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

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

- *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?

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:

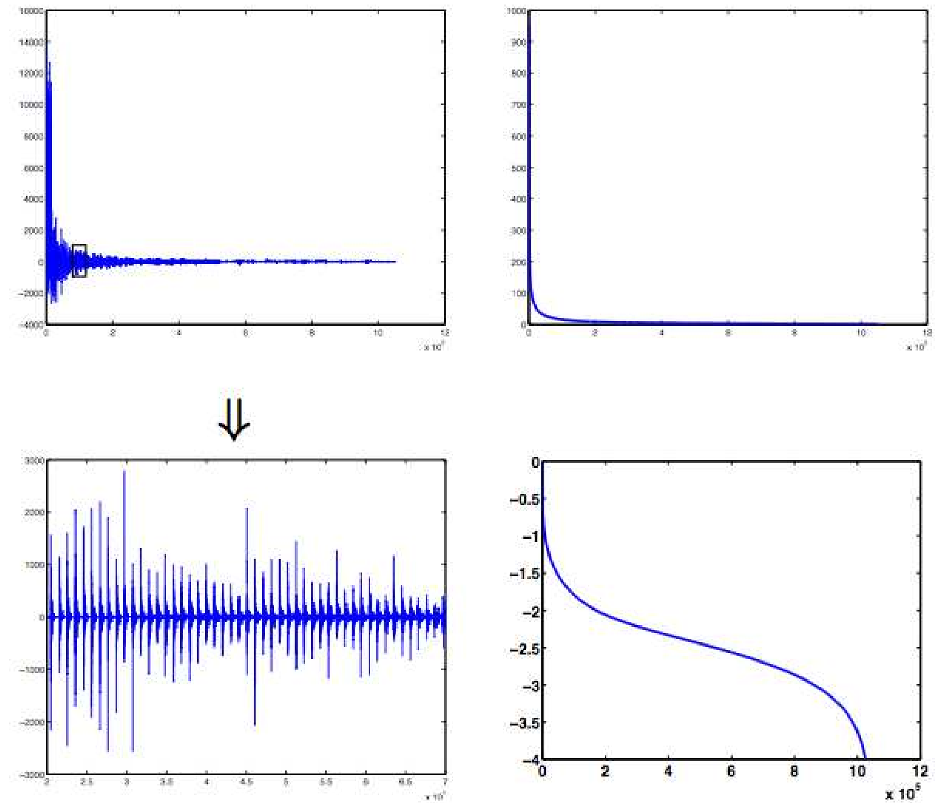
$$\boxed{\boldsymbol{W}\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{r}} \; : \; \begin{cases} \boldsymbol{W} & \text{unitary DWT operator,} \\ \boldsymbol{z} & \text{signal,} \\ \boldsymbol{x} & \text{sparse: only a few non-zero coefficients,} \\ \boldsymbol{r} & \text{small residual.} \end{cases}$$

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

DWT Example:

1 megapixel image

wavelet coeffs (sorted)



zoom in ($\log_{10}$ sorted)

Typically: MSE ≈ -20 dB from only 2.5% 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 $\boldsymbol{z}$, take $M \ll N$ linear measurements $\boxed{\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{z}}$ from which $\boldsymbol{z}$ can be accurately reconstructed.

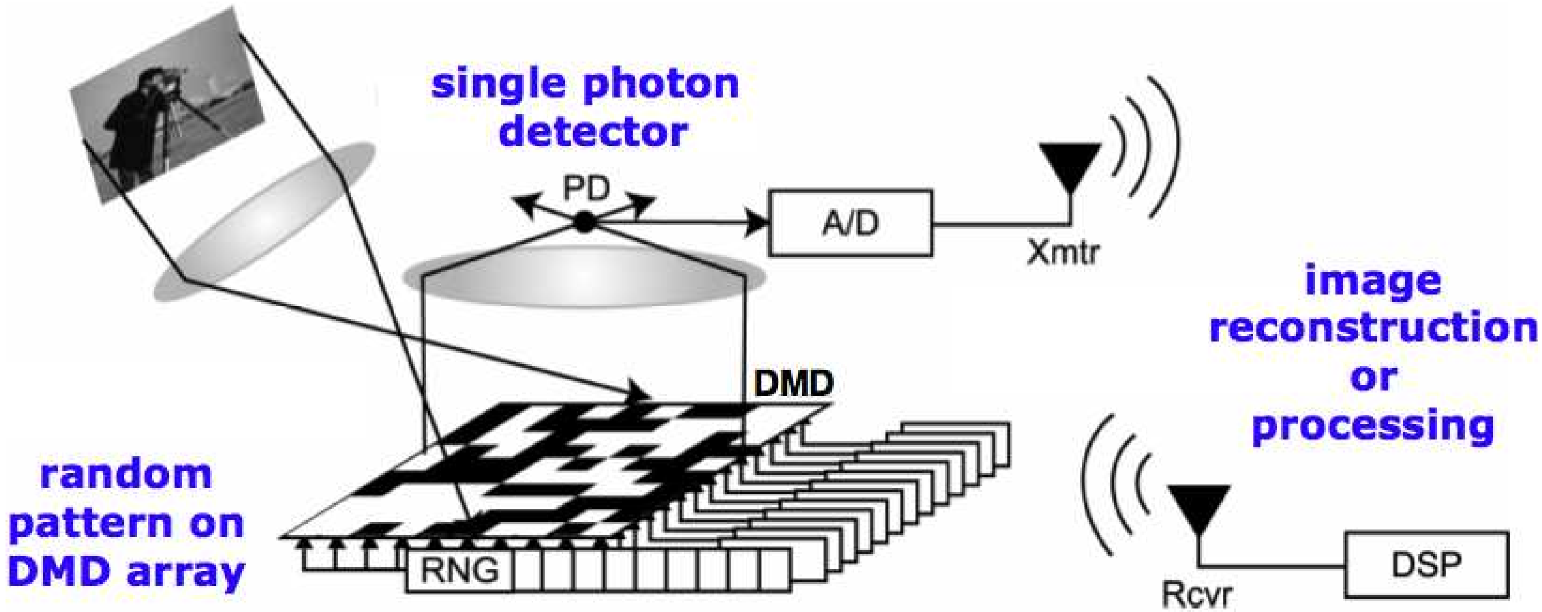
  Note: If the signal class has K significant degrees of freedom, then we need $M \geq K$ measurements. But the hope is that $M \not\gg K$.

- But how can we make the measurement scheme $\boldsymbol{\Phi}$ *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* $\boldsymbol{\Phi}$ *with random entries.*

Example — The Rice Univ. Single-Pixel Camera:



target
65536 pixels

11000 measurements
(16%)

1300 measurements
(2%)

This and the previous DWT figure courtesy of Rich Baraniuk, Rice University.

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

Putting these together, we get

$$\begin{aligned} \boldsymbol{y} &= \underbrace{\boldsymbol{\Phi W}^H}_{\boldsymbol{A}} \boldsymbol{x} + \underbrace{\boldsymbol{\Phi W}^H \boldsymbol{r} + \boldsymbol{\nu}}_{\boldsymbol{n}} \\ &= \boldsymbol{A x} + \boldsymbol{n} \end{aligned}$$

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

The remaining (and big!) challenge:

Reconstruct the sparse signal representation $\boldsymbol{x}$ from the "compressed" measurements $\boldsymbol{y} = \boldsymbol{A x} + \boldsymbol{n}$.

Other Applications of Sparse Reconstruction:

Sparse channel estimation:

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

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

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

- And sometimes over-determined:

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

$$\text{since} \quad \mathrm{E}\{\|\hat{\boldsymbol{x}}_{LS} - \boldsymbol{x}\|_2^2\} \geq \frac{K\sigma_n^2}{\|\boldsymbol{a}\|_2^2} \quad \text{for } K = \underbrace{\|\boldsymbol{x}\|_0}_{\#\text{ nonzero coefs}}.$$

Solving the Sparse Reconstruction Problem:

Key question:

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

Popular Techniques:

1. Constrained ℓ_1-minimization via convex optimization.
2. Matching pursuits.
3. Bayesian approaches.

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

Constrained ℓ_0-Minimization:

The classical problem formulation:

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

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

Unfortunately, this problem is NP-hard!

Let's think about this problem geometrically. . .

A Toy Problem:

Consider the setup

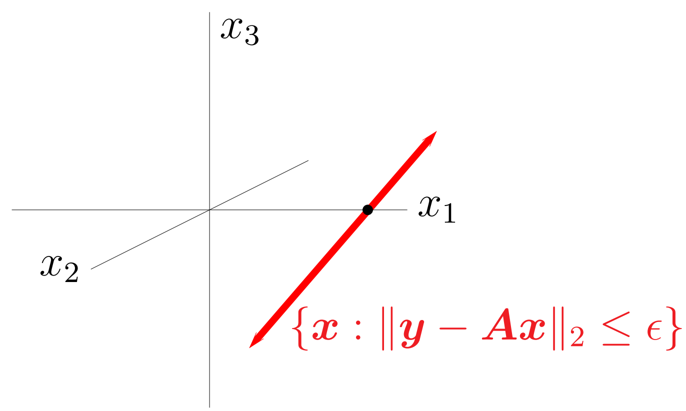
$$\overset{\boldsymbol{y}}{\begin{bmatrix} \bullet \\ \bullet \end{bmatrix}} = \overset{\boldsymbol{A}}{\begin{bmatrix} \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \end{bmatrix}} \overset{\boldsymbol{x}}{\begin{bmatrix} \bullet \\ \bullet \\ \bullet \end{bmatrix}} + \overset{\boldsymbol{n}}{\begin{bmatrix} \bullet \\ \bullet \end{bmatrix}} \qquad \begin{matrix} N = 3 \\ M = 2 \\ K = 1 \end{matrix}$$

Since $N = M + 1$,

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

Since $K = 1$,

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

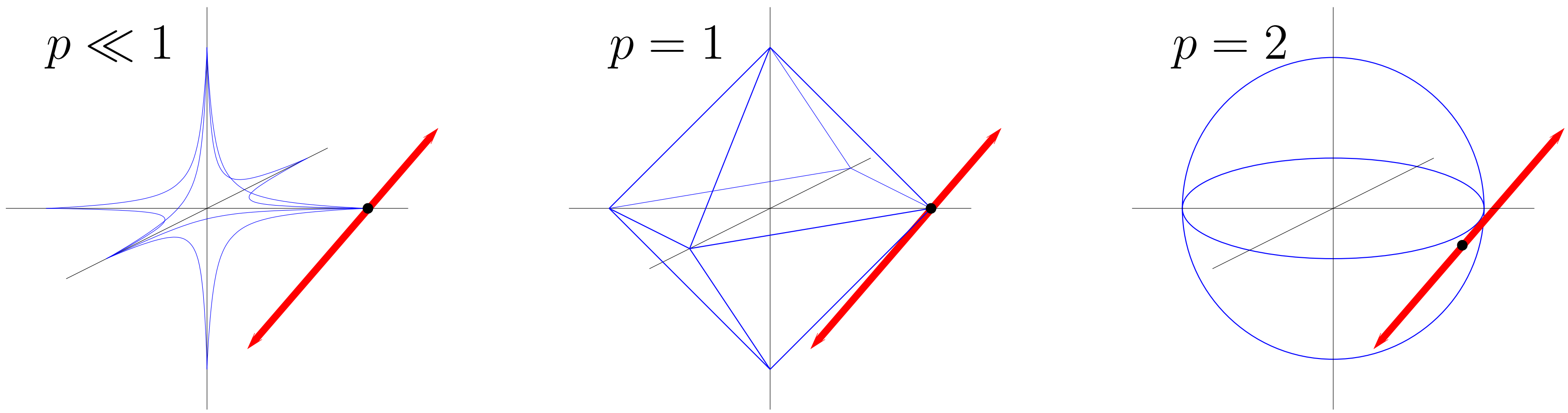


The Geometry of Constrained ℓ_p-Minimization:

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 ℓ_p-ball until it touches the ϵ-rod:



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

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

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

This suggests to use the ℓ_1 norm as a surrogate for the ℓ_0 norm.

1) Constrained ℓ_1-Minimization:

For the constrained-ℓ_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., $M \gtrsim K \log N$) 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.

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 $\boldsymbol{x}$.

  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 $\boldsymbol{x}$, then the second most important on, and so on...

    Usually: correlate $\boldsymbol{y}$ with each (normalized) column of $\boldsymbol{A}$.

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

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

$$i_\star = \arg\max_{i \notin \mathcal{I}} \frac{|\boldsymbol{a}_i^H \boldsymbol{e}|^2}{\|\boldsymbol{a}_i\|_2^2}, \quad \mathcal{I} \leftarrow \mathcal{I} \cup i_\star,$$

$$\hat{x}_{i_\star} = \frac{\boldsymbol{a}_{i_\star}^H \boldsymbol{e}}{\|\boldsymbol{a}_{i_\star}\|_2^2},$$

$$\boldsymbol{e} \leftarrow \boldsymbol{e} - \boldsymbol{a}_{i_\star} \hat{x}_{i_\star}$$

end.

Similar to "successive interference cancellation" for CDMA.

Orthogonal Matching Pursuit (OMP):

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

Initialize: $\boldsymbol{e} = \boldsymbol{y}$, $\mathcal{I} = \emptyset$, and $\boldsymbol{B} = [\,]$.
While $\|\boldsymbol{e}\|_2 >$ threshold,

$$i_\star = \arg\max_{i \notin \mathcal{I}} \frac{|\boldsymbol{a}_i^H \boldsymbol{e}|^2}{\|\boldsymbol{a}_i\|_2^2}, \quad \mathcal{I} \leftarrow \mathcal{I} \cup i_\star, \quad \boldsymbol{B} \leftarrow [\boldsymbol{B}, \boldsymbol{a}_{i_\star}],$$

$$\hat{\boldsymbol{x}}_{\mathcal{I}} = (\boldsymbol{B}^H \boldsymbol{B})^{-1} \boldsymbol{B}^H \boldsymbol{y},$$

$$\boldsymbol{e} \leftarrow \boldsymbol{e} - \boldsymbol{B}\hat{\boldsymbol{x}}_{\mathcal{I}}$$

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 $\boldsymbol{A}$ (e.g., nearly uncorrelated 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.*, Vol. 53, No. 12, Dec. 2007.

[2] D. Needell and R. Vershynin, "Uniform Uncertainty Principle and Signal Recovery via Regularized Orthogonal Matching Pursuit," 2007 (Preprint).

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 σ^2-variance AWGN and signal $\boldsymbol{x}$ such that

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

then the MAP estimate becomes

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

which is the constrained ℓ_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 λ?

Another Example — Sparse Bayesian Learning (SBL):

The so-called SBL method uses the signal model:

$$\begin{aligned} x_i|\sigma_i^2 &\sim \mathcal{N}(0,\sigma_i^2) \\ \{\sigma_i^{-2}\}_{i=1}^N &\sim \text{ i.i.d. } \Gamma(a,b) \quad \text{. . . for tractability!} \\ a,b &: \text{ chosen to induce sparsity,} \end{aligned}$$

and the noise model:

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

The EM algorithm can then be used to find the MAP estimate of $\boldsymbol{x}$.

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

[2] D. Wipf and B. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Processing*, vol. 52, 2004.

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:

$$\boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I}).$$

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

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

$$\lambda = \mathrm{E}[K]/N.$$

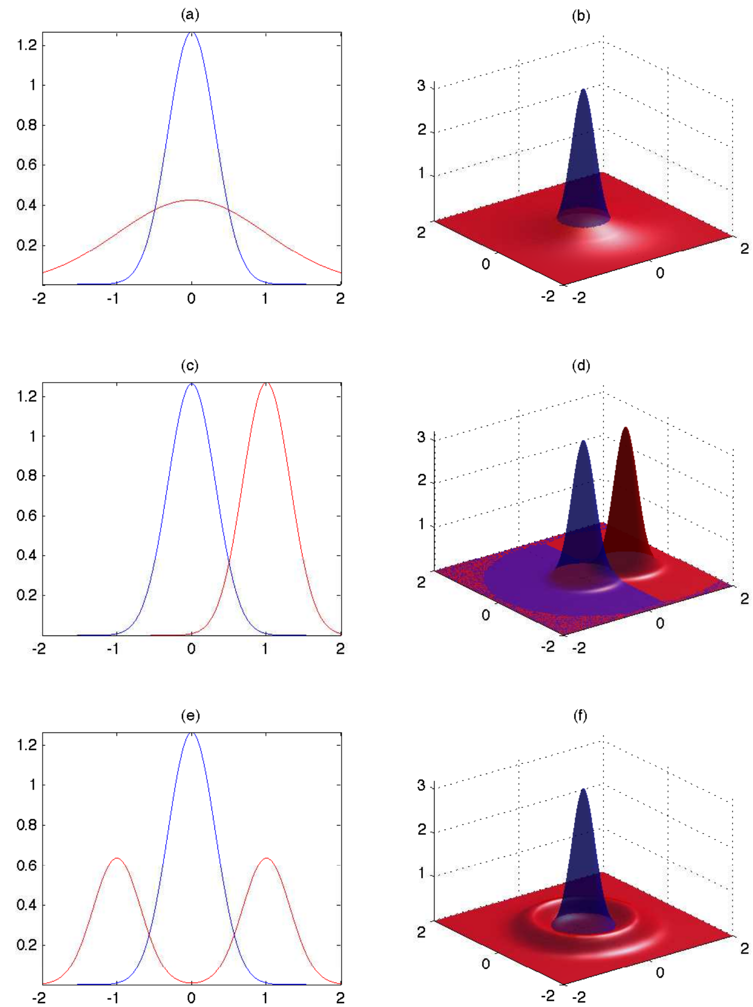
A Q-ary Signal Model:

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

$$x_n|s_n \sim \begin{cases} 0 & s_n = 0 \\ \mathcal{N}(\nu_1, \sigma_1^2) & s_n = 1 \\ \vdots & \vdots \\ \mathcal{N}(\nu_2, \sigma_2^2) & s_n = Q \end{cases}$$

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

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:

$$\hat{\boldsymbol{x}}_{\text{mmse}} = \mathrm{E}[\boldsymbol{x}|\boldsymbol{y}] = \sum_{\boldsymbol{s}\in\{0,1\}^N} p(\boldsymbol{s}|\boldsymbol{y})[\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}]$$

$$\approx \sum_{\boldsymbol{s}\in\mathcal{S}_\star} p(\boldsymbol{s}|\boldsymbol{y}) \underbrace{\mathrm{E}[\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}]}_{\hat{\boldsymbol{x}}_{\text{mmse}}|\boldsymbol{s}}$$

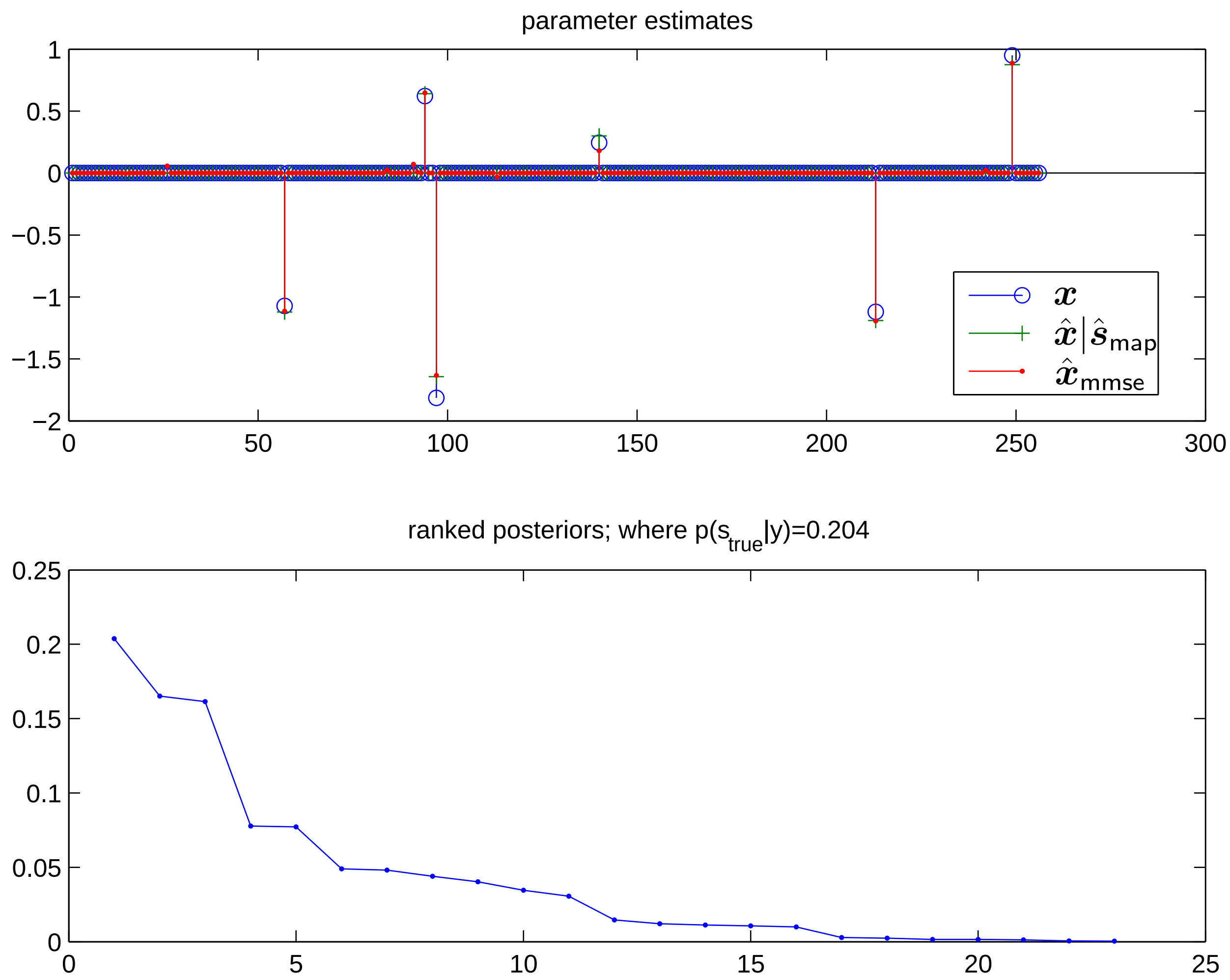
where $\mathcal{S}_\star$ is the set of *all probable configurations* $\boldsymbol{s} = [s_0, \ldots, s_{N-1}]^T$.

Comments:

- The MMSE estimate leverages the inherent uncertainty in estimating $\boldsymbol{s}$. An example of "Bayesian model averaging."
- We will need to search for all probable configurations $\mathcal{S}_\star$, rather than the single best configuration $\hat{\boldsymbol{s}}_{\text{map}} \triangleq \arg\max_{\boldsymbol{s}} p(\boldsymbol{s}|\boldsymbol{y})$.

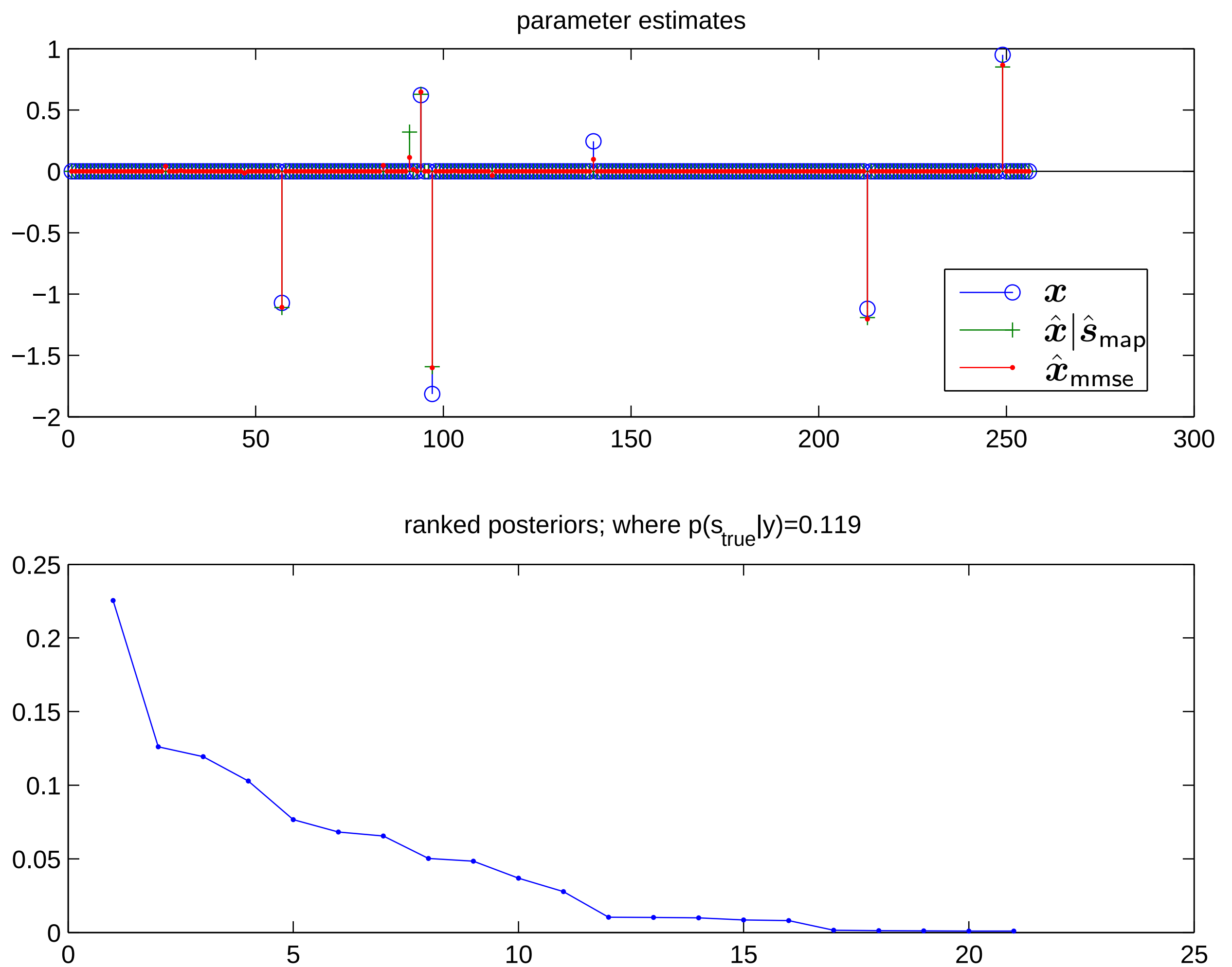
Example – Leveraging Uncertainty:

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



Example – Leveraging Uncertainty (cont.):

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



Configuration Metric:

- Posterior probability of configuration $\boldsymbol{s}$:

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

- Maximizing $p(\boldsymbol{s}|\boldsymbol{y}) \quad \Leftrightarrow \quad$ maximizing the "configuration metric" $\nu(\boldsymbol{s})$:

$$\nu(\boldsymbol{s}) \;\triangleq\; \log p(\boldsymbol{y}|\boldsymbol{s})p(\boldsymbol{s}).$$

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

$$\begin{aligned}\nu(\boldsymbol{s}) \;=\; & -\tfrac{M}{2}\log 2\pi - \tfrac{1}{2}\log\det(\boldsymbol{\Sigma}(\boldsymbol{s})) - \tfrac{1}{2}\boldsymbol{y}^T\boldsymbol{\Sigma}(\boldsymbol{s})^{-1}\boldsymbol{y} \\ & + \|\boldsymbol{s}\|_0 \log\tfrac{\lambda}{1-\lambda} + N\log(1-\lambda),\end{aligned}$$

where $\boldsymbol{\Sigma}(\boldsymbol{s}) \triangleq \mathrm{Cov}\{\boldsymbol{y}|\boldsymbol{s}\} = \boldsymbol{A}\,\mathcal{D}(\boldsymbol{s})\boldsymbol{A}^T + \sigma^2\boldsymbol{I}_M$.

Greedy Search:

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

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

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

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

  Bayesian model averaging yields $\hat{\boldsymbol{x}}_{\textsf{mmse}} \approx \sum_{\boldsymbol{s}\in\hat{\mathcal{S}}_\star} \hat{p}(\boldsymbol{s}|\boldsymbol{y})\ \hat{\boldsymbol{x}}_{\textsf{mmse}}|\boldsymbol{s}$.

Greedy Search (cont.):

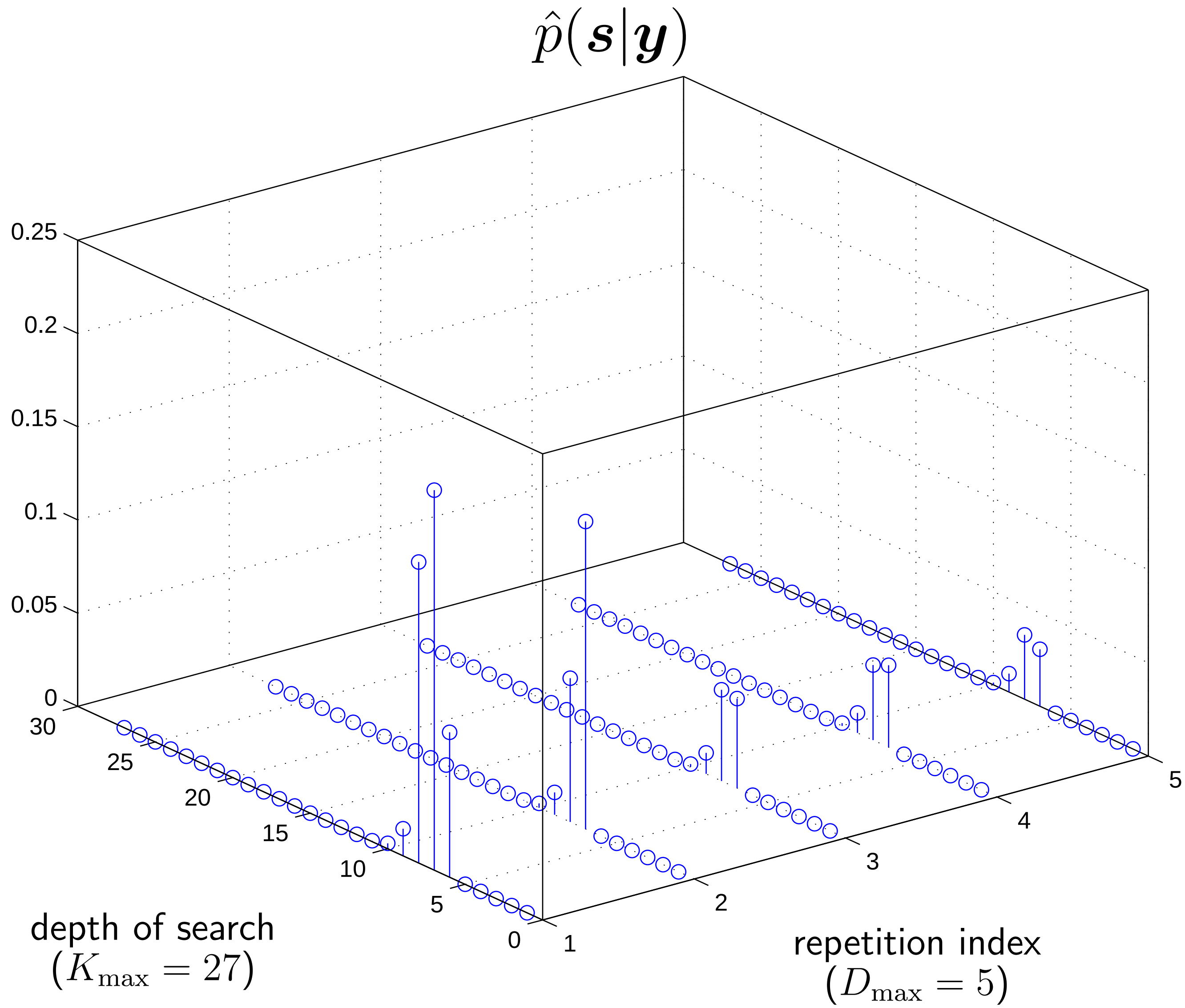
- Can choose the maximum search depth $K_{\max}$ so that $\Pr\{\|\boldsymbol{s}\|_0 > K_{\max}\} \triangleq \mathcal{P}_0$ is sufficiently small:

$$K_{\max} = \left\lceil \operatorname{erfc}^{-1}(2\mathcal{P}_0)\sqrt{2N\lambda(1-\lambda)} + N\lambda \right\rceil$$

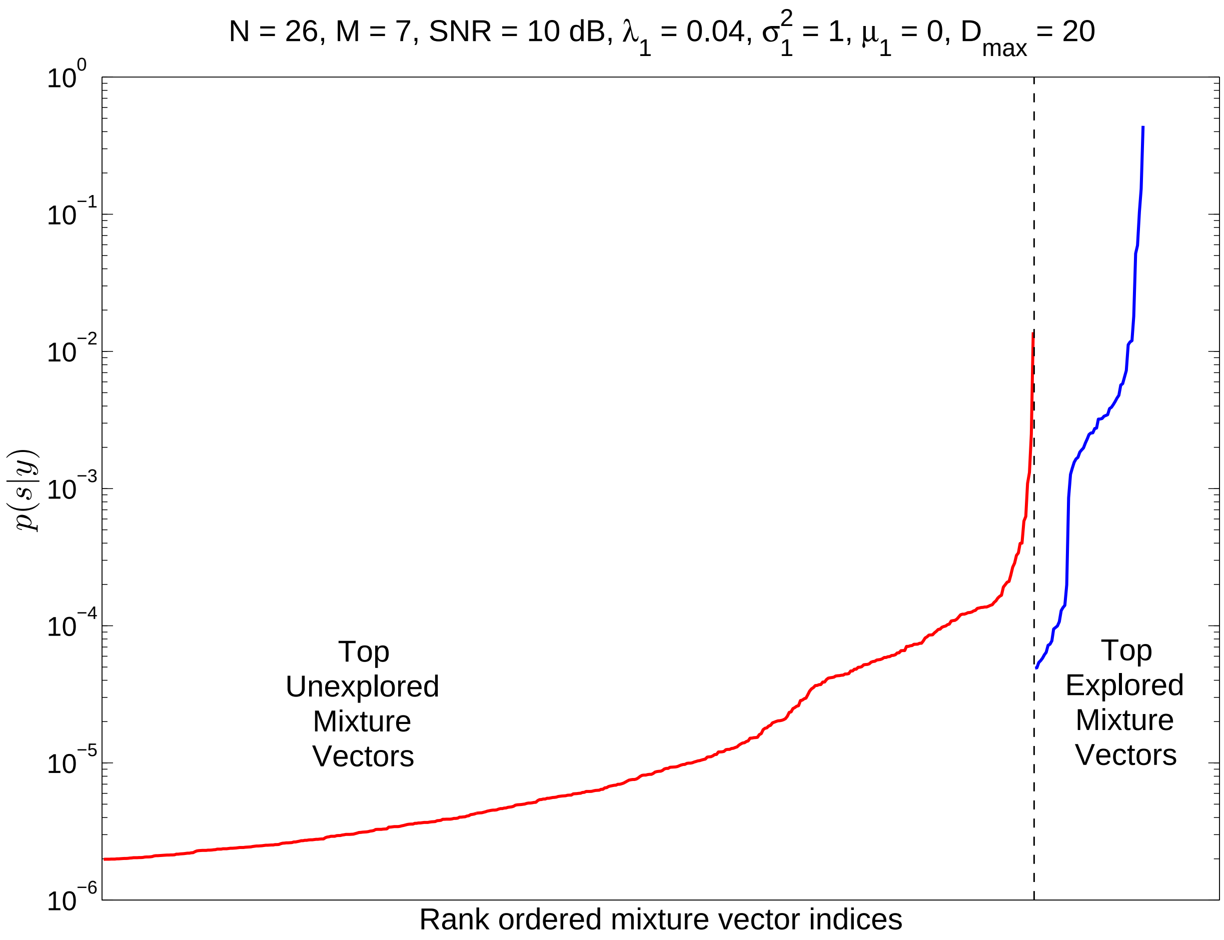
  via the Gaussian approximation of $\|\boldsymbol{s}\|_0 \sim$ Binomial(N, λ).

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

Example of repeated greedy search:



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.

    $\rightsquigarrow$ Total complexity $\mathcal{O}(NMK)$!

  But unlike other matching pursuits,
  - Our search is guided by maximization of the posterior $p(\boldsymbol{s}|\boldsymbol{y})$.
- Hence, the name *Fast Bayesian Matching Pursuit* (FBMP).

Estimation of Hyper-Parameters:

Until now we have assumed known hyper-parameters

$$\lambda = \Pr\{s_n = 1\}$$
$$\sigma^2 = \mathrm{var}\{x_n | s_n = 1\}$$

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

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

$$\hat{\boldsymbol{\theta}}^{(i+1)} = \arg\max_{\boldsymbol{\theta}} \sum_{\boldsymbol{s} \in \mathcal{S}_\star} p(\boldsymbol{s}|\boldsymbol{y}, \hat{\boldsymbol{\theta}}^{(i)}) \log p(\boldsymbol{y}, \boldsymbol{s}|\boldsymbol{\theta})$$

Connections to Noncoherent Decoding:

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

$$\boldsymbol{y} = \mathcal{T}(\boldsymbol{s})\boldsymbol{x} + \boldsymbol{n}$$

for Toeplitz matrix $\mathcal{T}(\boldsymbol{s})$.

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

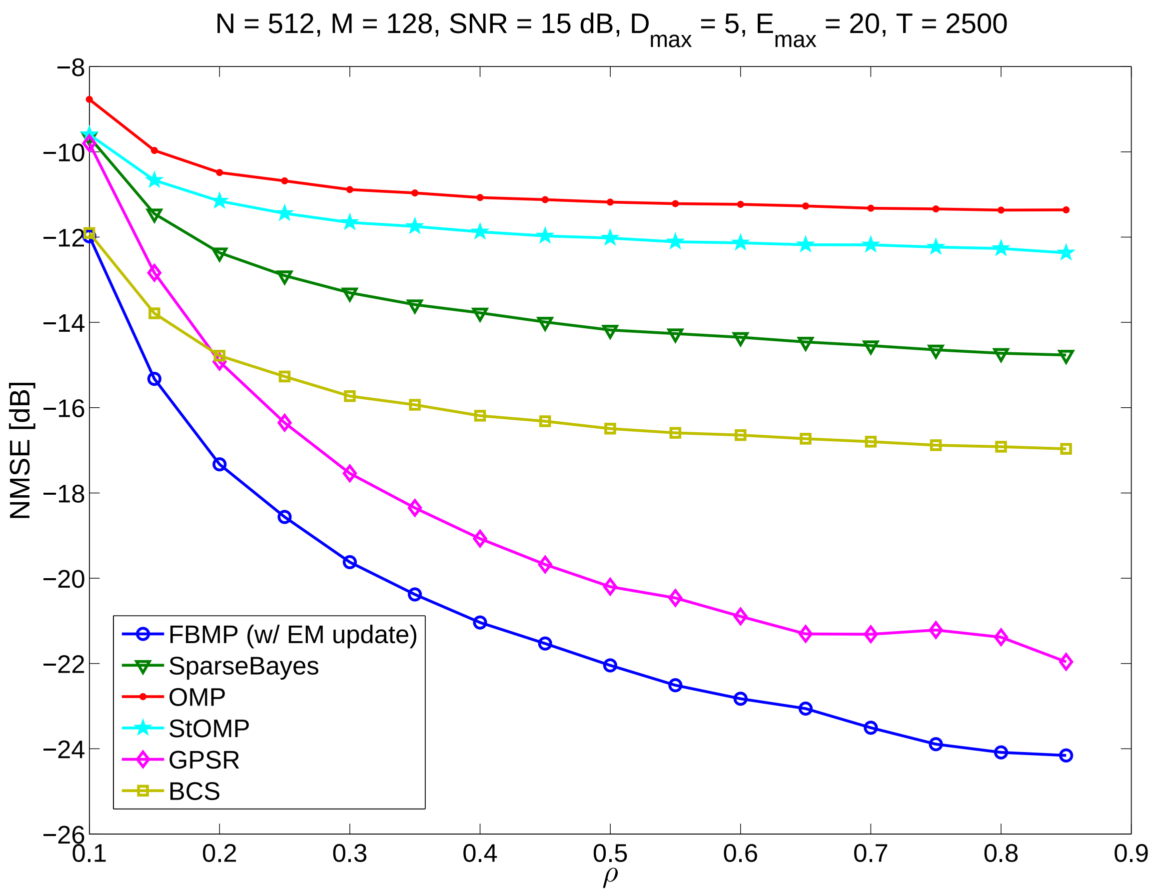
$$
\begin{aligned}
N &= 512 \\
M &= 128 \\
\boldsymbol{A} &: \text{ i.i.d. } \mathcal{N}(0,1) \text{ with columns scaled to unit norm} \\
\boldsymbol{x} &: x_n = e^{-\rho n} \quad \text{for decay rate } \rho \in (0,1) \\
\text{SNR} &= 15\,\text{dB}
\end{aligned}
$$

Algorithms:

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

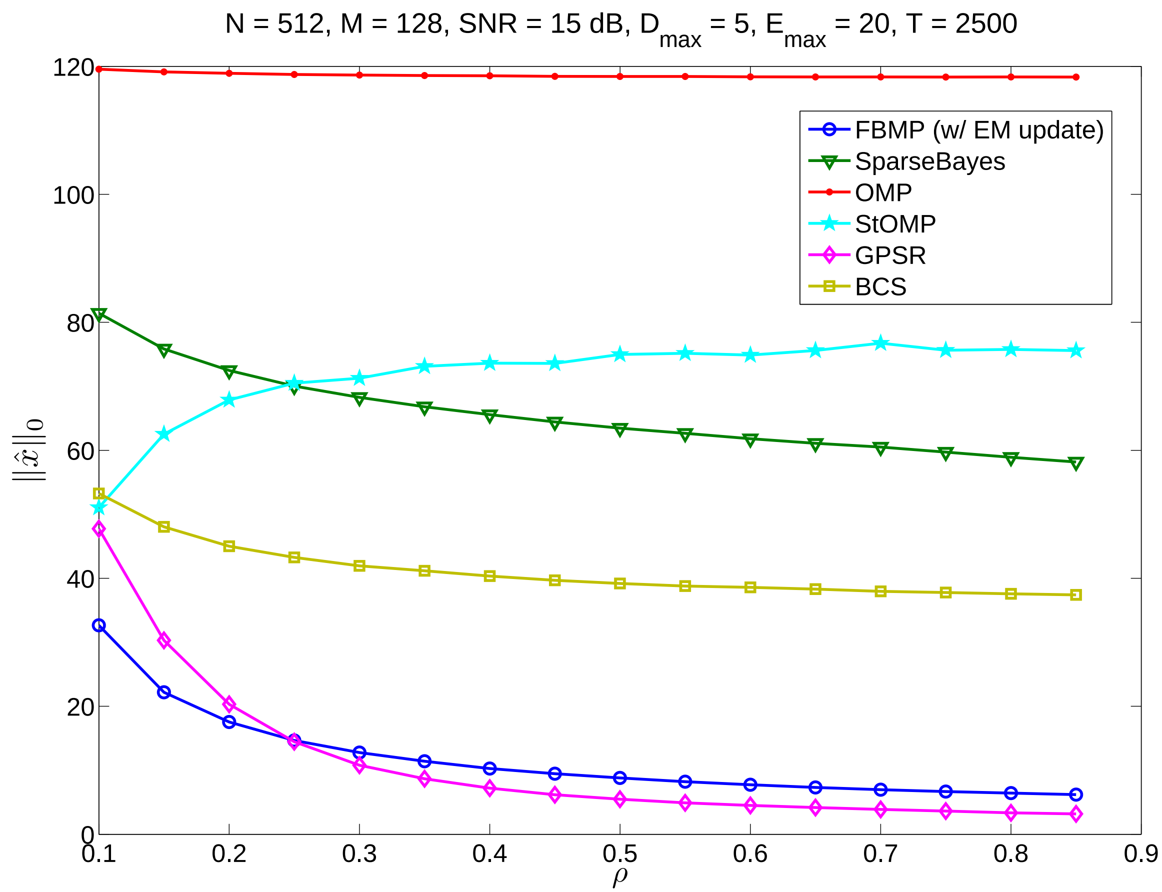
Performance: $\text{NMSE} \triangleq \text{Avg}\left\{\frac{\|\hat{\boldsymbol{x}}-\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2}\right\}$ over 2500 random trials.

NMSE versus decay rate ρ:



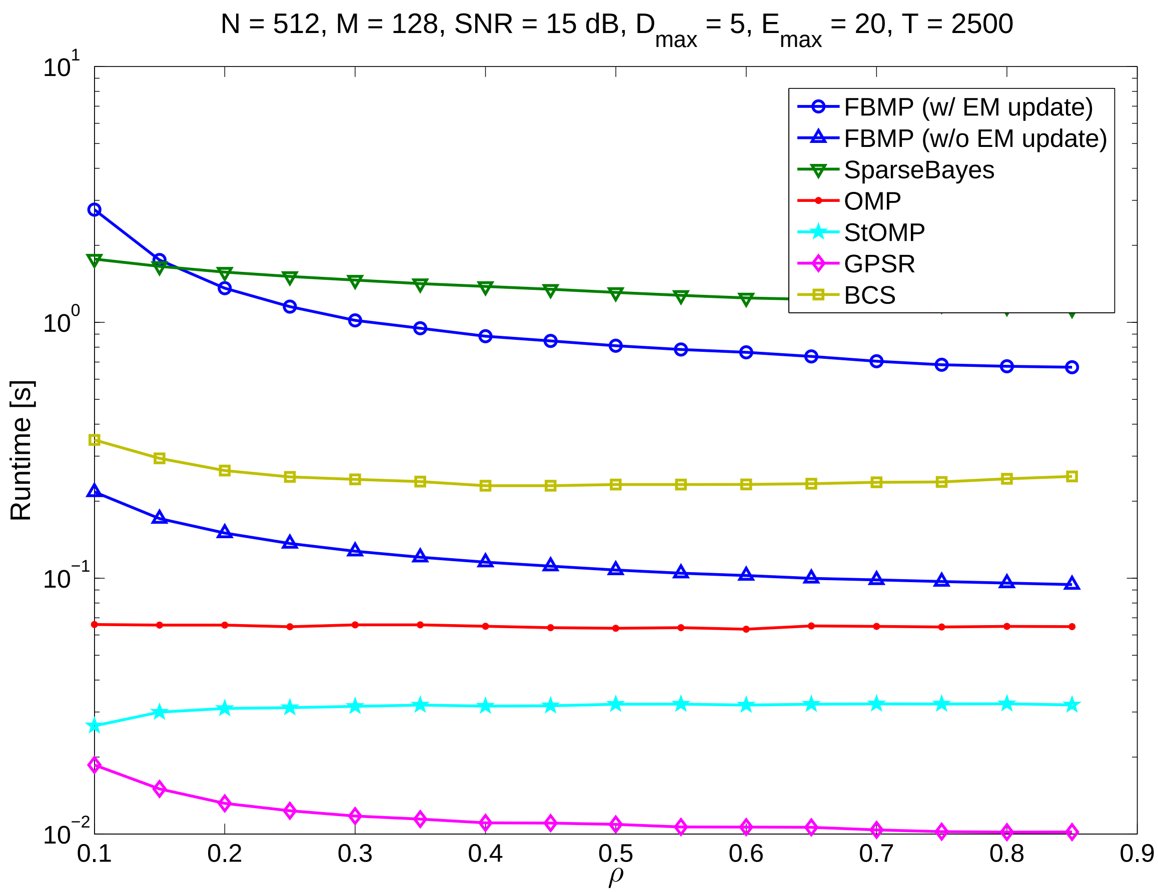
FBMP outperformed GPSR by 2 dB and others by much more. In general, NMSE performance suffers as ρ decreases (i.e., as $\boldsymbol{x}$ gets less sparse).

Sparsity of estimate versus decay rate ρ:



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

Runtime versus decay rate ρ:



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 $\mathcal{O}(NMK)$,
  - 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 $\hat{\boldsymbol{x}}_{\text{map}}$, $\hat{\boldsymbol{x}}_{\text{mmse}}$, and the greedy search.
  - How to track a sequence of correlated $\{\boldsymbol{x}\}$?

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:

$$N = 512$$

$$\lambda = 0.04 \qquad \text{...so } \mathrm{E}[K] = \lambda N = 20 \text{ active coefs}$$

$$M = 120$$

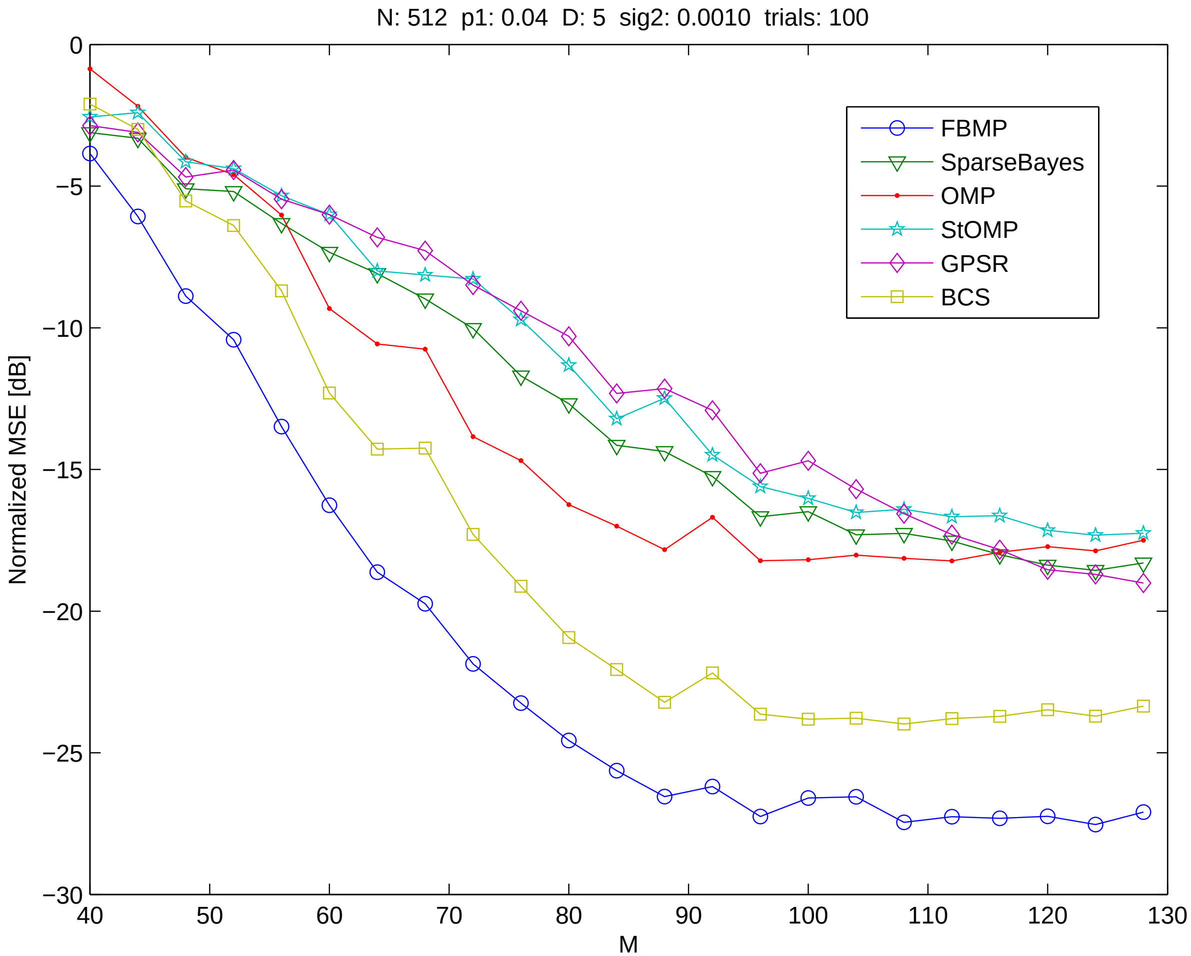
$$\text{SNR} = 19\,\text{dB} \qquad \text{...where SNR} \triangleq \frac{\mathrm{E}[K]}{\sigma^2 M}$$

$\boldsymbol{A}$: i.i.d. $\mathcal{N}(0,1)$ with columns scaled to unit norm.

Performance:

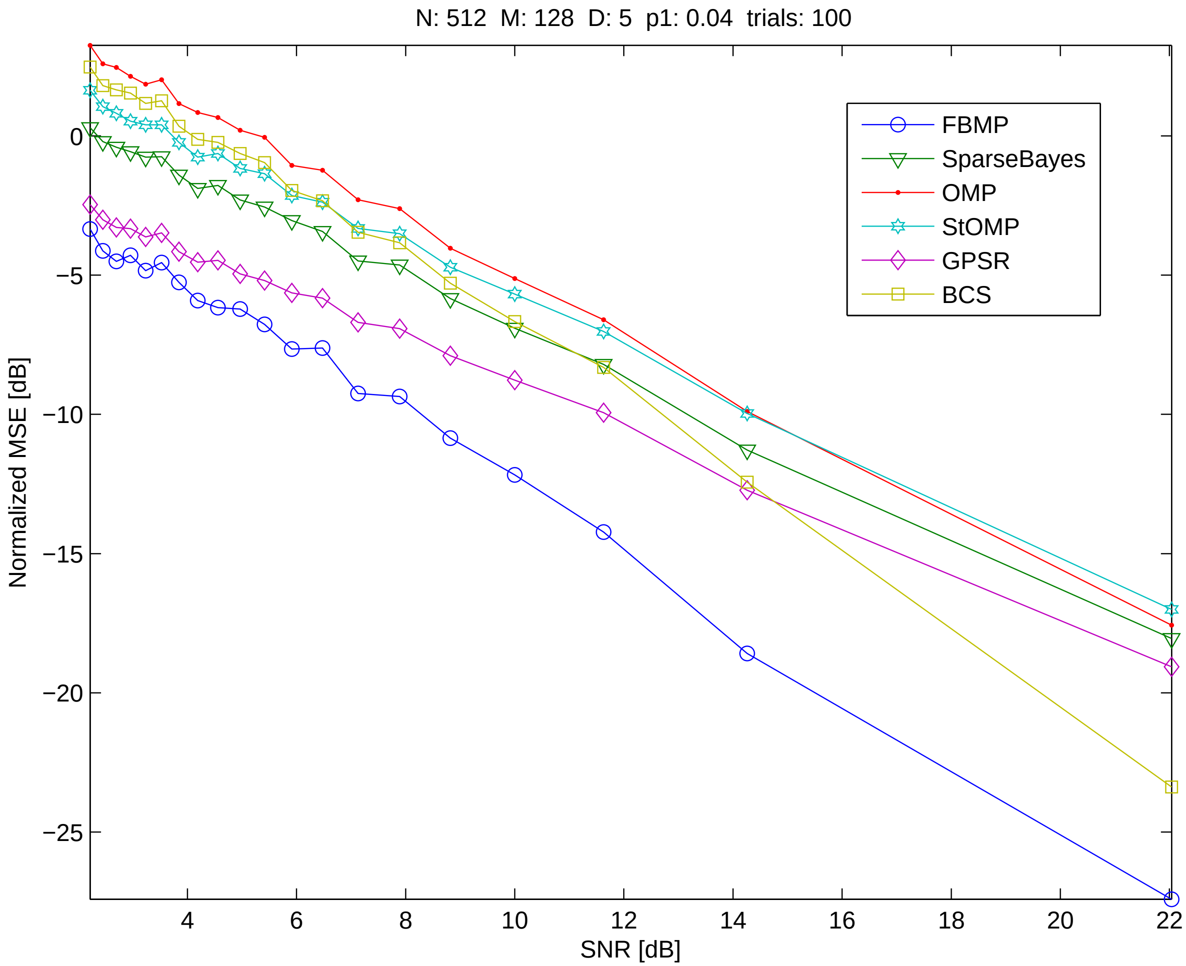
$$\text{NMSE} \triangleq \text{Avg}\left\{\frac{\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2}\right\} \text{ over } 100 \text{ random trials.}$$

NMSE versus observation length M:



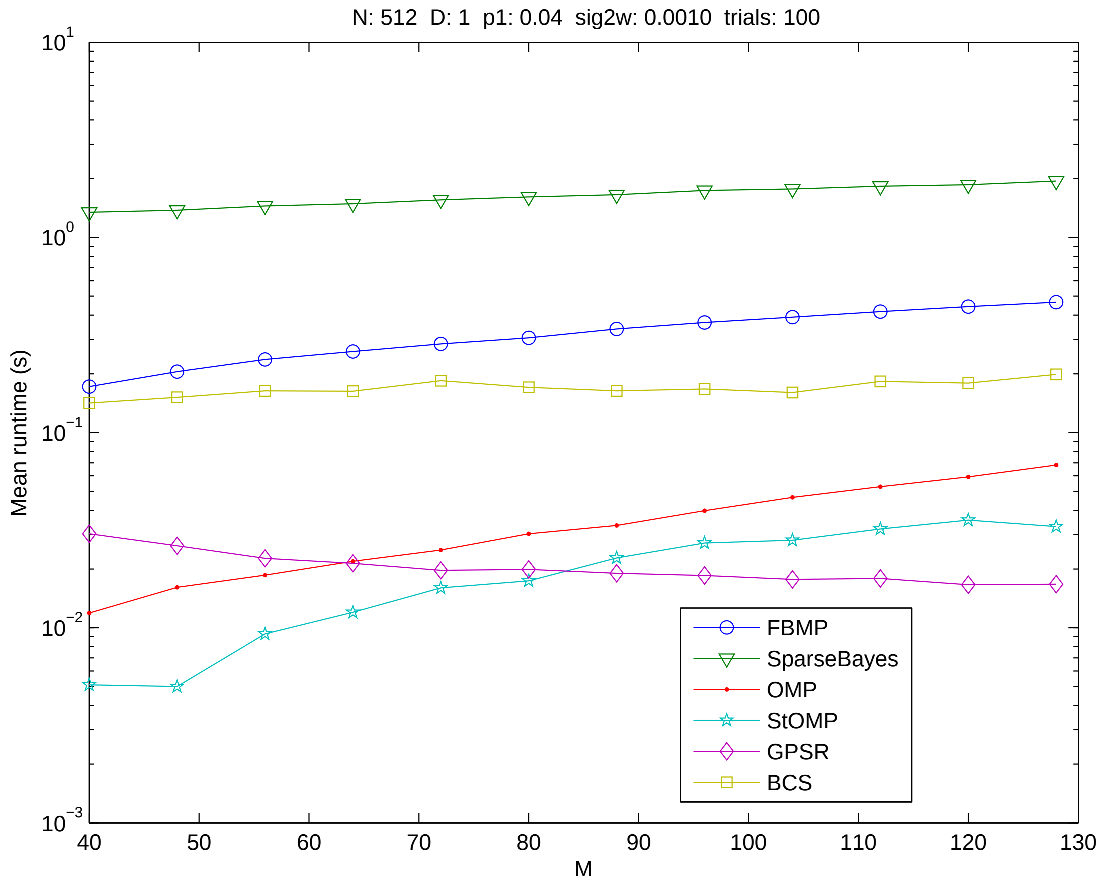
For $\frac{M}{\mathrm{E}[K]} > 5$, FBMP outperformed BCS by $3\,$dB and others by $\geq 10\,$dB. As $\frac{M}{\mathrm{E}[K]} \to 2$, all algorithms break down.

NMSE versus SNR:



At high SNR, FBMP outperformed BCS by 3 dB and others by ≥ 9 dB.
As SNR$\rightarrow 0$ dB, GPSR catches up.

Runtime versus observation length M:



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.)

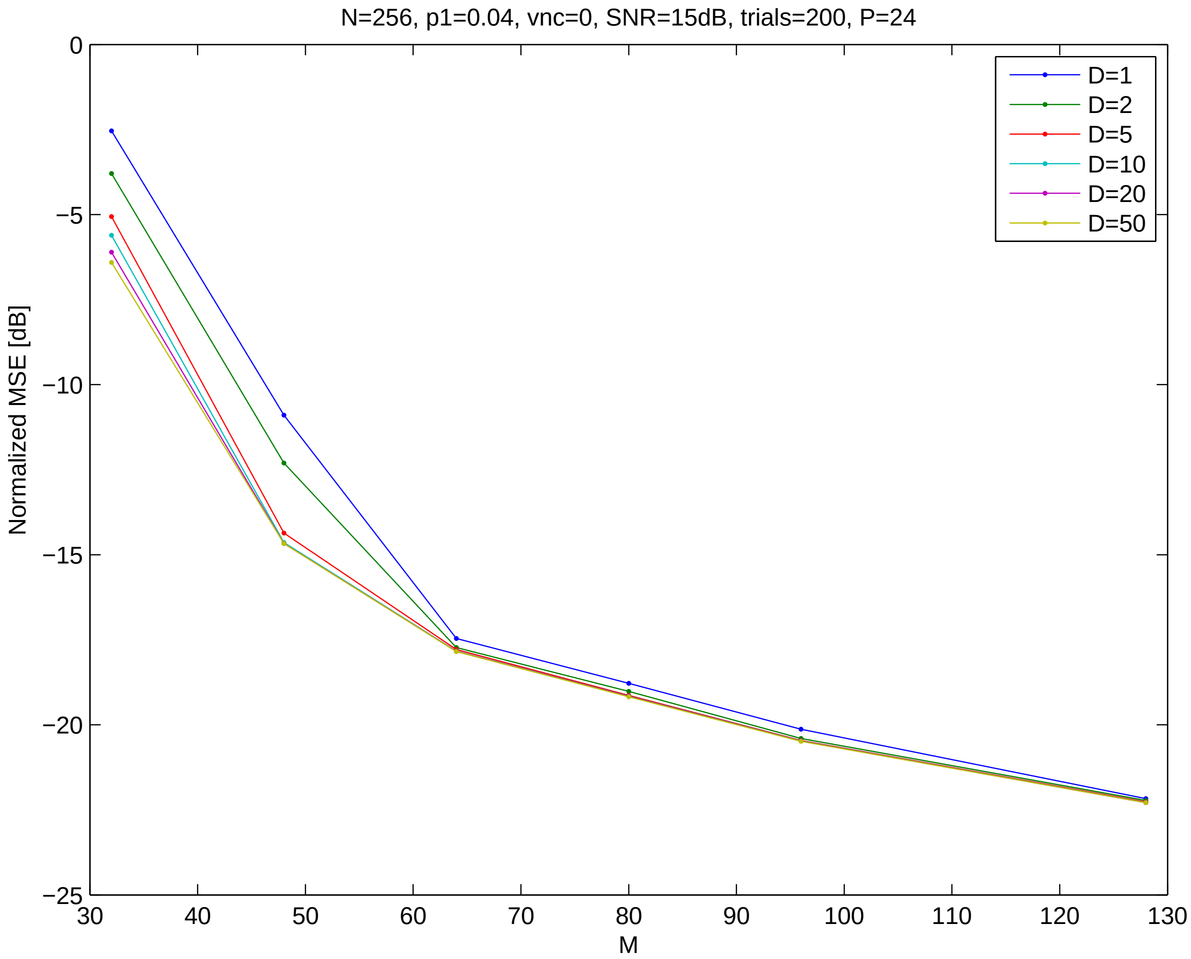
Nominal Signal Parameters:

$$N = 256$$
$$K = 10 \qquad \ldots \text{where FBMP uses } \lambda = K/N$$
$$M = 64$$
$$\boldsymbol{A} : \text{i.i.d. } \mathcal{N}(0,1) \text{ with columns scaled to unit norm.}$$
$$\text{SNR} = 15\,\text{dB}$$

Performance:

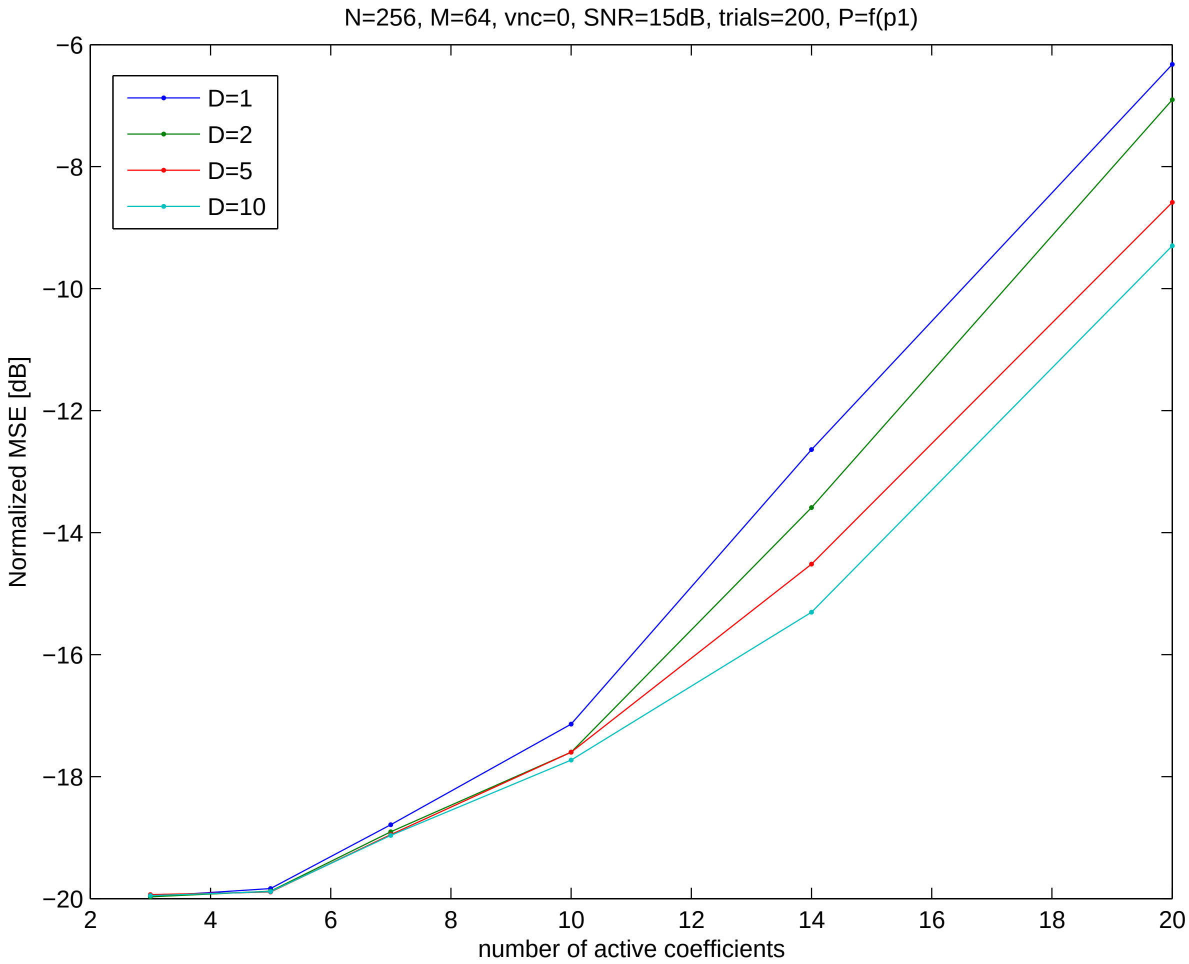
$$\text{NMSE} \triangleq \text{Avg}\left\{\frac{\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2}\right\} \text{ over 200 random trials.}$$

NMSE versus observation length M:



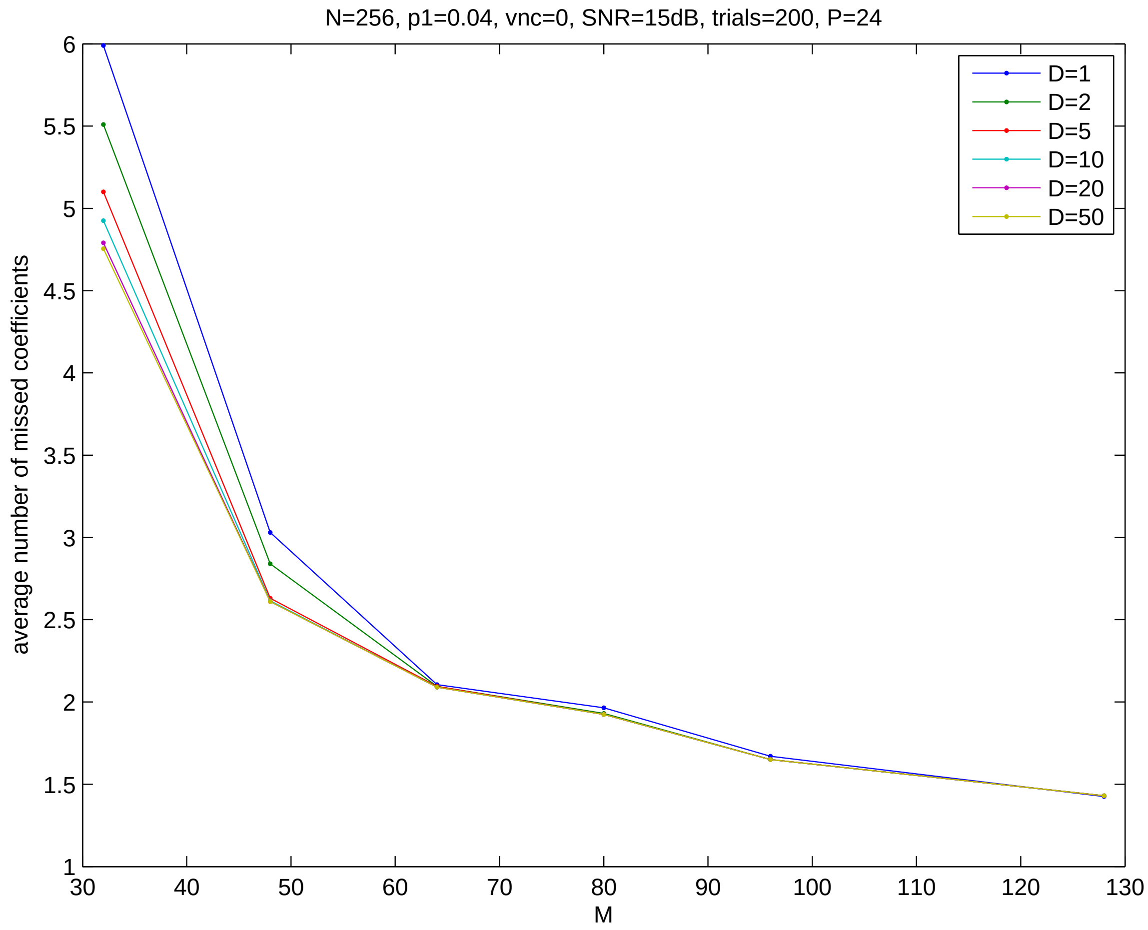
When $D = 1$, knee in curve at $\frac{M}{K} = \frac{64}{10} = 6.4 \ \frac{\text{measurements}}{\text{active coef}}$.
For larger D, knee moves to $5 \ \frac{\text{measurements}}{\text{active coef}}$ and NMSE improves by $3\,\text{dB}$.

NMSE versus # active coefs K:



When $D = 1$, knee in curve at $\frac{M}{K} = \frac{64}{10} = 6.4\ \frac{\text{measurements}}{\text{active coef}}$.
For $D = 10$, knee at $\frac{M}{K} = \frac{64}{13} = 4.9\ \frac{\text{measurements}}{\text{active coef}}$ and NMSE $3\,$dB improved.

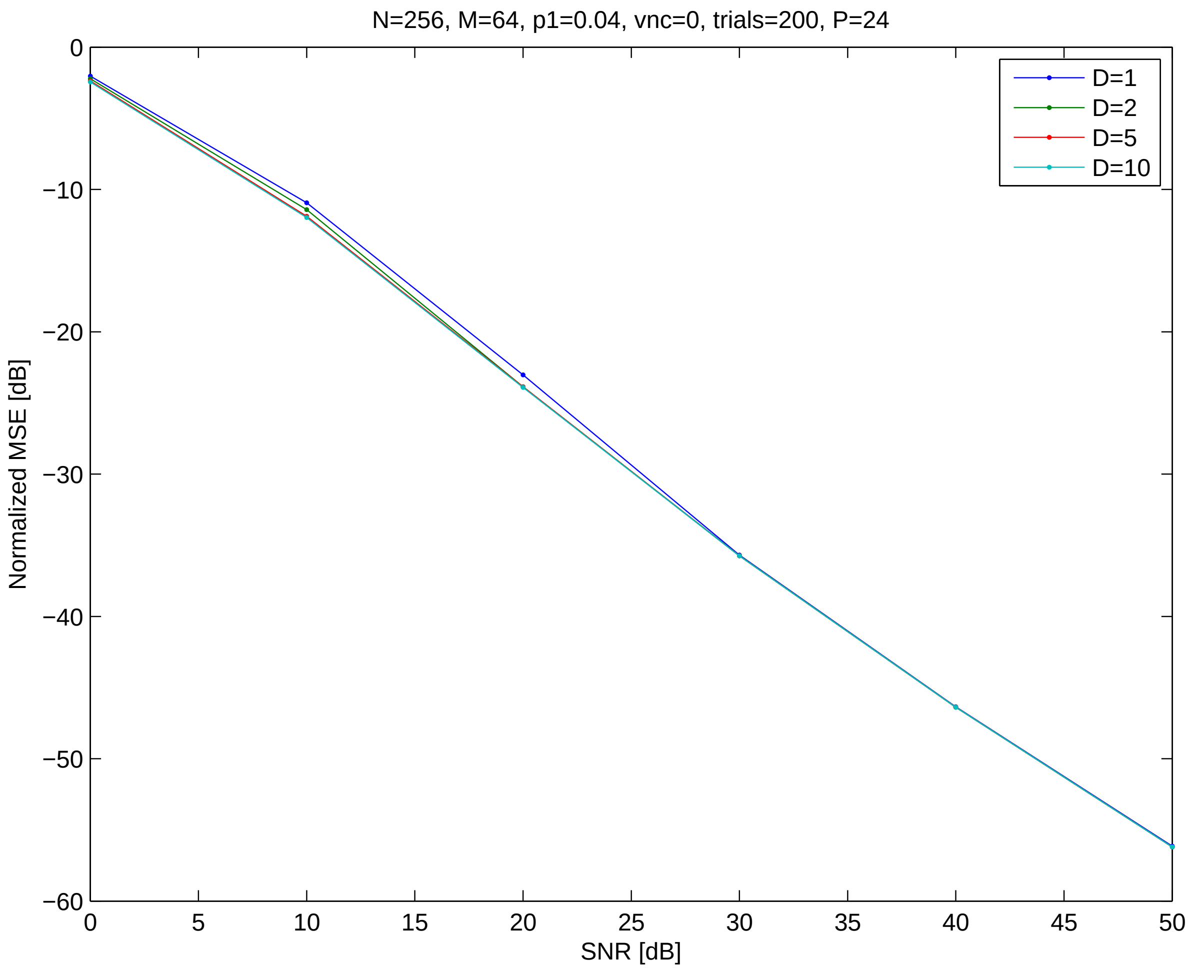
Active coefs missing from $\hat{s}_{\text{map}}$:



Again, knee in curve at $\frac{M}{K} \approx 5 \, \frac{\text{measurements}}{\text{active coef}}$.

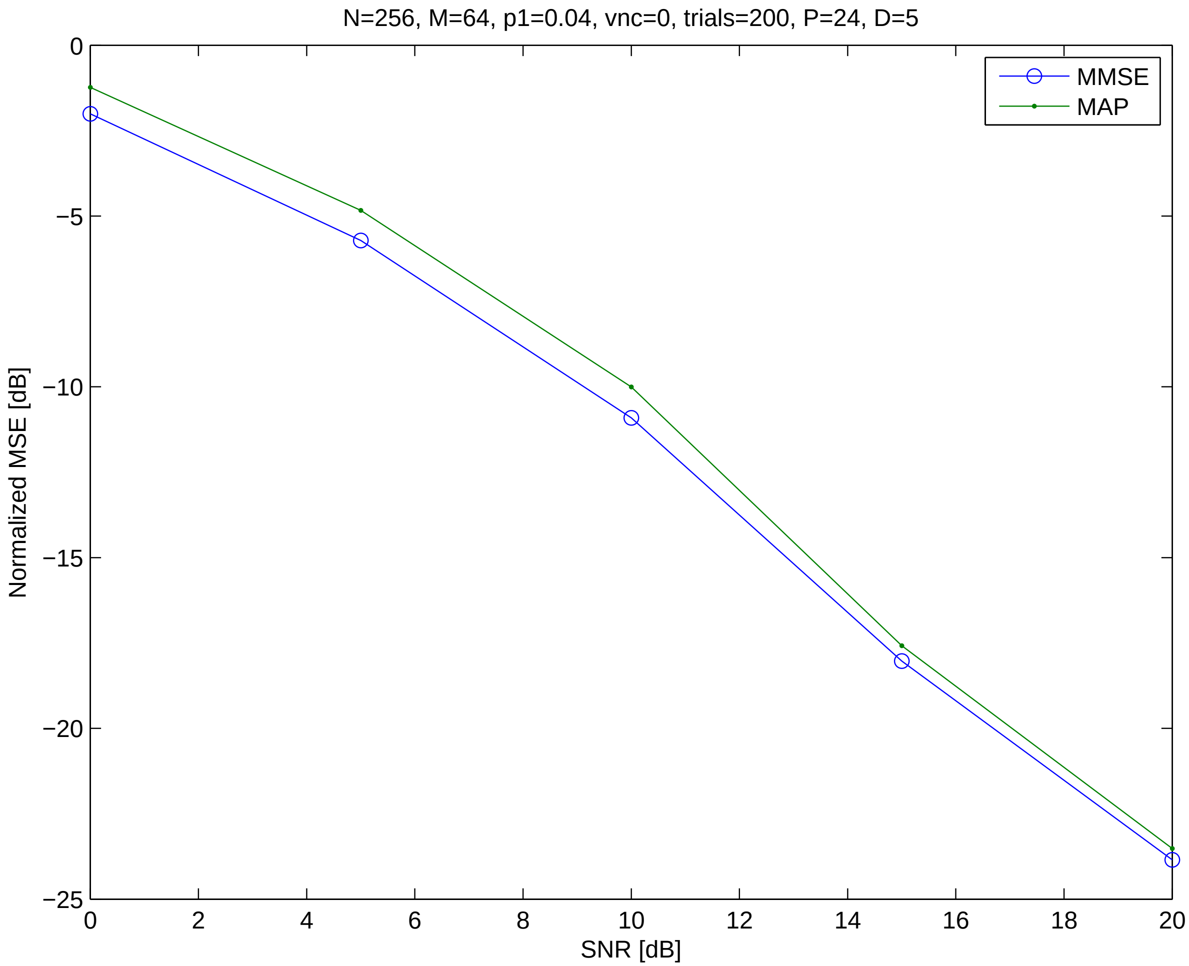
(Note: Expect improvement from generalized signal model.)

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 $\hat{\boldsymbol{x}}_{\mathsf{mmse}}$ and $\hat{\boldsymbol{x}}_{\mathsf{mmse}}\big|\hat{\boldsymbol{s}}_{\mathsf{map}}$:



Exploiting configuration *uncertainty* gives ≈ 1 dB gain in NMSE.