Sparse Reconstruction via Bayesian Model Averaging

Phil Schniter

July 2009

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

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

$$y = Ax + e$$ 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?
Standard Two-Step Data Acquisition:

1. Nyquist-rate sampling,

2. Lossy compression:
   (a) take the discrete wavelet transform (DWT),
   (b) keep only the large DWT coefficients.

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

\[ Wz = x + r \]

- \( W \): unitary DWT operator,
- \( z \): structured signal,
- \( x \): large wavelet coefficients... sparse,
- \( r \): small residual... non-sparse.

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

1 megapixel image

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

Sampling at the Nyquist rate produces many samples!

This poses problems when sampling is expensive, e.g.,
1. Magnetic resonance imaging (MRI): sampling is very time-intensive.
2. Sensor networks with “dumb” nodes: before compression, samples are communicated through a network.

Is there a more efficient way to sample?
An Alternative — Compressive Sampling/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.

How should we construct $\Phi$?

• Since we make no prior assumptions on the signal’s eigen-structure, the measurement scheme $\Phi$ must be universal.

• Intuitively, measurements should collect approximately equal energy from all subspaces of $\mathbb{C}^N$.

  Use a measurement matrix $\Phi \in \mathbb{C}^{M \times N}$ with random entries!
Example — Single-Pixel Camera (Rice University):

This and the previous DWT figure courtesy of Rich Baraniuk, Rice University.
An Alternative — Compressive Sampling/Sensing:

How many measurements \( M \) do we need?

- Say \( z \) belongs to a class of signals with \( K \) dominant eigenspaces.
- These \( K \) subspaces can be arranged \( \binom{N}{K} \) ways within the \( N \) dimensional space, so the signal \( z \) carries at least

\[
\log_2 \left( \binom{N}{K} \right) \geq K \log_2 \left( \frac{N}{K} \right) \text{ bits of information.}
\]

- For measurements \( y = \Phi z + \nu \) in AWGN, we learn at most

\[
\frac{1}{2} \log_2 (1 + \text{SNR}) \text{ bits per measurement,}
\]

suggesting that we need at least

\[
M \geq \frac{2}{\log_2 (1 + \text{SNR})} K \log_2 \left( \frac{N}{K} \right) \text{ measurements.}
\]
Compressive Sensing:

Essential components:

**Measurement:** \( y = \Phi z + \nu \) for \( \Phi \in \mathbb{C}^{M \times N} \) and noise \( \nu \).

**Compressibility:** \( Wz = x + r \) for \( K \)-sparse \( x \), small residual \( r \), and unitary transform \( W \).

Putting these together, we get

\[
y = \Phi W^H x + \Phi W^H r + \nu = A x + e\]

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

The remaining challenge:

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

Sparse channel estimation:

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

- When the problem is under-determined \((M < N)\):
  \[ \rightsquigarrow \text{Sparsity is needed to solve the problem!} \]
  - Span of \( r \) limited to ensure that channel is time-invariant over block.
  - Span of \( r \) limited due to small number of pilot symbols.

- When the problem is not under-determined \((M \geq N)\):
  \[ \rightsquigarrow \text{Sparsity can be leveraged to improve performance!} \]
  since \( E\{|\hat{x}_{\text{LS}} - x|^2\} \approx \frac{K\sigma_e^2}{M\sigma_a^2} \) for \( K = \|x\|_0 \).
  \# nonzero coefs
Solving the Sparse Reconstruction Problem:

Key question:

How do we use the sparsity of $x$ to solve the noisy under-determined inverse problem $y = Ax + e$?

Popular Techniques:

1. Convex optimization of constrained $\ell_1$ criteria.
3. Bayesian approaches.

Notation: $\|x\|_p = \sqrt[p]{\sum_n |x_n|^p}$ is the “$\ell_p$ norm.”
The Canonical “Sparse-Approximation” Problem:

Find the sparsest $x$ which explains $y$ up to a specified tolerance of $\epsilon$:

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

Key points:

1. NP-hard: need to check all configurations of non-zero coefficients!
2. May not be the “right” objective.

Let’s think about this sparse-approximation problem geometrically…
A Toy Example:

Consider the setup

$$
\begin{bmatrix}
\hat{y}
\end{bmatrix} = \begin{bmatrix}
\hat{A}
\end{bmatrix} \begin{bmatrix}
x
\end{bmatrix} + \begin{bmatrix}
e
\end{bmatrix}
$$

Since $N = M + 1$,

- the set $\{x: y = Ax\}$ is described by a line (via $\text{null}(A)$), and
- the set $\{x: \|y - Ax\|_2 \leq \varepsilon\}$ is described by an $\varepsilon$-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:

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

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

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

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

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


But, $A$ may not be well-behaved, especially when $M$ and $N$ are not huge!

Also, generally incompatible with complex-valued $x$. 

2) Matching-Pursuit Algorithms:

- Basic “matching pursuit” algorithm:

  Similar to “successive interference cancellation” for CDMA:

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

- More sophisticated versions, like “orthogonal matching pursuit” (OMP), are more robust to correlation among the columns of $A$. 
Matching-Pursuit Theory:

For some matching-pursuit algorithms, one can prove that, with

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

reconstruction error is small with high probability.


*Practical, but how well does it work when $A$ is not well-behaved?*
3) **Bayesian Approaches to Sparse Reconstruction:**

Say that we have some prior statistical knowledge of

- the pattern of active coefficients,
- the values of active coefficients,
- the noise.

Can we take advantage of this knowledge for sparse reconstruction?

Three of the most popular Bayesian strategies are...

a) Laplacian signal prior,

b) The relevance vector machine.

c) **Bayesian variable selection and Bayesian model averaging,**
3a) Laplacian Signal Prior:

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

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

then the MAP estimate becomes

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

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

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

which is the constrained $\ell_p$-optimization problem we saw earlier.

- Choosing the Laplacian prior (i.e., $p = 1$), we know that $x$ will be sparse and that $\hat{x}$ can be obtained via convex programming.

_Interesting, but… Physical meaning of Laplace prior? Choice of $\tau$?
3b) The Relevance Vector Machine (RVM):

- To model coefficient activity, use “precisions” $\alpha \in (\mathbb{R}^+)^N$:
  \[
  x|\alpha \sim \prod_n \mathcal{N}(0, \alpha_n^{-1}) \quad \text{and} \quad \alpha \sim \text{iid } \Gamma(0, 0)
  \]
  \[
  e|\beta \sim \prod_n \mathcal{N}(0, \beta_n^{-1}) \quad \text{and} \quad \beta \sim \Gamma(0, 0)
  \]

  As $\alpha_n \to \infty$, the coefficient $x_n$ is effectively “turned off”.

- The use of gamma hyperpriors leads to the convenient posterior

  \[
  p(x|y, \alpha, \beta) \sim \mathcal{N}(\mu, \Sigma) \quad \text{for} \quad \begin{cases} 
  \mu = \beta \Sigma A^T y \\
  \Sigma = (\beta A^T A + D(\alpha))^{-1}
  \end{cases}
  \]

  and thus the convenient estimate $\hat{x}_{\text{MMSE}} = \mu$.

- The EM algorithm can be used to estimate $\{\alpha, \beta\}$ jointly with $\{\mu, \Sigma\}$.

  Can implement with an $O(NK^2)$ recursion after an $O(N^2M)$ initialization.

3c) Variable Selection:

- Using $S$ to denote the set of active-coefficient indices, we can write

$$ y = A_S x_S + e. $$

- With $S$ known, estimation of the nonzero coefficients $x_S$ is easy:

$$ \hat{x}_{\text{LS}|S} = (A_S^T A_S)^{-1} A_S^T y $$
$$ \hat{x}_{\text{MMSE}|S} = (A_S^T A_S + \sigma_e^2 I)^{-1} A_S^T y $$

This motivates the variable-selection problem:

*From $y = Ax + e$, estimate the active-coefficient set $S$.*

a long-standing problem in statistics!

Non-Bayesian Variable Selection:

Consider the GLRT detector:

$$
\hat{S}_{\text{GLRT}} = \arg \max_S p(y|S, \hat{x}_{\text{ML}|S})
$$

$$
= \arg \min_S \| y - A_S \left( A_S^T A_S \right)^{-1} A_S^T y \|_2^2
$$

$$
= \arg \min_S \| P_{A_S}^\perp y \|_2^2 \quad \hat{x}_{\text{ML}|S}
$$

$$
= \text{any } S : |S| \geq M.
$$

It fails!

This happens whenever the models are “nested”. 
Bayesian Variable Selection:

Consider the MAP model estimate:

\[
\hat{S}_{MAP} = \arg \max_S p(S|y) = \arg \max_S p(y|S)p(S) = \arg \max_S \int_{\mathcal{N}} p(y|S,x)p(x|S)dx \cdot p(S). 
\]

Need to specify the priors \( p(x|S) \) and \( p(S) \).

Popular BVS Priors:

- iid Bernoulli coefficient-activity:
  \[ p(S) = \lambda^{|S|}(1 - \lambda)^{(N - |S|)} \] where \( \lambda < 0.5 \) induces sparsity,

- Gaussian active-coefficients \( x_S \):
  \[ p(x_S|S) \sim \mathcal{N}(\mu 1_{|S|}, R_S) \]
  for \[ R_S = \sigma^2 x I_{|S|}, \; \mu \in \mathbb{R} \] “iid”
  \[ R_S = \sigma^2 x (A_S^T A_S)^{-1}, \; \mu = 0 \] “Zellner g-prior”

where the hyperparameters \( \{ \lambda, \mu, \sigma^2 x, \sigma^2 e \} \) could be treated as . . .

1. random: assign non-informative conjugate priors & integrate out unknowns,
2. deterministic: use the EM-algorithm to estimate hyperparameters.

BVS Posteriors:

Fixing \( \{ \lambda, \mu, \sigma^2_x, \sigma^2_e \} \), we get

- the model’s posterior log-density:
  \[
  \ln p(S|y) = -\frac{1}{2} \|y - \mu A_S 1_{|S|}\|^2_{\Phi_S^{-1}} - \frac{1}{2} \ln \det(\Phi_S) - |S| \ln(1 - \lambda) + C,
  \]
  where \( \Phi_S \) denotes the observation covariance matrix given \( S \):
  \[
  \Phi_S = \begin{cases} 
  \sigma^2_x A_S A_S^T + \sigma^2_e I_{|S|} & (\text{iid}) \\
  \sigma^2_x A_S (A_S^T A_S)^{-1} A_S^T + \sigma^2_e I_{|S|} & (\text{Zellner})
  \end{cases}
  \]

- the \((S\text{-conditional})\) coefficient density:
  \[
  p(x_S|y, S) \sim \mathcal{N}(\mu_S, \Sigma_S)
  \]
  where
  \[
  \begin{align*}
  \mu_S &= \mu 1_{|S|} + R_S A_S^T \Phi_S^{-1} (y - \mu A_S 1_{|S|}) = \hat{x}_{\text{MMSE}|S} \\
  \Sigma_S &= R_S - R_S A_S^T \Phi_S^{-1} A_S R_S = \text{cov}(\hat{x}_{\text{MMSE}|S}).
  \end{align*}
  \]
Connections to Model-Order Selection:

Under the Zellner prior, it can be shown that

\[ \hat{S}_{\text{MAP}} = \arg \min_S \left\{ \frac{1}{\sigma_e^2} \| y - A_S \hat{x}_{\text{LS}|S} \|^2_2 + |S| \cdot \eta \right\} \]

for \( \eta = \frac{\sigma_x^2 + \sigma_e^2}{\sigma_x^2} \ln \left( (1 + \frac{\sigma_x^2}{\sigma_e^2})(1 - \frac{1}{\lambda})^2 \right) \).

Note the close connections to “information theoretic” model-order selectors:

\[ \hat{S}_{\text{AIC}} = \arg \min_S \left\{ \frac{1}{\sigma_e^2} \| y - A_S \hat{x}_{\text{LS}|S} \|^2_2 + |S| \cdot 2 \right\} \]
\[ \hat{S}_{\text{BIC}} = \arg \min_S \left\{ \frac{1}{\sigma_e^2} \| y - A_S \hat{x}_{\text{LS}|S} \|^2_2 + |S| \cdot \ln M \right\} \]
\[ \hat{S}_{\text{RIC}} = \arg \min_S \left\{ \frac{1}{\sigma_e^2} \| y - A_S \hat{x}_{\text{LS}|S} \|^2_2 + |S| \cdot 2 \ln N \right\}. \]

Connections to Noncoherent Decoding:

- Consider a generic communication system with vectorized observations

\[ y = A_j h + e, \]

where \( j \) = codeword index, \( h \) = channel gains, and \( e = \text{AWGN} \).

- In non-coherent decoding, \( h \) is known only in distribution, e.g.,

\[ h \sim \mathcal{CN}(0, I_N) \]

for Rayleigh fading. Then

\[ \hat{j}_{\text{MAP}} = \arg \max_{j=1,\ldots,J} p(j|y). \]

- This can be interpreted in our sparse reconstruction framework via

\[ A \triangleq [A_1 \cdots A_J] \quad \text{and} \quad x \triangleq [0^T, \ldots, h^T, \ldots, 0^T]^T \]

with a constraint on the admissible models \( S \in \mathcal{S} \):

\[ \mathcal{S} \triangleq \{(1,\ldots,K), (K+1,\ldots,2K), \ldots, (JK-K+1,\ldots,JK)\}, \]

with the result that

\[ \hat{S}_{\text{MAP}} = \arg \max_{S \in \mathcal{S}} p(S|y) \iff \hat{j}_{\text{MAP}} \]
Bayesian Model Averaging:

- Previously we motivated Bayesian variable selection, e.g.,

\[ \hat{S}_{\text{MAP}} = \arg \max_S p(S|\mathbf{y}) \]

for subsequent use in a conditional estimation strategy, e.g.,

\[ \hat{x}_{\text{MMSE}|\hat{S}_{\text{MAP}}} = \mathbb{E}\{x|\mathbf{y}, \hat{S}_{\text{MAP}}\} \]

- But having access to the “soft information” \( \{p(S|\mathbf{y})\}_{\forall S} \) allows more sophisticated unconditional estimates, e.g.,

\[ \hat{x}_{\text{MMSE}} = \sum_S \hat{x}_{\text{MMSE}|S} p(S|\mathbf{y}) \]

that are well approximated by summing over the few most probable \( S \).

Implementation of BVS/BMA:

- The conventional approach to determining the set of high-probability \( S \) is based on random search (e.g., Gibbs Sampling or Markov Chain Monte Carlo).

- Recently, a (non-exhaustive) tree search has been proposed to learn the set of high-probability \( S \) (and their \( \hat{x}_{\text{MMSE}|S} \)) with very low complexity:
  - iid Gaussian \( x_S \): “Fast Bayesian matching pursuit”
  - Zellner Gaussian \( x_S \): “Fast Zellner matching pursuit”

In fact, the complexity order equals that of OMP: \( O(MNK) \).

- An expectation/maximization (EM) approach can be used to optimally tune the hyperparameters \( \{\lambda, \mu, \sigma^2_x, \sigma^2_e\} \) wrt the data.


**Turbo Implementation:**

Recall that

- BVS is based on *binary* indicators of coefficient activity.
- BVS is a *soft-input soft-output* form of sparse reconstruction (i.e., it takes $p(S)$ as input and generates $p(S|y)$ as output).

With complicated $p(S)$, exact BVS implication may be difficult. Thus one might consider an iterative approach:

![Diagram of BVS-based sparse reconstruction](image)

that passes extrinsic log-likelihood ratios on the indicator variables.

BMA versus RVM:

- To parameterize coefficient activity, RVM uses the *continuous* variables $\alpha$, while BMA uses the *discrete* set $S$.
- RVM has trade-off parameters that must be tuned using *cross-validation*, whereas the BMA hyperparameters can be tuned via the *EM algorithm*.
- RVM infers *marginal* coefficient activity, whereas BMA infers *joint* coefficient activity.
- Upon termination, the RVM posterior is Gaussian
  \[ p(x|y) \sim N(\mu, \Sigma) \]
  whereas the BMA posterior is a Gaussian mixture:
  \[ p(x|y) \sim \sum_S N(\mu_S, \Sigma_S) \ p(S|y) \quad \ldots \text{more informative} \]
- Fast implementations of RVM and BMA have roughly the same complexity.
- Simulation results suggest better performance for BMA.
Numerical Experiments — “Compressible” Signal:

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

Algorithms:
- OMP – Tropp & Gilbert
- StOMP – Donoho, Tsaig, Drori & Starck
- GPSR-Basic – Figueiredo, Nowak & Wright \( \text{min}_x \|y - Ax\|_2^2 + \tau \|x\|_1 \)
- SparseBayes – Wipf & Rao (RVM)
- BCS – Ji & Carin (RVM)
- FBMP – Schniter, Potter & Ziniel (BMA)

Performance: \( \text{NMSE} \triangleq \text{Avg} \left\{ \frac{\|\hat{x} - x\|_2^2}{\|x\|_2^2} \right\} \) over 2500 random trials.
**NMSE versus decay rate $\rho$:**

FBMP outperformed GPSR and OMP by 2 dB and others by much more. Note: The signal priors favor GPSR!
Sparsity of estimate versus decay rate $\rho$:

The estimates returned by FBMP are among the sparsest. (While BMA is generally not sparse, FBMP is, due to non-exhaustive search.)
Runtime versus decay rate $\rho$:

FBMP is on par with other Bayesian algorithms, but slower than OMP and convex programming algorithms.
Performance Guarantees for MAP Variable Selection:

Under the iid-Beroulli/Gaussian signal model and a Restricted Isometry Property (RIP) for $A$:

There exists $\delta_K \in (0,1)$ such that

$$\forall x \text{ s.t. } \|x\|_0 = K, \quad (1 - \delta_K)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_K)\|x\|_2^2,$$

it has been recently shown that the following properties hold with high probability for reasonably small constants $K_1, K_2, K_3, K_4$:

1. The energy of the missed signal coefficients is upper bounded by $K_1 M \sigma_e^2$.
2. No active coefficients are missed when $|\mu| > 4\sigma_1 + K_2 \sqrt{M} \sigma_e^2$.
3. No coefficients are falsely detected when $|\mu| > K_3 \sqrt{M} \sigma_1 + K_4 \sqrt{M} \sigma_e^2$.

Lemma 1  For generic $S$, the pairwise error probability of

$$
\hat{S} = \arg \min_{S \in S} \left\{ \frac{1}{\sigma^2_e} \| y - A_S \hat{x}_{LS|S} \|_2^2 + \eta |S| \right\} \quad \text{under} \quad x_S|S \sim \mathcal{N}(0, \gamma \sigma^2_e I_{|S|})
$$

has the upper bound (tight as $\gamma \to \infty$):

$$
P_{\hat{S}|S} \leq (\alpha_{\hat{S},S} \gamma)^{-K_m} C_{K_m,K_f}(\eta),
$$

where $K_m$ and $K_f$ denote the # of missed and false-alarm coefficients, and

$$
C_{K_m,K_f}(\eta) = \begin{cases} 
\sum_{k=0}^{K_f-1} \frac{(K_f-K_m)^k \eta^k}{k!} \left( \frac{K_m+K_f-1-k}{K_m} \right) & K_m \leq K_f, \\
\sum_{k=0}^{K_m} \frac{(K_m-K_f)^k \eta^k}{k!} \left( \frac{K_m+K_f-1-k}{K_f-1} \right) & K_m > K_f.
\end{cases}
$$

$$
\alpha_{\hat{S},S} = \lambda_{\min}(A_m^T \Pi_{A_S}^\perp A_m) \quad \ldots \text{Restricted Isometry Property}
$$

Conclusions:

- Bayesian variable selection (BVS) and Bayesian model averaging (BMA) are well-established statistical methods for sparse reconstruction.
- While BVS & BMA were previously considered to be “too expensive,” modern tree-search implementations have reasonable complexity.
- There are close connections between BVS and AIC/BIC/RIC, as well as between BVS and noncoherent decoding.
- Numerical experiments suggest that BMA yields NMSE superior to that of other state-of-the-art algorithms.
- Current work includes BVS/BMA performance analysis, turbo implementation, and applications in
  - medical imaging,
  - through-wall radar,
  - underwater acoustic channel tracking,
  - decoding with intermittent and degraded side-information.