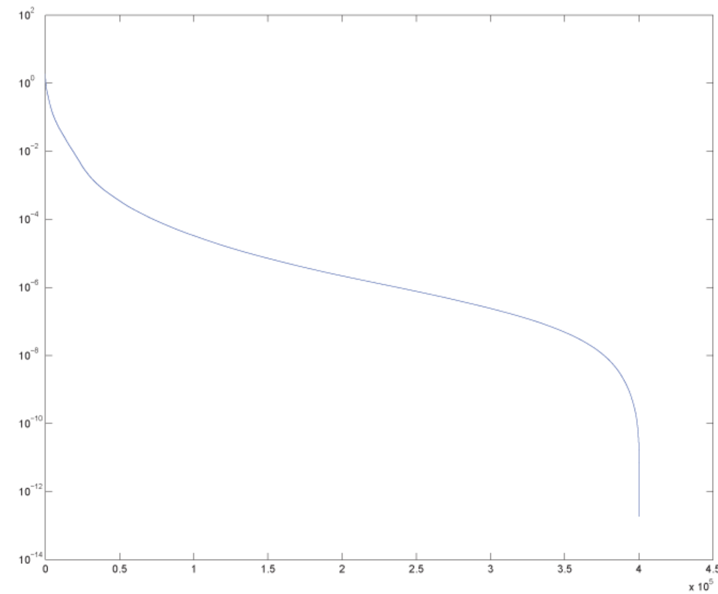


ℓ_1-analysis: Experimental Setup

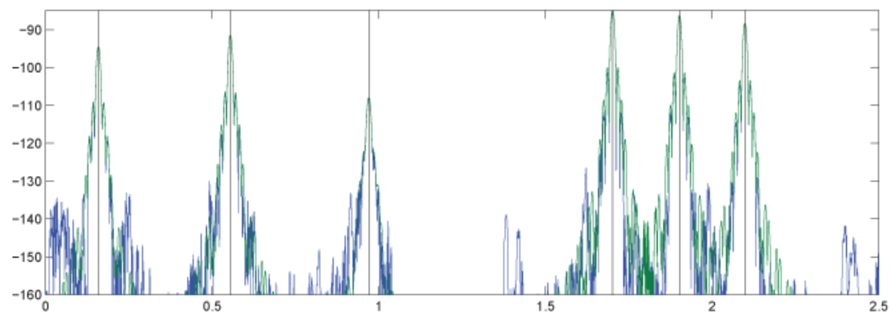
$n = 8192, m = 400, d = 491,520$

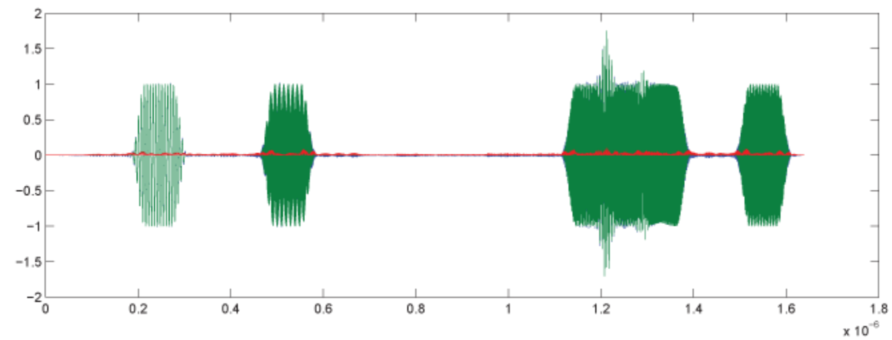
A: $m \times n$ Gaussian, D: $n \times d$ Gabor





ℓ_1-analysis: Experimental Results





Other algorithms

- ℓ_1-analysis is very accurate when $D^* f$ has rapidly decaying coefficients and D is a tight frame. This is precisely because this method operates in "analysis" space.
- *Open:* analysis methods for non-tight frames, without decaying analysis coefficients (concatenations), other models
- What about operating in signal or coefficient space?

Is it really a pipe?

(Thanks to M. Davenport for this clever analogy.)

CoSaMP

COSAMP (N-Tropp)

---

input: Sampling operator A, measurements y, sparsity level s
initialize: Set $x^0 = 0$, $i = 0$.
repeat
  signal proxy: Set $p = A^*(y - Ax^i)$, $\Omega = \mathrm{supp}(p_{2s})$, $T = \Omega \cup \mathrm{supp}(x^i)$.
  signal estimation: Using least-squares, set $b|_T = A_T^\dagger y$ and $b|_{T^c} = 0$.
  prune and update: Increment i and to obtain the next approximation, set $x^i = b_s$.
  output: s-sparse reconstructed vector $\widehat{x} = x^i$

Signal Space CoSaMP

SIGNAL SPACE COSAMP (Davenport-N-Wakin)

input: $\boldsymbol{A}$, $\boldsymbol{D}$, $\boldsymbol{y}$, s, stopping criterion
initialize: $\boldsymbol{r} = \boldsymbol{y}$, $\boldsymbol{x}^0 = 0$, $\ell = 0$, $\Gamma = \emptyset$
repeat

proxy: $\boldsymbol{h} = \boldsymbol{A}^* \boldsymbol{r}$
identify: $\Omega = \mathscr{S}_{\boldsymbol{D}}(\boldsymbol{h}, 2s)$
merge: $T = \Omega \cup \Gamma$
update: $\widetilde{\boldsymbol{x}} = \operatorname{argmin}_{\boldsymbol{z}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}\|_2 \quad \text{s.t.} \quad \boldsymbol{z} \in \mathscr{R}(\boldsymbol{D}_T)$
$\Gamma = \mathscr{S}_{\boldsymbol{D}}(\widetilde{\boldsymbol{x}}, s)$
$\boldsymbol{x}^{\ell+1} = \mathscr{P}_{\Gamma} \widetilde{\boldsymbol{x}}$
$\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^{\ell+1}$
$\ell = \ell + 1$
output: $\widehat{\boldsymbol{x}} = \boldsymbol{x}^{\ell}$

Signal Space CoSaMP

- Here we must contend with

$$\Lambda_{\mathsf{opt}}(\boldsymbol{z}, s) := \operatorname*{argmin}_{\Lambda:|\Lambda|=s} \|\boldsymbol{z} - \mathscr{P}_\Lambda \boldsymbol{z}\|_2, \quad \mathscr{P}_\Lambda : \mathbb{C}^n \to \mathscr{R}(\boldsymbol{D}_\Lambda).$$

- Estimate by $\mathscr{S}_{\boldsymbol{D}}(\boldsymbol{z}, s)$ with $|\mathscr{S}_{\boldsymbol{D}}(\boldsymbol{z}, s)| = s$, that satisfies

$$\left\|\mathscr{P}_{\Lambda_{\mathsf{opt}}(\boldsymbol{z},s)}\boldsymbol{z} - \mathscr{P}_{\mathscr{S}_{\boldsymbol{D}}(\boldsymbol{z},s)}\boldsymbol{z}\right\|_2 \le \min\left(\epsilon_1 \left\|\mathscr{P}_{\Lambda_{\mathsf{opt}}(\boldsymbol{z},s)}\boldsymbol{z}\right\|_2, \epsilon_2 \left\|\boldsymbol{z} - \mathscr{P}_{\Lambda_{\mathsf{opt}}(\boldsymbol{z},s)}\boldsymbol{z}\right\|_2\right)$$

for some constants $\epsilon_1, \epsilon_2 \ge 0$.

Approximate Projection

- Practical choices for $\mathscr{S}_{\boldsymbol{D}}(\boldsymbol{z}, s)$:

  - Any sparse recovery algorithm!

  - OMP

  - CoSaMP

  - ℓ_1-minimization followed by hard thresholding

Signal Space CoSaMP

✦ Theorem [Davenport-N-Wakin] Let D be an arbitrary tight frame, A be a measurement matrix satisfying D-RIP, and f a sparse signal with respect to D. Then the solution $\hat{f}$ from *Signal Space CoSaMP* satisfies

$$\|\hat{f} - f\|_2 \lesssim \varepsilon.$$

(And similar results for approximate sparsity, depending on the approximation method used for $\Lambda_{\text{opt}}(z, s)$.)

✦ *Open:* Design approximation methods that satisfy necessary recovery bounds (sparse approximation).

Signal Space CoSaMP: Experimental Results

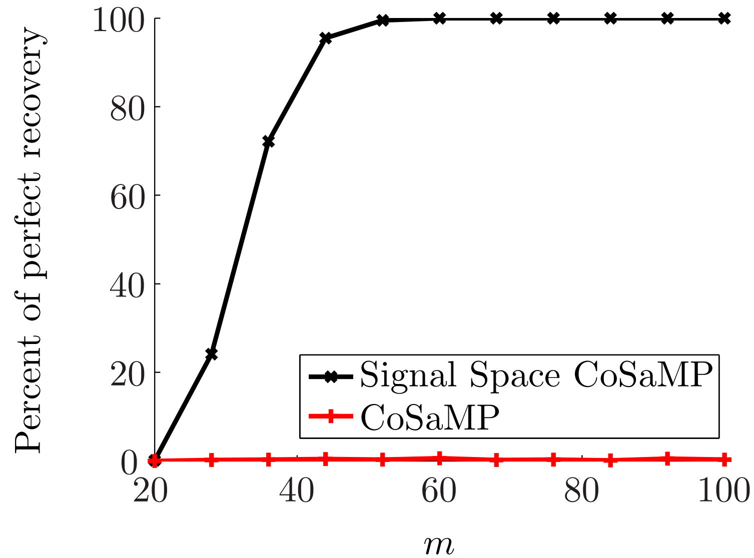


Figure 2: *Performance in recovering signals having a $s = 8$ sparse representation in a dictionary $\boldsymbol{D}$ with orthogonal, but not normalized, columns.*

Signal Space CoSaMP: Experimental Results

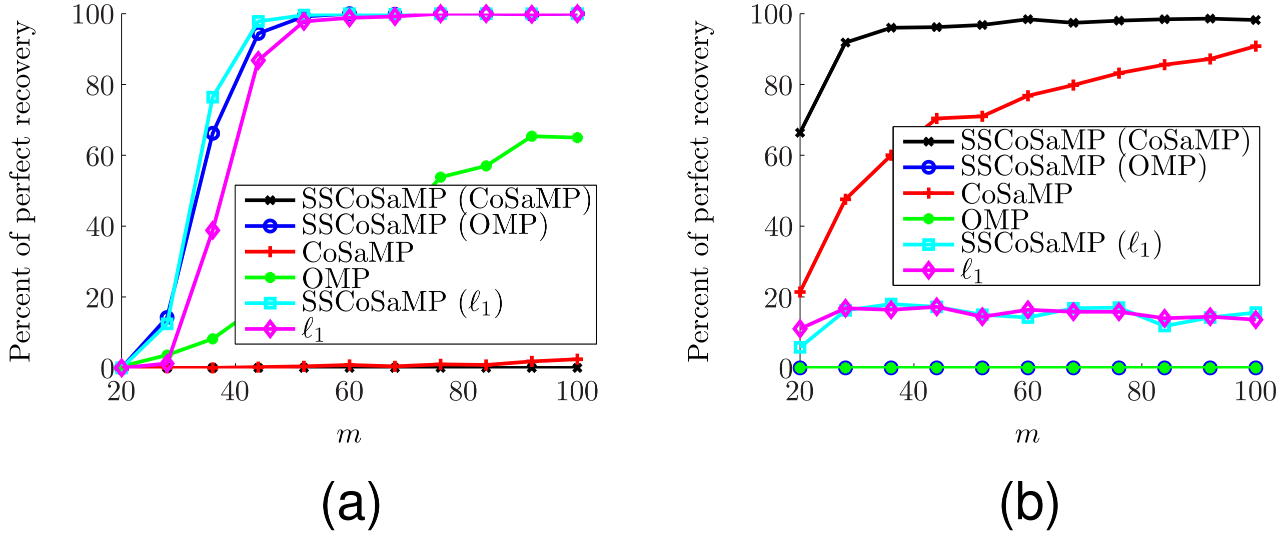


Figure 3: *Results with $s = 8$ sparse representation in a 4× overcomplete DFT dictionary: (a) well-separated coefficients, (b) clustered coefficients.*

Signal Space CoSaMP: Experimental Results

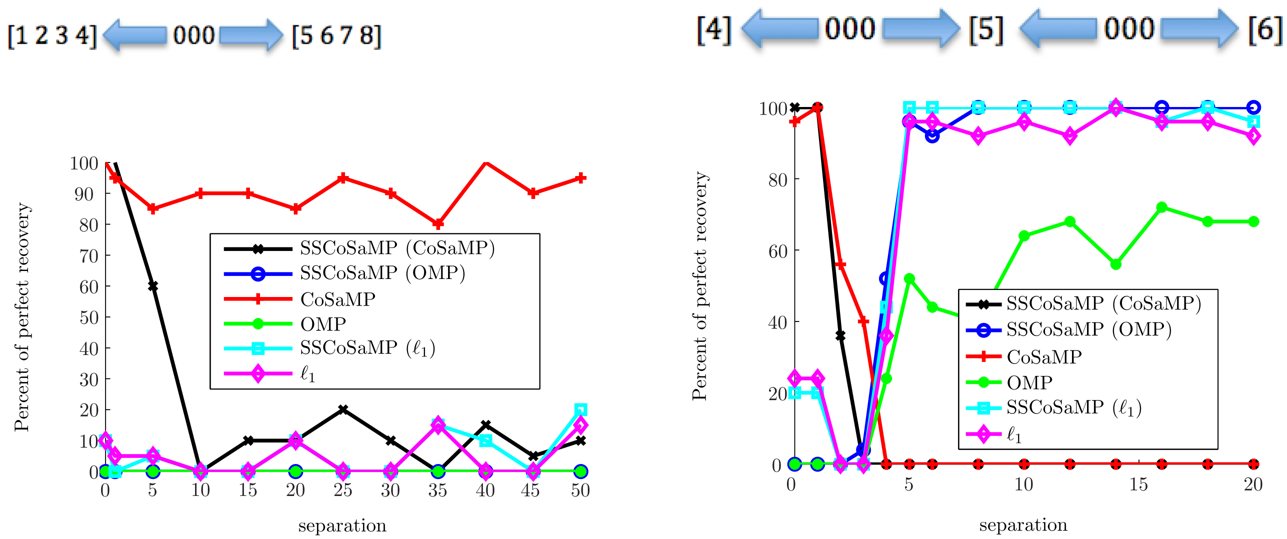


Figure 4: *Left: separations represent the number of zeros between two clusters size $s/2$. Right: separations represent the number of zeros between each nonzero entry. Measurements and sparsity are $m = 100$ and $s = 8$, respectively with a $4\times$ overcomplete DFT dictionary.*

Signal Space CoSaMP: Experimental Results

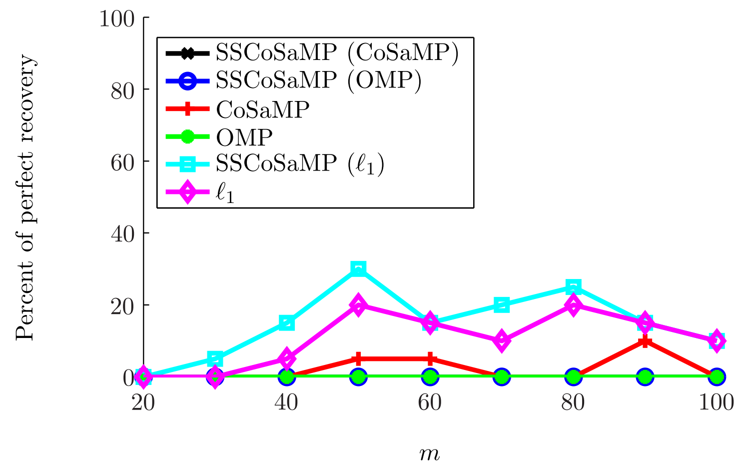


Figure 5: *SSCoSaMP recovering a sparse vector with a* hybrid sparse support*: a block of $s/2$ nonzeros with the remaining $s/2$ nonzeros spaced at least 8 slots apart from all other nonzeros.*

Signal Space CoSaMP: Experimental Results

"Ad-hoc Neighborly Methods" (NOMP, ϵ-OMP)

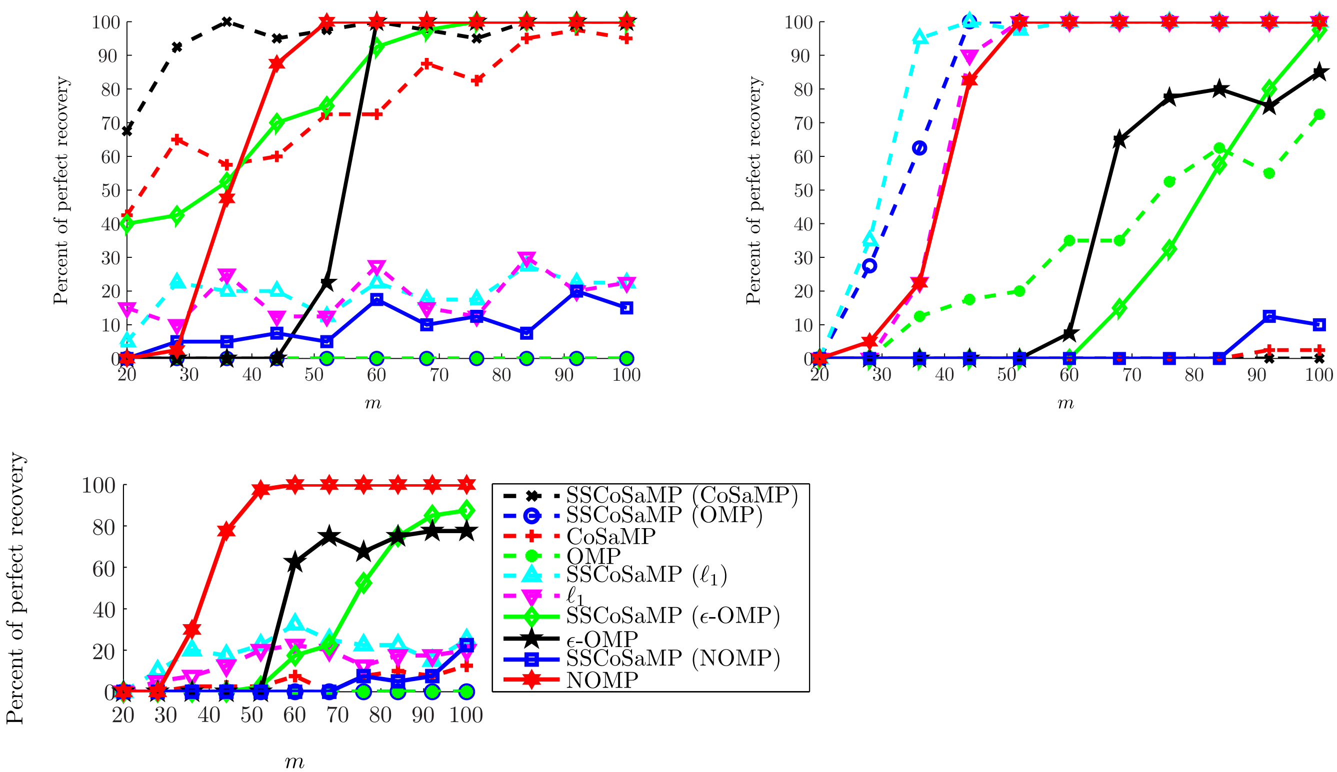


Figure 6: *Percent perfect recovery of clustered signals (left) and well separated signals (right) and hybrid signals (bottom).*

Signal Space CoSaMP: Experimental Results

Ad-hoc “Union” Methods (USSCoSaMP)

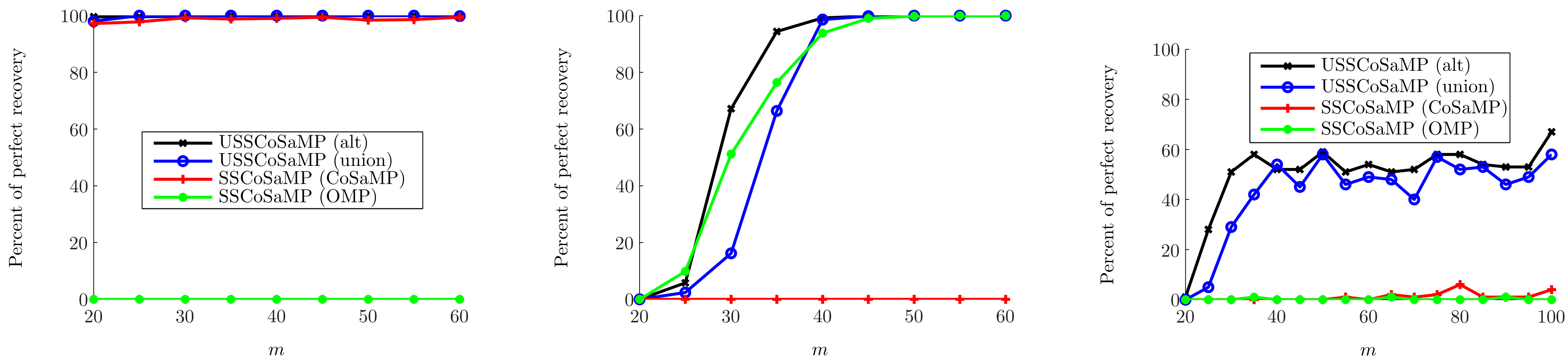


Figure 7: *Left: Clustered. Right: Well-Separated. Bottom: Hybrid signal.*

Signal Space CoSaMP: Recent improvements

✦ Recently improved results [Giryes-N and Hegde-Indyk-Schmidt] which relax the assumptions on the approximate projections.

✦ These results show that at least for RIP/incoherent dictionaries, standard algorithms like CoSaMP/OMP/IHT suffice for the approximate projections.

Open:

✦ The interesting/challenging case is when the dictionary does *not* satisfy such a condition. Are there methods which provide these approximate projections? Or are they not even necessary?

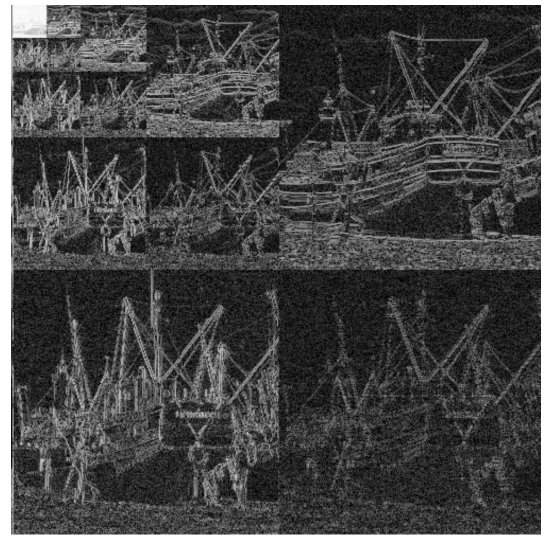
Natural images

Sparse...

256×256 “Boats" image

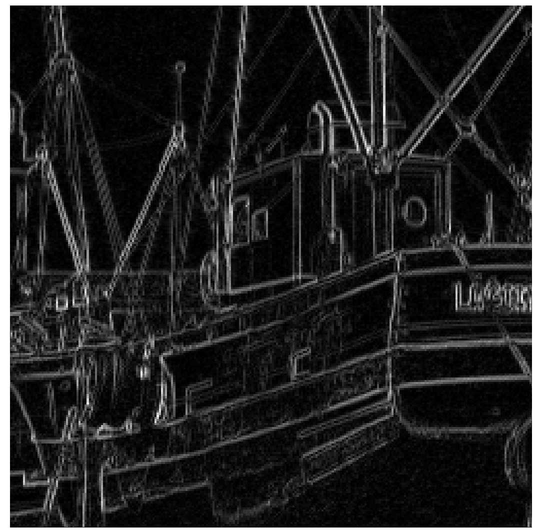
Natural images

Sparse wavelet representation...



Natural images

Images are compressible in *discrete gradient*.



Natural images

Images are compressible in *discrete gradient*.

The discrete directional derivatives of an image $f \in \mathbb{C}^{N\times N}$ are

$$f_x : \mathbb{C}^{N\times N} \to \mathbb{C}^{(N-1)\times N}, \qquad (f_x)_{j,k} = f_{j,k} - f_{j-1,k},$$

$$f_y : \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times (N-1)}, \qquad (f_y)_{j,k} = f_{j,k} - f_{j,k-1},$$

the discrete gradient operator is

$$\nabla[f] = (f_x, f_y)$$

Sparsity in gradient

✦ CS Theory

The gradient operator ∇ is not an orthonormal basis or a tight frame. In fact, it is extremely ill-conditioned!

Comparison of two compressed sensing reconstruction algorithms

- Haar-minimization (L_1-Haar)

$\hat{f}_{Haar} = \operatorname{argmin} \|H(Z)\|_1$ subject to $\|\mathscr{A} Z - y\|_2 \leq \varepsilon$

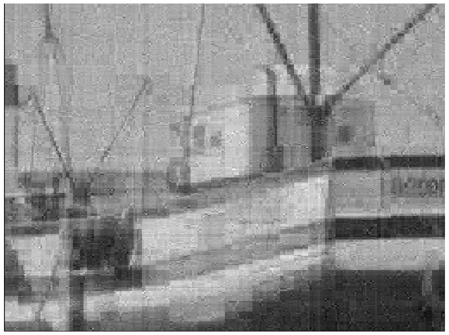
- Total Variation minimization (TV)

$\hat{f}_{TV} = \operatorname{argmin} \|\nabla[Z]\|_1 \quad \text{subject to} \quad \|\mathscr{A} Z - y\|_2 \leq \varepsilon$, where $\|Z\|_{TV} = \|\nabla[Z]\|_1$ is the *total-variation norm*.

Imaging via compressed sensing

(a) Original

(b) TV



(c) L_1-Haar

Figure 8: Reconstruction using $m = .2N^2$

Imaging via compressed sensing

(a) Original

(b) TV

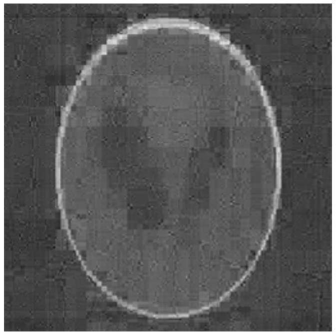
(c) L_1-Haar

Figure 9: Reconstruction using $m = .2N^2$ measurements

Imaging via compressed sensing

(a) Original

(b) TV



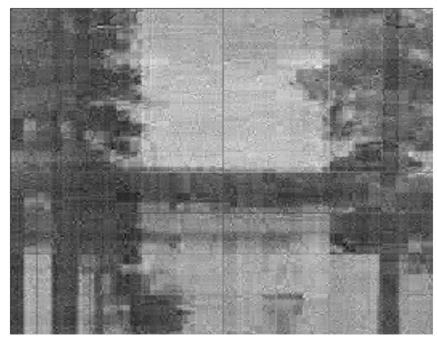
(c) L_1-Haar

Figure 10: Reconstruction using $m = .2N^2$ measurements.

Imaging via compressed sensing

(a) (Quantization)

(b) TV



(c) L_1-Haar

Figure 11: Reconstruction using $m = .2N^2$ measurements

Imaging via compressed sensing

InView (Austin TX)

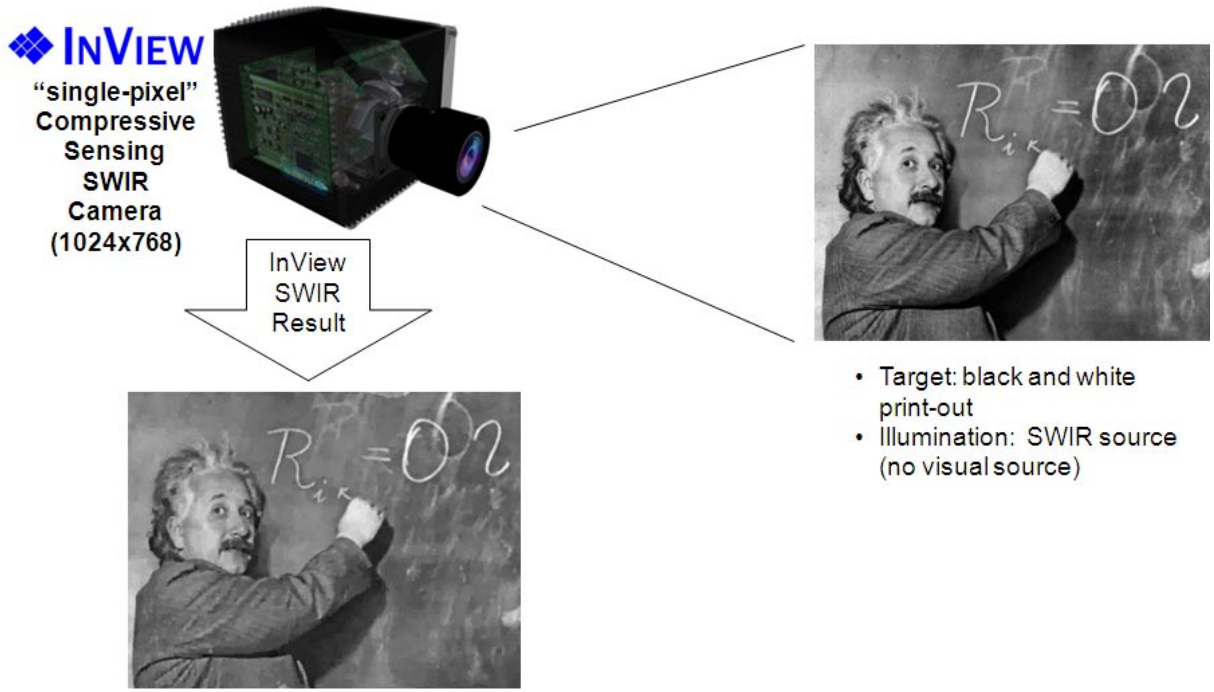


Figure 12: SWIR Reconstruction using $m = .5N^2$ measurements

Imaging via compressed sensing

InView (Austin TX)

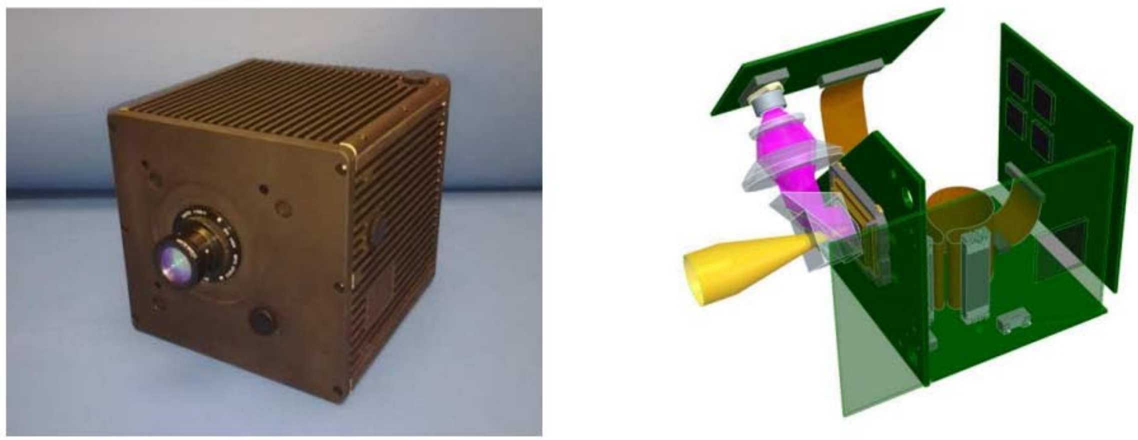


Figure 13: InView SWIR camera

Empirical → Theoretical?

✦ TV Works

Empirically, it has been well known that

$$\hat{f}_{TV} = \operatorname{argmin} \|Z\|_{TV} \quad \text{subject to} \quad \|\mathscr{A}Z - y\|_2 \leq \varepsilon, \qquad (TV)$$

provides quality, stable image recovery.

✦ No provable stability guarantees.

Stable signal recovery using total-variation minimization

Theorem 1. [N-Ward] *From $m \gtrsim s\log(N)$ linear RIP measurements, for any $f \in \mathbb{C}^{N\times N}$,*

$$\hat{f} = \mathrm{argmin}\,\|Z\|_{TV} \quad \textit{such that} \quad \|\mathscr{A}(Z) - y\|_2 \le \varepsilon,$$

satisfies

$$\|f - \hat{f}\|_{TV} \lesssim \|\nabla[f] - \nabla[f]_s\|_1 + \sqrt{s}\varepsilon \qquad \text{(gradient error)}$$

and

$$\|f - \hat{f}\|_2 \lesssim \log(N) \cdot \left[\frac{\|\nabla[f] - \nabla[f]_s\|_1}{\sqrt{s}} + \varepsilon\right] \qquad \text{(signal error)}$$

This error guarantee is optimal up to the $\log(N)$ factor

Higher dimensional objects

Movies, higher dimensional objects?

Theorem 2. [N-Ward] *From $m \gtrsim s\log(N^d)$ linear RIP measurements, for any $f \in \mathbb{C}^{N^d}$,*

$$\hat{f} = \operatorname{argmin} \|Z\|_{TV} \quad \textit{such that} \quad \|\mathscr{A}(Z) - y\|_2 \le \varepsilon,$$

satisfies

$$\|f - \hat{f}\|_{TV} \lesssim \|\nabla[f] - \nabla[f]_s\|_1 + \sqrt{s}\varepsilon \qquad \text{(gradient error)}$$

and

$$\|f - \hat{f}\|_2 \lesssim \log(N^d/s) \cdot \left[\frac{\|\nabla[f] - \nabla[f]_s\|_1}{\sqrt{s}} + \varepsilon\right] \qquad \text{(signal error)}$$

This error guarantee is optimal up to the $\log(N^d/s)$ factor

Stable signal recovery using total-variation minimization

Method of proof:

✧ First prove stable *gradient* recovery

✧ Translate stable *gradient* recovery to stable *signal* recovery using the strengthened Sobolev inequality.

Open:

✦ Remove logarithmic factors, design more efficient measurement schemes.

✦ Incorporate wavelets, Laplacian, etc. for optimal performance.

✦ Prove for 1-d signals!

Re-visiting the D-RIP

$$(1-\delta)\|Dx\|_2^2 \le \|\mathscr{A}Dx\|_2^2 \le (1+\delta)\|Dx\|_2^2$$

- Required for most recovery guarantees using frames
- If a matrix A satisfies RIP then $\tilde{A}$ obtained by applying random signs to the columns satisfies D-RIP
- This implies (sub)Gaussian matrices, Bernoulli matrices, etc. still satisfy the D-RIP
- But for structured matrices (e.g. Fourier), we need to apply column signs...
- Not always feasible in practice!

Uniform sampling

The *mutual coherence* of two bases $\{\varphi_k\}$ and $\{b_j\}$ is defined to be

$$\mu = \sup_{j,k} |\langle b_j, \varphi_k \rangle|.$$

✦ Theorem [Rudelson-Vershynin '06, Rauhut '07]

Consider the matrix $A = \Phi_\Omega B^* \in \mathbb{C}^{m \times N}$ with entries

$$A_{\ell,k} = \langle \varphi_{j_\ell}, b_k \rangle, \quad \ell \in [m], k \in [N], \tag{1}$$

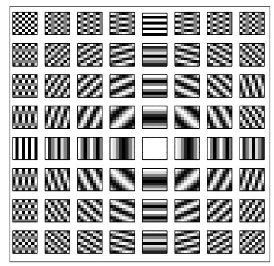
where the φ_{j_ℓ} are independent samples drawn uniformly at random from an ONB $\{\varphi_j\}_{j=1}^N$ incoherent with the sparsity basis $\{b_j\}$ in the sense that $\mu \le K N^{-1/2}$. Then once, for some $s \gtrsim \log(N)$,

$$m \ge C \delta^{-2} K^2 s \log^3(s) \log(N), \tag{2}$$

with probability at least $1 - N^{-\gamma \log^3(s)}$, the restricted isometry constant δ_s of $\frac{1}{\sqrt{m}} A$ satisfies $\delta_s \le \delta$. The constants $C, \gamma > 0$ are universal.

Variable density sampling

[Lustig-Donoho-Pauly ’07]: “For a better performance with real images, one should be undersampling less near the k-space origin and more in the periphery of k-space. For example, one may choose samples randomly with sampling density scaling according to a power of distance from the origin."



- Idea by Puy-Vandergheynst-Wiaux ’11:
  - Variable density sampling can reduce coherence.
  - Strategy: Find optimal weights using convex optimization.
  - Work with problem specific discretization level.
  - No theoretical recovery guarantees.

Local coherence

- *Empirical observation of Puy et al.:*
  Often only few Fourier basis vectors have high coherence with the sparsity basis. Changing the weights can compensate for this inhomogeneity.

- We introduce the *local coherence* to address this issue.

- The *local coherence* of an ONB $\{\varphi_j\}_{j=1}^N$ of $\mathbb{C}^N$ with respect to another ONB $\{\psi_k\}_{k=1}^N$ of $\mathbb{C}^N$ is the function $\mu_{loc}(j) = \sup_{1 \le k \le N} |\langle \varphi_j, \psi_k \rangle|$.

RIP for variable density subsampling

✦ Theorem [Consequence of Rauhut-Ward '12]

Assume the local coherence of an ONB $\Phi = \{\varphi_j\}_{j=1}^N$ with respect to an ONB $\Psi = \{\psi_k\}_{k=1}^N$ is pointwise bounded by the function κ, that is, $\sup_{1\le k\le N} |\langle \varphi_j, \psi_k\rangle| \le \kappa_j$. Consider the matrix $A \in \mathbb{C}^{m\times N}$ with entries

$$A_{\ell,k} = \langle \varphi_{j_\ell}, \psi_k\rangle, \quad j \in [m], k \in [N], \tag{3}$$

where the j_ℓ are drawn independently according to $\nu_\ell = \mathbb{P}(\ell_j = \ell) = \frac{\kappa_\ell^2}{\|\kappa\|_2^2}$. Suppose that

$$m \ge C\delta^{-2}\|\kappa\|_2^2 s \log^3(s)\log(N), \tag{4}$$

and let $D = diag(d_{j,j})$, where $d_{j,j} = \|\kappa\|_2/\kappa_j$. Then with probability at least $1 - N^{-\gamma\log^3(s)}$, the preconditioned matrix $\frac{1}{\sqrt{m}}DA$ has a restricted isometry constant $\delta_s \le \delta$. The constants $C, \gamma > 0$ are universal.

✦ Theorem [Krahmer-N-Ward '15]

Fix a sparsity level $s < N$, and constant $0 < \delta < 1$. Let $D \in \mathbb{C}^{n \times N}$ be a tight frame, let $\mathscr{A} = \{a_1, \ldots, a_n\}$ be an ONB of $\mathbb{C}^n$, and $\kappa \in \mathbb{C}^n_+$ an entrywise upper bound of the local coherence, that is,

$$\mu_i^{loc}(\mathscr{A}, D) = \sup_{j \in [N]} |\langle a_i, d_j \rangle| \le \kappa_i.$$

Consider the unrecoverable energy $\varepsilon^\star$. Construct $\tilde{\mathscr{A}} \in \mathbb{C}^{m \times n}$ by sampling vectors from $\mathscr{A}$ at random according to the probability distribution ν given by $\nu(i) = \frac{\kappa_i^2}{\|\kappa\|_2^2}$ and normalizing by $\sqrt{n/m}$. Then as long as

$$\begin{aligned} m &\ge C\delta^{-2} s \|\kappa\|_2^2 (1+\varepsilon^\star)^2 \log^3(s(1+\varepsilon^\star)^2) \log(N), \quad \text{and} \\ m &\ge C\delta^{-2} s \|\kappa\|_2^2 (1+\varepsilon^\star)^2 \log(1/\gamma) \end{aligned} \tag{5}$$

then with probability $1-\gamma$, $\tilde{\mathscr{A}}$ satisfies the D-RIP with parameters s and δ.

Consequences

- Recovery guarantees for Fourier measurements and Haar wavelet frames of redundancy 2 by previous local coherence analysis.
- Constant local coherence: Implies incoherence based guarantees (for example for oversampled Fourier dictionary).
- No need to apply random column signs anymore.

Thank you!

E-mail:

✧ `dneedell@cmc.edu`

Web:

✧ `www.cmc.edu/pages/faculty/DNeedell`

References:

- E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. Communications on Pure and Applied Mathematics, 59(8):1207Öś223, 2006.
- E. J. Candès, Y. C. Eldar, D. Needell and P. Randall. Compressed sensing with coherent and redundant dictionaries. Applied and Computational Harmonic Analysis, 31(1):59-73, 2010.
- M. A. Davenport, D. Needell and M. B. Wakin. Signal Space CoSaMP for Sparse Recovery with Redundant Dictionaries, submitted.
- P. Indyk, E. Price and D. Woodruff. On the Power of Adaptivity in Sparse Recovery, FOCS 2011.
- D. Needell and R. Ward. Stable image reconstruction using total variation minimization. J. Fourier Analysis and Applications, to appear.
- D. Needell and R. Ward. Total variation minimization for stable multidimensional signal recovery, submitted.
- F. Krahmer, D. Needell and R. Ward. Compressed sensing with redundant dictionaries and structured measurements, in preparation.

An Intro. to Dictionary Learning

– AMS Short Course on Finite Frame Theory

San Antonio, TX

Guangliang Chen

San Jose State University

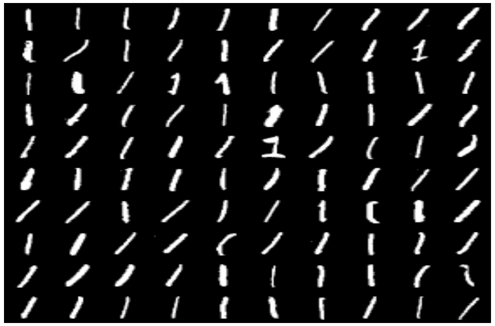
January 9, 2015

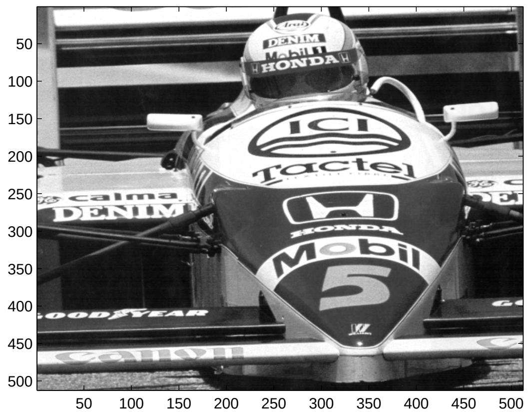
Outline of the presentation

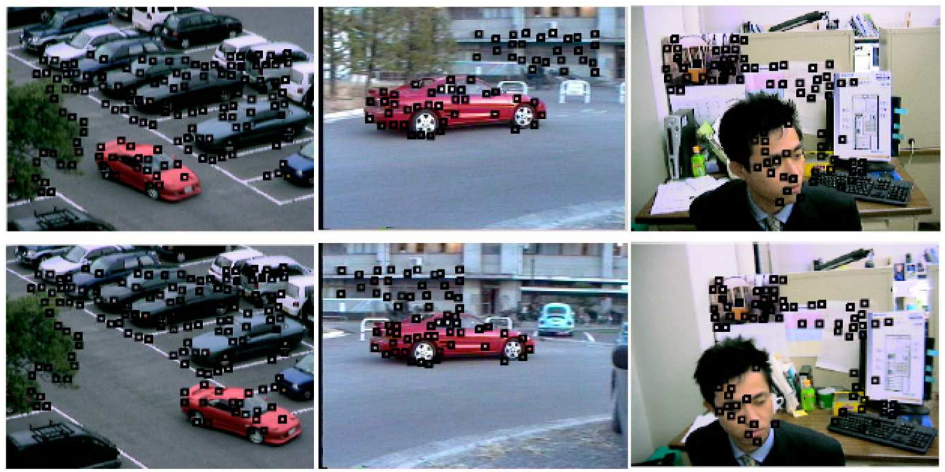
- The problem
  - Background
  - Problem formulations
  - Applications
- Existing approaches
  - K-SVD
  - GMRA
- Summary

Problem setting

- Data sets are normally represented as point clouds $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^\ell$, for ℓ large;



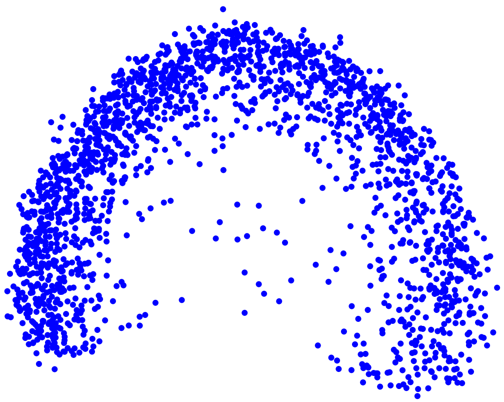




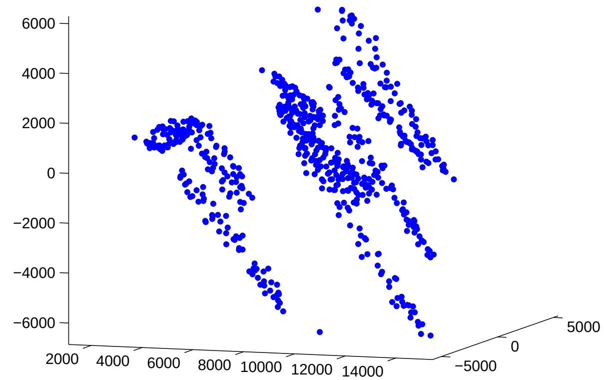
Problem setting

- Data sets are normally represented as point clouds $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^\ell$, for ℓ large;
- but they are often intrinsically low dimensional.

The handwritten digit 1



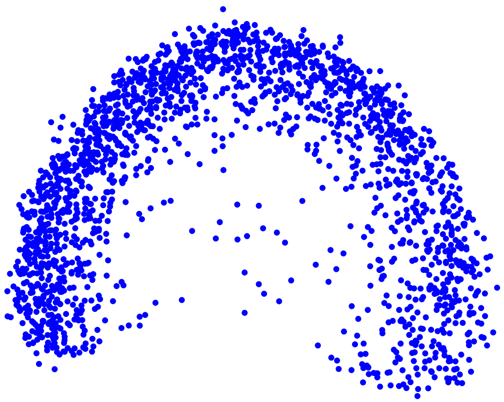
The human facial images



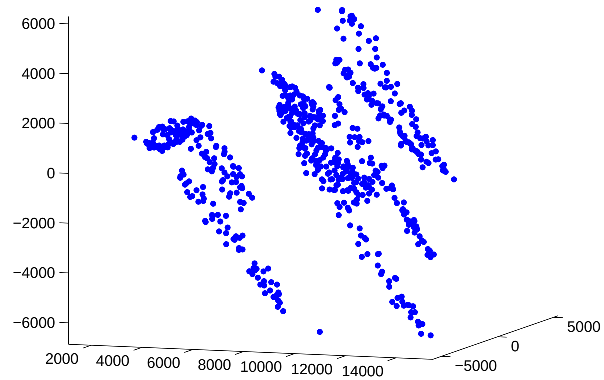
Problem setting

- Data sets are normally represented as point clouds $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^\ell$, for ℓ large;
- but they are often intrinsically low dimensional.

The handwritten digit 1



The human facial images



- This observation/assumption is commonly exploited for effectively modeling real data.

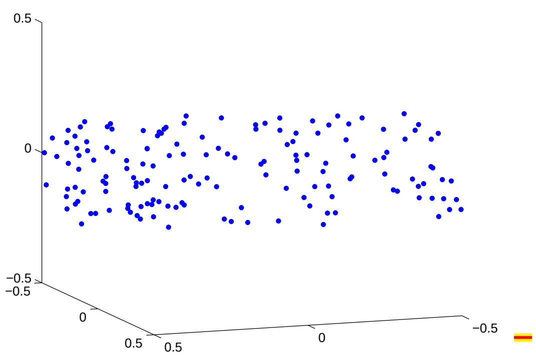
Orthogonal basis modeling

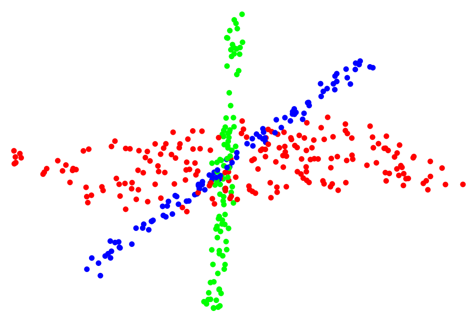
The simplest way is to use an orthogonal basis $\mathcal{B} = \{\mathbf{b}_i\}$:

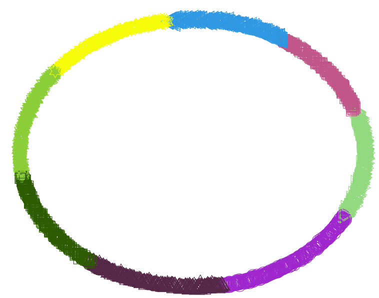
$$\mathbf{x} = \sum c_i \mathbf{b}_i, \qquad \text{where} \quad c_i = \frac{(\mathbf{x}, \mathbf{b}_i)}{\|\mathbf{b}_i\|_2^2}$$

that is

- either designed analytically: Fourier, Wavelet, etc.
- or learned from data:
  - PCA: Model data by a single subspace;
  - Hybrid linear modeling: Use a union of subspaces







Orthogonal basis modeling

Pros:

- Mathematically very simple to operate
- Good performance (when assumption is satisfied)

Cons:

- Very limited expressiveness
- No clear interpretation for the basis
- Too simplistic for real data

Overcomplete basis modeling

- The idea is to represent data using a (large) redundant collection of (linearly dependent) vectors $\mathbf{d}_i$:

$$\mathbf{x} = \sum \alpha_i \mathbf{d}_i = \mathbf{D}\alpha.$$

Overcomplete basis modeling

- The idea is to represent data using a (large) redundant collection of (linearly dependent) vectors $\mathbf{d}_i$:

$$\mathbf{x} = \sum \alpha_i \mathbf{d}_i = \mathbf{D}\alpha.$$

- This relaxes the linear independence condition for bases and gives us great flexibility in choosing which subset of $\mathbf{d}_i$ to represent $\mathbf{x}$.

Overcomplete basis modeling

- The idea is to represent data using a (large) redundant collection of (linearly dependent) vectors $\mathbf{d}_i$:

$$\mathbf{x} = \sum \alpha_i \mathbf{d}_i = \mathbf{D}\alpha.$$

- This relaxes the linear independence condition for bases and gives us great flexibility in choosing which subset of $\mathbf{d}_i$ to represent $\mathbf{x}$.
- What kind of $\mathbf{D}$ should we use?

Overcomplete basis modeling

- Two different types of requirements for $\mathbf{D}$:
  - Impose the frame condition on the collection $\mathbf{D}$:

$$A\|\mathbf{x}\|_2^2 \leq \|\mathbf{D}^T\mathbf{x}\|_2^2 \leq B\|\mathbf{x}\|_2^2, \quad \text{for all } \mathbf{x}$$

    to make it a spanning set with good theoretical properties, or
  - Add a sparsity constraint to the coefficient α (only for given signals $\mathbf{x}$) in order to promote simplicity and easy interpretability.

Overcomplete basis modeling

- Two different types of requirements for $\mathbf{D}$:
  - Impose the frame condition on the collection $\mathbf{D}$:

$$A\|\mathbf{x}\|_2^2 \leq \|\mathbf{D}^T\mathbf{x}\|_2^2 \leq B\|\mathbf{x}\|_2^2, \quad \text{for all } \mathbf{x}$$

    to make it a spanning set with good theoretical properties, or
  - Add a sparsity constraint to the coefficient α (only for given signals $\mathbf{x}$) in order to promote simplicity and easy interpretability.
- These two considerations lead to, respectively, frames and dictionaries.

Ways to produce dictionaries

- Depending on how they are obtained, dictionaries can be divided into two categories:
  - Analytically designed: frames (Xlets for digital images)
  - Learned from data: trained dictionaries

Ways to produce dictionaries

- Depending on how they are obtained, dictionaries can be divided into two categories:
  - Analytically designed: frames (Xlets for digital images)
  - Learned from data: trained dictionaries
- Both have pros and cons:
  - Analytic: supported by theory, fast transform, but works only when assumptions are satisfied
  - Trained: adapts better to data, better performance, but computationally intensive and hard to analyze due to lack of structure

Ways to produce dictionaries

- Depending on how they are obtained, dictionaries can be divided into two categories:
  - Analytically designed: frames (Xlets for digital images)
  - Learned from data: trained dictionaries
- Both have pros and cons:
  - Analytic: supported by theory, fast transform, but works only when assumptions are satisfied
  - Trained: adapts better to data, better performance, but computationally intensive and hard to analyze due to lack of structure
- We focus on trained dictionaries; such research is called *data-dependent dictionary learning*.

Data-dependent DL

Problem definition. Given training signals $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^\ell$, we learn a dictionary $\mathbf{D}$ consisting of *atomic* signals $\mathbf{d}_1, \ldots, \mathbf{d}_m$, in order to represent each given signal as a *linear* combination of *few* atoms:

$$\min_{\mathbf{D},\alpha_1,\ldots,\alpha_n} \sum ||\alpha_i||_0 \quad \text{subject to} \quad ||\mathbf{x}_i - \mathbf{D}\alpha_i||_2 \leq \epsilon$$

in which

- $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_m] \in \mathbb{R}^{\ell \times m}$: (long) dictionary matrix;
- $||\alpha_i||_0$: # nonzeros in the coefficient vector α_i;
- ϵ: desired precision

Data-dependent DL

Problem definition. Given training signals $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^\ell$, we learn a dictionary $\mathbf{D}$ consisting of *atomic* signals $\mathbf{d}_1, \ldots, \mathbf{d}_m$, in order to represent each given signal as a *linear* combination of *few* atoms:

$$\min_{\mathbf{D}, \alpha_1, \ldots, \alpha_n} \sum ||\alpha_i||_0 \quad \text{subject to} \quad ||\mathbf{x}_i - \mathbf{D}\alpha_i||_2 \leq \epsilon$$

in which

- $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_m] \in \mathbb{R}^{\ell \times m}$: (long) dictionary matrix;
- $||\alpha_i||_0$: # nonzeros in the coefficient vector α_i;
- ϵ: desired precision

This is the fixed-precision, minimal-cost formulation.

Alternative formulations

- Fixed cost, minimal error

$$\min_{\mathbf{D},\alpha_1,\ldots,\alpha_n} \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{D}\alpha_i||_2^2 \quad \text{subject to} \quad ||\alpha_i||_0 \leq s$$

Alternative formulations

- Fixed cost, minimal error

$$\min_{\mathbf{D},\alpha_1,\ldots,\alpha_n} \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{D}\alpha_i||_2^2 \quad \text{subject to} \quad ||\alpha_i||_0 \leq s$$

- Unified version

$$\min_{\mathbf{D},\alpha_1,\ldots,\alpha_n} \sum ||\mathbf{x}_i - \mathbf{D}\alpha_i||_2^2 + \lambda||\alpha_i||_0$$

Alternative formulations

- Fixed cost, minimal error

$$\min_{\mathbf{D},\alpha_1,\ldots,\alpha_n} \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{D}\alpha_i||_2^2 \quad \text{subject to} \quad ||\alpha_i||_0 \leq s$$

- Unified version

$$\min_{\mathbf{D},\alpha_1,\ldots,\alpha_n} \sum ||\mathbf{x}_i - \mathbf{D}\alpha_i||_2^2 + \lambda||\alpha_i||_0$$

- Convex relaxation

$$\min_{\mathbf{D},\alpha_1,\ldots,\alpha_n} \sum ||\mathbf{x}_i - \mathbf{D}\alpha_i||_2^2 + \lambda||\alpha_i||_1$$

Alternative formulations

- Fixed cost, minimal error

$$\min_{\mathbf{D},\alpha_1,\ldots,\alpha_n} \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{D}\alpha_i||_2^2 \quad \text{subject to} \quad ||\alpha_i||_0 \leq s$$

- Unified version

$$\min_{\mathbf{D},\alpha_1,\ldots,\alpha_n} \sum ||\mathbf{x}_i - \mathbf{D}\alpha_i||_2^2 + \lambda||\alpha_i||_0$$

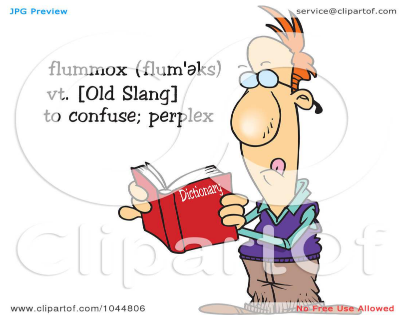
- Convex relaxation

$$\min_{\mathbf{D},\alpha_1,\ldots,\alpha_n} \sum ||\mathbf{x}_i - \mathbf{D}\alpha_i||_2^2 + \lambda||\alpha_i||_1$$

- Matrix form

$$\min_{\mathbf{D},\mathbf{A}} ||\mathbf{X} - \mathbf{D}\mathbf{A}||_F^2 + \lambda||A||_{1,1}, \quad ||A||_{1,1} = \sum ||\alpha_i||_1$$

Analogy to natural languages



Take the English language as an example:

- There is a dictionary which is a large collection of words (atoms)
- Each sentence, an ordered list of words, can be regarded as a signal
- There is normally more than one way to express something, but the most concise sentence is preferred
- The DL task can be thought of as reconstructing the English dictionary from many sentences

Related fields

- **PCA** and **hybrid linear modeling**: The dictionary model is the most flexible because it uses all possible combinations of the atoms

Related fields

- **PCA** and **hybrid linear modeling**: The dictionary model is the most flexible because it uses all possible combinations of the atoms
- **Frame design**: Dictionary combines the sparsity notion, and is more application oriented (much less theory)

Related fields

- **PCA** and **hybrid linear modeling**: The dictionary model is the most flexible because it uses all possible combinations of the atoms
- **Frame design**: Dictionary combines the sparsity notion, and is more application oriented (much less theory)
- **Sparse coding** (D known):

$$\min_{\alpha} ||\mathbf{x} - \mathbf{D}\alpha||_2^2 + \lambda||\alpha||_0$$

Can be efficiently solved by pursuit algorithms (greedy methods or convex optimization)

Related fields

- **PCA** and **hybrid linear modeling**: The dictionary model is the most flexible because it uses all possible combinations of the atoms
- **Frame design**: Dictionary combines the sparsity notion, and is more application oriented (much less theory)
- **Sparse coding** ($\mathbf{D}$ known):

$$\min_{\alpha} ||\mathbf{x} - \mathbf{D}\alpha||_2^2 + \lambda||\alpha||_0$$

  Can be efficiently solved by pursuit algorithms (greedy methods or convex optimization)

- **Compressive sensing** ($\mathbf{D}$: designed sensing matrix): The CS problem is to recover the sparse signal α from its compressed measurements $\mathbf{x} = \mathbf{D}\alpha$.

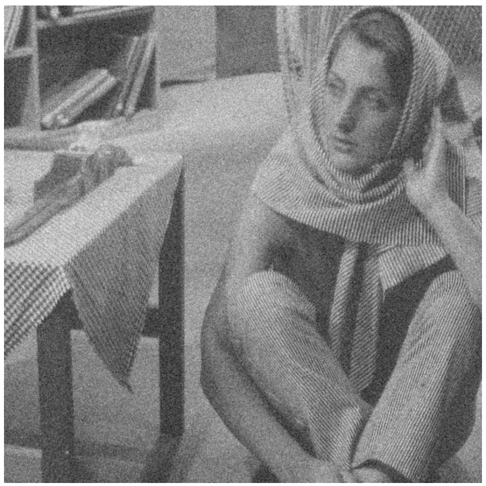
Application to image processing

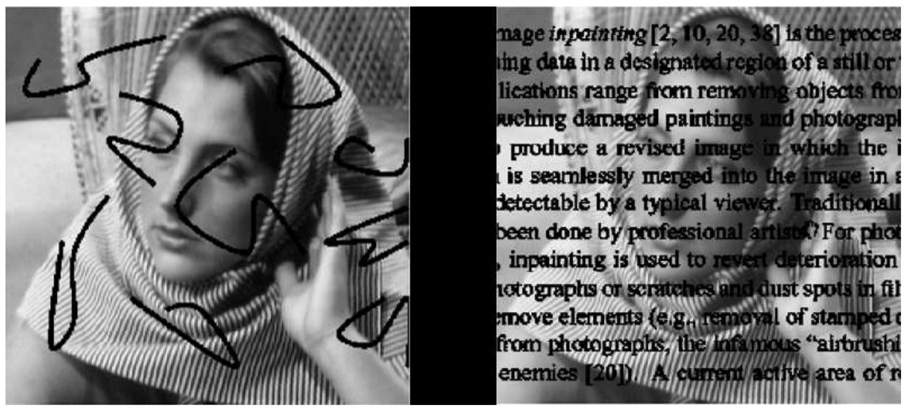
Assume that we observe a noisy, degraded version of a clean image $\mathbf{t}$:

$$\mathbf{x} = \mathbf{H}\mathbf{t} + \mathbf{e},$$

in which

- $\mathbf{e}$: additive noise;
- $\mathbf{H}$: identity or a linear degradation operator representing a blur, downsampling, or masking.

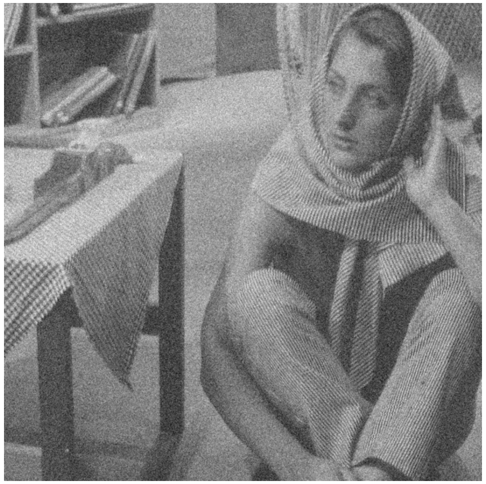


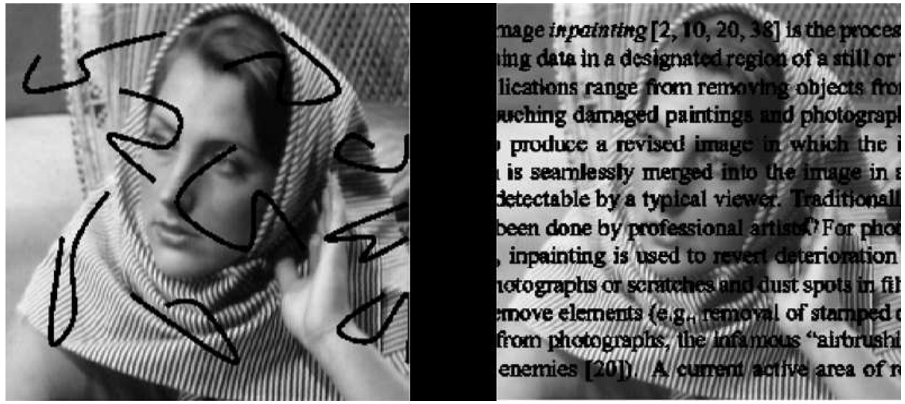


Application to image processing

Assume that we observe a noisy, degraded version of a clean image $\mathbf{t}$:

$$\mathbf{x} = \mathbf{H}\mathbf{t} + \mathbf{e}.$$





We would like to recover the clean image $\mathbf{t}$ from $\mathbf{x}$.

The problem is correspondingly referred to as *image denoising*, *deblurring*, *super-resolution*, and *inpainting*.

The dictionary model for images

Assuming that all images $\mathbf{t}$ are generated from a large dictionary $\mathbf{D}$ (i.e. $\mathbf{t} = \mathbf{D}\alpha$), we rewrite

$$\mathbf{x} = \mathbf{HD}\alpha + \mathbf{e}$$

The clean image $\mathbf{t}$ is recovered from its noisy, degraded version $\mathbf{x}$ by first solving

$$\hat{\alpha} = \arg\min_{\alpha} \|\mathbf{x} - (\mathbf{HD})\alpha\|_2^2 + \lambda||\alpha||_0$$

and then using $\hat{\mathbf{t}} = \mathbf{D}\hat{\alpha}$.

The dictionary model for images

Assuming that all images $\mathbf{t}$ are generated from a large dictionary $\mathbf{D}$ (i.e. $\mathbf{t} = \mathbf{D}\alpha$), we rewrite

$$\mathbf{x} = \mathbf{HD}\alpha + \mathbf{e}$$

The clean image $\mathbf{t}$ is recovered from its noisy, degraded version $\mathbf{x}$ by first solving

$$\hat{\alpha} = \arg\min_{\alpha} \|\mathbf{x} - (\mathbf{HD})\alpha\|_2^2 + \lambda||\alpha||_0$$

and then using $\hat{\mathbf{t}} = \mathbf{D}\hat{\alpha}$.

$\mathbf{H}$ is known, but $\mathbf{D}$ is unknown:

- Typically, it is trained on many (similar) natural images
- Sometimes, it can be self-learned (by operating on the patches of $\mathbf{t}$)

Existing DL methods

Algorithms that have been developed:

- Iterative methods (based on optimization using previous formulations)
  - Method of optimal directions (Engan et al., ICASSP 99')
  - K-SVD (Elad et al., SPARSE 05')
  - Online dictionary learning (Mairal et al., ICML 09')
- Bayesian method (Carin et al., NIPS 09')
- Geometric method (with Maggioni, CISS 10')

Existing DL methods

Algorithms that have been developed:

- Iterative methods (based on optimization using previous formulations)
  - Method of optimal directions (Engan et al., ICASSP 99')
  - **K-SVD** (Elad et al., SPARSE 05')
  - Online dictionary learning (Mairal et al., ICML 09')
- Bayesian method (Carin et al., NIPS 09')
- **Geometric method** (with Maggioni, CISS 10')

The K-SVD algorithm

- K-SVD represents the current state of the art.

The K-SVD algorithm

- K-SVD represents the current state of the art.
- It is a sort of generalization of the K-means algorithm, and thus has many similar properties.

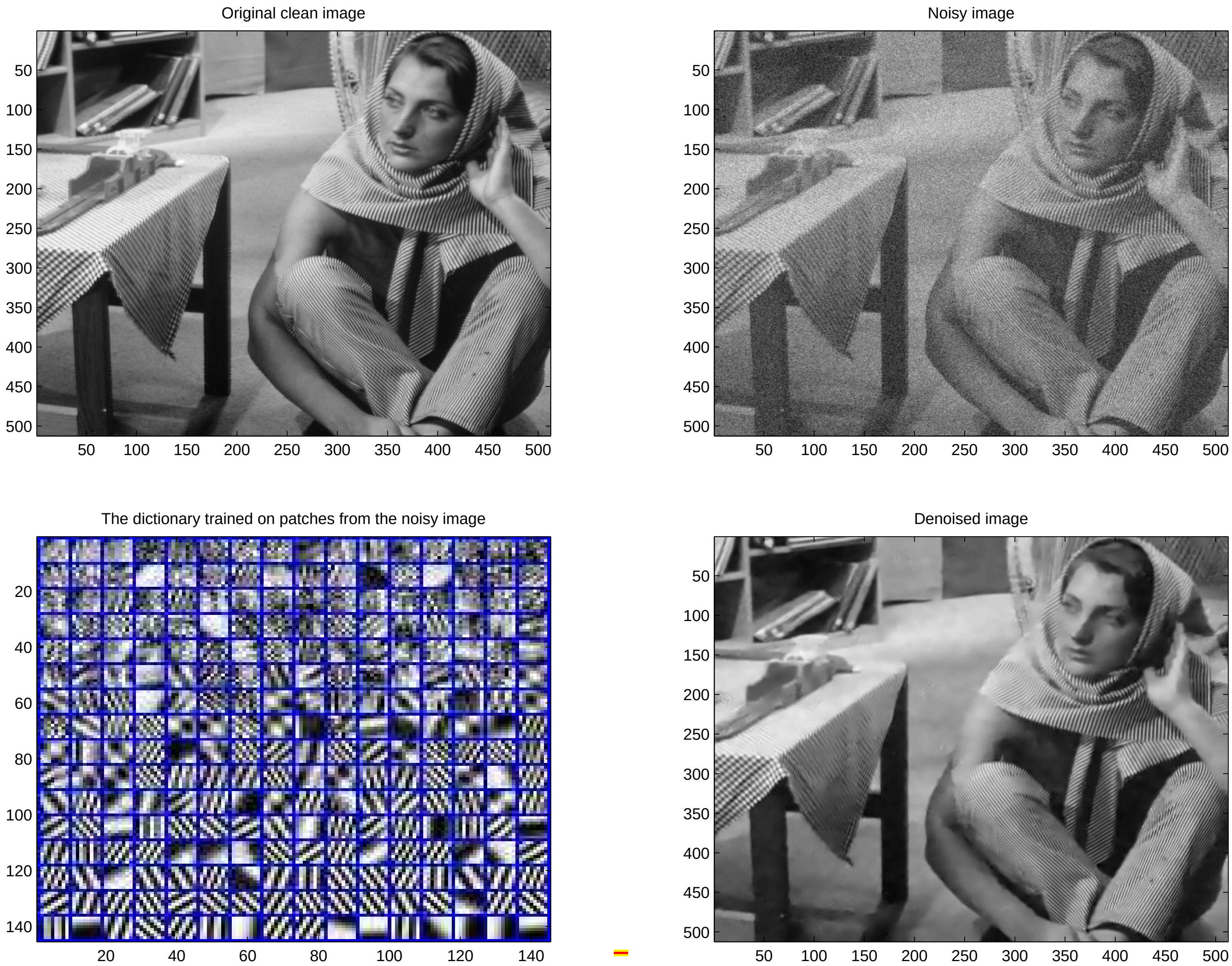
The K-SVD algorithm

- K-SVD represents the current state of the art.
- It is a sort of generalization of the K-means algorithm, and thus has many similar properties.
- With an initial guess of the dictionary, it solves

$$\min_{\mathbf{D},\alpha_1,\ldots,\alpha_n} \sum \|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2 \quad \text{subject to} \quad \|\alpha_i\|_0 \leq s$$

  by alternating between two steps:
  - **Sparse coding** (given $\mathbf{D}$): use pursuit algorithms such as OMP
  - **Dictionary update** (given all α_i): update one atom each time, using only those training signals that need this atom at that step

Example: K-SVD denoising



Example: K-SVD inpainting

Strengths & weaknesses of K-SVD

Advantages:

- Simple to implement and relatively fast to run
- Some sort of local convergence is expected
- Applied successfully to many imaging tasks

Strengths & weaknesses of K-SVD

Advantages:

- Simple to implement and relatively fast to run
- Some sort of local convergence is expected
- Applied successfully to many imaging tasks

Disadvantages:

- Convergence depends on the initial dictionary used
- Dictionary size and sparsity are often arbitrarily picked
- Output dictionary is completely unconstrained and unstructured (making sparse coding very costly)

Overview of GMRA

We build data-dependent dictionaries that are without the previously-mentioned disadvantages.

- We assume that the data follows a manifold model.

Overview of GMRA

We build data-dependent dictionaries that are without the previously-mentioned disadvantages.

- We assume that the data follows a manifold model.
- We construct the dictionary explicitly from data.

Overview of GMRA

We build data-dependent dictionaries that are without the previously-mentioned disadvantages.

- We assume that the data follows a manifold model.
- We construct the dictionary explicitly from data.
- We extend wavelets for 1D signals and PCA for subspaces to nonlinear manifolds in higher dimensions.

Overview of GMRA

We build data-dependent dictionaries that are without the previously-mentioned disadvantages.

- We assume that the data follows a manifold model.
- We construct the dictionary explicitly from data.
- We extend wavelets for 1D signals and PCA for subspaces to nonlinear manifolds in higher dimensions.
- We obtain a structured dictionary which is hierarchically organized.

Overview of GMRA

We build data-dependent dictionaries that are without the previously-mentioned disadvantages.

- We assume that the data follows a manifold model.
- We construct the dictionary explicitly from data.
- We extend wavelets for 1D signals and PCA for subspaces to nonlinear manifolds in higher dimensions.
- We obtain a structured dictionary which is hierarchically organized.
- We show that with our dictionary sparse coding becomes a trivial task.

Overview of GMRA

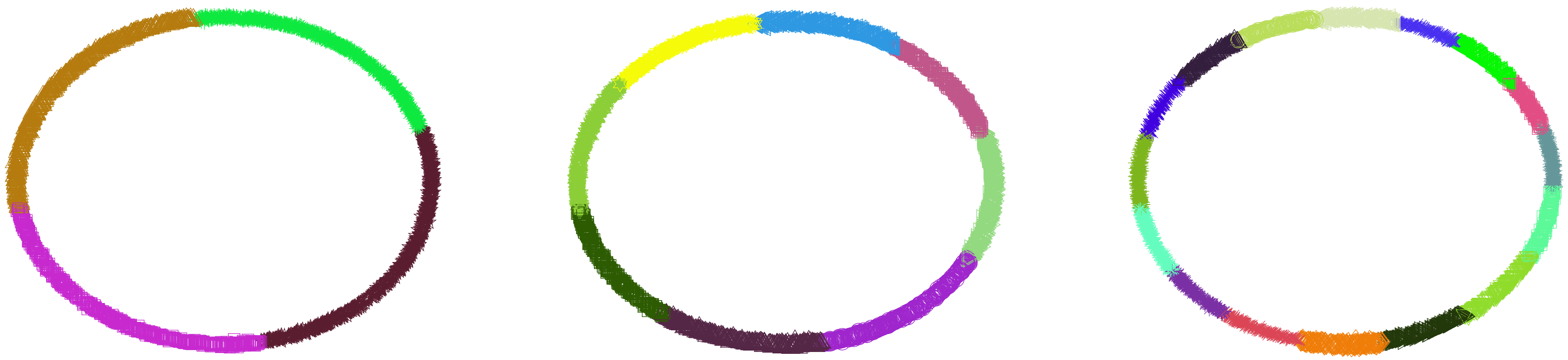
We build data-dependent dictionaries that are without the previously-mentioned disadvantages.

- We assume that the data follows a manifold model.
- We construct the dictionary explicitly from data.
- We extend wavelets for 1D signals and PCA for subspaces to nonlinear manifolds in higher dimensions.
- We obtain a structured dictionary which is hierarchically organized.
- We show that with our dictionary sparse coding becomes a trivial task.
- We derive theoretical guarantees on the dictionary size and coefficient sparsity.

Main steps

Our construction is based on a *geometric multiresolution analysis (GMRA)* of the data:

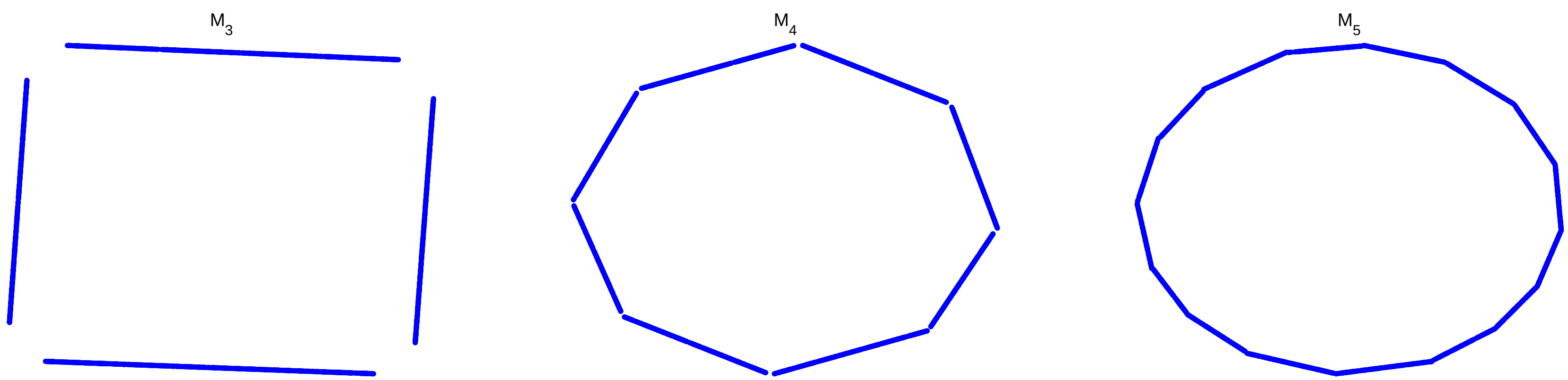
1. A multiscale (nested) spatial decomposition of $\mathcal{M}$ into dyadic cubes at a total of J scales



Main steps

Our construction is based on a *geometric multiresolution analysis (GMRA)* of the data:

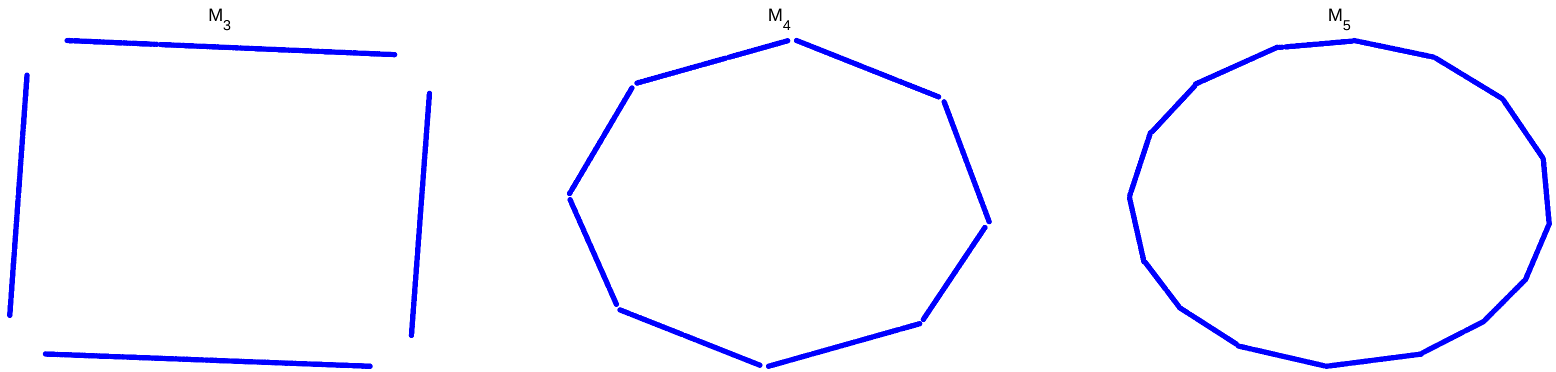
1. A multiscale (nested) spatial decomposition of $\mathcal{M}$ into dyadic cubes at a total of J scales
2. A d-dimensional affine approximation in each cube, yielding a sequence of piecewise linear sets $\mathcal{M}_j$



Main steps

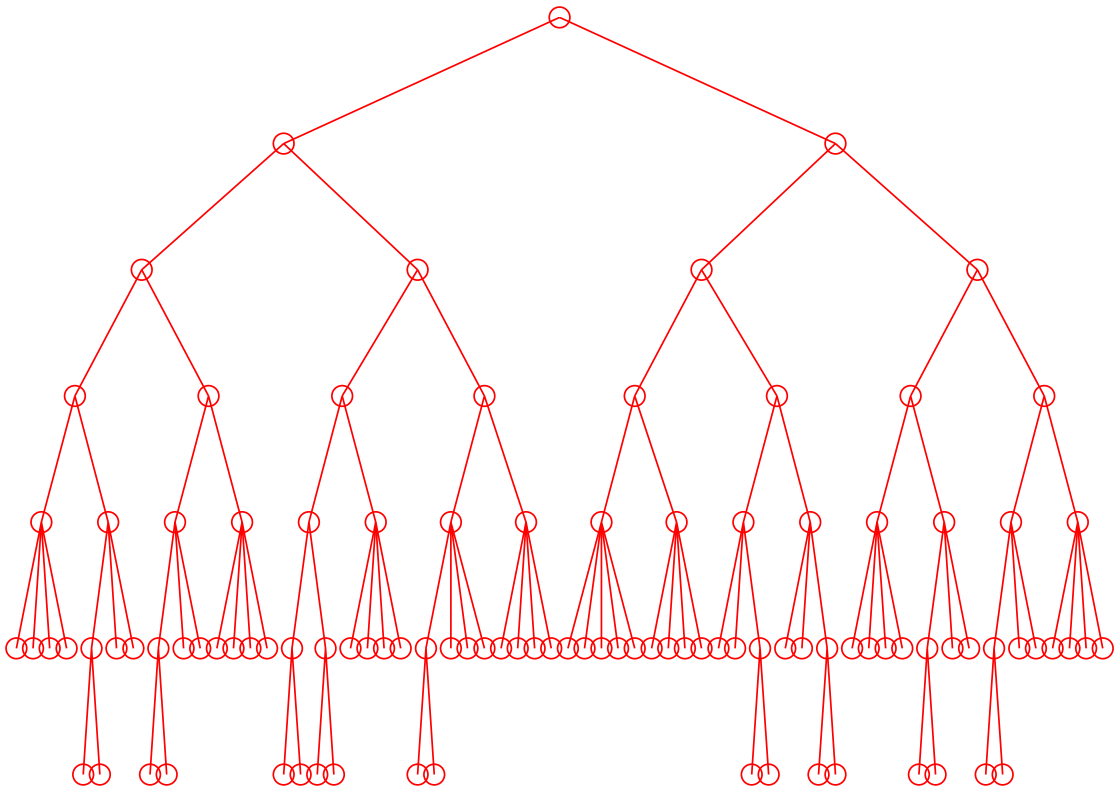
Our construction is based on a *geometric multiresolution analysis (GMRA)* of the data:

1. A multiscale (nested) spatial decomposition of $\mathcal{M}$ into dyadic cubes at a total of J scales
2. A d-dimensional affine approximation in each cube, yielding a sequence of piecewise linear sets $\mathcal{M}_j$
3. A construction of dictionary atoms encoding differences between $\mathcal{M}_j$ and $\mathcal{M}_{j+1}$



The partition tree

There is a natural tree structure associated to the family of dyadic cubes $\{C_{j,k}\}$, with each node representing a cube.



Scaling & wavelet bases

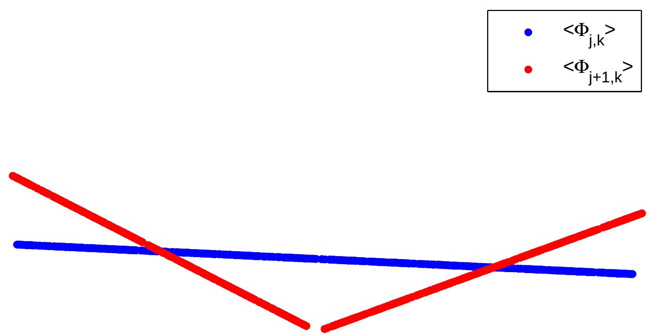
- In every node $C_{j,k}$ of the tree we perform rank-d SVD (after removing the local mean $\overline{c}_{j,k}$). The resulting basis, denoted by $\Phi_{j,k}$, is called the scaling basis.

Scaling & wavelet bases

- In every node $C_{j,k}$ of the tree we perform rank-d SVD (after removing the local mean $\bar{c}_{j,k}$). The resulting basis, denoted by $\Phi_{j,k}$, is called the scaling basis.
- For any $C_{j+1,k'} \subset C_{j,k}$, define

$$W_{j+1,k'} := (I - \Phi_{j,k}\Phi_{j,k}^T)\ \mathrm{colspan}(\Phi_{j+1,k'})$$

and let $\Psi_{j+1,k'}$ be an orthonormal basis for $W_{j+1,k'}$. The $\Psi_{j+1,k'}$ is the "wavelet basis" associated to $C_{j+1,k'}$.



Encoding the differences

Let $\mathbf{x}_i$ represent the projection of $\mathbf{x}$ at scale i, for all i. We can show that

$$\underbrace{\mathbf{x}_{j+1} - \mathbf{x}_j}_{\text{diff across scales}} = \underbrace{\Psi_{j+1,k'}}_{\text{wavelet basis}} \cdot \underbrace{q_{j+1}}_{\text{wavelet coeff}} + \underbrace{\omega_{j+1,k'}}_{\text{wavelet const}},$$

in which

$$q_{j+1} := \Psi^T_{j+1,k'} \left(\mathbf{x}_{j+1} - \overline{c}_{j+1,k'} \right)$$

Encoding the differences

Let $\mathbf{x}_i$ represent the projection of $\mathbf{x}$ at scale i, for all i. We can show that

$$\underbrace{\mathbf{x}_{j+1} - \mathbf{x}_j}_{\text{diff across scales}} = \underbrace{\Psi_{j+1,k'}}_{\text{wavelet basis}} \cdot \underbrace{q_{j+1}}_{\text{wavelet coeff}} + \underbrace{\omega_{j+1,k'}}_{\text{wavelet const}},$$

in which

$$q_{j+1} := \Psi^T_{j+1,k'} \left(\mathbf{x}_{j+1} - \overline{c}_{j+1,k'} \right)$$

This defines a discrete geometric wavelet transform (GWT):

$$x \in \mathcal{M} \longmapsto (q_J, q_{J-1}, \ldots, q_1, q_0) \in \mathbb{R}^{\leq (J+1)d}$$

GMRA (X, d, ϵ)

1) Construct the dyadic cubes $C_{j,k}$ and form a tree $\mathcal{T}$.
2) $J \leftarrow$ finest scale with the ϵ-approximation property.
3) Compute the scaling bases $\Phi_{j,k}$ for all leaf nodes
4) **for** $j = J - 1$ **down to** 1

    for each nonleaf node $C_{j,k}$

        Calculate the associated scaling basis $\Phi_{j,k}$.

        For each child $C_{j+1,k'} \subset C_{j,k}$, find the wavelet basis $\Psi_{j+1,k'}$ and constant $\omega_{j+1,k'}$.

    end

end

5) Set $\Psi_{0,1} = \Phi_{0,1}$ at the root node.
6) Return GMRA = $\{\Psi_{j,k}, \overline{c}_{j,k}, \omega_{j,k}\}$.

Geometric wavelet transforms

$\{q_j\} =$ `ForwardGWT(GMRA,` x`)`

$k \leftarrow$ index of "nearest" leaf node to x

for $j = J$ **down to** 0

$$q_j = \Psi_{j,k}^T \cdot (x - \overline{c}_{j,k})$$

$$x = x - (\Psi_{j,k} \cdot q_j + w_{j,k})$$

$k \leftarrow$ parent(k)

end

$\hat{x} =$ `InverseGWT(GMRA,` $\{q_j\}$`)`

Initialization: $\hat{x} = 0$

for $j = 0 : J$

$$\hat{x} = \hat{x} + (\Psi_{j,k} \cdot q_j + w_{j,k})$$

end

Theoretical guarantees

Theorem. Let $(\mathcal{M}, g)$ be a compact $\mathcal{C}^2$ manifold of dimension d in $\mathbb{R}^D$. Assume $\mathrm{vol}(\mathcal{M}) = 1$ such that there is only one cube at scale 0. Suppose we sample n points from $\mathcal{M}$, and fix a precision $\epsilon > 0$. Then

- The number of scales needed is $J \leq \frac{1}{2} \log_2 \frac{1}{\epsilon}$.
- The size of the GWT for each $\mathbf{x}$ is $\leq (J+1)d$.
- The total cost for storing all coefficients is $\leq nd(J+1)$.
- The dictionary size is $\leq 2d\epsilon^{-\frac{d}{2}}$.

Theoretical guarantees

Theorem. Let $(\mathcal{M}, g)$ be a compact $\mathcal{C}^2$ manifold of dimension d in $\mathbb{R}^D$. Assume $\mathrm{vol}(\mathcal{M}) = 1$ such that there is only one cube at scale 0. Suppose we sample n points from $\mathcal{M}$, and fix a precision $\epsilon > 0$. Then

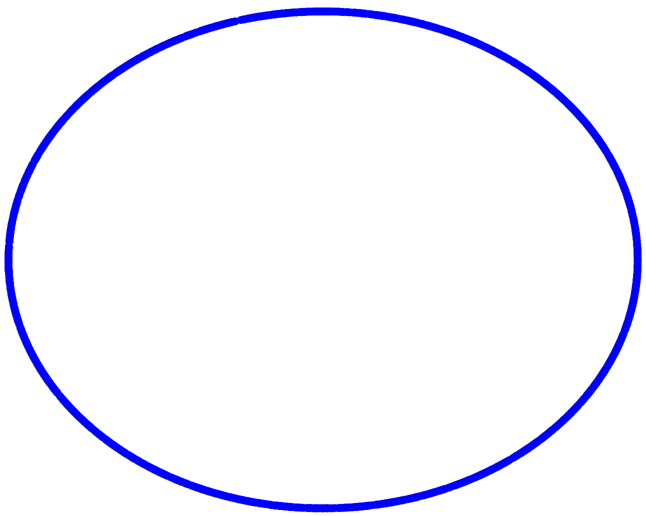
- The number of scales needed is $J \leq \frac{1}{2} \log_2 \frac{1}{\epsilon}$.
- The size of the GWT for each $\mathbf{x}$ is $\leq (J+1)d$.
- The total cost for storing all coefficients is $\leq nd(J+1)$.
- The dictionary size is $\leq 2d\epsilon^{-\frac{d}{2}}$.

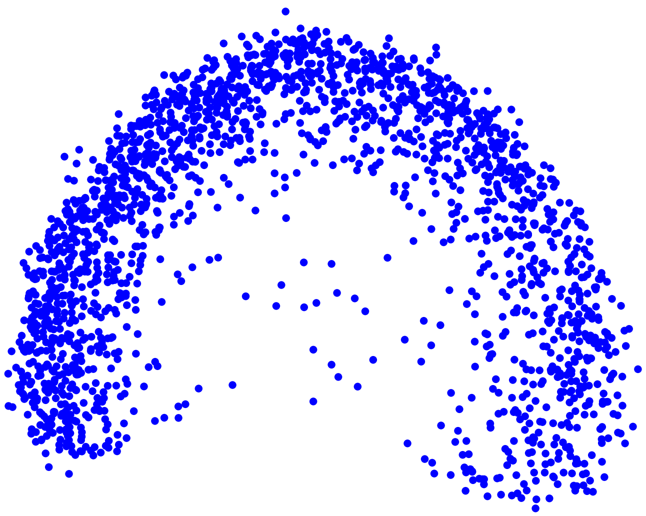
Compare with PCA for linear subspaces:

- The dictionary size is d;
- The total cost for storing all coefficients is nd.

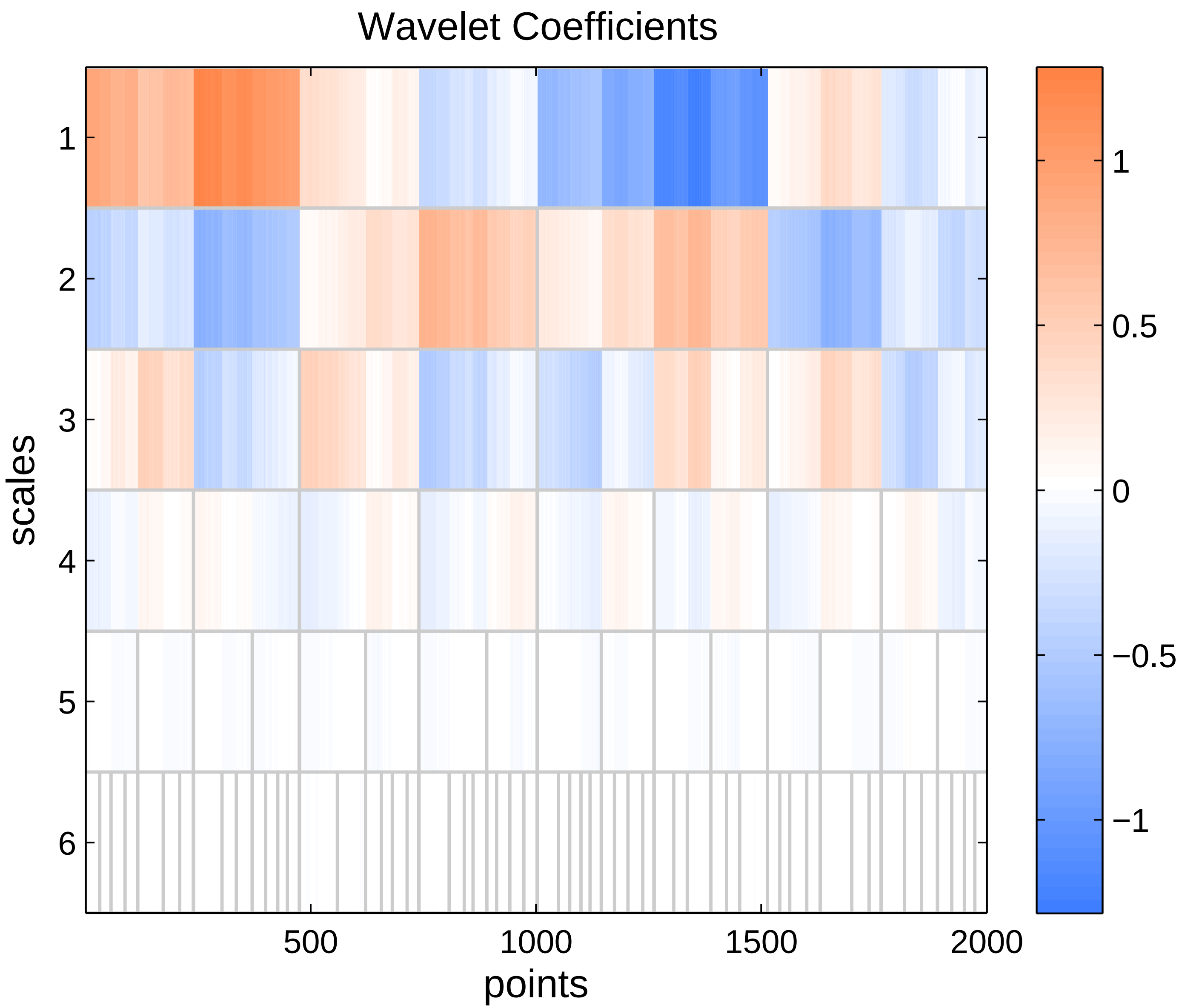
Demonstrations

- 1D circle in $\mathbb{R}^{50}$, 3000 samples, without noise
- 5000 images of the MNIST digit 1, each of size 28×28

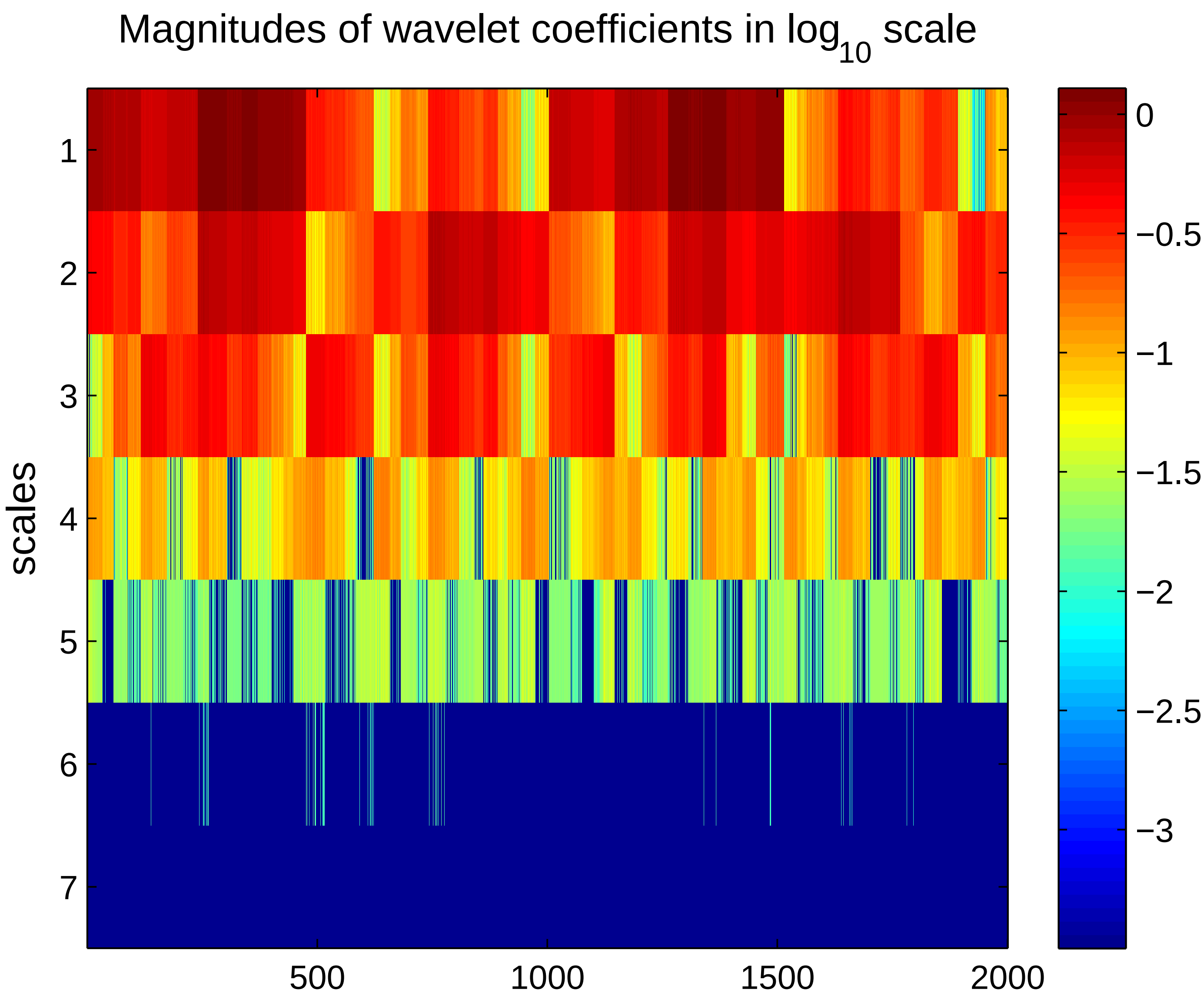




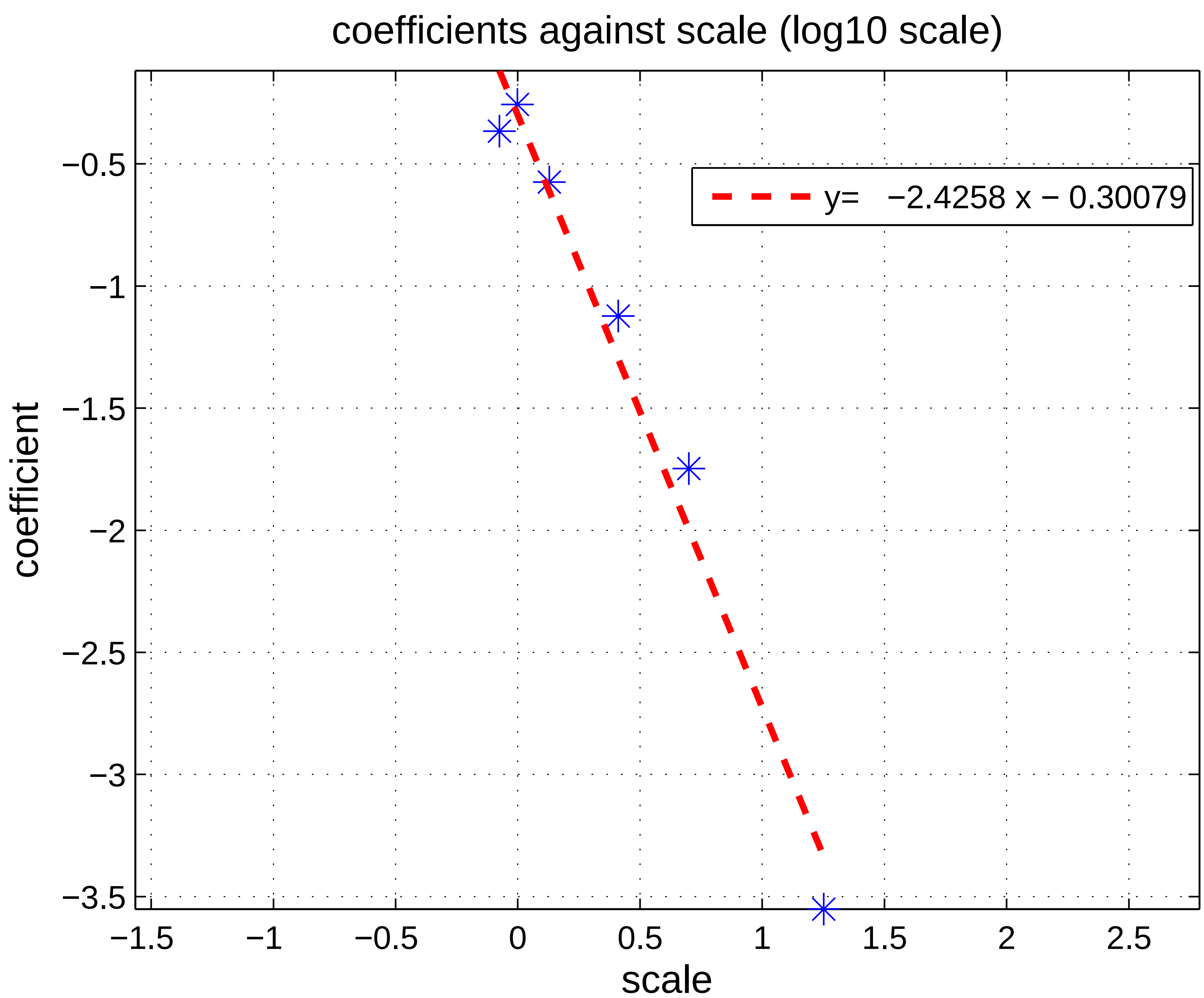
Circle: Wavelet coefficients



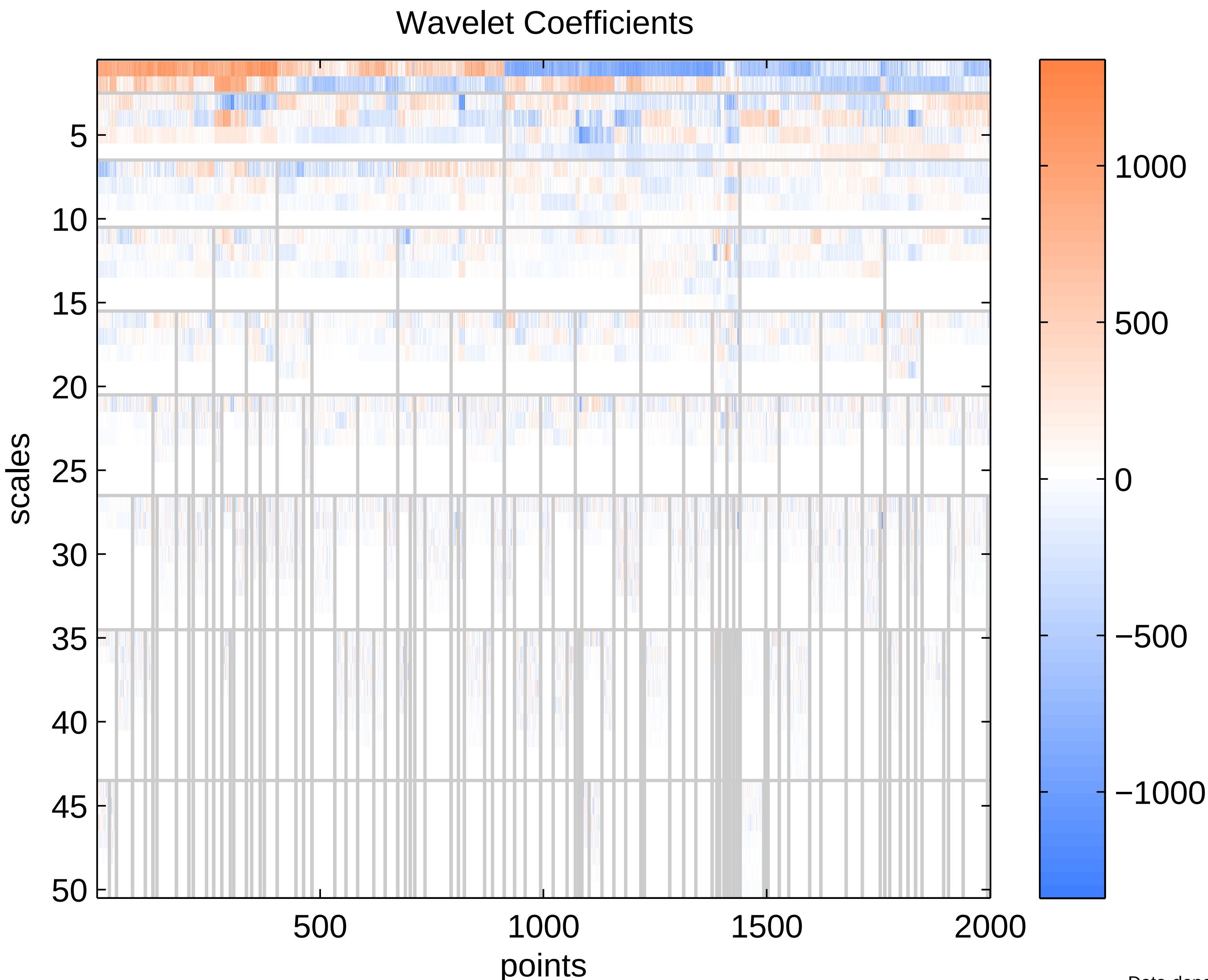
Circle: Wavelet coefficients



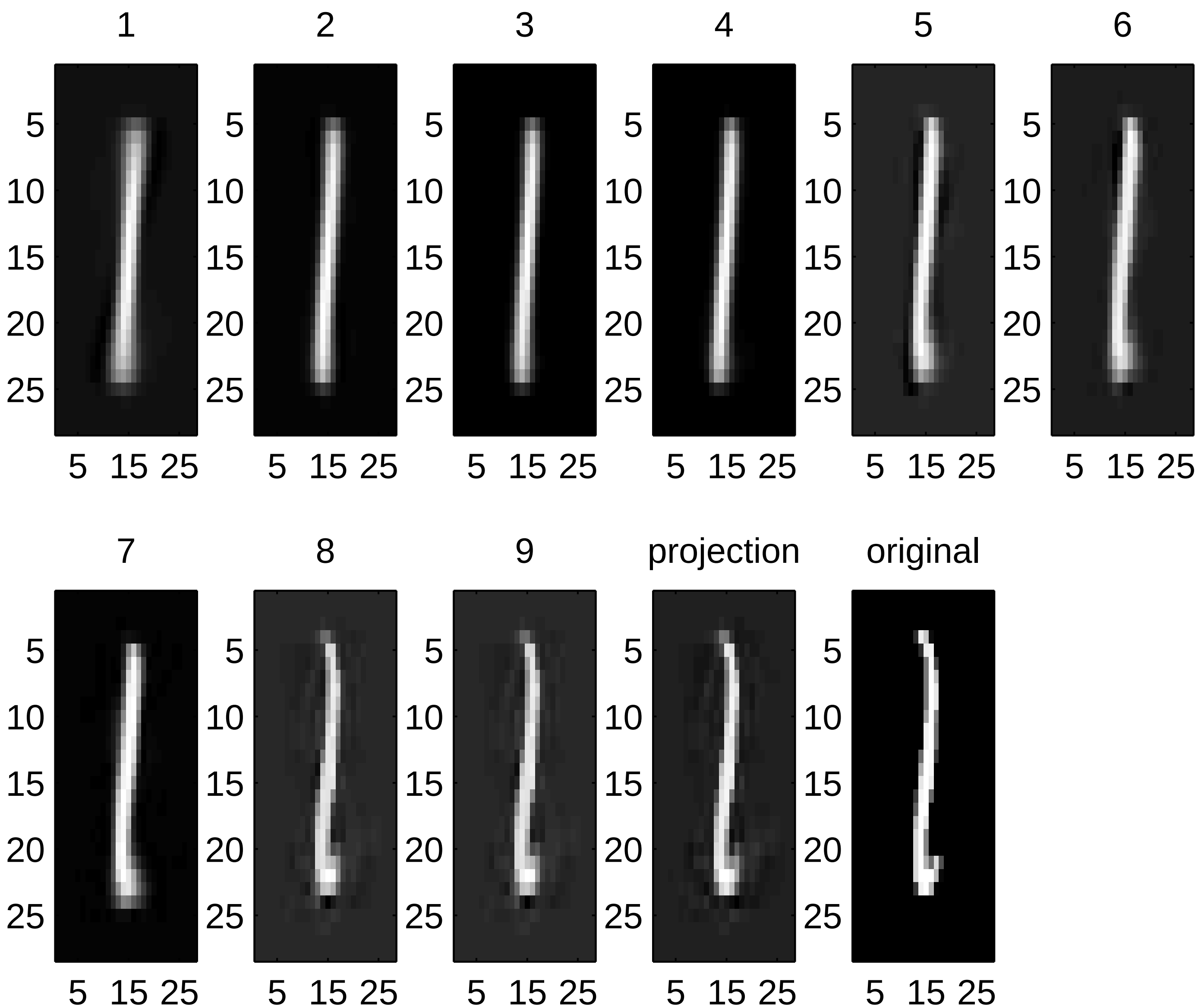
Circle: Wavelet coefficients



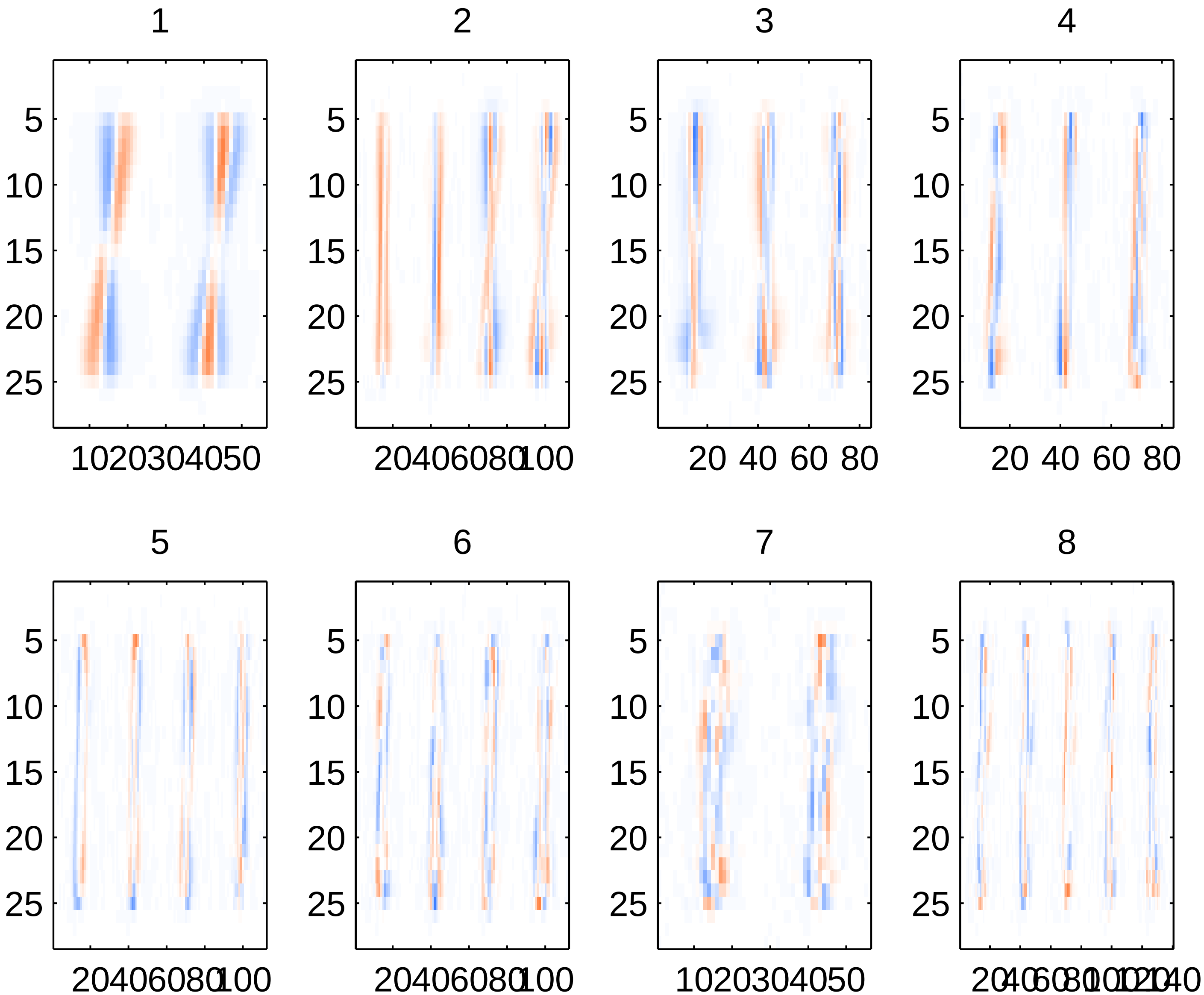
Digit1: Wavelet coefficients



Digit1: Reconstruction of a point



Digit1: Atoms used



Summary and beyond

- Introduced the DL problem + two algorithms
- What is next step in DL?
  - Theoretical justification of DL algorithms
  - Introducing structure to dictionary atoms
  - Imposing structure to representation coefficients
  - Developing next-generation models

Thank you for your attention

- References:
  - Lecture notes (available online)
  - **Dictionary learning**: *Dictionaries for sparse representation modeling. Elad et al., Proceedings of the IEEE, 2010*.
  - **Applications to image processing**: *On the role of sparse and redundant representations in image processing. Elad et al., Proceedings of the IEEE, 2010*.
  - **K-SVD**: *K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation. Elad et al., IEEE Transactions on Signal Processing, 2006*.
  - **GMRA**: *Multiscale Geometric Methods for Data Sets II: Geometric Multi-Resolution Analysis, Appl. Comput. Harmon. Analysis, 2013*
  - **Future directions**: *Sparse and redundant representation modeling – what next? M. Elad, IEEE Signal Processing Letters, 2012.*
- Website: `www.math.sjsu.edu/~gchen`
- Email: guangliang.chen@sjsu.edu