Compressed Sensing: Extending CLEAN and NNLS

Ludwig Schwardt

SKA South Africa (KAT Project)

Calibration & Imaging Workshop
Socorro, NM, USA
31 March 2009

Outline

1. Compressed Sensing (CS)
   - Introduction
   - Reconstruction Algorithms
2. Compressed Sensing for Radio Astronomy
   - Formulation
   - The CLEAN and NNLS Algorithms
3. Experiments and Future Work

Motivation

- Modern data acquisition systems have to cope with ever-increasing data rates (think SKA!)
- Shannon sampling theorem: perfect reconstruction for bandlimited signals if sampling rate $\geq 2\times$ bandwidth
- **BUT** we frequently compress the data straight after sampling... (e.g. JPEG in a digital camera)
- Is it possible to sample at the signal's **"information rate"** instead of the usually much higher Nyquist rate?
- **YES** (Shannon was a pessimist)

The CS Sampling Process

- Generalise sampling to a series of **linear measurements**
- Each measurement contains contributions from each part of signal, instead of being localised, in the same way that a visibility sample is affected by all parts of image domain
- Each measurement of N-dimensional discrete signal $\boldsymbol{x}$ has the form $y_i = \langle \boldsymbol{x}, \boldsymbol{\phi}_i \rangle$, where $\boldsymbol{\phi}_i$ is an $N \times 1$ basis vector
- A set of M measurements (one per basis vector) can be written in matrix form as

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x},$$

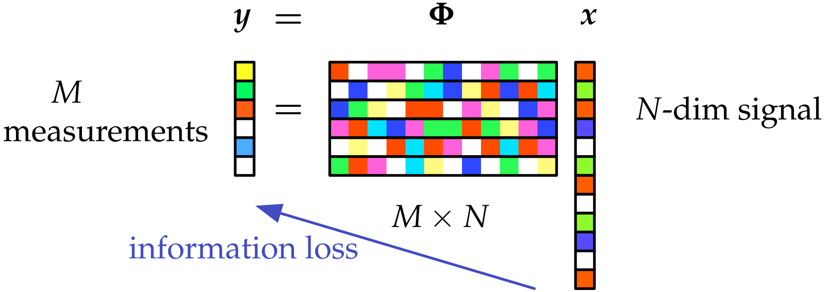
  where the $M \times N$ **measurement matrix $\boldsymbol{\Phi}$** has $\boldsymbol{\phi}_i$ as rows
- For standard (Dirac) sampling, $M = N$ and $\boldsymbol{\Phi} = \boldsymbol{I}$ (i.e. the basis vectors are the standard unit vectors)

Sensing Sparse Signals

- **Question:** Can $\boldsymbol{x}$ be reconstructed from $\boldsymbol{y}$ even if $M \ll N$?
- **Surprising answer:** Yes, with high probability, as long as $\boldsymbol{x}$ is **S-sparse** (i.e. it has exactly S non-zero entries), and $S < M$, and $\boldsymbol{\Phi}$ has certain properties
- Compressible signals are sparse in some domain — they can therefore always be subsampled!
- In theory, M should be more than $O(S \log N)$ for perfect reconstruction
- In practice, $3S$ to $5S$ measurements are sufficient

Why Does It Work?

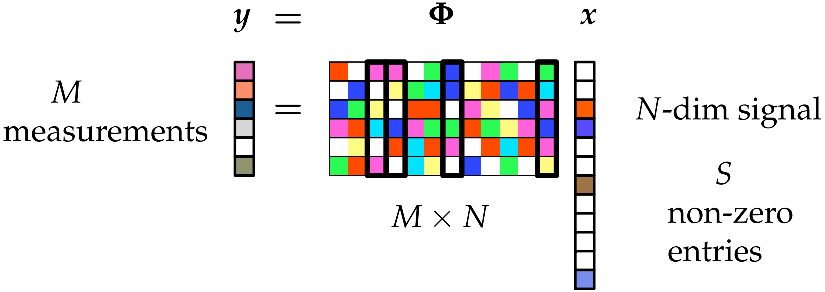
- For arbitrary x and $M \ll N$, recovery is impossible



- An S-sparse signal selects S columns at random from $\mathbf{\Phi}$
- **Restricted isometry property (RIP)**: If any S columns of $\mathbf{\Phi}$ are approximately mutually orthogonal, perfect reconstruction is possible for arbitrary S-sparse x

Why Does It Work?

- For arbitrary $\boldsymbol{x}$ and $M \ll N$, recovery is impossible

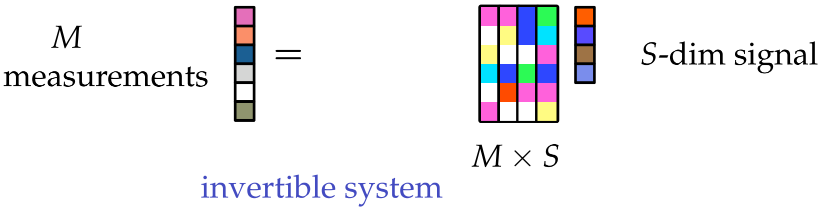


- An S-sparse signal selects S columns at random from $\boldsymbol{\Phi}$
- **Restricted isometry property (RIP)**: If any S columns of $\boldsymbol{\Phi}$ are approximately mutually orthogonal, perfect reconstruction is possible for arbitrary S-sparse $\boldsymbol{x}$

Why Does It Work?

- For arbitrary $\boldsymbol{x}$ and $M \ll N$, recovery is impossible

$$\boldsymbol{y} = \boldsymbol{\Phi}_S \, \boldsymbol{x}_S$$

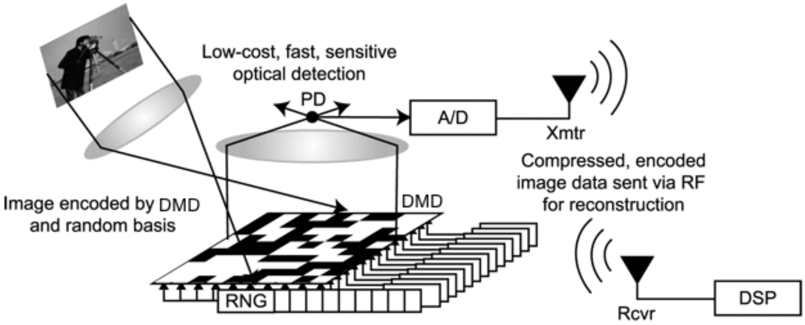


- An S-sparse signal selects S columns at random from $\boldsymbol{\Phi}$
- **Restricted isometry property (RIP)**: If any S columns of $\boldsymbol{\Phi}$ are approximately mutually orthogonal, perfect reconstruction is possible for arbitrary S-sparse $\boldsymbol{x}$

The Measurement Matrix $\mathbf{\Phi}$

- **Random** matrices (with zero mean) typically have RIP with high probability
- Compressed sensing therefore projects the desired signal onto a **few random basis functions**, instead of many shifted impulses
- Good choices for $\mathbf{\Phi}$ include:
  - **Gaussian matrix** with i.i.d. normal random entries
  - **Bernoulli matrix** with i.i.d. Bernoulli random entries
  - **Partial Fourier matrix** with rows drawn at random from DFT matrix (random frequencies)
- Orthogonal change of signal basis preserves RIP of $\mathbf{\Phi}$: CS still works if signal x is not sparse itself, but its orthogonal transform is (consider natural images and wavelet domain)

CS Showcase: The Single-Pixel Camera



Picture courtesy of `http://www.dsp.ece.rice.edu/cscamera/`

- How to make a megapixel image using only a single-pixel sensor and no raster scanning …

Reconstruction Algorithms

Various classes of CS algorithms exist, of which the most popular are:

- **Convex relaxation** (BP, ...)
- **Greedy methods** (MP, OMP, ROMP, StOMP, CoSaMP, ...)
- Iterative thresholding (IST, TwIST, ...)
- Combinatorial algorithms (chaining pursuit, Heavy-Hitters on Steroids (HHS), ...)
- Bayesian methods (CS $\equiv$ MAP with Laplacian prior...)

Convex Relaxation

- Sparsity measured by ℓ_0 **norm** $\|\boldsymbol{x}\|_0$, which is the number of non-zero elements of $\boldsymbol{x}$
- Ideal sparse reconstruction minimises $\|\boldsymbol{x}\|_0$ while being consistent with the measurements $\boldsymbol{\Phi x} = \boldsymbol{y}$
- This is intractable, so use next best norm instead, which is the ℓ_1 **norm** $\|\boldsymbol{x}\|_1 = \sum_{i=1}^{N} |x_i|$ (**convex relaxation** of ℓ_0)
- ℓ_1 minimisation promotes sparsity, while ℓ_2 minimisation (least-squares) discourages it... That is why pseudoinverse solution to CS problem, $\boldsymbol{x}_{\mathrm{PI}} = \boldsymbol{\Phi}^{\dagger}\boldsymbol{y}$, is a **bad idea**

Convex Relaxation: Basis Pursuit (BP)

- This directly solves the convex optimisation problem

$$\texttt{[eq]} \qquad \min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \boldsymbol{\Phi}\boldsymbol{x} = \boldsymbol{y}$$

- For noisy measurements, change to one of

$$\texttt{[ineq]} \qquad \min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2 \le \epsilon$$

$$\texttt{[regls]} \qquad \min_{\boldsymbol{x}} \left(\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + \gamma\|\boldsymbol{x}\|_1\right)$$

$$\texttt{[dantzig]} \qquad \min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \|\boldsymbol{\Phi}^T(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x})\|_\infty \le \epsilon$$

Convex Relaxation: Basis Pursuit (BP)

- The `regls` version is least-squares with ℓ_1 regularisation
- Convex optimisation is easy: single global optimum and efficient algorithms available (e.g. interior-point methods)
- Easy to add constraints such as non-negativity of $\boldsymbol{x}$ **(BP+)**, as long as convexity is preserved
- Computational complexity $O(M^2 N^{1.5})$ — slow in practice, which opens the door for greedy methods

Greedy: Matching Pursuit (MP)

- Views recovery problem as finding a sparse representation for the $M \times 1$ **measurement vector** $\boldsymbol{y} = \sum_{j=1}^{N} x_j \boldsymbol{\varphi}_j$, based on the **columns** $\boldsymbol{\varphi}_j$ of $\boldsymbol{\Phi}$ (i.e. only a few x_j terms are non-zero)
- MP terminology: $\boldsymbol{\Phi}$ is **dictionary** of **atoms** $\boldsymbol{\varphi}_j$
- Greedy, iterative approach:
  - Initialise **residual** $\boldsymbol{r}^{(0)} = \boldsymbol{y}$
  - At kth iteration, select atom which fits residual best, as $\boldsymbol{\varphi}^{(k)} = \arg\max_{\boldsymbol{\varphi}} |\langle \boldsymbol{r}^{(k)}, \boldsymbol{\varphi} \rangle|$, which amounts to picking the peak of $|\boldsymbol{\Phi}^T \boldsymbol{r}^{(k)}|$
  - Update residual to $\boldsymbol{r}^{(k+1)} = \boldsymbol{r}^{(k)} - a_k \boldsymbol{\varphi}^{(k)}$, with $a_k = \langle \boldsymbol{r}^{(k)}, \boldsymbol{\varphi} \rangle$
  - Stop when residual becomes small enough
  - Recovered signal has non-zero entries a_k at locations of selected atoms

Orthogonal Matching Pursuit (OMP)

- This is identical to MP, but adds a least-squares fit step after selecting a new atom, which readjusts the amplitudes of all atoms to best fit the data
- In practice, OMP is preferred to plain MP, as it converges faster
- OMP has computational complexity $O(SMN)$, which is faster than BP, and it is simpler to code than BP
- Relatively easy to add non-negativity constraint **(OMP+)**
- A downside is that naive implementation of least-squares fit requires an $M \times S$ matrix, which can be prohibitively large for large data sets

CS for Radio Astronomy

- Consider simplified imaging equation expressing visibilities V in terms of image brightness I,

$$V(u_j, v_j) = \sum_{k=1}^{N} I(l_k, m_k) \mathrm{e}^{-i2\pi(u_j l_k + v_j m_k)}$$

- In matrix form it becomes $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}$, with M visibilities $y_j = V(u_j, v_j)$, N image pixels $x_k = I(l_k, m_k)$ and measurement matrix entries $\Phi_{jk} = \exp\{-i2\pi(u_j l_k + v_j m_k)\}$
- Natural fit to CS: the interferometer does random projections for you! (similar situation in MRI)
- $\boldsymbol{\Phi}$ related to partial Fourier matrix
- An S-sparse image $\boldsymbol{x}$ containing S point sources can be recovered via CS as long as there are **enough baselines** ($M \geq \text{Const} \cdot S \log N \approx 5S$) and the uv plane is sampled **randomly enough**

Relating RA Terms to CS Notation

- **Measurement equation** (simplified!):

$$\boldsymbol{y} = \boldsymbol{\Phi x}$$

- **Dirty image** (where $\boldsymbol{\Phi}^T$ is conjugate transpose of $\boldsymbol{\Phi}$):

$$\boldsymbol{x}_{\mathrm{DI}} = \frac{1}{M}\boldsymbol{\Phi}^T\boldsymbol{y}$$

- **Dirty beam** (where $\mathbf{1}$ is $M \times 1$ vector of ones):

$$\boldsymbol{x}_{\mathrm{DB}} = \frac{1}{M}\boldsymbol{\Phi}^T\mathbf{1},$$

- **Measurement equation in the image domain:**

$$\boldsymbol{\Phi}^T\boldsymbol{y} = \boldsymbol{\Phi}^T\boldsymbol{\Phi x}$$

Relating RA Terms to CS Notation

- **Measurement equation** (simplified!):

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}$$

- **Dirty image** (where $\boldsymbol{\Phi}^T$ is conjugate transpose of $\boldsymbol{\Phi}$):

$$\boldsymbol{x}_{\mathrm{DI}} = \frac{1}{M}\boldsymbol{\Phi}^T\boldsymbol{y}$$

- **Dirty beam** (where $\mathbf{1}$ is $M \times 1$ vector of ones):

$$\boldsymbol{x}_{\mathrm{DB}} = \frac{1}{M}\boldsymbol{\Phi}^T\mathbf{1},$$

- **Measurement equation in the image domain:**

$$\boldsymbol{\Phi}^T\boldsymbol{y} = \boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{x}$$

Relating RA Terms to CS Notation

- **Measurement equation** (simplified!):

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}$$

- **Dirty image** (where $\boldsymbol{\Phi}^T$ is conjugate transpose of $\boldsymbol{\Phi}$):

$$\boldsymbol{x}_{\text{DI}} = \frac{1}{M}\boldsymbol{\Phi}^T\boldsymbol{y}$$

- **Dirty beam** (where $\mathbf{1}$ is $M \times 1$ vector of ones):

$$\boldsymbol{x}_{\text{DB}} = \frac{1}{M}\boldsymbol{\Phi}^T\mathbf{1},$$

- **Measurement equation in the image domain:**

$$\boldsymbol{\Phi}^T y = \boldsymbol{\Phi}^T\boldsymbol{\Phi}x$$

Relating RA Terms to CS Notation

- **Measurement equation** (simplified!):

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}$$

- **Dirty image** (where $\boldsymbol{\Phi}^T$ is conjugate transpose of $\boldsymbol{\Phi}$):

$$\boldsymbol{x}_{\mathrm{DI}} = \frac{1}{M}\boldsymbol{\Phi}^T\boldsymbol{y}$$

- **Dirty beam** (where $\mathbf{1}$ is $M \times 1$ vector of ones):

$$\boldsymbol{x}_{\mathrm{DB}} = \frac{1}{M}\boldsymbol{\Phi}^T\mathbf{1},$$

- **Measurement equation in the image domain:**

$$\boldsymbol{\Phi}^T\boldsymbol{y} = \boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{x}$$

Relating CLEAN to CS

- Högbom CLEAN is identical to MP, but forms residual in image space instead of in measurement (uv) space
- The idea of Clark CLEAN to subtract multiple components in one iteration is echoed in many MP variants, such as ROMP and StOMP, where it regularises problem and speeds up algorithm, respectively
- Cotton-Schwab CLEAN actually operates in measurement (uv) space, like standard MP, but also adds ideas from Clark CLEAN
- CLEAN loop gain idea not prevalent in MP literature

Relating NNLS to CS

- NNLS is identical to OMP with non-negativity constraint, but operates in the image domain instead of uv domain, solving
$$\boldsymbol{\Phi}^T \boldsymbol{y} = \boldsymbol{\Phi}^T \boldsymbol{\Phi} \boldsymbol{x} \quad \text{subject to} \quad \boldsymbol{x} \geq 0$$
- This explains the tendency of NNLS to compact flux
- The CS version improves on standard NNLS by operating directly in uv domain: improved accuracy and greatly reduced memory usage ($M \times S$ instead of $N \times N$)
- Standard OMP is something between CLEAN and NNLS

Experiment

Small-scale simulation:

- Four point sources 4.5 pixels apart in 32x32 image, i.e. $S = 4$ and $N = 1024$ (sources are off-grid...)
- Image pixel size chosen to have 5 pixels across main lobe of dirty beam (sources are merged...)
- Single VLA A snapshot at declination $\delta = 35^\circ$ provides $M = 351$ measurements
- Noise with SNR of 0 dB added to visibilities (16 dB image SNR)

Methods tested

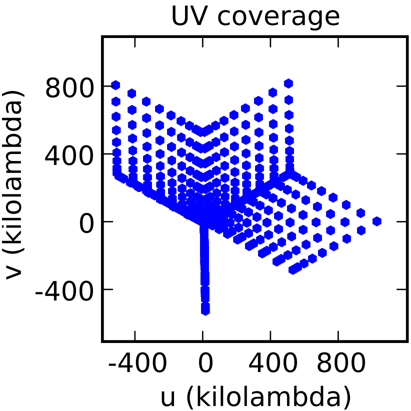
Standard algorithms:

- Högbom CLEAN (CASA) with 400 iterations, loop gain 0.1
- Cotton-Schwab CLEAN (CASA) with 400 iterations, loop gain 0.1
- NNLS (CASA struggled, implemented via OMP+ instead)

CS algorithms:

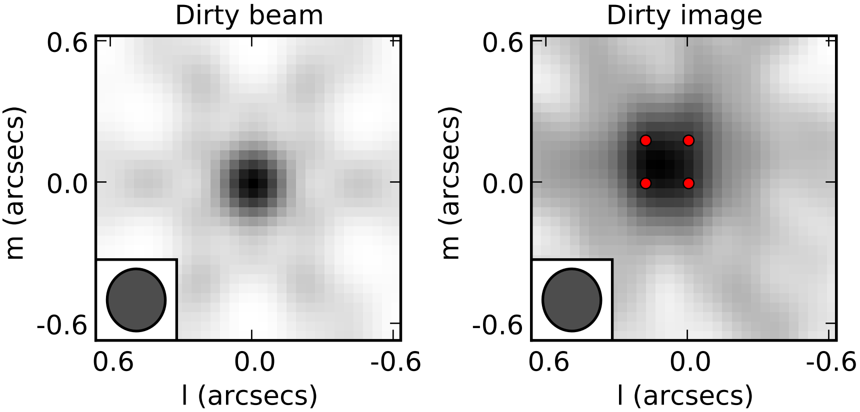
- OMP, OMP+ searches for 40 components
- BP, BP+ (`ineq`, `regls`, `dantzig` variants)
- Algorithms have **explicit measurement matrix** version (fast and correct but memory-limited) and **operator** version (uses FFT to jump between image and uv domains)
- Algorithm parameters are tuned based on expected SNR

Results: UV Coverage



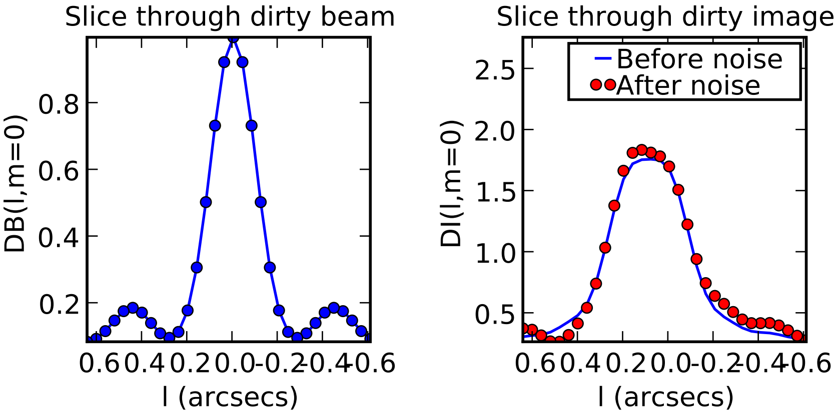
- Typical VLA snapshot coverage (mirroring unnecessary)

Results: Dirty Beam and Image



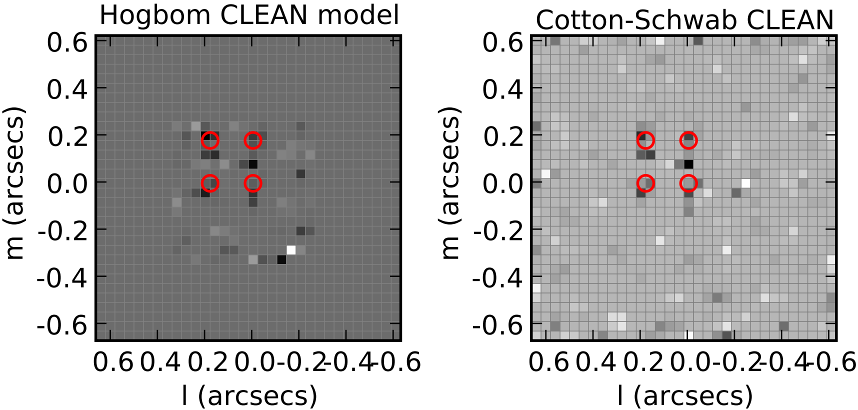
- Red circles show positions of true point sources

Results: Close-up of Dirty Beam and Image



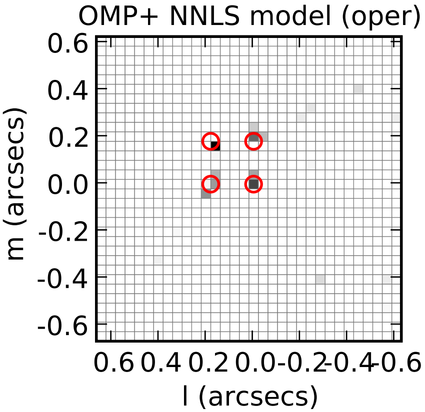
- A single peak in dirty image . . .

Results: Standard CLEAN



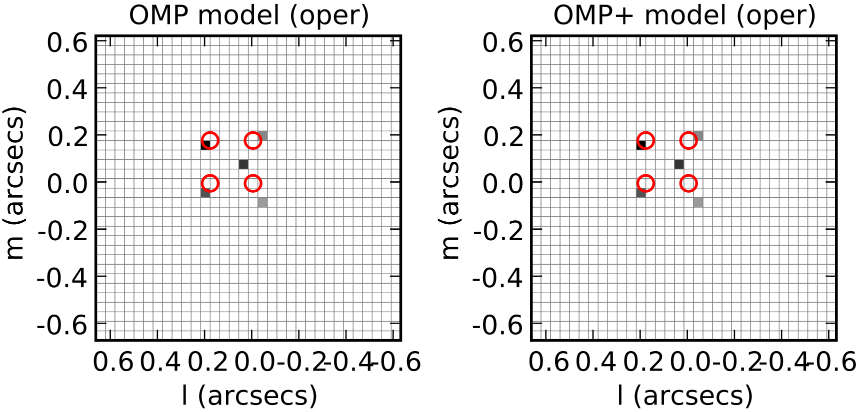
- Many spurious components, Högbom only cleans inner quarter

Results: NNLS



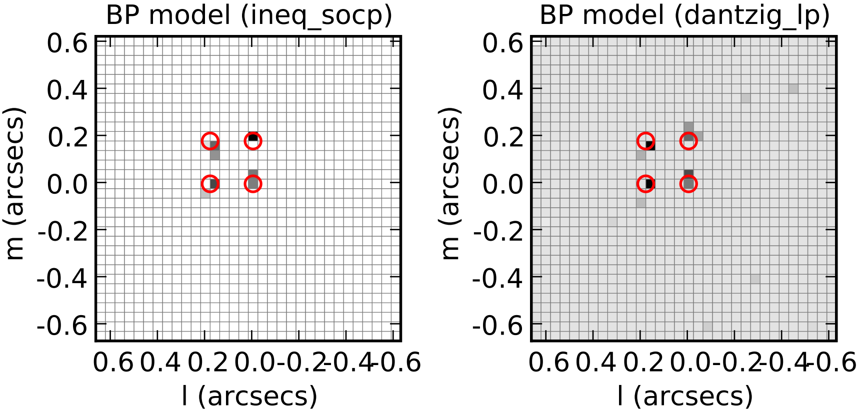
- Already a big improvement

Results: OMP vs. OMP+



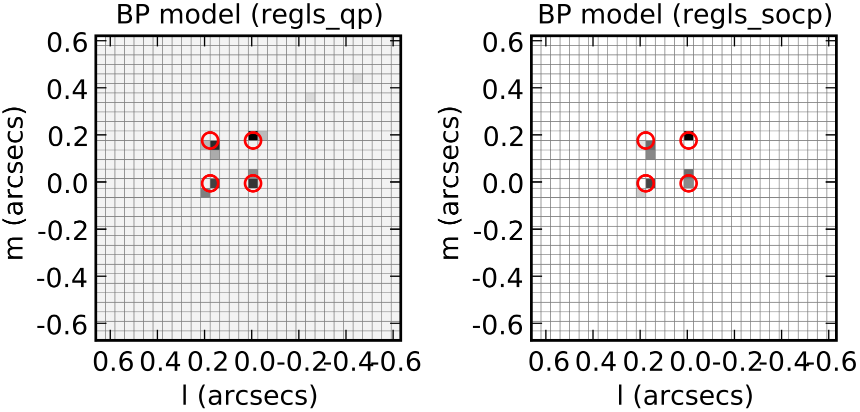
- Non-negativity constraint makes little difference

Results: BP



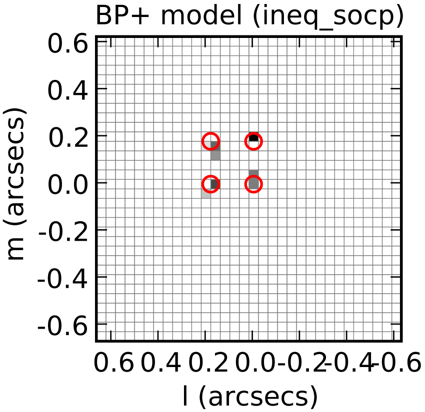
BP model (ineq_socp)
BP model (dantzig_lp)
m (arcsecs)
l (arcsecs)
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6

Results: BP



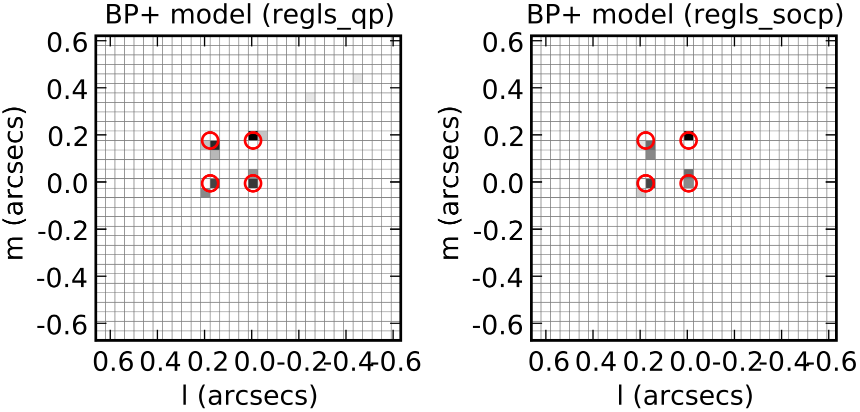
- Two different methods to solve `regls` problem

Results: BP+



- The `dantzig` version struggled here

Results: BP+



- Two different methods to solve `regls` problem

Conclusions

- Compressed sensing forms a mathematical framework for both CLEAN and NNLS, and allows for many extensions
- NNLS apparently improves on CLEAN due to the least-squares step of OMP, rather than due to non-negativity constraint
- CS looks promising for VLBI (small N, big D)

Future Work

- Improve convergence of BP solvers based on operators
- Perform larger-scale experiments (already possible)
- Integrate with curvelet transform for extended sources
- Extend CS ideas to w-correction and full measurement equation