SPARSITY:
COMPRESSED SENSING

Rebecca Willett

DUKE
EDMUND T. PRATT, JR.
SCHOOL OF ENGINEERING

SENSORS, SENSORS EVERYWHERE

Sensing systems limited by constraints: physical size, time, cost, energy

Reduce the number of measurements needed for reconstruction

$\Downarrow$

Higher accuracy data subject to constraints

Original Scene | Downsampled | Reconstruction from ¼ as many measurements

Original Scene | Downsampled | Reconstruction from ¼ as many measurements

Conventional Imaging

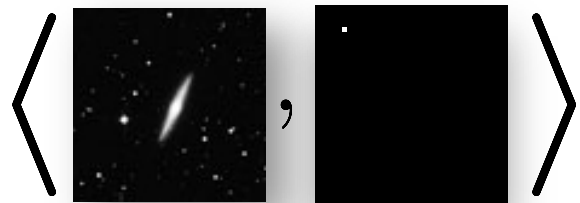
 =  + 

$$y = f + n$$

Each observation is a measurement of ONE pixel

Conventional Imaging

$$y_1 = \langle f, I_1 \rangle =$$



$$y_2 = \langle f, I_2 \rangle =$$

$$\vdots$$

$$y_N = \langle f, I_N \rangle =$$

Each observation is a measurement of ONE pixel

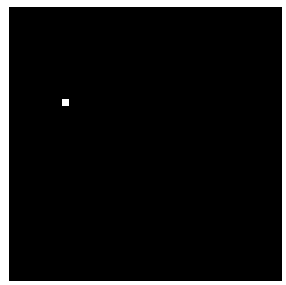
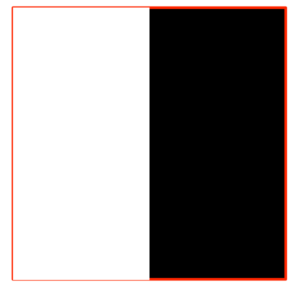
Images are compressible

⇓

Measuring all pixels inherently wasteful

NEW PARADIGM FOR SENSING

$$y_1 = \langle f, r_1 \rangle$$

$= \Big\langle$  ,  $\Big\rangle$

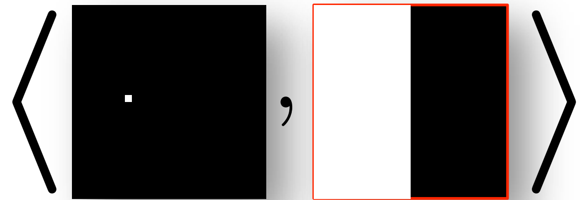
Measure sum of half the pixels

⇓

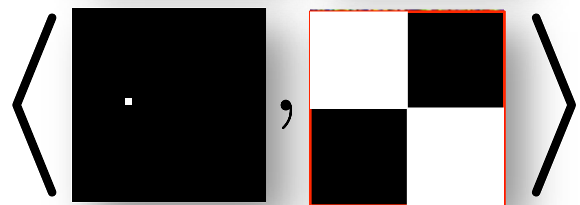
Narrow down star location

NEW PARADIGM FOR SENSING

$$y_1 = \langle f, r_1 \rangle =$$

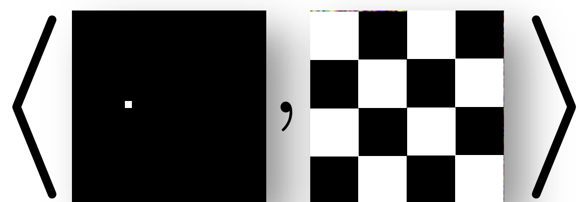


$$y_2 = \langle f, r_2 \rangle =$$



$$\vdots$$

$$y_M = \langle f, r_M \rangle =$$



Each observation is a measurement of half the pixels

NEW PARADIGM FOR SENSING

$y_1 = \langle f, r_1 \rangle =$ $\langle$  ,  $\rangle$

$y_2 = \langle f, r_2 \rangle =$ $\langle$  ,  $\rangle$

$\vdots$

$y_M = \langle f, r_M \rangle =$ $\langle$  ,  $\rangle$

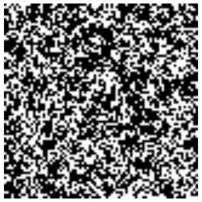
These ideas extend to multiple stars and random combinations of pixels

NEW OBSERVATION MODEL

$$y = Rf + n$$

Observations

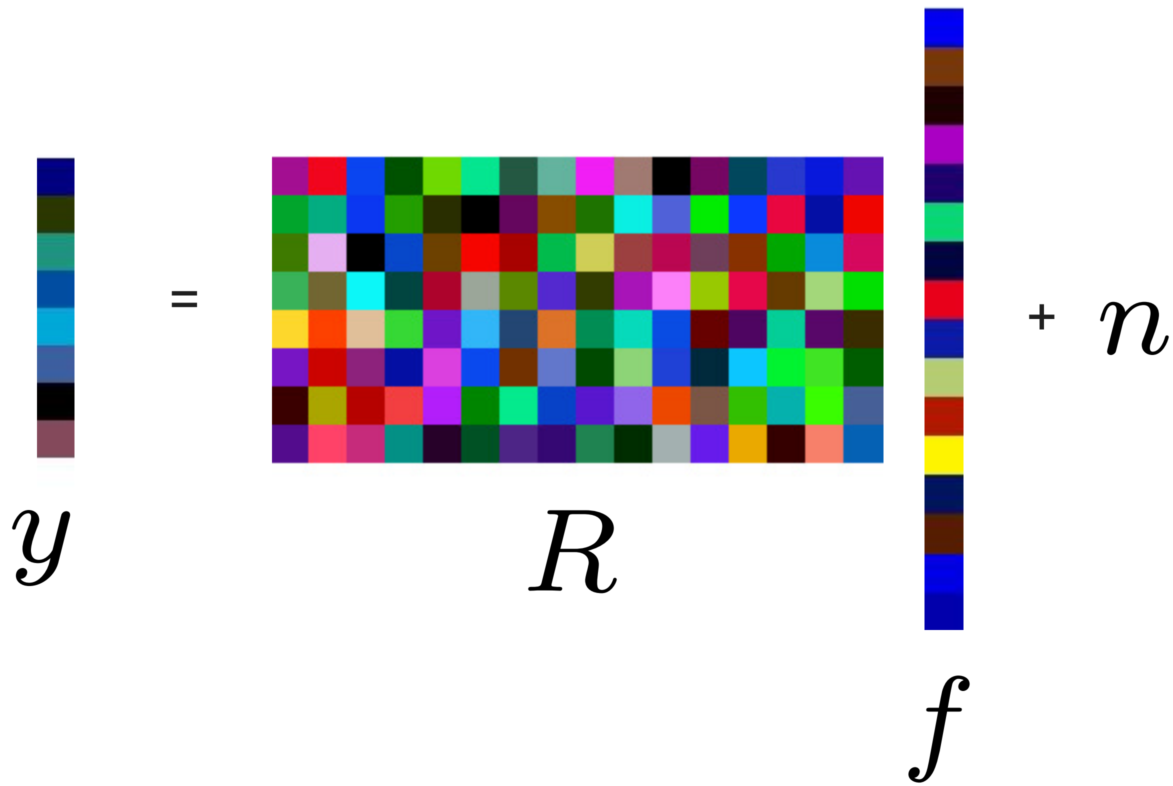
All k random projections



True image

Noise

ILL-POSED PROBLEM



System is underdetermined:
infinitely many solutions

SPARSITY

Assume f is K-sparse or β-compressible in some basis Ψ. That is,

$$f = \sum_{i=1}^{N} \theta_i \psi_i$$

and either

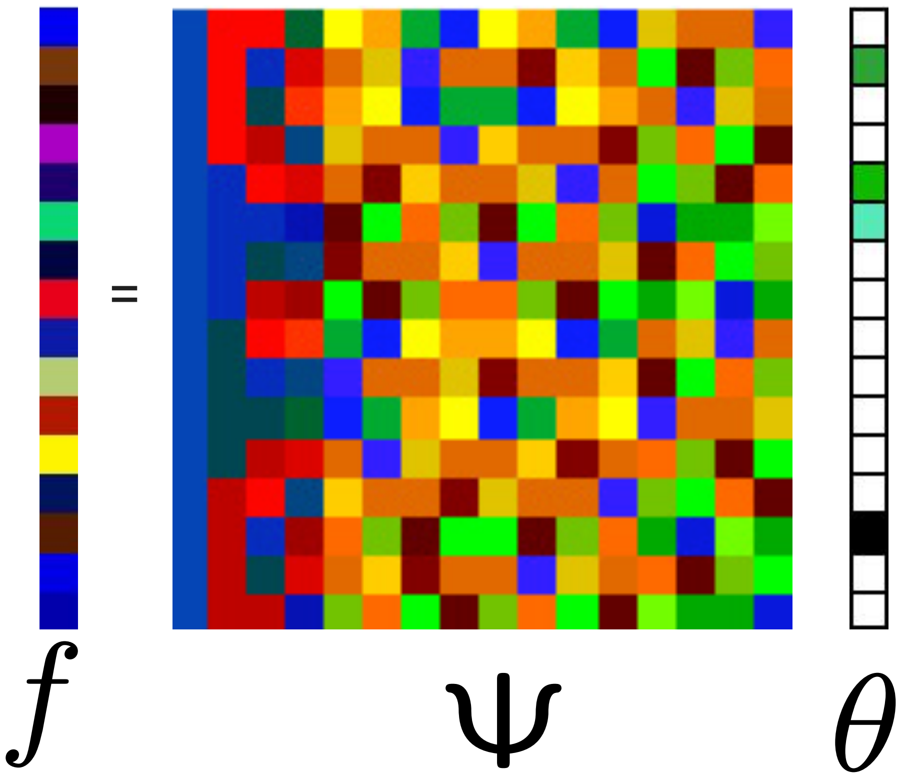
$$\|\theta\|_0 \leq K$$

or

$$\|f - f_K\| \preceq K^{-\beta}$$

where f_K is the best K-term approximation of f in the basis Ψ.

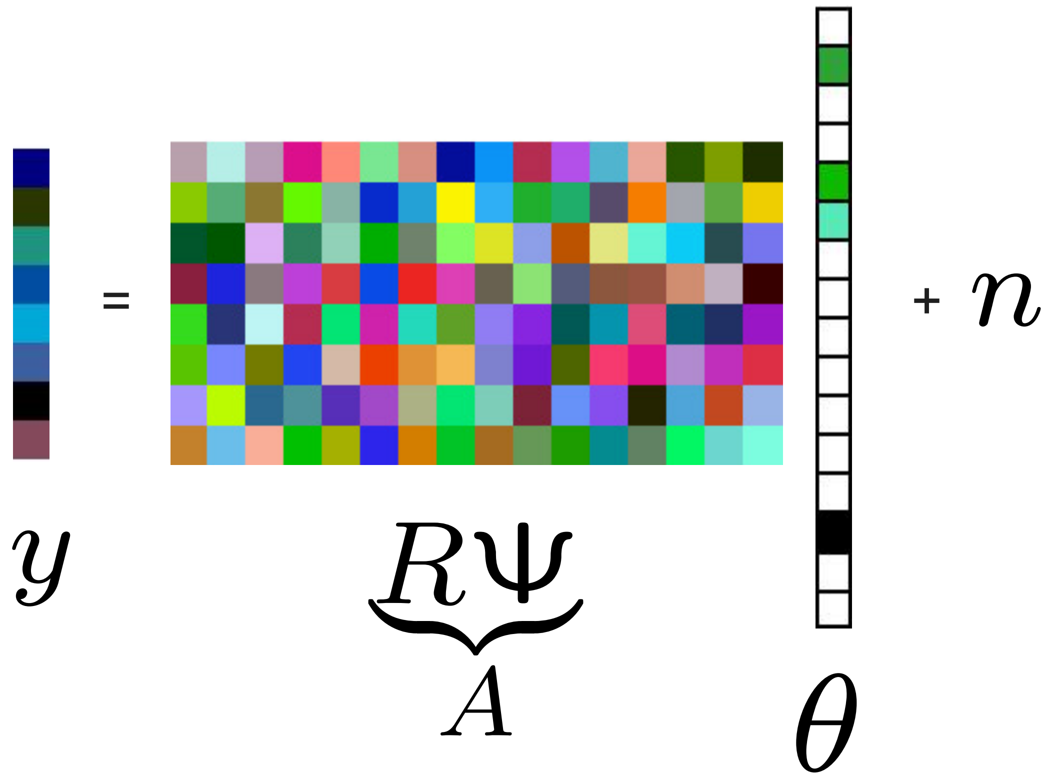
Sparsity



$$f = \Psi\theta$$

Sparse Inverse Problem

Combining $y = Rf + n$ with $f = \Psi\theta$:



Compressed Sensing

$$\widehat{\theta} = \arg\min_{\theta} \underbrace{\|y - R\Psi\theta\|_2^2}_{\text{data fit}} + \underbrace{\tau\|\theta\|_1}_{\text{sparsity}}$$

$$\widehat{f} = \Psi\widehat{\theta}$$

Key theory: If R meets certain conditions and f is sparse or compressible in Ψ, then $\widehat{f}$ will be very accurate even when the number of measurements is small relative to N.

Conventional Sensing

Noisy Image

Compressed Sensing

Random Projections
Smaller
Less Data
Cheaper

Restricted Isometry Property

Definition: Restricted Isometry Property. The matrix A satisfies the Restricted Isometry Property of order K with parameter $\delta_K \in [0, 1)$ if

$$(1 - \delta_K)\|\theta\|_2^2 \leq \|A\theta\|_2^2 \leq (1 + \delta_K)\|\theta\|_2^2$$

holds simultaneously for all K-sparse vectors θ. Matrices with this property are denoted RIP(K, δ_K).

Candes and Tao (2006)

RIP EXAMPLE

For example, if the entries of A are independent and identically distributed according to

$$A_{i,j} \sim \mathcal{N}\left(0, \frac{1}{M}\right) \quad \text{or} \quad A_{i,j} = \begin{cases} M^{-1/2} & \text{with probability} \\ -M^{-1/2} & \text{with probability} \end{cases}$$

then A satisfies RIP(K, δ_K) with high probability for any integer $K = O(M/\log N)$.

Sparse recovery

Matrices which satisfy the RIP combined with sparse recovery algorithms are guaranteed to yield accurate estimates of the underlying function f, as specified by the following theorem.

Theorem: Noisy Sparse Recovery with RIP Matrices. Let A be a matrix satisfying RIP$(2K, \delta_{2K})$ with $\delta_{2K} < \sqrt{2} - 1$, and let $y = A\theta + n$ be a vector of noisy observations of any signal $\theta \in \mathbb{R}^N$, where the n is a noise or error term with $\|n\|_2 \leq \epsilon$. Let θ_K be the best K-sparse approximation of θ. Then the estimate

$$\widehat{\theta} = \arg\min_{\theta} \|\theta\|_1 \text{ subject to } \|y - A\theta\|_2 \leq \epsilon$$

obeys

$$\|\theta - \widehat{\theta}\|_2 \leq C_{1,K}\epsilon + C_{2,K}\frac{\|\theta - \theta_K\|_1}{\sqrt{K}},$$

where $C_{1,K}$ and $C_{2,K}$ are constants which depend on K but not on N or M.

Candes (2006)

ALGORITHMS $\widehat{\theta} = \arg\min_{\widetilde{\theta}} \|y - A\widetilde{\theta}\|_2^2 + \tau\|\widetilde{\theta}\|_1.$

- This estimate can be computed in a variety of ways.
- Many off-the-shelf optimization software packages are unsuitable
  - Can't handle large N
  - Our objective isn't differentiable
  - Don't exploit fast transforms (e.g. Fourier and wavelet)
- Gradient projection methods
  - Introduce additional variables and recast problem as constrained optimization with differentiable objective
  - Projection onto constraint set can be done with thresholding
  - More robust to noise
- Orthogonal matching pursuits (OMP)
  - Start with estimate = 0
  - Greedily choose elements of estimate to have non-zero magnitude by iteratively processing residual errors
  - Very fast when little noise

Iterative Hard/Soft Thresholding

Our objective is

$$\widehat{\theta} = \arg\min_{\widetilde{\theta}} \|y - A\widetilde{\theta}\|_2^2 + \tau\|\widetilde{\theta}\|_1.$$

The first term can be re-written as

$$y^T y - 2\widetilde{\theta}^T A^T y + \widetilde{\theta}^T A^T A\widetilde{\theta}$$

and its gradient is

$$-2A^T(y - A\widetilde{\theta}).$$

This suggests a simple strategy for computing $\widehat{\theta}$: start with an initial estimate $\widetilde{\theta}$, update it by adding a step in the negative gradient direction, then apply thresholding!

Iterative Hard/Soft Thresholding

Start with some initial estimate $\widehat{\theta}^{(0)}$; see how well it fits y:

$$y - A\widehat{\theta}^{(0)}.$$

Use this residual to update the initial estimate:

$$\widehat{\theta}^{(0)} + A^T \left(y - A\widehat{\theta}^{(0)} \right).$$

Impose sparsity via thresholding this estimate:

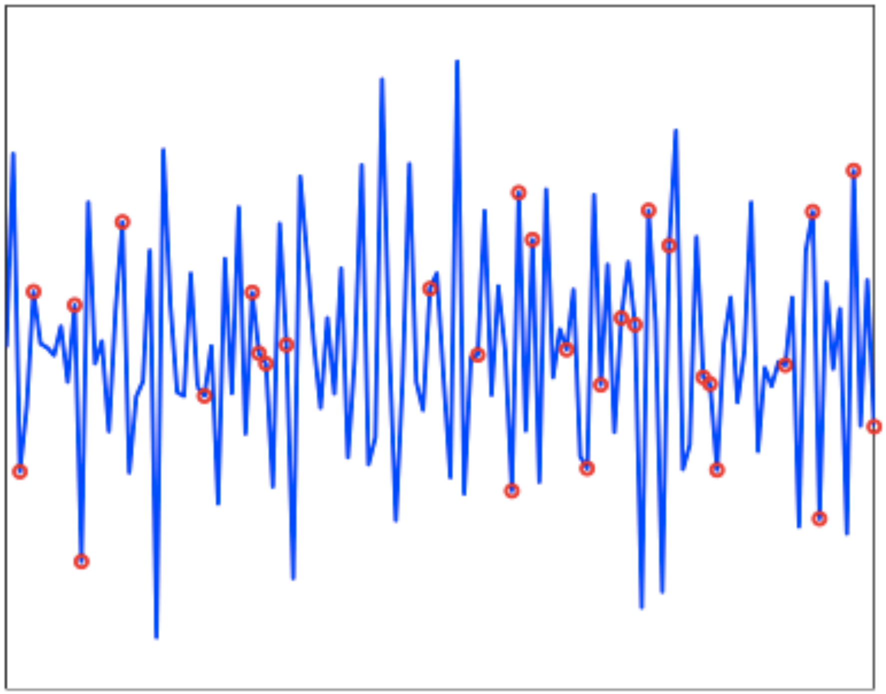
$$\widehat{\theta}^{(1)} = \text{threshold} \left[\widehat{\theta}^{(0)} + A^T \left(y - A\widehat{\theta}^{(0)} \right) \right]$$

Repeat until $\|y - A\widehat{\theta}^{(i)}\|$ is small:

$$\widehat{\theta}^{(i+1)} = \text{threshold} \left[\widehat{\theta}^{(i)} + A^T \left(y - A\widehat{\theta}^{(i)} \right) \right].$$

EXAMPLE

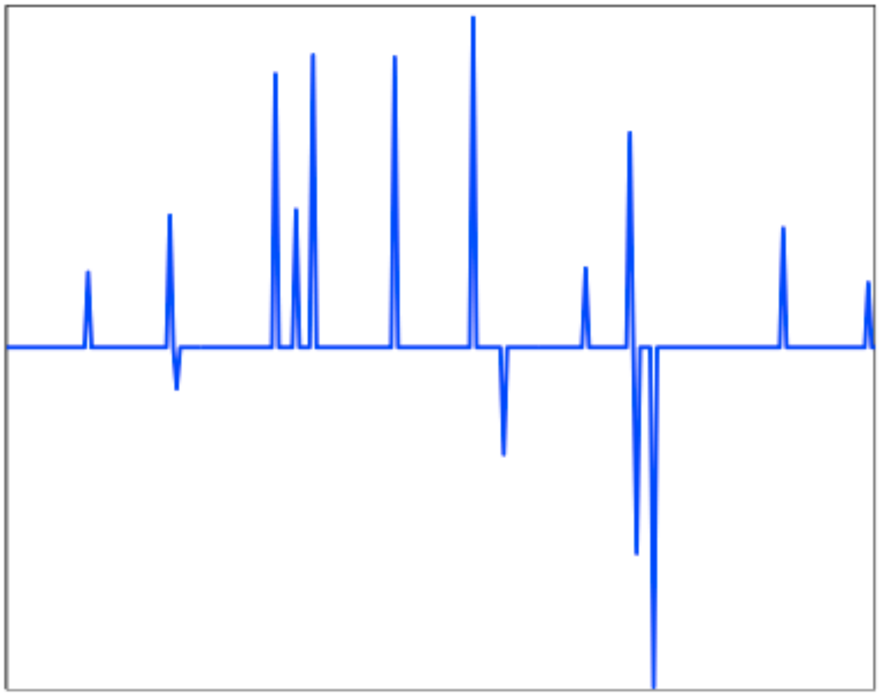
Time domain $f(t)$



Measure M samples

(red circles = samples)

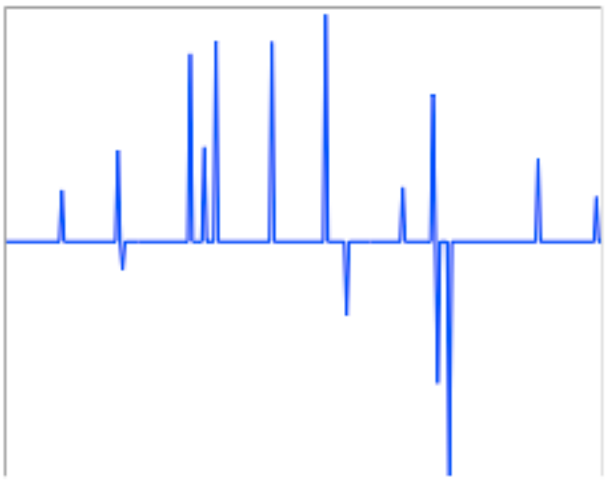
Frequency domain $\hat{f}(\omega)$



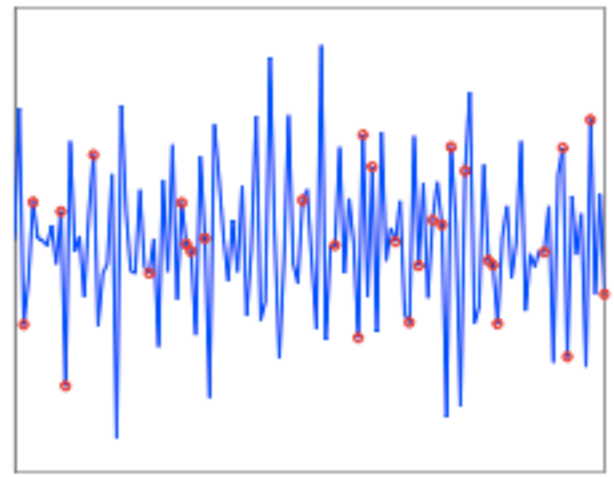
K nonzero components

$\#\{\omega : \hat{f}(\omega) \neq 0\} = K$

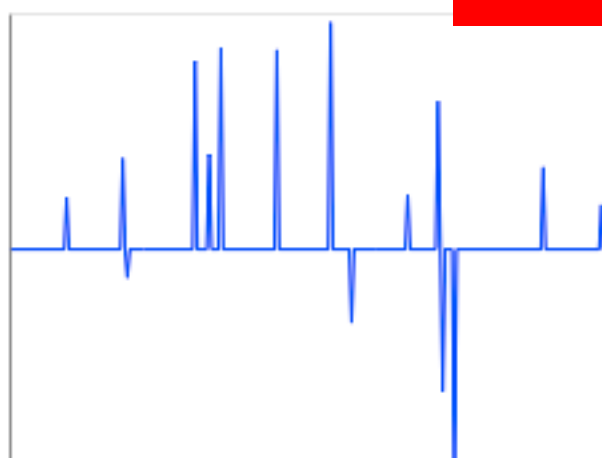
EXAMPLE



Original θ, with $K = 15$



f (blue) and y (red circles); $M = 30$



perfect reconstruction!

SPARSE RECOVERY

Matrices which satisfy the RIP combined with sparse recovery algorithms are guaranteed to yield accurate estimates of the underlying function f, as specified by the following theorem.

Theorem: Noisy Sparse Recovery with RIP Matrices. Let A be a matrix satisfying RIP$(2K, \delta_{2K})$ with $\delta_{2K} < \sqrt{2} - 1$, and let $y = A\theta + n$ be a vector of noisy observations of any signal $\theta \in \mathbb{R}^N$, where the n is a noise or error term with $\|n\|_2 \leq \epsilon$. Let θ_K be the best K-sparse approximation of θ. Then the estimate

$$\widehat{\theta} = \arg\min_{\theta} \|\theta\|_1 \text{ subject to } \|y - A\theta\|_2 \leq \epsilon$$

obeys

$$\|\theta - \widehat{\theta}\|_2 \leq C_{1,K}\epsilon + C_{2,K}\frac{\|\theta - \theta_K\|_1}{\sqrt{K}},$$

where $C_{1,K}$ and $C_{2,K}$ are constants which depend on K but not on N or M.

Candes (2006), Candes, Romberg & Tao (2006)

Proof

Let $h \triangleq \widehat{\theta} - \theta$ be our error vector.

Let T_0 be the indicies of the largest K elements of θ, T_1 be the indicies of the largest K elements of $h_{T_0^c}$, T_2 be the indicies of the next K largest elements of $h_{T_0^c}$, and so on. For a vector x, let x_{T_j} be defined via

$$x_{T_j,i} \triangleq \begin{cases} x_i, & i \in T_j \\ 0, & i \notin T_j \end{cases}.$$

Then $h = h_{T_0} + h_{T_1} + h_{T_2} + \ldots$

There are two main steps to our proof:

$$\begin{aligned} \|\widehat{\theta} - \theta\|_2 = \|h\|_2 &\leq \|h_{T_0 \cup T_1}\|_2 + \|h_{(T_0 \cup T_1)^c}\|_2 \\ \text{(STEP 1)} \quad &\leq C\|h_{T_0 \cup T_1}\|_2 + CK^{-1/2}\|\theta - \theta_K\|_1 \\ \text{(STEP 2)} \quad &\leq C\epsilon + CK^{-1/2}\|\theta - \theta_K\|_1 \end{aligned}$$

C will represent constants which may depend on K but not N or M.

Step 1

$$
\begin{aligned}
\|h_{(T_0 \cup T_1)^c}\|_2 &= \left\| \sum_{j \geq 2} h_{T_j} \right\|_2 \\
&\leq \sum_{j \geq 2} \|h_{T_j}\|_2 \qquad \text{(remember me later!!)} \\
&\leq \sum_{j \geq 2} K^{1/2} \|h_{T_j}\|_\infty \\
&\leq \sum_{j \geq 2} K^{1/2} \|h_{T_{j-1}}\|_1 / K \\
&= K^{-1/2} \left(\|h_{T_1}\|_1 + \|h_{T_2}\|_1 + \ldots \right) \\
&= K^{-1/2} \underbrace{\|h_{T_0^c}\|_1}_{\text{how big??}}
\end{aligned}
$$

STEP 1

First note

$$\begin{aligned}\|\theta\|_1 \geq \|\widehat{\theta}\|_1 &= \|\theta + h\|_1 \\ &\geq \|\theta_{T_0}\|_1 - \|h_{T_0}\|_1 + \|h_{T_0^c}\| - \|\theta_{T_0^c}\|_1\end{aligned}$$

Rearranging terms we find

$$\begin{aligned}\|h_{T_0^c}\|_1 &\leq \|h_{T_0}\|_1 + 2\|\theta_{T_0^c}\|_1 \\ &= \|h_{T_0}\|_1 + 2\|\theta - \theta_K\|_1\end{aligned}$$

Putting everything together we have

$$\begin{aligned}\|h_{(T_0 \cup T_1)^c}\|_2 &\leq K^{-1/2}(\|h_{T_0}\|_1 + 2\|\theta - \theta_K\|_1) \\ &\leq \|h_{T_0 \cup T_1}\|_2 + 2K^{-1/2}\|\theta - \theta_K\|_1\end{aligned}$$

as desired for Step 1.

Step 2

We now need to bound $\|h_{T_0 \cup T_1}\|_2$. Note

$$
\begin{aligned}
(1-\delta_{2K})\|h_{T_0 \cup T_1}\|_2^2 &\leq \|Ah_{T_0 \cup T_1}\|_2^2 \\
&= \langle Ah_{T_0 \cup T_1}, Ah \rangle - \langle Ah_{T_0 \cup T_1}, \sum_{j \geq 2} Ah_{T_j} \rangle
\end{aligned}
$$

For the first term

$$
\begin{aligned}
\langle Ah_{T_0 \cup T_1}, Ah \rangle &\leq \|Ah_{T_0 \cup T_1}\|_2 \|Ah\|_2 \\
&\leq (\sqrt{1+\delta_{2K}}\|h_{T_0 \cup T_1}\|_2)\|A(\widehat{\theta}-\theta)\|_2 \\
&\leq (\sqrt{1+\delta_{2K}}\|h_{T_0 \cup T_1}\|_2)(\|A\widehat{\theta}-y\|_2 + \|y - A\theta\|_2) \\
&\leq (\sqrt{1+\delta_{2K}}\|h_{T_0 \cup T_1}\|_2)2\epsilon
\end{aligned}
$$

The second term is bounded similarly by

$$
-\langle Ah_{T_0 \cup T_1}, \sum_{j \geq 2} Ah_{T_j} \rangle \leq \sqrt{2}\delta_{2K} \sum_{j \geq 2} \|h_{T_j}\|_2 \|h_{T_0 \cup T_1}\|_2
$$

STEP 2

Thus

$$(1-\delta_{2K})\|h_{T_0\cup T_1}\|_2^2 \leq \|h_{T_0\cup T_1}\|_2 \left(2\epsilon\sqrt{1+\delta_{2K}} + \sqrt{2}\delta_{2K}\sum_{j\geq 2}\|h_{T_j}\|_2\right)$$

$$\|h_{T_0\cup T_1}\|_2 \leq C\epsilon + CK^{-1/2}\|\theta-\theta_K\|_1.$$

Putting it all together we have

$$\|\widehat{\theta}-\theta\|_2 \leq C\epsilon + CK^{-1/2}\|\theta-\theta_K\|_1$$

as desired.

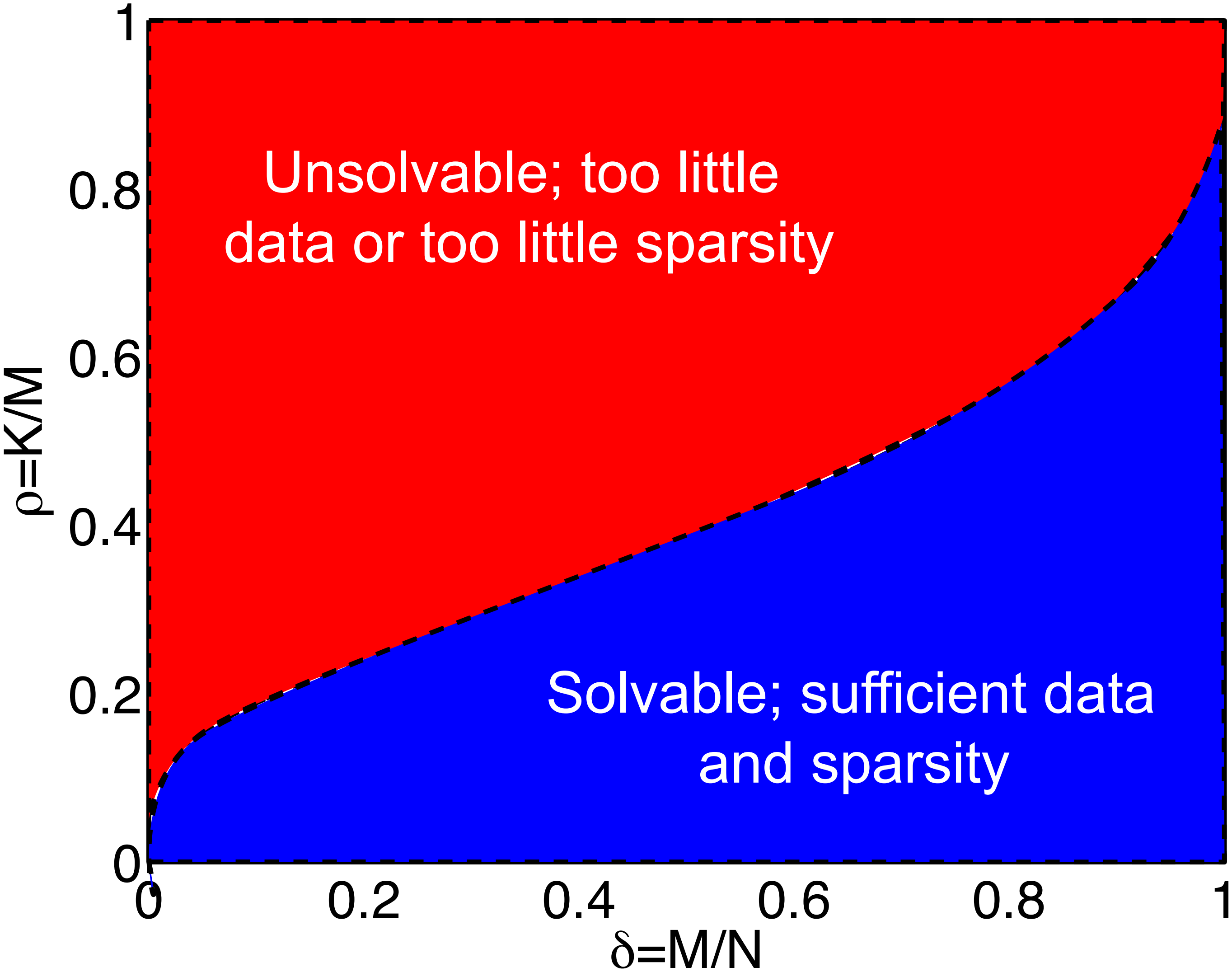
In other words, the accuracy of the reconstruction of a general image f from measurements collected using a system which satisfies the RIP depends on (a) the amount of noise present and (b) how well f may be approximated by an image sparse in Ψ.

If we have no noise ($\epsilon = 0$) and our signal is K-sparse, then we have

$$\theta = \widehat{\theta};$$

i.e., we can perfectly reconstruct the original signal!

SOLVABILITY BOUNDARY



Donoho and Tanner (2010)

ANOTHER PERSPECTIVE

Consider the worst-case coherence of $A \equiv R\Psi$. Formally, one denotes the Gram matrix $G \triangleq A^T A$ and let

$$\mu(A) \triangleq \max_{\substack{1 \le i,j \le \\ i \neq j}} |\langle G_{i,j} \rangle|$$

be the largest off-diagonal element of the Gram matrix. A good goal in designing a sensing matrix is to therefore choose R and Ψ so that μ is as close as possible to $N^{-1/2}$.

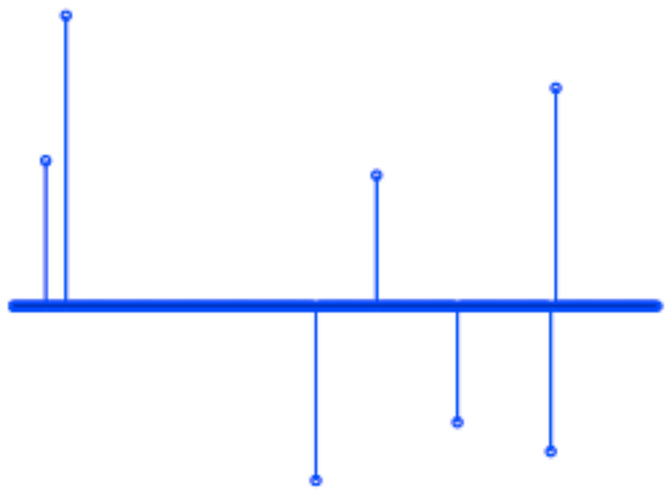
Theorem: Noisy Sparse Recovery with Incoherent Matrices. Let $y = A\theta + n$ be a vector of noisy observations of any K-sparse signal $\theta \in \mathbb{R}^N$, where $K \le (\mu(A)^{-1} + 1)/4$ and the n is a noise or error term with $\|n\|_2 \le \epsilon$. Then our estimate obeys

$$\|\theta - \widehat{\theta}\|_2^2 \le \frac{4\epsilon^2}{1 - \mu(A)(4K - 1)}.$$

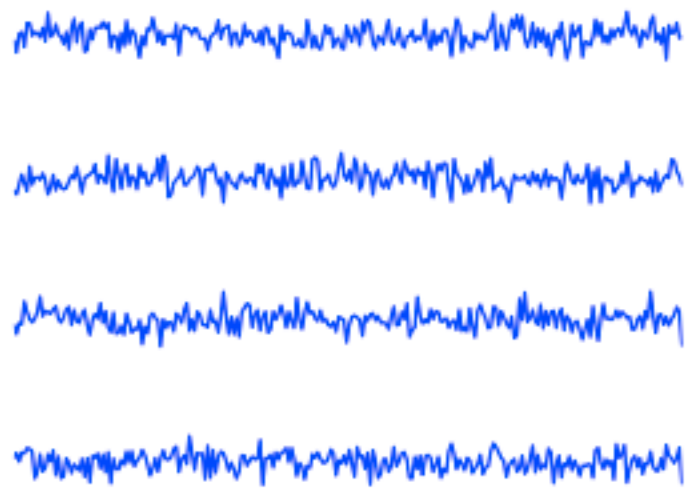
Tropp (2004), Donoho, Elad & Temlyakov (2006),
Donoho & Huo (2001), Gilbert, Muthukrishnan & Strauss (2003)

INCOHERENT MEASUREMENT

Sparse vector



Projection vectors

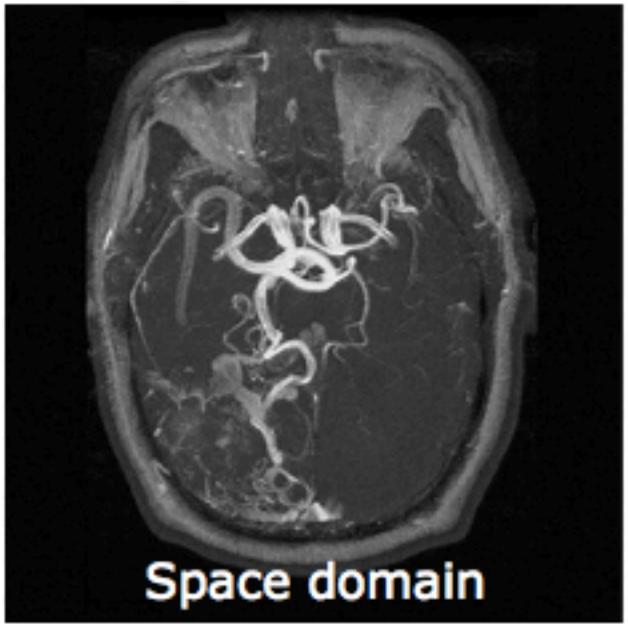


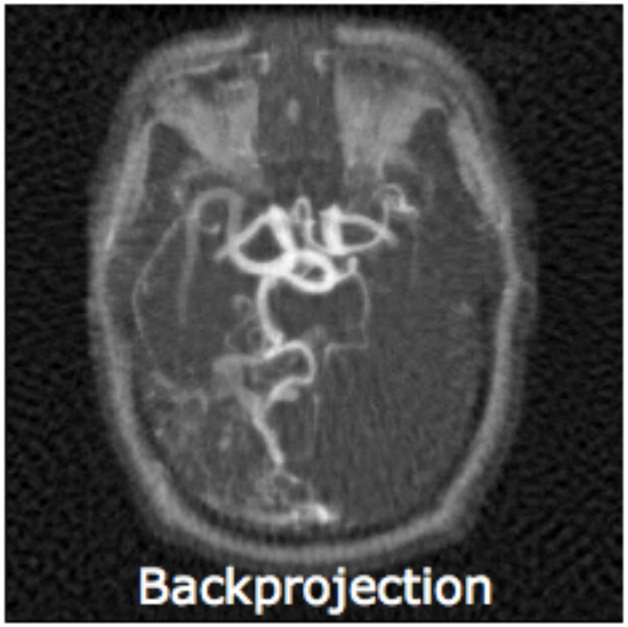
Signal is locally concentrated, measurements are global

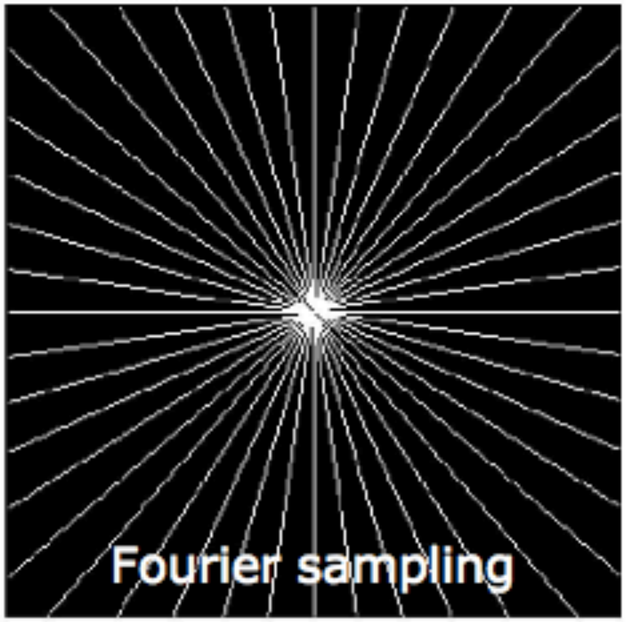
⇓

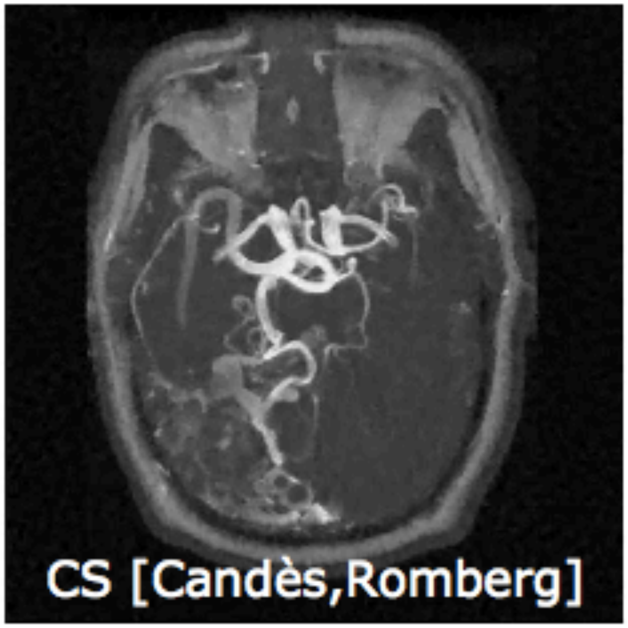
Each measurement contains a little information about each component

MAGNETIC RESONANCE IMAGING









NEXT TIME…

- What are the major open problems and areas of research?
- In what ways can these concepts be generalized to other problem domains?