Sparse Reconstruction via 
Fast Bayesian Matching Pursuit

Phil Schniter

July 2008

(Joint work with Dr. Lee Potter and Mr. Justin Ziniel)
Sparse Reconstruction:

Estimate sparse $x$ from the under-determined noisy linear mixture:

$$y = Ax + n$$ for known $A \in \mathbb{C}^{M \times N}$, with $M \ll N$.

- Under-determined means that the number of measurements ($M$) is less than the number of unknowns ($N$).
  
  In general, there is no unique solution!

- Sparse means that the number of nonzero coefs ($K$) is relatively few.
  
  This may help to solve the problem...

But why do we care about this problem?
Modern Data Acquisition:

A two-step approach:

1. Nyquist-rate sampling,

2. Lossy compression:
   (a) take the discrete wavelet transform (DWT),
   (b) discard all but the large DWT coefficients.

Why? Because the DWT of a “structured” signal is sparse:

\[ Wz = x + r \]

- \( W \) unitary DWT operator,
- \( z \) signal,
- \( x \) sparse: only a few non-zero coefficients,
- \( r \) small residual.

Furthermore, the DWT is universal; it doesn’t need to know the particular “structure” of \( x \)!
DWT Example:

Typically: $\text{MSE} \approx -20 \text{ dB from only } 2.5\% \text{ of DWT coefficients!}$
What’s Wrong With Sampling-then-Compressing:

*Nyquist-rate sampling generates a lot of intermediate data!*

In some applications, this poses problems, e.g.,

1. Magnetic resonance imaging (MRI): sampling is time-intensive,
2. Sensor networks: joint compression requires a lot of inter-sensor communication.

*Is there a smarter way to sample?*
An Alternative — Compressive Sensing:

• Main idea:

For an $N$-dimensional signal $z$, take $M \ll N$ linear measurements $y = \Phi z$ from which $z$ can be accurately reconstructed.

Note: If the signal class has $K$ significant degrees of freedom, then we need $M \geq K$ measurements. But the hope is that $M \gg K$.

• But how can we make the measurement scheme $\Phi$ universal?

Intuition: Measurements should be spread uniformly over all times and frequencies, since we can’t make assumptions on temporal/frequency support. More generally, measurements should be uniformly spread across all subspaces of $\mathbb{C}^N$.

Idea: *Use a measurement matrix $\Phi$ with random entries.*
Example — The Rice Univ. Single-Pixel Camera:

This and the previous DWT figure courtesy of Rich Baraniuk, Rice University.
**Compressive Sensing:**

Essential components:

- **Measurement:** \( y = \Phi z + \nu \) for \( \Phi \in \mathbb{C}^{M \times N} \) and noise \( \nu \).
- **Compressibility:** \( W z = x + r \) for \( K \)-sparse \( x \), small residual \( r \), and unitary transform \( W \).

Putting these together, we get

\[
y = \underbrace{\Phi W^H x}_{A} + \underbrace{\Phi W^H r + \nu}_{n} = A x + n\]

where \( A \in \mathbb{C}^{M \times N} \) and \( K < M \ll N \).

The remaining (and big!) challenge:

*Reconstruct the sparse signal representation \( x \) from the “compressed” measurements \( y = A x + n \).*
Other Applications of Sparse Reconstruction:

Sparse channel estimation:

\[ r = Ax + n \text{ for } \begin{cases} x: \text{sparse channel impulse response}, \\ A: \text{symbol matrix}. \end{cases} \]

- Sometimes under-determined:
  - Time-span limited so that channel is time-invariant over block.
  - \( r \) and \( A \) might be constructed only from pilot symbols/observations.

\[ \Rightarrow \text{Sparsity is needed to solve the problem!} \]

- And sometimes over-determined:

\[ \Rightarrow \text{Sparsity can be leveraged to improve performance!} \]

since \( \mathbb{E}\{\|\hat{x}_{LS} - x\|_2^2\} \geq \frac{K\sigma_n^2}{\|a\|_2^2} \text{ for } K = \|x\|_0 \).
Solving the Sparse Reconstruction Problem:

Key question:

*_How do we use the sparsity of* \( x \) _to solve the noisy under-determined inverse problem* \( y = Ax + n \)?

Popular Techniques:

1. Constrained \( \ell_1 \)-minimization via convex optimization.
3. Bayesian approaches.

**Notation:** \( \| x \|_p = \sqrt[p]{\sum_n |x_n|^p} \) is known as the “\( \ell_p \) norm.”
Constrained $\ell_0$-Minimization:

The classical problem formulation:

*Find the sparsest $\hat{x}$ which explains $y$ up to a given tolerance $\epsilon$, i.e.,*

$$
\hat{x} = \arg\min_x \|x\|_0 \quad \text{s.t.} \quad \|y - Ax\|_2 \leq \epsilon.
$$

Unfortunately, this problem is NP-hard!

Let’s think about this problem geometrically...
A Toy Problem:

Consider the setup
\[
\begin{bmatrix}
\bullet \\
\bullet \\
\bullet \\
\bullet
\end{bmatrix} = \begin{bmatrix}
\bullet & \bullet & \bullet & \bullet & \bullet \\
\bullet & \bullet & \bullet & \bullet & \bullet
\end{bmatrix}
\begin{bmatrix}
\bullet \\
\bullet
\end{bmatrix} + \begin{bmatrix}
\bullet
\end{bmatrix}
\]

Since \( N = M + 1 \),
- the set \( \{x : y = Ax\} \) is described by a line (via \( \text{null}(A) \)), and
- the set \( \{x : \|y - Ax\|_2 \leq \epsilon\} \) is described by an \( \epsilon \)-thin rod.

Since \( K = 1 \),
- the true \( x \) intersects one of the coordinate axes. (But which one?)
The Geometry of Constrained $\ell_p$-Minimization:

Now consider, for some general $p > 0$, the optimization problem:

$$\hat{x} = \arg \min_x \|x\|_p \text{ s.t. } \|y - Ax\|_2 \leq \epsilon.$$ 

$\hat{x}$ can be found by growing the $\ell_p$-ball until it touches the $\epsilon$-rod:

- $p \ll 1$: Solution is sparse but problem is \textbf{NP hard}.
- $p = 1$: Solution is sparse and problem is \textbf{convex}!
- $p = 2$: Solution is \textbf{not sparse}; $\leftrightarrow$ LS when $\epsilon = 0$.

\textit{This suggests to use the $\ell_1$ norm as a surrogate for the $\ell_0$ norm.}
1) Constrained $\ell_1$-Minimization:

For the constrained-$\ell_1$ approach (known as “LASSO”)

$$\hat{x} = \arg \min_x \|x\|_1 \text{ s.t. } \|y - Ax\|_2 \leq \epsilon,$$

there exist elegant theorems which say that, given

- enough measurements (e.g., $M \gtrsim K \log N$) and
- sufficiently well-behaved $A$ (e.g., nearly uncorrelated columns),

$\|\hat{x} - x\|_2$ will be very small with very high probability.


*In practice, though, we can’t guarantee that these conditions hold!*
2) Matching Pursuits:

- These are basically “greedy” searches.

- Intuition:
  
  Focus on finding the locations of the $K$ non-zero coefficients in $x$.
  
  If you know where the non-zero coefficients are, estimating their $K$ values from $M \geq K$ measurements is easy.

- Operation:
  
  - Do a simple search for the most important coefficient in $x$, then the second most important on, and so on...
    
    Usually: correlate $y$ with each (normalized) column of $A$.
    
    Search simplicity is key since the number of coefficients ($N$) is large.

  - Estimation can be more sophisticated since the number of active coefficients ($K$) and observations ($M$) are much smaller.
Matching Pursuit (MP):

1. Find the column $a_i$ of $A$ that is most correlated with $y$.
2. Estimate the corresponding signal coefficient $x_i$ using LS.
3. Compute the residual: $e = y - a_i \hat{x}_i$.
4. Repeat with $e$ in place of $y$.

Initialize: $e = y$, $\mathcal{I} = \emptyset$, and $\hat{x} = 0$.
While $\|e\|_2 > \text{threshold}$,

\[
\hat{x}_{i_*} = \frac{a_{i_*}^H e}{\|a_{i_*}\|_2^2}, \quad \mathcal{I} \leftarrow \mathcal{I} \cup i_*,
\]
\[
e \leftarrow e - a_{i_*} \hat{x}_{i_*},
\]
end.

Similar to “successive interference cancellation” for CDMA.
Orthogonal Matching Pursuit (OMP):

1. Find the column $a_i$ of $A$ that is most correlated with $y$.
2. Add $i$ to the list of active-coefficient indices, $\mathcal{I}$.
3. Jointly estimate all active coefficients using LS.
4. Compute the residual: $e = y - A\hat{x}$.
5. Repeat with $e$ in place of $y$.

Initialize: $e = y$, $\mathcal{I} = \emptyset$, and $B = [\,]$.

While $\|e\|_2 >$ threshold,

\[
\begin{align*}
    i_* &= \arg \max_{i \notin \mathcal{I}} \frac{|a_i^H e|^2}{\|a_i\|^2}, \\
    \mathcal{I} &\leftarrow \mathcal{I} \cup i_*, \\
    B &\leftarrow [B, a_{i_*}], \\
    \hat{x}_\mathcal{I} &= (B^H B)^{-1} B^H y, \\
    e &\leftarrow e - B\hat{x}_\mathcal{I}
\end{align*}
\]

end.

Note: Can implement recursively to avoid explicit matrix inversion.
Matching Pursuit Theory:

For certain matching pursuits, e.g.,
- stage-wise orthogonal matching pursuit (StOMP),
- regularized orthogonal matching pursuit (ROMP), . . .

can prove that, with
- enough measurements (i.e., $M \gtrsim K \log N$), and
- sufficiently well-behaved $A$ (e.g., nearly uncorrelated columns),

reconstruction error is small with high probability.


*In practice, though, we can’t guarantee that these conditions hold!*
3) Bayesian Approaches to Sparse Reconstruction:

What if we have some prior statistical knowledge of

- the number of active coefficients $K$,
- the noise distribution,
- the distribution of active coefficients.

Can this knowledge be leveraged in sparse reconstruction?
Example — a Laplacian Signal Prior:

If we assume $\sigma^2$-variance AWGN and signal $x$ such that

$$p(x) \propto e^{-\lambda \|x\|_p^p},$$

then the MAP estimate becomes

$$\hat{x} = \arg \max_x p(x|y) = \arg \max_x \frac{p(y|x)p(x)}{p(y)}$$

$$= \arg \max_x p(y|x)p(x) = \arg \max_x \log p(y|x) + \log p(x)$$

$$= \arg \min_x \sigma^{-2}\|y - Ax\|_2^2 + \lambda \|x\|_p^p$$

$$= \arg \min_x \|x\|_p \text{ s.t. } \|y - Ax\|_2 \leq \epsilon(\lambda\sigma^2)$$

which is the constrained $\ell_p$-optimization problem we saw earlier. Choosing $p = 1$ yields sparse estimates and facilitates convex programming.

*But this is artificial!*

What’s the physical meaning of the Laplacian prior? How do we choose $\lambda$?
Another Example — Sparse Bayesian Learning (SBL):

The so-called SBL method uses the signal model:

\[ x_i | \sigma_i^2 \sim \mathcal{N}(0, \sigma_i^2) \]

\[ \{\sigma_i^{-2}\}_{i=1}^N \sim \text{i.i.d. } \Gamma(a, b) \quad \text{...for tractability!} \]

\[ a, b : \text{ chosen to induce sparsity,} \]

and the noise model:

\[ n | \sigma^2 \sim \text{i.i.d. } \mathcal{N}(0, \sigma^2) \quad \text{where } \sigma^{-2} \sim \Gamma(c, d) \text{ for some } c, d. \]

The EM algorithm can then be used to find the MAP estimate of \( x \).


*Again, artificial! Physical meaning of the \( \Gamma(\cdot, \cdot) \) prior? Choice of \( a, b, c, d \)?*
Our Approach:

A simple binary Gaussian mixture model:

- **Signal model:**

  \[ x_n | s_n \sim \begin{cases} 
  \mathcal{N}(0, 1) & s_n = 1 \quad \text{“active”} \\
  0 & s_n = 0 \quad \text{“inactive”} 
  \end{cases} \]

  \[ \{s_n\}_{n=1}^N \sim \text{i.i.d. Bernoulli}(\lambda). \]

- **Noise model:**

  \[ n \sim \mathcal{N}(0, \sigma^2 I). \]

To model sparsity, we would set \( \lambda \ll 1 \).

Note: priors are *physically meaningful*, e.g.,

\[ \lambda = \text{E}[K]/N. \]
A $Q$-ary Signal Model:

An extended signal model could incorporate $Q$ types of active coefficient, each with distinct (mean, variance, prior probability):

$$x_n | s_n \sim \begin{cases} 
0 & s_n = 0 \\
\mathcal{N}(\nu_1, \sigma_1^2) & s_n = 1 \\
\vdots & \\
\mathcal{N}(\nu_Q, \sigma_Q^2) & s_n = Q
\end{cases}$$

$$\Pr\{s_n = q\} = \lambda_q.$$  

This could enable a clearer distinction between signal and noise.

For simplicity, we assume the real binary zero-mean model in the sequel.
Main Objective:

MMSE estimation:

\[
\hat{x}_{\text{mmse}} = \mathbb{E}[x|y] = \sum_{s \in \{0,1\}^N} p(s|y)[x|y,s] \\
\approx \sum_{s \in S_*} p(s|y) \mathbb{E}[x|y,s] \hat{x}_{\text{mmse}}|s
\]

where \( S_* \) is the set of all probable configurations \( s = [s_0, \ldots, s_{N-1}]^T \).

Comments:

- The MMSE estimate leverages the inherent uncertainty in estimating \( s \). An example of “Bayesian model averaging.”

- We will need to search for all probable configurations \( S_* \), rather than the single best configuration \( \hat{s}_{\text{map}} \triangleq \arg \max_s p(s|y) \).
Example – Leveraging Uncertainty:

At SNR=14dB, \( \hat{x}_{\text{mmse}}|\hat{s}_{\text{map}} \approx \hat{x}_{\text{mmse}} \approx \hat{x} \):

![Graph showing parameter estimates and ranked posteriors]

ranked posteriors; where \( p(s_{\text{true}}|y) = 0.204 \)
Example – Leveraging Uncertainty (cont.):

But at SNR=13dB, $\hat{x}_{\text{mmse}}|\hat{s}_{\text{map}} \not\approx \hat{x}_{\text{mmse}} \approx x!!$

![Parameter estimates graph]

![Ranked posteriors graph; where $p(s_{\text{true}}|y)=0.119$]
Configuration Metric:

- Posterior probability of configuration $s$:

$$p(s|y) = \frac{p(y|s)p(s)}{p(y)} = \frac{p(y|s)p(s)}{\sum_{s'\in\{0,1\}^N} p(y|s')p(s')}.$$  

- Maximizing $p(s|y)$ $\iff$ maximizing the “configuration metric” $\nu(s)$:

$$\nu(s) \triangleq \log p(y|s)p(s).$$  

- Recalling that $p(y|s)$ is Gaussian and $p(s)$ is i.i.d Bernoulli,

$$\nu(s) = -\frac{M}{2} \log 2\pi - \frac{1}{2} \log \det(\Sigma(s)) - \frac{1}{2} y^T \Sigma(s)^{-1} y$$

$$+ \|s\|_0 \log \frac{\lambda}{1-\lambda} + N \log(1-\lambda),$$

where $\Sigma(s) \triangleq \text{Cov}\{y|s\} = A D(s) A^T + \sigma^2 I_M.$
Greedy Search:

- We propose to find $S_\star$ by greedy inflation search:
  - Find the single-tap $s$ maximizing $\nu(s)$. (Search over $N$ possibilities.)
  - By adding one additional tap, find the 2-tap $s$ maximizing $\nu(s)$. (Search over $N - 1$ possibilities.)
  - By adding one additional tap, find the 3-tap $s$ maximizing $\nu(s)$. (Search over $N - 2$ possibilities.)
  - Continue until $K_{\text{max}}$ taps have been explored.

If needed, repeat the greedy search, but forcing a different path.

- Having obtained $\hat{S}_\star$, we can compute

$$\hat{p}(s|y) = \frac{e^{\nu(s)}}{\sum_{s' \in \hat{S}_\star} e^{\nu(s')}}$$ and $\hat{x}_{\text{mmse}|s}$ for each $s \in \hat{S}_\star$.

Bayesian model averaging yields

$$\hat{x}_{\text{mmse}} \approx \sum_{s \in \hat{S}_\star} \hat{p}(s|y) \hat{x}_{\text{mmse}|s}.$$
Greedy Search (cont.):

- Can choose the maximum search depth $K_{\text{max}}$ so that 
  \[ \Pr\{\|s\|_0 > K_{\text{max}}\} \triangleq P_0 \] 
  is sufficiently small:
  \[
  K_{\text{max}} = \lceil \text{erfc}^{-1}(2P_0) \sqrt{2N\lambda(1 - \lambda)} + N\lambda \rceil
  \]
  via the Gaussian approximation of $\|s\|_0 \sim \text{Binomial}(N, \lambda)$.

- Can choose the maximum number of repeated greedy searches $D_{\text{max}}$ as a tradeoff between accuracy and complexity.
Example of repeated greedy search:

\[ \hat{p}(s|y) \]

- Depth of search \( (K_{\text{max}} = 27) \)
- Repetition index \( (D_{\text{max}} = 5) \)
Example of repeated greedy search:

Note that most of the probable configurations have been explored.
“Fast Bayesian Matching Pursuit”

• Like most matching pursuits,
  – Our search is greedy, exploring one additional tap each time.
  – There exists a fast recursive implementation.
    \[ \sim \text{Total complexity } O(NMK)! \]

But unlike other matching pursuits,
  – Our search is guided by maximization of the posterior \( p(s|y) \).

• Hence, the name *Fast Bayesian Matching Pursuit* (FBMP).
Estimation of Hyper-Parameters:

Until now we have assumed known hyper-parameters

\[ \lambda = \Pr\{s_n = 1\} \]
\[ \sigma^2 = \text{var}\{x_n|s_n = 1\} \]

In practice they are unknown, so what can we do?

Expectation-maximization (EM) for iterative ML estimation of \( \theta \triangleq [\lambda, \sigma^2] \):

\[ \hat{\theta}^{(i+1)} = \arg \max_{\theta} \sum_{s \in S_x} p(s|y, \hat{\theta}^{(i)}) \log p(y, s|\theta) \]
Connections to Noncoherent Decoding:

- Consider soft noncoherent decoding with coded binary symbols $s$ and (non-sparse unknown) Gaussian channel impulse response $x$. Can write

$$y = T(s)x + n$$

for Toeplitz matrix $T(s)$.

- Tree-search (e.g., sphere decoding) techniques can be used to find $S_*$, the set of most probable $s$, and in doing so will generate conditional MMSE channel estimates $\hat{x}_{\text{mmse}}|s$. From $S_*$, LLRs can be computed and sent to the decoder.

- Tree-search can be implemented using a fast recursive update of the noncoherent MAP metric.
Numerical Experiments — Deterministic Signal:

Setup: \( N = 512 \)
\( M = 128 \)
\( A : \) i.i.d. \( \mathcal{N}(0, 1) \) with columns scaled to unit norm
\( x : \) \( x_n = e^{-\rho n} \) for decay rate \( \rho \in (0, 1) \)
\( \text{SNR} = 15 \text{ dB} \)

Algorithms:
- SparseBayes - Wipf & Rao
- OMP - Tropp & Gilbert
- StOMP - Donoho, Tsaig, Drori & Starck
- GPSR-Basic - Figueiredo, Nowak & Wright
- BCS - Ji & Carin
- FBMP - \ldots with 5 repeated searches

Performance: \( \text{NMSE} \triangleq \text{Avg} \left\{ \frac{\| \hat{x} - x \|^2}{\| x \|^2} \right\} \) over 2500 random trials.
NMSE versus decay rate $\rho$:

FBMP outperformed GPSR by 2 dB and others by much more. In general, NMSE performance suffers as $\rho$ decreases (i.e., as $x$ gets less sparse).
Sparsity of estimate versus decay rate $\rho$:

The MMSE estimates returned by FBMP are among the sparsest. (FBMP’s $\hat{x}_{mmse}|s_{map}$ would be even sparser!)

$$N = 512, M = 128, \text{SNR} = 15 \text{ dB}, D_{\max} = 5, E_{\max} = 20, T = 2500$$
Runtime versus decay rate $\rho$:

FBMP (without EM iterations) is faster than other Bayesian algorithms, but slower than other matching pursuit and convex programming algs.
Conclusion:

- Sparse reconstruction is critical to compressive sensing & other apps.
- Using a Bayesian approach to sparse reconstruction, we proposed
  - a simple and physically meaningful signal model,
  - greedy search for the set of all high-probability configurations,
  - a fast recursive algorithm with complexity $O(NMK)$,
  - an EM-based method to estimate hyperparameters.
- Comparisons against other state-of-the-art algorithms showed NMSE improvements of several dB over a wide range of parameters.
- Current Work:
  - Performance guarantees for $\hat{x}_{\text{map}}$, $\hat{x}_{\text{mmse}}$, and the greedy search.
  - How to track a sequence of correlated $\{x\}$?
Thanks for listening!
Numerical Experiments — Sparse Gaussian Signal:

Now we use the same type of signal assumed for the derivation of FBMP. (One might say that this gives FBMP an unfair advantage!)

Nominal Params: \( N = 512 \)
\( \lambda = 0.04 \) \( \Rightarrow \) so \( \mathbb{E}[K] = \lambda N = 20 \) active coefs
\( M = 120 \)
\( \text{SNR} = 19 \text{ dB} \) \( \Rightarrow \) where \( \text{SNR} \triangleq \frac{\mathbb{E}[K]}{\sigma^2 M} \)
\( A \) : i.i.d. \( \mathcal{N}(0, 1) \) with columns scaled to unit norm.

Performance: \( \text{NMSE} \triangleq \text{Avg} \left\{ \frac{\| \hat{x} - x \|_2^2}{\| x \|_2^2} \right\} \) over 100 random trials.
NMSE versus observation length $M$:

For $\frac{M}{E[K]} > 5$, FBMP outperformed BCS by 3 dB and others by $\geq 10$ dB. As $\frac{M}{E[K]} \to 2$, all algorithms break down.
NMSE versus SNR:

At high SNR, FBMP outperformed BCS by 3 dB and others by ≥ 9 dB. As SNR → 0 dB, GPSR catches up.
FBMP is an order of magnitude faster than SparseBayes, about the same speed as BCS, and an order of magnitude slower than OMP, StOMP, GPSR.  
*Note:* This is an earlier (slower) version of FBMP!
Numerical Experiments – Gaussian Signal with fixed $K$:

Now we use the same type of signal assumed for the derivation of FBMP, except with a fixed number of active coefficients $K$. (Again, FBMP might have an unfair advantage.)

Nominal Signal Parameters:

\[
\begin{align*}
N &= 256 \\
K &= 10 \\
M &= 64 \\
A &\sim \text{i.i.d. } \mathcal{N}(0, 1) \text{ with columns scaled to unit norm.} \\
\text{SNR} &= 15 \text{ dB}
\end{align*}
\]

Performance:

\[
\text{NMSE} \triangleq \text{Avg} \left\{ \frac{\| \hat{x} - x \|_2^2}{\| x \|_2^2} \right\} \quad \text{over 200 random trials.}
\]
**NMSE versus observation length** $M$:

When $D = 1$, knee in curve at $\frac{M}{K} = \frac{64}{10} = 6.4 \frac{\text{measurements}}{\text{active coef}}$. For larger $D$, knee moves to $5 \frac{\text{measurements}}{\text{active coef}}$ and NMSE improves by 3 dB.
NMSE versus # active coefs $K$:

When $D = 1$, knee in curve at $\frac{M}{K} = \frac{64}{10} = 6.4$ measurements/active coef.  
For $D = 10$, knee at $\frac{M}{K} = \frac{64}{13} = 4.9$ measurements/active coef. and NMSE 3 dB improved.
Active coefs missing from $\hat{s}_{map}$:

Again, knee in curve at $\frac{M}{K} \approx 5 \frac{\text{measurements}}{\text{active coef}}$.
(Note: Expect improvement from generalized signal model.)
NMSE versus SNR:

Note linear relationship between NMSE [dB] & SNR [dB]. (No benefit from $D$-increase anticipated because $\frac{M}{K} = 6.4.$)
NMSE for $\hat{x}_{\text{mmse}}$ and $\hat{x}_{\text{mmse}}|\hat{s}_{\text{map}}$:

Exploiting configuration uncertainty gives $\approx 1$ dB gain in NMSE.