# 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 $\boldsymbol{x}$ from the under-determined noisy linear mixture:*

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e} \quad \text{for known } \boldsymbol{A} \in \mathbb{C}^{M \times N}, \text{ with } M \ll N.$$

- *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.

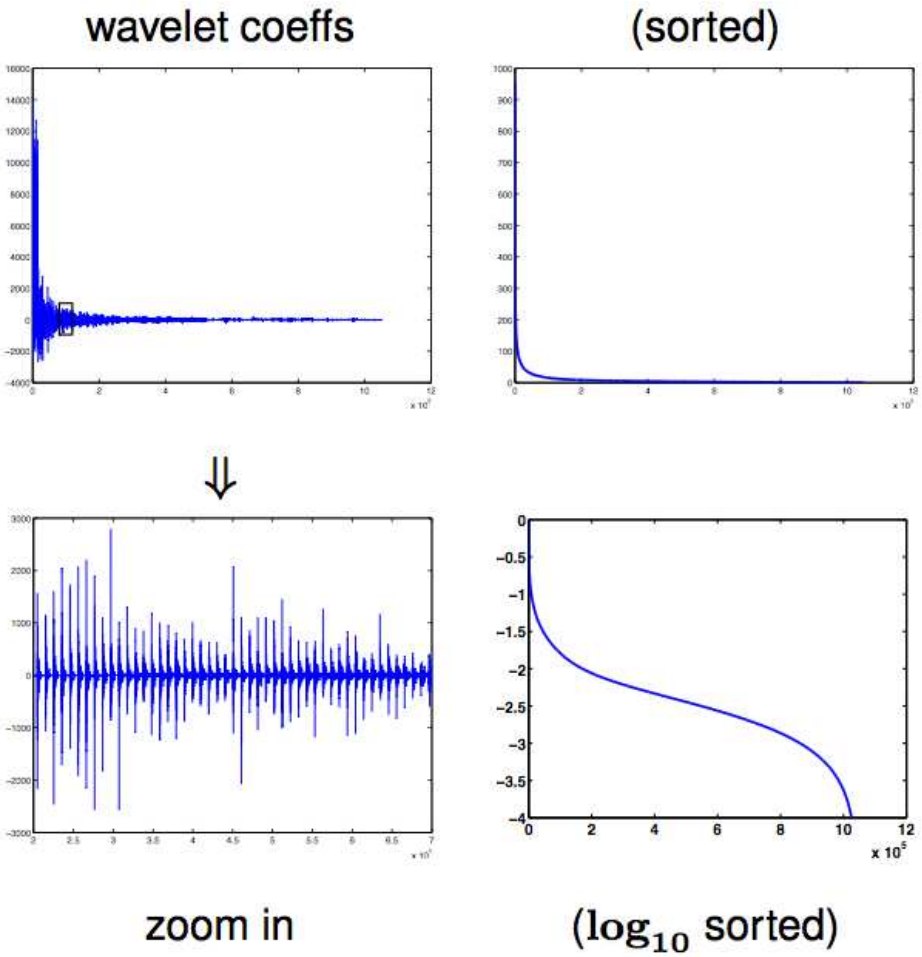
Why? Because the DWT of a "structured" signal is sparse:

$$\boxed{\boldsymbol{W}\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{r}} \; : \begin{cases} \boldsymbol{W} & \text{unitary DWT operator,} \\ \boldsymbol{z} & \text{structured signal,} \\ \boldsymbol{x} & \text{large wavelet coefficients...sparse,} \\ \boldsymbol{r} & \text{small residual...non-sparse.} \end{cases}$$

Furthermore, the DWT is *universal*; it doesn't need to know the particular "structure" of $\boldsymbol{x}$!

## DWT Example:

1 megapixel image



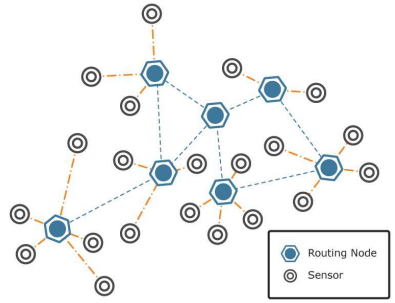

Typically: MSE $\approx -20$ dB from only $2.5\%$ of DWT coefficients!

## 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:

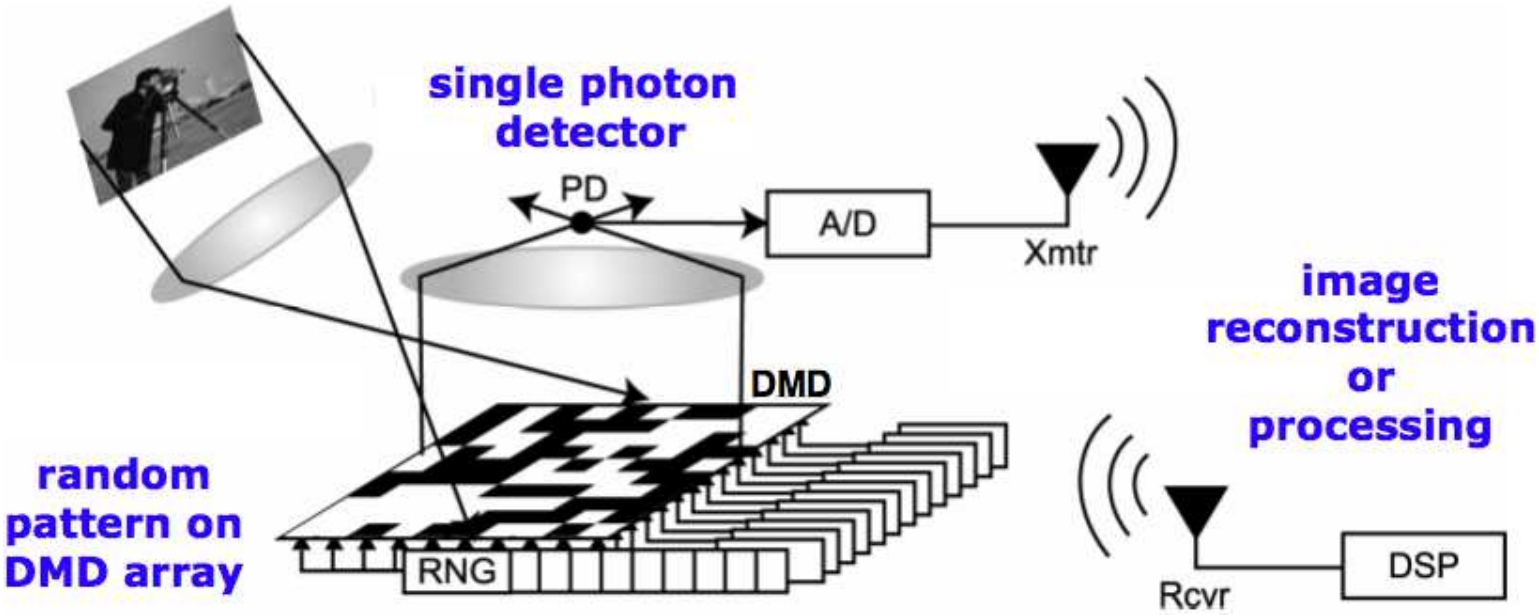
For an $N$-dimensional signal $\boldsymbol{z}$, take $M \ll N$ *linear* measurements $\boxed{\boldsymbol{y} = \boldsymbol{\Phi z}}$ from which $\boldsymbol{z}$ can be accurately reconstructed.

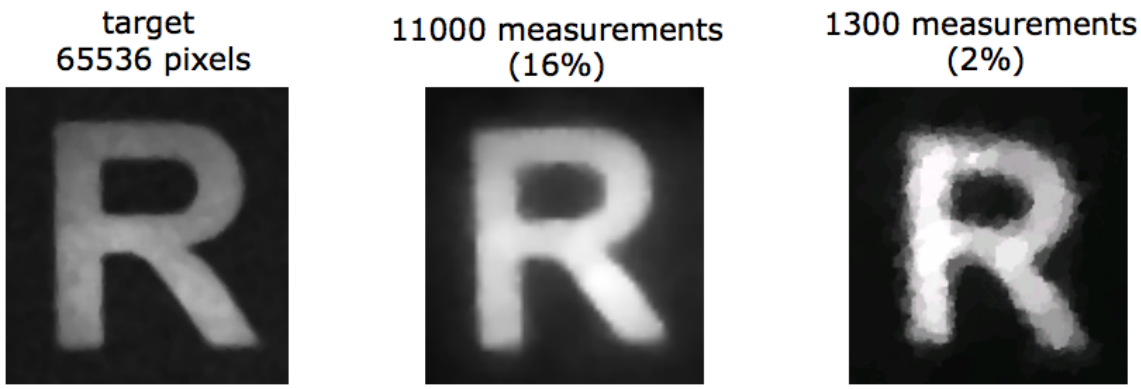
How should we construct $\boldsymbol{\Phi}$?

- Since we make no prior assumptions on the signal's eigen-structure, the measurement scheme $\boldsymbol{\Phi}$ must be *universal*.
- Intuitively, measurements should collect approximately equal energy from *all* subspaces of $\mathbb{C}^N$.

  *Use a measurement matrix $\boldsymbol{\Phi} \in \mathbb{C}^{M \times N}$ 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 $\boldsymbol{z}$ 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 $\boldsymbol{z}$ carries at least

$$\log_2 \tbinom{N}{K} \;\geq\; K \log_2(\tfrac{N}{K}) \;\text{ bits of information.}$$

- For measurements $\boldsymbol{y} = \boldsymbol{\Phi z} + \boldsymbol{\nu}$ in AWGN, we learn at most

$$\tfrac{1}{2}\log_2(1+\mathsf{SNR}) \;\text{ bits per measurement,}$$

suggesting that we need at least

$$M \;\geq\; \tfrac{2}{\log_2(1+\mathsf{SNR})} K \log_2(\tfrac{N}{K}) \;\text{ measurements.}$$

## Compressive Sensing:

Essential components:

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

*Compressibility*: $\boldsymbol{Wz} = \boldsymbol{x} + \boldsymbol{r}$ for $K$-sparse $\boldsymbol{x}$, small residual $\boldsymbol{r}$, and unitary transform $\boldsymbol{W}$.

Putting these together, we get

$$\begin{aligned}\boldsymbol{y} &= \underbrace{\boldsymbol{\Phi W}^H}_{\boldsymbol{A}}\boldsymbol{x} + \underbrace{\boldsymbol{\Phi W}^H\boldsymbol{r} + \boldsymbol{\nu}}_{\boldsymbol{e}} \\ &= \boldsymbol{Ax} + \boldsymbol{e} \qquad \text{where } \boldsymbol{A} \in \mathbb{C}^{M \times N} \text{ and } K < M \ll N.\end{aligned}$$

The remaining challenge:

*Reconstruct the sparse signal representation $\boldsymbol{x}$ from the "compressed" measurements $\boldsymbol{y} = \boldsymbol{Ax} + \boldsymbol{e}$.*

## Other Applications of Sparse Reconstruction:

Sparse channel estimation:

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

- When the problem is under-determined ($M < N$):

  $\rightsquigarrow$ *Sparsity is needed to solve the problem!*

  - Span of $\boldsymbol{r}$ limited to ensure that channel is time-invariant over block.
  - Span of $\boldsymbol{r}$ limited due to small number of pilot symbols.

- When the problem is not under-determined ($M \geq N$):

  $\rightsquigarrow$ *Sparsity can be leveraged to improve performance!*

  $$\text{since} \quad \mathrm{E}\{\|\hat{\boldsymbol{x}}_{\mathsf{LS}} - \boldsymbol{x}\|_2^2\} \approx \frac{K\sigma_e^2}{M\sigma_a^2} \quad \text{for } K = \underbrace{\|\boldsymbol{x}\|_0}_{\#\text{ nonzero coefs}}.$$

## Solving the Sparse Reconstruction Problem:

Key question:

> *How do we use the sparsity of $\boldsymbol{x}$ to solve the noisy under-determined inverse problem $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$?*

Popular Techniques:

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

*Notation:* $\|\boldsymbol{x}\|_p = \sqrt[p]{\sum_n |x_n|^p}$ is the "$\ell_p$ norm."

## The Canonical "Sparse-Approximation" Problem:

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

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \underbrace{\|\boldsymbol{x}\|_0}_{\#\text{ nonzero coefs}} \text{ s.t. } \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \leq \epsilon.$$

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 $\underset{\boldsymbol{y}}{\begin{bmatrix}\bullet\\ \bullet\end{bmatrix}} = \underset{\boldsymbol{A}}{\begin{bmatrix}\bullet & \bullet & \bullet\\ \bullet & \bullet & \bullet\end{bmatrix}} \underset{\boldsymbol{x}}{\begin{bmatrix}\bullet\\ \bullet\\ \bullet\end{bmatrix}} + \underset{\boldsymbol{e}}{\begin{bmatrix}\bullet\\ \bullet\end{bmatrix}}$

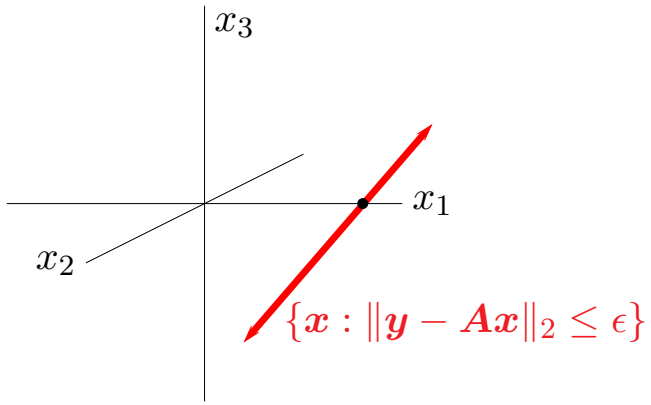
$N = 3$
$M = 2$
$K = 1$

Since $N = M + 1$,

- the set $\{\boldsymbol{x} : \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}\}$ is described by a line (via $\mathrm{null}(\boldsymbol{A})$), and
- the set $\{\boldsymbol{x} : \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \leq \epsilon\}$ is described by an $\epsilon$-thin rod.

Since $K = 1$,

- the true $\boldsymbol{x}$ intersects one of the coordinate axes. (But which one?)

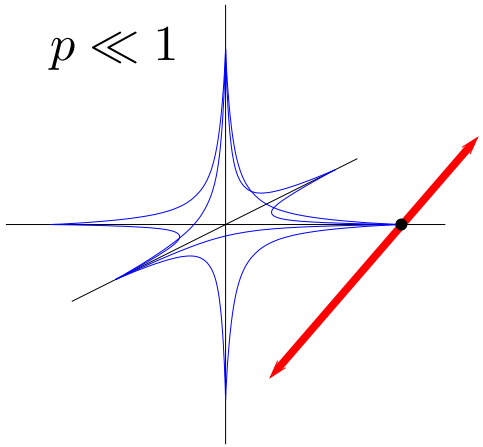


## The Geometry of Constrained $\ell_p$-Minimization:

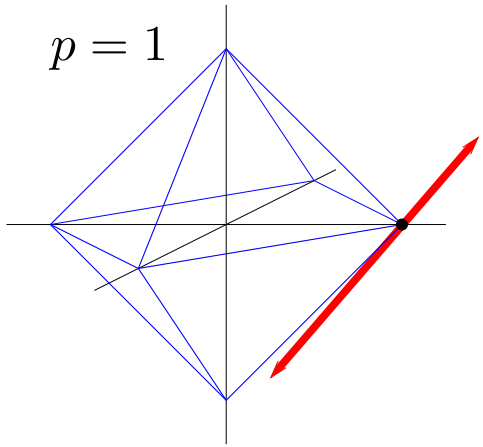
Now consider, for some general $p > 0$, the optimization problem:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_p \;\; \text{s.t.} \;\; \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \leq \epsilon.$$

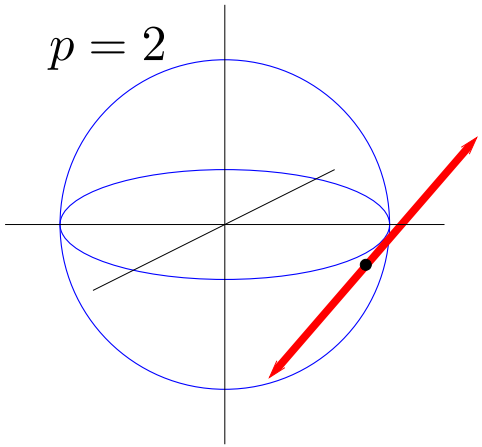
$\hat{\boldsymbol{x}}$ can be found by growing the $\ell_p$-ball until it touches the $\epsilon$-rod:




Solution definitely sparse but problem is **NP hard**.




Solution usually sparse and problem is **convex**!




Solution is **not sparse**; $\Leftrightarrow$ LS when $\epsilon = 0$.

*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")

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

there exist elegant theorems which say that, given

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

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

[1] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inform. Theory*, vol. 52, no. 1, 2006.

[2] J. A. Tropp, "Just relax: Convex programming methods for identifying sparse signal," *IEEE Trans. Info. Theory*, vol. 51, no. 3, 2006.

[3] E. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. on Pure and Applied Math.*, vol. 59, no. 8, 2006.

*But, $\boldsymbol{A}$ may not be well-behaved, especially when $M$ and $N$ are not huge!*
*Also, generally incompatible with complex-valued $\boldsymbol{x}$.*

## 2) Matching-Pursuit Algorithms:

- Basic "matching pursuit" algorithm:

  *Similar to "successive interference cancellation" for CDMA:*

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

- More sophisticated versions, like "orthogonal matching pursuit" (OMP), are more robust to correlation among the columns of $\boldsymbol{A}$.

## 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 $\boldsymbol{A}$ (e.g., nearly orthogonal columns),

reconstruction error is small with high probability.

[1] J. A. Tropp and A. C. Gilbert, "Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit," *IEEE Trans. Inform. Thy.*, Dec. 2007.

[2] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Harmon. Anal.*, 2008.

[3] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing: Closing the gap between performance and complexity," Preprint 2008.

*Practical, but how well does it work when $\boldsymbol{A}$ 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 $\boldsymbol{x}$ such that

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

then the MAP estimate becomes

$$\begin{aligned}\hat{\boldsymbol{x}} &= \arg\max_{\boldsymbol{x}} \; p(\boldsymbol{x}|\boldsymbol{y}) \\ &= \arg\min_{\boldsymbol{x}} \; \sigma^{-2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 + \tau\|\boldsymbol{x}\|_p^p \\ &= \arg\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_p \;\; \text{s.t.} \;\; \|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2 \le \epsilon(\tau\sigma^2)\end{aligned}$$

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

- Choosing the Laplacian prior (i.e., $p=1$), we know that $\boldsymbol{x}$ will be sparse and that $\hat{\boldsymbol{x}}$ 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" $\boldsymbol{\alpha} \in (\mathbb{R}^+)^N$:

$$\boldsymbol{x}|\boldsymbol{\alpha} \sim \prod_n \mathcal{N}(0, \alpha_n^{-1}) \quad \text{and} \quad \boldsymbol{\alpha} \sim \text{iid } \Gamma(0,0)$$

$$\boldsymbol{e}|\beta \sim \prod_n \mathcal{N}(0, \beta^{-1}) \quad \text{and} \quad \beta \sim \Gamma(0,0)$$

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

- The use of gamma hyperpriors leads to the convenient posterior

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

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

- The EM algorithm can be used to estimate $\{\boldsymbol{\alpha}, \beta\}$ jointly with $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$.
  Can implement with an $\mathcal{O}(NK^2)$ recursion after an $O(N^2M)$ initialization.

[1] Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Machine Learning Res.*, 2001.

[2] Wipf and Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Processing*, 2004.

[3] Ji, Xue, and Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Processing*, 2008.

## 3c) Variable Selection:

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

$$\boldsymbol{y} \;=\; \boldsymbol{A}_S \boldsymbol{x}_S + \boldsymbol{e}.$$

- With $S$ known, estimation of the nonzero coefficients $\boldsymbol{x}_S$ is easy:

$$\begin{aligned}\hat{\boldsymbol{x}}_{\mathsf{LS}|S} &= (\boldsymbol{A}_S^T \boldsymbol{A}_S)^{-1} \boldsymbol{A}_S^T \boldsymbol{y} \\ \hat{\boldsymbol{x}}_{\mathsf{MMSE}|S} &= (\boldsymbol{A}_S^T \boldsymbol{A}_S + \sigma_e^2 \boldsymbol{I})^{-1} \boldsymbol{A}_S^T \boldsymbol{y}\end{aligned}$$

  This motivates the variable-selection problem:

  *From $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$, estimate the active-coefficient set $S$.*

  a long-standing problem in statistics!

[1] Hocking, "The analysis and selection of variables in linear regression," *Biometrics*, 1976.

## Non-Bayesian Variable Selection:

Consider the GLRT detector:

$$\begin{aligned}
\hat{S}_{\mathsf{GLRT}} &= \arg\max_S p(\boldsymbol{y}|S, \hat{\boldsymbol{x}}_{\mathsf{ML}|S}) \\
&= \arg\min_S \|\boldsymbol{y} - \boldsymbol{A}_S \underbrace{(\boldsymbol{A}_S^T \boldsymbol{A}_S)^{-1} \boldsymbol{A}_S^T \boldsymbol{y}}_{\hat{\boldsymbol{x}}_{\mathsf{ML}|S}} \|_2^2 \\
&= \arg\min_S \|\mathcal{P}^{\perp}_{\boldsymbol{A}_S} \boldsymbol{y}\|_2^2 \\
&= \text{any } S : |S| \geq M.
\end{aligned}$$

It fails!

This happens whenever the models are "nested".

## Bayesian Variable Selection:

Consider the MAP model estimate:

$$\begin{aligned}\hat{S}_{\mathsf{MAP}} &= \arg\max_S p(S|\boldsymbol{y}) \\ &= \arg\max_S p(\boldsymbol{y}|S)p(S) \\ &= \arg\max_S \int_{\boldsymbol{x}} \underbrace{p(\boldsymbol{y}|S,\boldsymbol{x})}_{\mathcal{N}} p(\boldsymbol{x}|S)d\boldsymbol{x} \cdot p(S).\end{aligned}$$

Need to specify the priors $p(\boldsymbol{x}|S)$ and $p(S)\ldots$

[1] Lempers, *Posterior probabilities of alternative linear models*, Rotterdam: Rotterdam Univ. Press, 1971

[2] Mitchell & Beauchamp, "Bayesian variable selection in linear regression," *J. Amer. Statist. Assoc.*, 1988.

[3] George & McCulloch, "Variable selection via Gibbs sampling," *J. Amer. Statist. Assoc.*, 1993.

[4] Smith & Kohn, "Nonparametric regression using Bayesian variable selection," *J. Econometrics*, 1996.

[5] George & McCulloch, "Approaches for Bayesian variable selection," *Statist. Sinica*, 1997.

[6] George, "The variable selection problem," *J. Amer. Statist. Assoc.*, 2000.

## Popular BVS Priors:

- iid Bernoulli coefficient-activity:

$$p(S) \;=\; \lambda^{|S|}(1-\lambda)^{(N-|S|)} \quad \text{where } \lambda < 0.5 \text{ induces sparsity,}$$

- Gaussian active-coefficients $\boldsymbol{x}_S$:

$$p(\boldsymbol{x}_S|S) \;\sim\; \mathcal{N}(\mu\mathbf{1}_{|S|}, \boldsymbol{R}_S)$$

$$\text{for} \begin{cases} \boldsymbol{R}_S = \sigma_x^2 \boldsymbol{I}_{|S|}, \;\; \mu \in \mathbb{R} & \text{“iid”} \\ \boldsymbol{R}_S = \sigma_x^2 (\boldsymbol{A}_S^T \boldsymbol{A}_S)^{-1}, \; \mu = 0 & \text{“Zellner } g\text{-prior”} \end{cases}$$

where the hyperparameters $\{\lambda, \mu, \sigma_x^2, \sigma_e^2\}$ could be treated as...

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

[1] Cui & George, “Empirical Bayes vs. fully Bayes variable selection,” *J. Statist. Planning Infer.*, 2008.

## BVS Posteriors:

Fixing $\{\lambda, \mu, \sigma_x^2, \sigma_e^2\}$, we get

- the model's posterior log-density:

$$\ln p(S|\boldsymbol{y}) \;=\; -\tfrac{1}{2}\big\|\boldsymbol{y} - \mu\boldsymbol{A}_S\boldsymbol{1}_{|S|}\big\|^2_{\boldsymbol{\Phi}_S^{-1}} - \tfrac{1}{2}\ln\det(\boldsymbol{\Phi}_S) - |S|\ln(\tfrac{1-\lambda}{\lambda}) + C,$$

  where $\boldsymbol{\Phi}_S$ denotes the observation covariance matrix given $S$:

$$\boldsymbol{\Phi}_S \;=\; \begin{cases} \sigma_x^2\boldsymbol{A}_S\boldsymbol{A}_S^T + \sigma_e^2\boldsymbol{I}_{|S|} & \text{(iid)} \\ \sigma_x^2\boldsymbol{A}_S(\boldsymbol{A}_S^T\boldsymbol{A}_S)^{-1}\boldsymbol{A}_S^T + \sigma_e^2\boldsymbol{I}_{|S|} & \text{(Zellner)} \end{cases}.$$

- the ($S$-conditional) coefficient density:

$$p(\boldsymbol{x}_S|\boldsymbol{y}, S) \;\sim\; \mathcal{N}\big(\boldsymbol{\mu}_S, \boldsymbol{\Sigma}_S\big)$$

  where

$$\begin{aligned} \boldsymbol{\mu}_S \;&=\; \mu\boldsymbol{1}_{|S|} + \boldsymbol{R}_S\boldsymbol{A}_S^T\boldsymbol{\Phi}_S^{-1}(\boldsymbol{y} - \mu\boldsymbol{A}_S\boldsymbol{1}_{|S|}) \;=\; \hat{\boldsymbol{x}}_{\mathsf{MMSE}|S} \\ \boldsymbol{\Sigma}_S \;&=\; \boldsymbol{R}_S - \boldsymbol{R}_S\boldsymbol{A}_S^T\boldsymbol{\Phi}_S^{-1}\boldsymbol{A}_S\boldsymbol{R}_S \;=\; \operatorname{cov}(\hat{\boldsymbol{x}}_{\mathsf{MMSE}|S}). \end{aligned}$$

## Connections to Model-Order Selection:

Under the Zellner prior, it can be shown that

$$\hat{S}_{\mathsf{MAP}} \;=\; \arg\min_{S}\left\{\tfrac{1}{\sigma_e^2}\big\|\boldsymbol{y}-\boldsymbol{A}_S\hat{\boldsymbol{x}}_{\mathsf{LS}|S}\big\|_2^2+|S|\cdot\eta\right\}$$
$$\text{for}\quad \eta=\tfrac{\sigma_x^2+\sigma_e^2}{\sigma_x^2}\ln\big((1+\tfrac{\sigma_x^2}{\sigma_e^2})(\tfrac{1-\lambda}{\lambda})^2\big).$$

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

$$\hat{S}_{\mathsf{AIC}} \;=\; \arg\min_{S}\left\{\tfrac{1}{\sigma_e^2}\big\|\boldsymbol{y}-\boldsymbol{A}_S\hat{\boldsymbol{x}}_{\mathsf{LS}|S}\big\|_2^2+|S|\cdot 2\right\}$$
$$\hat{S}_{\mathsf{BIC}} \;=\; \arg\min_{S}\left\{\tfrac{1}{\sigma_e^2}\big\|\boldsymbol{y}-\boldsymbol{A}_S\hat{\boldsymbol{x}}_{\mathsf{LS}|S}\big\|_2^2+|S|\cdot \ln M\right\}$$
$$\hat{S}_{\mathsf{RIC}} \;=\; \arg\min_{S}\left\{\tfrac{1}{\sigma_e^2}\big\|\boldsymbol{y}-\boldsymbol{A}_S\hat{\boldsymbol{x}}_{\mathsf{LS}|S}\big\|_2^2+|S|\cdot 2\ln N\right\}.$$

[1] George & Foster, "Calibration and empirical Bayes variable selection," *Biometrika*, 2000.

## Connections to Noncoherent Decoding:

- Consider a generic communication system with vectorized observations

$$\boldsymbol{y} \;=\; \boldsymbol{A}_j \boldsymbol{h} + \boldsymbol{e},$$

  where $j =$ codeword index, $\boldsymbol{h} =$ channel gains, and $\boldsymbol{e} =$ AWGN.

- In *non-coherent decoding*, $\boldsymbol{h}$ is known only in distribution, e.g., $\boldsymbol{h} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I}_N)$ for Rayleigh fading. Then

$$\hat{\jmath}_{\mathsf{MAP}} \;=\; \arg \max_{j=1,\ldots,J} p(j|\boldsymbol{y}).$$

- This can be interpreted in our sparse reconstruction framework via

$$\boldsymbol{A} \triangleq \big[\boldsymbol{A}_1 \cdots \boldsymbol{A}_J\big] \;\text{ and }\; \boldsymbol{x} \triangleq [\boldsymbol{0}^T, \ldots, \boldsymbol{h}^T, \ldots, \boldsymbol{0}^T]^T$$

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

$$\mathcal{S} \;\triangleq\; \big\{(1,\ldots,K), (K{+}1,\ldots,2K), \ldots, (JK{-}K{+}1,\ldots,JK)\big\},$$

  with the result that

$$\hat{S}_{\mathsf{MAP}} \;=\; \arg \max_{S \in \mathcal{S}} p(S|\boldsymbol{y}) \;\Leftrightarrow\; \hat{\jmath}_{\mathsf{MAP}}$$

## Bayesian Model Averaging:

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

$$\hat{S}_{\text{MAP}} = \arg\max_{S} p(S|\boldsymbol{y})$$

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

$$\hat{\boldsymbol{x}}_{\text{MMSE}|\hat{S}_{\text{MAP}}} = \text{E}\{\boldsymbol{x}|\boldsymbol{y}, \hat{S}_{\text{MAP}}\}.$$

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

$$\hat{\boldsymbol{x}}_{\text{MMSE}} = \sum_{S} \hat{\boldsymbol{x}}_{\text{MMSE}|S}\ p(S|\boldsymbol{y})$$

  that are well approximated by summing over the *few* most probable $S$.

[1] Leamer, *Specification Searches*, New York: Wiley 1978.

[2] Raftery, Madigan, & Hoeting, "Bayesian model averaging for linear regression models," *J. Amer. Statist. Assoc.*, 1997.

[3] Clyde and George, "Model Uncertainty," *Statist. Sci.*, 2004.

## 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 $\hat{\boldsymbol{x}}_{\mathsf{MMSE}|S}$) with very low complexity:
  - iid Gaussian $\boldsymbol{x}_S$: "Fast Bayesian matching pursuit"
  - Zellner Gaussian $\boldsymbol{x}_S$: "Fast Zellner matching pursuit"

  In fact, the complexity order equals that of OMP: $\mathcal{O}(MNK)$.
- An expectation/maximization (EM) approach can be used to optimally tune the hyperparameters $\{\lambda, \mu, \sigma_x^2, \sigma_e^2\}$ wrt the data.

[1] Schniter, Potter, and Ziniel, "Fast Bayesian matching pursuit," *ITA*, 2008.

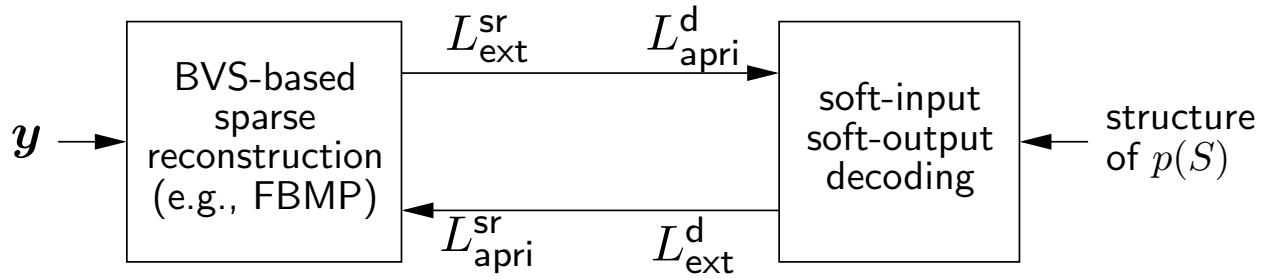
[2] Schniter, Potter, and Ziniel, "Fast Bayesian matching pursuit: Model uncertainty and parameter estimation for sparse linear models," Preprint, 2008.

## 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 $p(S|\boldsymbol{y})$ as output).

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




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

[1] Schniter, "Bayesian Sparse Reconstruction and Tracking using the Turbo Principle," *Preprint*, 2009.

## BMA versus RVM:

- To parameterize coefficient activity, RVM uses the *continuous* variables $\boldsymbol{\alpha}$, 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

$$p(\boldsymbol{x}|\boldsymbol{y}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

  whereas the BMA posterior is a Gaussian mixture:

$$p(\boldsymbol{x}|\boldsymbol{y}) \sim \sum_S \mathcal{N}\big(\boldsymbol{\mu}_S, \boldsymbol{\Sigma}_S\big)\; p(S|\boldsymbol{y}) \quad \ldots \text{more informative}$$

- Fast implementations of RVM and BMA have roughly the same complexity.
- Simulation results suggest better performance for BMA.

## Numerical Experiments — "Compressible" Signal:

Setup:

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

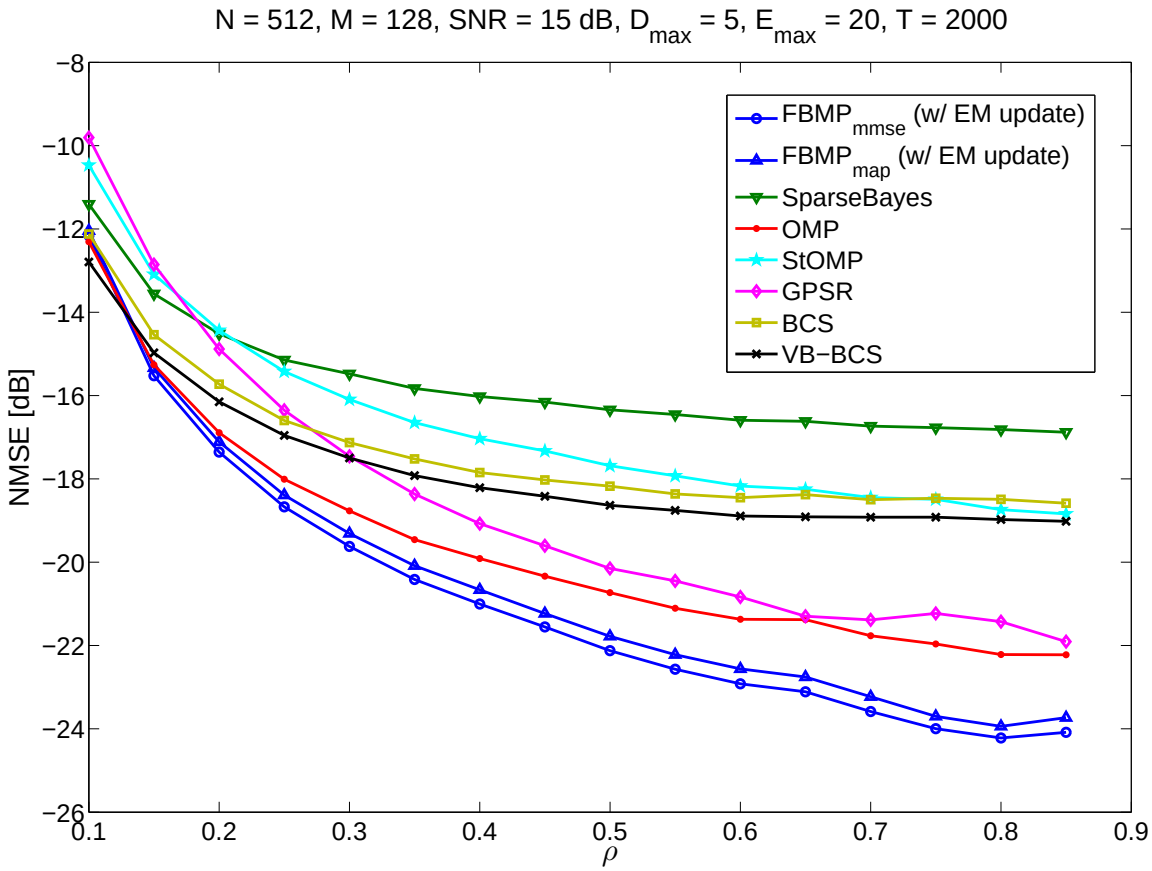
Algorithms:

| | |
|---|---|
| OMP | Tropp & Gilbert |
| StOMP | Donoho, Tsaig, Drori & Starck |
| GPSR-Basic | Figueiredo, Nowak & Wright ($\min_{\boldsymbol{x}} \|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 + \tau\|\boldsymbol{x}\|_1$) |
| SparseBayes | Wipf & Rao (RVM) |
| BCS | Ji & Carin (RVM) |
| FBMP | Schniter, Potter & Ziniel (BMA) |

Performance:

$$\text{NMSE} \triangleq \text{Avg}\left\{\frac{\|\hat{\boldsymbol{x}}-\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2}\right\} \text{ over } 2500 \text{ random trials.}$$

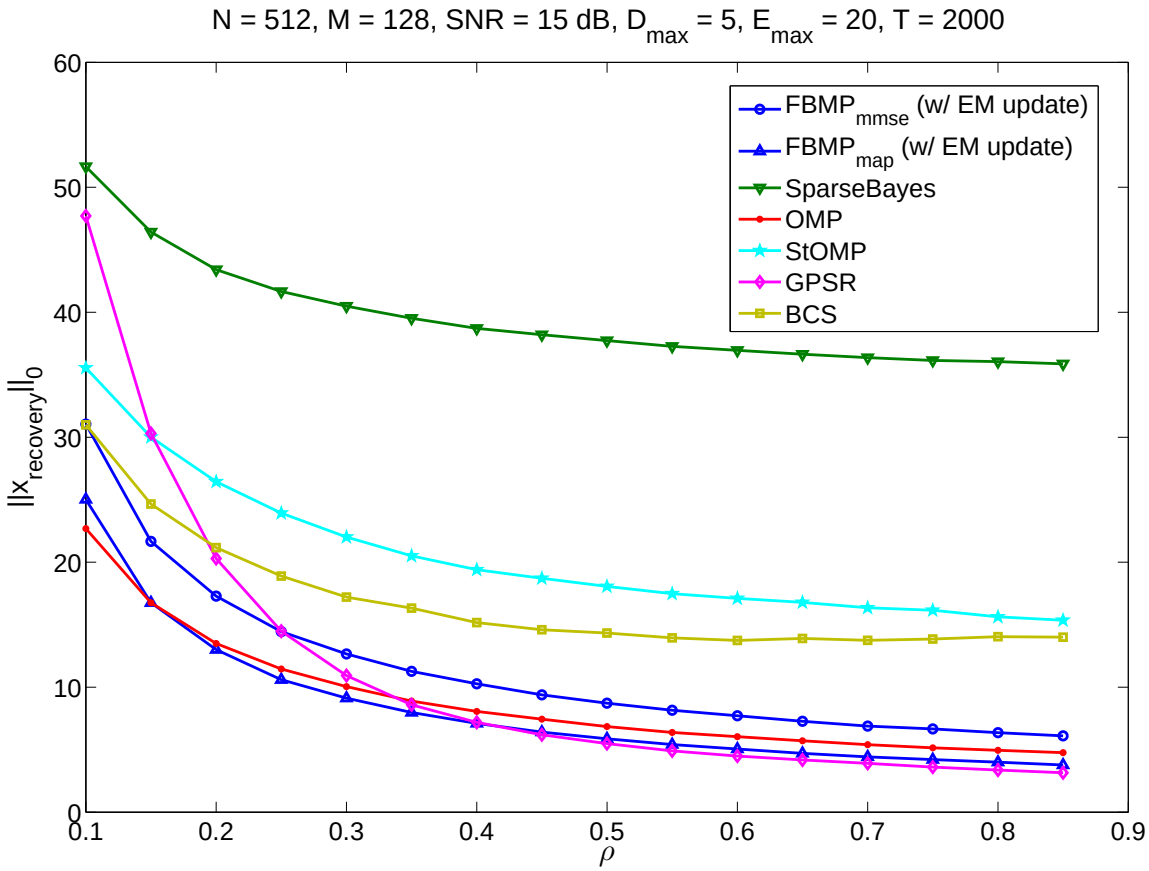
## NMSE versus decay rate $\rho$:




FBMP outperformed GPSR and OMP by 2 dB and others by much more.

Note: The signal priors favor GPSR!

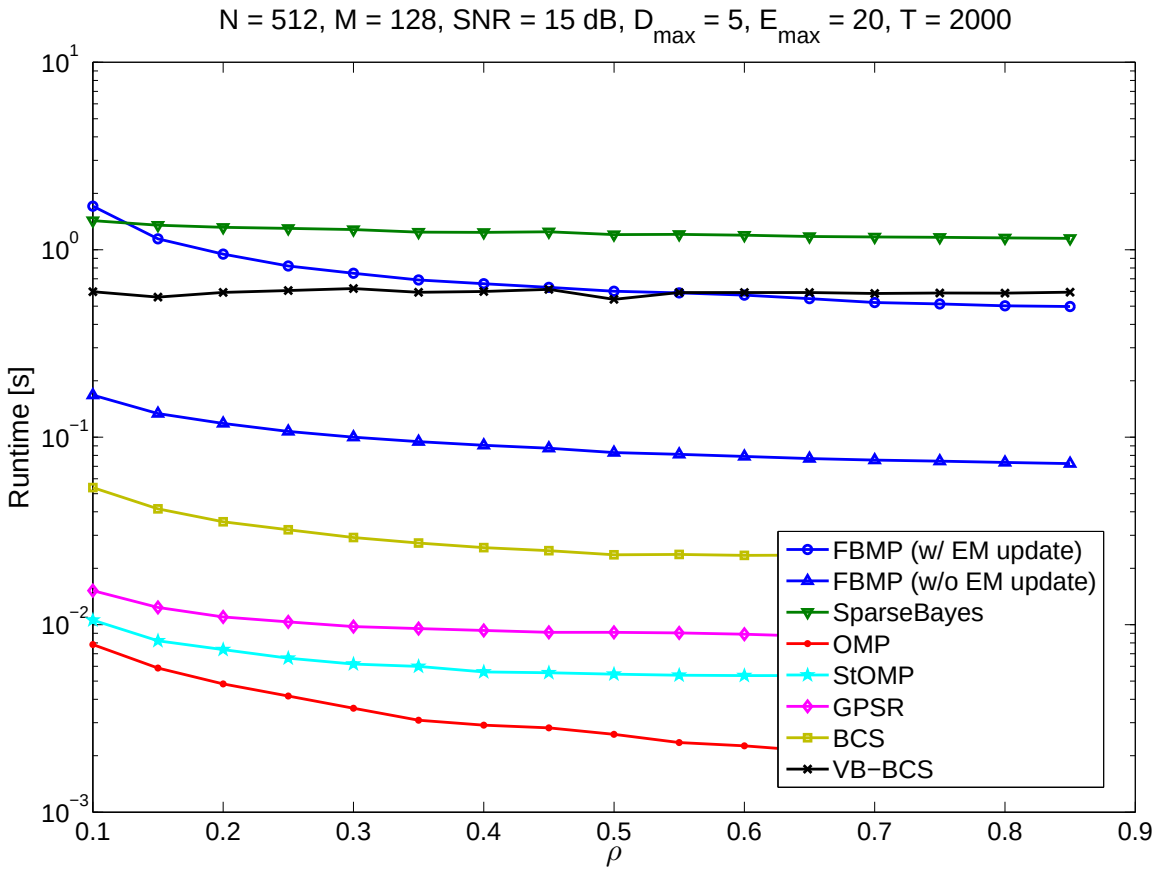
## Sparsity of estimate versus decay rate $\rho$:




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 $\boldsymbol{A}$:

> There exists $\delta_K \in (0,1)$ such that
> $\forall \boldsymbol{x}$ s.t. $\|\boldsymbol{x}\|_0 = K, \quad (1-\delta_K)\|\boldsymbol{x}\|_2^2 \;\le\; \|\boldsymbol{A}\boldsymbol{x}\|_2^2 \;\le\; (1+\delta_K)\|\boldsymbol{x}\|_2^2$ ,

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 $|\mu| > K_3\sqrt{M}\sigma_1 + K_4\sqrt{M}\sigma_e^2$.

[1] Som and Potter, "MAP Estimation of Support Set in Sparse Signal Recovery," *Preprint*, 2009.

## PWEP of Model Selection under general "Information Criteria":

**Lemma 1** *For generic $\mathcal{S}$, the pairwise error probability of*

$$\hat{S} \;=\; \arg\min_{S\in\mathcal{S}} \left\{ \tfrac{1}{\sigma_e^2} \left\| \boldsymbol{y} - \boldsymbol{A}_S \hat{\boldsymbol{x}}_{\mathsf{LS}|S} \right\|_2^2 + \eta |S| \right\} \quad \textit{under} \quad \boldsymbol{x}_S|S \sim \mathcal{N}(\boldsymbol{0}, \gamma\sigma_e^2 \boldsymbol{I}_{|S|})$$

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

$$P_{\hat{S}|S} \;\le\; (\alpha_{\hat{S},S}\gamma)^{-K_{\mathsf{m}}} C_{K_{\mathsf{m}},K_{\mathsf{f}}}(\eta),$$

*where $K_{\mathsf{m}}$ and $K_{\mathsf{f}}$ denote the # of missed and false-alarm coefficients, and*

$$C_{K_{\mathsf{m}},K_{\mathsf{f}}}(\eta) \;=\; \begin{cases} e^{(K_{\mathsf{m}}-K_{\mathsf{f}})\eta} \displaystyle\sum_{k=0}^{K_{\mathsf{f}}-1} \frac{(K_{\mathsf{f}}-K_{\mathsf{m}})^k \eta^k}{k!} \binom{K_{\mathsf{m}}+K_{\mathsf{f}}-1-k}{K_{\mathsf{m}}} & K_{\mathsf{m}} \le K_{\mathsf{f}}, \\ \displaystyle\sum_{k=0}^{K_{\mathsf{m}}} \frac{(K_{\mathsf{m}}-K_{\mathsf{f}})^k \eta^k}{k!} \binom{K_{\mathsf{m}}+K_{\mathsf{f}}-1-k}{K_{\mathsf{f}}-1} & K_{\mathsf{m}} > K_{\mathsf{f}}. \end{cases}$$

$$\alpha_{\hat{S},S} \;=\; \lambda_{\min}(\boldsymbol{A}_{\mathsf{m}}^T \boldsymbol{\Pi}^{\perp}_{\boldsymbol{A}_{\hat{S}}} \boldsymbol{A}_{\mathsf{m}}) \qquad \ldots \textit{Restricted Isometry Property}$$

[1] Schniter, "On Model Selection under the Generalized Information Criteria," *Preprint*, 2009.

## 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 inludes 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.