Fast Bayesian Matching Pursuit

Phil Schniter, Lee Potter, and Justin Ziniel

THE OHIO STATE UNIVERSITY

January 2008

Introduction:

The linear regression problem

$$\text{“estimate sparse } \boldsymbol{x} \text{ from measurements } \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\nu}\text{”}$$

is often posed as a penalized least-squares (LS) problem:

$$\hat{\boldsymbol{x}} \;=\; \operatorname*{argmin}_{\boldsymbol{x}} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{x}\|_p^p$$

$$\text{or} \quad \hat{\boldsymbol{x}} \;=\; \operatorname*{argmin}_{\boldsymbol{x}} \|\boldsymbol{x}\|_p^p \text{ s.t. } \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 < \epsilon$$

since the choice $p = 1$ provides a unique solution to the non-convex task

$$\hat{\boldsymbol{x}} = \operatorname*{argmin}_{\boldsymbol{x}} \|\boldsymbol{x}\|_0 \text{ s.t. } \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 < \epsilon.$$

The LS approach can be interpreted as seeking the Bayesian MAP estimate of $\boldsymbol{x}$ under the sparsity-inducing prior

$$p(\boldsymbol{x}) \sim \exp\big\{ -\tfrac{\lambda}{2}\|\boldsymbol{x}\|_p^p \big\}.$$

"Sparse Bayesian Learning":

The method of "sparse Bayesian learning" explicitly adopts a Bayesian framework in which $\{x_i\}$ are independent such that

$x_i \sim \mathcal{N}(0, \sigma_i^2)$

σ_i^2: unknown variance with a Gamma conjugate prior.

The EM algorithm is then used to find the MAP estimate.

However, the physical interpretation of the prior is not clear...

For example,

M. E. Tipping "Sparse Bayesian learning and relevance vector machine," *J. Machine Learning Res.*, 2001.

D. Wipf and B. Rao, "Sparse Bayesian learning for basis selection," *IEEE TSP*, 2004.

Our Problem Formulation:

Measurements:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\nu} \in \mathbb{R}^M$$

$\boldsymbol{A}$: known mixing matrix

$\boldsymbol{\nu}$: AWGN with variance σ^2

$\boldsymbol{x}$: unknown sparse signal $\in \mathbb{R}^N$, (typically $N \gg M$).

Sparse Signal:

$$x_n | s_n \sim \begin{cases} \mathcal{N}(0,1) & s_n = 1 \\ 0 & s_n = 0 \end{cases}$$

$$s_n \sim \text{Bernoulli}(p_1) \quad \text{where typically } p_1 \ll 1$$

$$\{x_n\}_{n=0}^{N-1}, \{s_n\}_{n=0}^{N-1} \; : \; \text{i.i.d.}$$

Reminiscent of the model adopted in

E. Larsson and Y. Selén, "Linear regression with a sparse parameter vector," *IEEE TSP*, Feb. 2007.

Objectives:

1. Basis Estimation:

   Note that $\boldsymbol{s} = [s_0, \ldots, s_{N-1}]^T$ specifies one of 2^N basis hypotheses.
   Want to find a (small) subset $\mathcal{S}_\star \subset \{0,1\}^N$ of basis hypotheses with *non-negligible* probability $p(\boldsymbol{s}|\boldsymbol{y})$.

2. Signal Estimation:

   The MMSE estimate,

$$\hat{\boldsymbol{x}}_{\mathsf{mmse}} = \sum_{\boldsymbol{s}\in\{0,1\}^N} p(\boldsymbol{s}|\boldsymbol{y}) \underbrace{\mathrm{E}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\}}_{\hat{\boldsymbol{x}}_{\mathsf{mmse}}|\boldsymbol{s}} \approx \sum_{\boldsymbol{s}\in\mathcal{S}_\star} p(\boldsymbol{s}|\boldsymbol{y})\,\mathrm{E}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\},$$

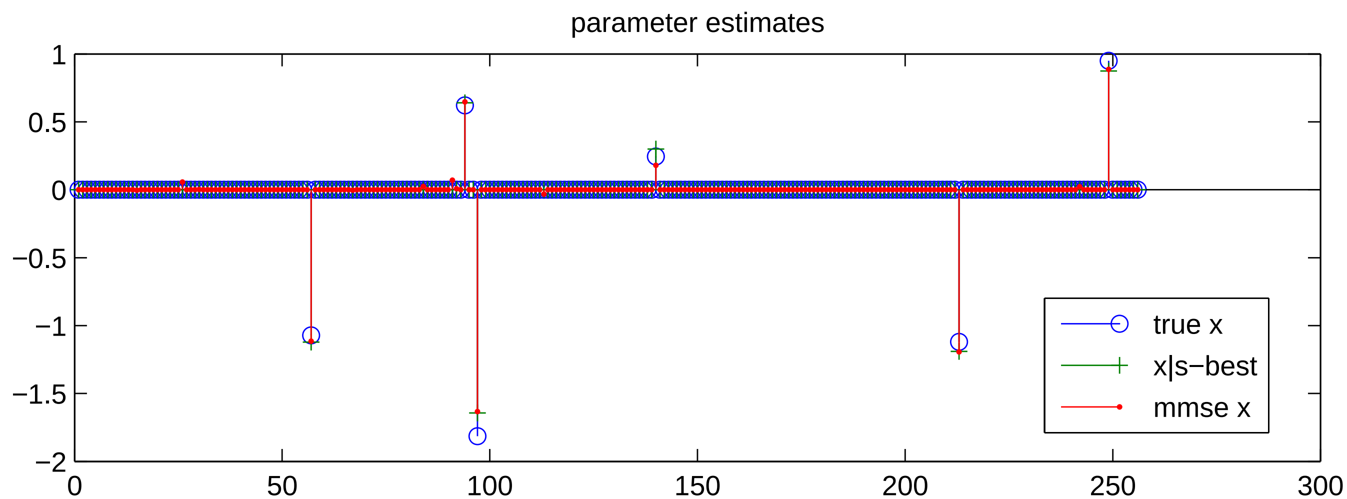
   leverages the *inherent uncertainty* in basis estimation.

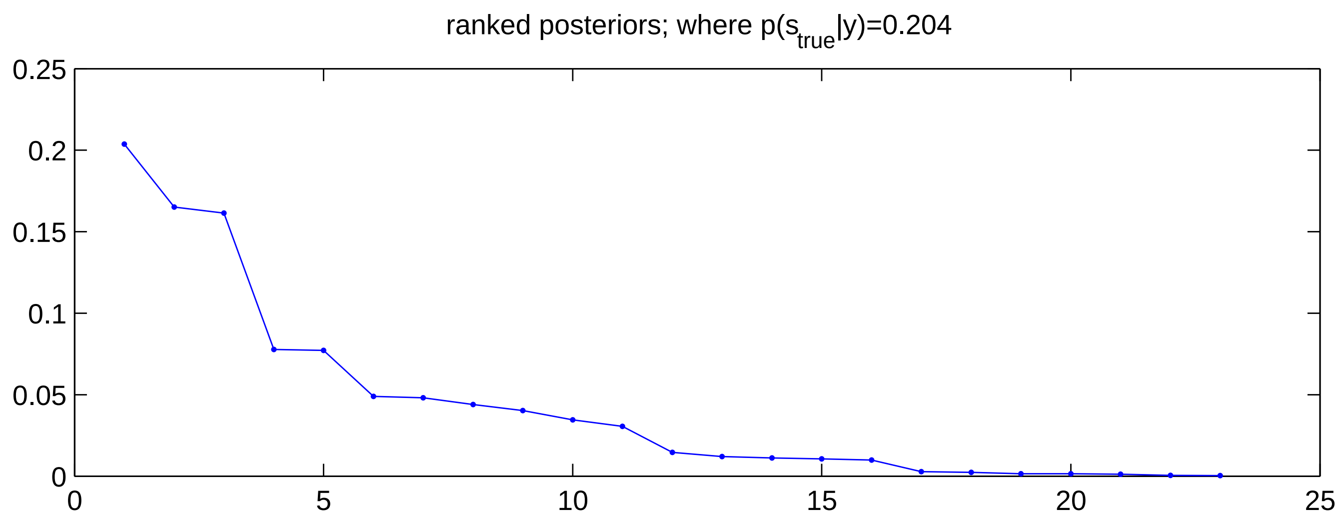
   Can contrast $\hat{\boldsymbol{x}}_{\mathsf{mmse}}$ with the MMSE estimate conditioned on the most probable basis ($\hat{\boldsymbol{s}}_{\mathsf{map}}$):

$$\hat{\boldsymbol{x}}_{\mathsf{mmse}}|\hat{\boldsymbol{s}}_{\mathsf{map}} = \mathrm{E}\{\boldsymbol{x}|\boldsymbol{y},\hat{\boldsymbol{s}}_{\mathsf{map}}\}.$$

Example – Leveraging Basis Uncertainty:

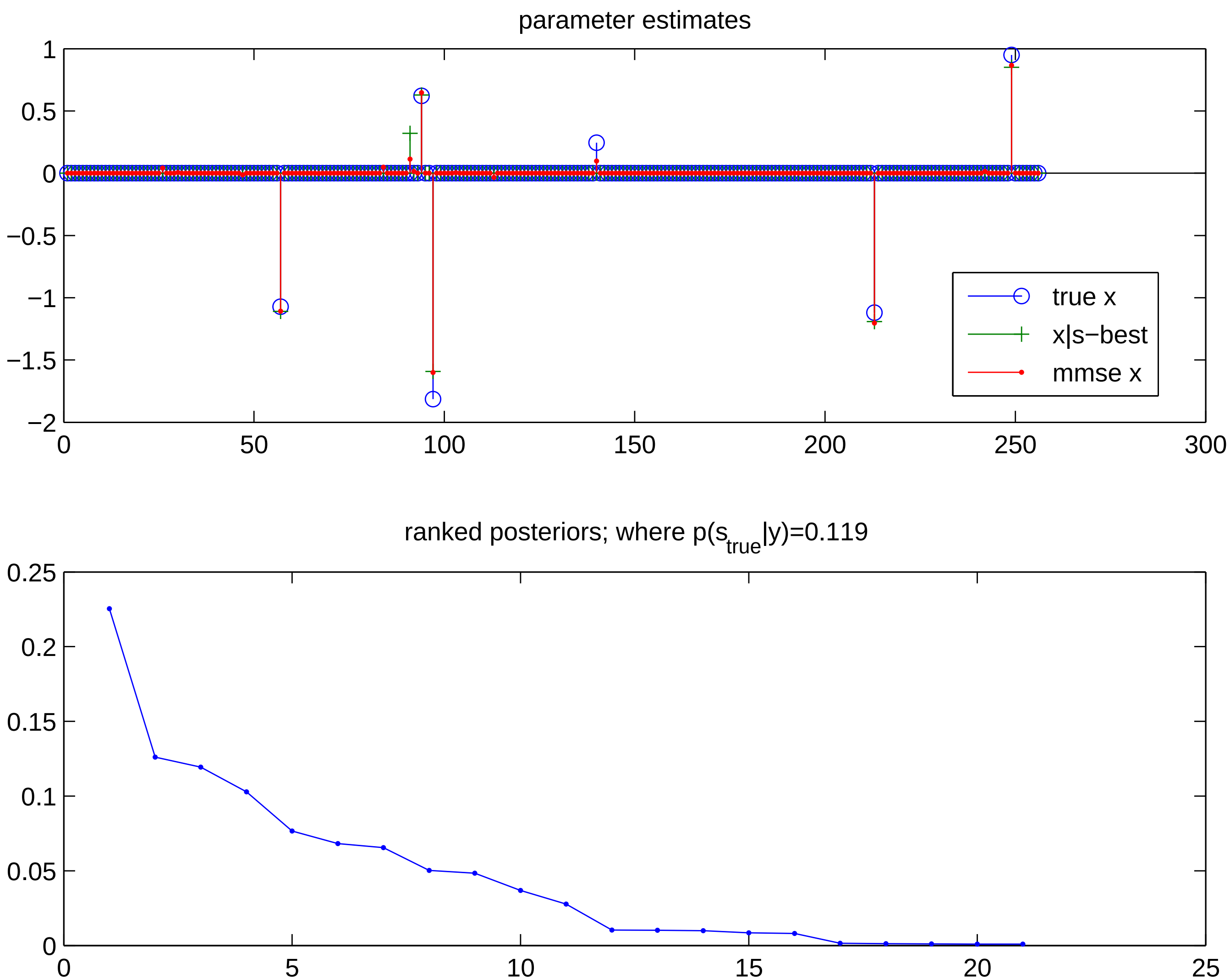
At SNR=14dB, $\hat{\boldsymbol{x}}_{\text{mmse}}|\hat{\boldsymbol{s}}_{\text{map}} \approx \hat{\boldsymbol{x}}_{\text{mmse}} \approx \boldsymbol{x}$:





Example – Leveraging Basis Uncertainty (cont.):

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



A Basis Selection Metric:

From Bayes rule, we know

$$p(\boldsymbol{s}|\boldsymbol{y}) \;=\; \frac{p(\boldsymbol{y}|\boldsymbol{s})p(\boldsymbol{s})}{p(\boldsymbol{y})}.$$

The "basis selection metric"

$$\mu(\boldsymbol{s}) \;:=\; \log p(\boldsymbol{y}|\boldsymbol{s})p(\boldsymbol{s}),$$

can be expanded as

$$\begin{aligned}\mu(\boldsymbol{s}) \;=\; & -\tfrac{M}{2}\ln 2\pi - \tfrac{1}{2}\ln\det(\boldsymbol{\Sigma}(\boldsymbol{s})) - \tfrac{1}{2}\boldsymbol{y}^T\boldsymbol{\Sigma}(\boldsymbol{s})^{-1}\boldsymbol{y} \\ & + \|\boldsymbol{s}\|_0 \ln\tfrac{p_1}{1-p_1} + N\ln(1-p_1),\end{aligned}$$

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

So how can we find the basis hypotheses $\{\boldsymbol{s}\}$ with large $\mu(\boldsymbol{s})$?

Simple Bayesian Matching Pursuit (BMP-1):

Consider matching pursuit (MP), but using metric $\mu(\boldsymbol{s})$ rather than magnitude of projection-onto-residual:

$$
\begin{aligned}
&\boldsymbol{s}_\star^{(0)} = \mathbf{0};\\
&\text{for } i = 1:P,\\
&\quad \mathcal{S}^{(i)} = \big\{\text{extensions of } \boldsymbol{s}_\star^{(i-1)} \text{ to one more active element}\big\};\\
&\quad \boldsymbol{s}_\star^{(i)} = \operatorname{argmax}_{\boldsymbol{s}\in\mathcal{S}^{(i)}} \mu(\boldsymbol{s});\\
&\text{end};\\
&\hat{\mathcal{S}}_\star = \big\{\boldsymbol{s}_\star^{(i)}\big\}_{i=0}^{P};\\
&\hat{\boldsymbol{x}} = \textstyle\sum_{\boldsymbol{s}\in\hat{\mathcal{S}}_\star} \hat{p}(\boldsymbol{s}|\boldsymbol{y})\,\mathrm{E}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\} \quad \text{for} \quad \hat{p}(\boldsymbol{s}|\boldsymbol{y}) = \dfrac{\exp\{\mu(\boldsymbol{s})\}}{\sum_{\boldsymbol{s}'\in\hat{\mathcal{S}}_\star} \exp\{\mu(\boldsymbol{s}')\}}
\end{aligned}
$$

Can choose P via Gaussian approx of $\|\boldsymbol{s}\|_0 \sim \text{Binomial}(N, p_1)$:

$$P \;=\; \Big\lceil \operatorname{erfc}^{-1}(2\mathcal{P}_0)\sqrt{2Np_1(1-p_1)} + Np_1 \Big\rceil \text{ for } \mathcal{P}_0 := \Pr\{\|\boldsymbol{s}\|_0 > P\}$$

Reminiscent of the technique proposed in

E. Larsson and Y. Selén, “Linear regression with a sparse parameter vector,” *IEEE TSP*, Feb. 2007.

Bayesian Matching Pursuit (BMP-D):

Extension of simple BMP to D simultaneous hypotheses:

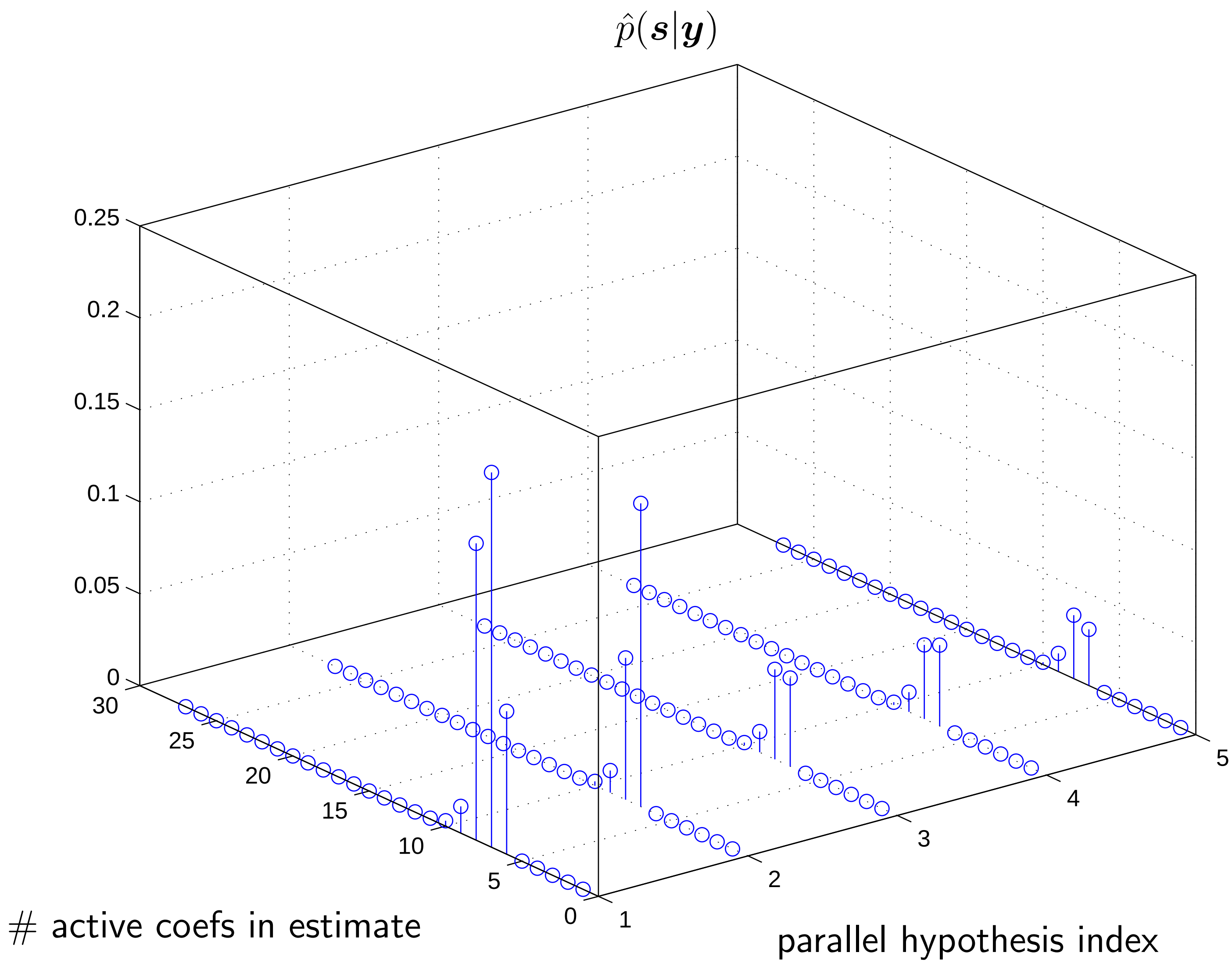
$$
\begin{aligned}
&\mathcal{S}_\star^{(0)} = \{\mathbf{0}\};\\
&\text{for } i = 1 : P,\\
&\quad \mathcal{S}^{(i)} = \big\{\text{unique 1-tap extensions of } \mathcal{S}_\star^{(i-1)}\big\};\\
&\quad \mathcal{S}_\star^{(i)} = \big\{\text{the } D \text{ elements of } \mathcal{S}^{(i)} \text{ with largest } \mu(\boldsymbol{s})\big\};\\
&\text{end;}\\
&\hat{\mathcal{S}}_\star = \textstyle\bigcup_{i=0}^{P} \mathcal{S}_\star^{(i)};\\
&\hat{\boldsymbol{x}} = \sum_{\boldsymbol{s}\in\hat{\mathcal{S}}_\star} \hat{p}(\boldsymbol{s}|\boldsymbol{y})\,\mathrm{E}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\} \quad \text{for} \quad \hat{p}(\boldsymbol{s}|\boldsymbol{y}) = \frac{\exp\{\mu(\boldsymbol{s})\}}{\sum_{\boldsymbol{s}'\in\hat{\mathcal{S}}_\star}\exp\{\mu(\boldsymbol{s}')\}}
\end{aligned}
$$

D effects a tradeoff between search accuracy and complexity.

A similar extension of basic MP was proposed in

S. F. Cotter and B. D. Rao, "Application of tree-based searches to matching pursuit," *Proc. IEEE ICASSP*, 2001.

Example – Basis-hypothesis posteriors found by BMP-D:



Fast Bayesian Matching Pursuit (FBMP):

- The BMP-like approach proposed by Larsson and Selén consumes $\mathcal{O}(N^3M^2)$ multiplications.

  $\rightsquigarrow$Too expensive!!

- Using a recursive metric update, we propose a *fast* BMP-1 that consumes only $\mathcal{O}(NMP)$ multiplications.

  $\rightsquigarrow$ Savings of $\mathcal{O}(N^2\frac{P}{M})$; many orders of magnitude!

- Can straightforwardly extend to a *fast* BMP-D with complexity $\mathcal{O}(NMPD)$.

Fast Recursive Evaluation of Metric $\mu(\boldsymbol{s})$:

FBMP's fast update is based on two key properties:

1. $\Delta_n(\boldsymbol{s})$, the *change* in metric $\mu(\boldsymbol{s})$ that results from activating the n^{th} tap in $\boldsymbol{s}$, can be expressed as

$$\begin{aligned} \Delta_n(\boldsymbol{s}) &= \tfrac{1}{2}\log\beta_n^{(i)} + \tfrac{1}{2}\beta_n^{(i)}\big(\boldsymbol{y}^T\boldsymbol{b}_n^{(i)}\big)^2 + \log\tfrac{p_1}{1-p_1} \\ \beta_n^{(i)} &= (1+\boldsymbol{a}_n^T\boldsymbol{b}_n^{(i)})^{-1} \\ \boldsymbol{b}_n^{(i)} &= \boldsymbol{\Sigma}(\boldsymbol{s})^{-1}\boldsymbol{a}_n \qquad \longleftarrow \text{ but still } \mathcal{O}(M^2)? \end{aligned}$$

2. The structure of $\boldsymbol{\Sigma}(\boldsymbol{s})^{-1}$ permits $\mathcal{O}(M)$ calculation of $\boldsymbol{b}_n^{(i)}$:

$$\boldsymbol{b}_n^{(i)} = \boldsymbol{b}_n^{(i-1)} - \boldsymbol{b}_{n_\star}^{(i-1)}\beta_{n_\star}^{(i-1)}\boldsymbol{b}_{n_\star}^{(i-1)T}\boldsymbol{a}_n,$$

   for $n_\star$ the index of the $\mu(\boldsymbol{s})$-maximizing element of $\mathcal{S}^{(i-1)}$. Here, $i := \|\boldsymbol{s}\|_0$, and so $^{(i-1)}$ refers to the parent node.

FBMP-1:

$\mu^{(0)} = N\log(1-p_1) - \frac{M}{2}\log 2\pi - \frac{M}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\|\boldsymbol{y}\|_2^2$;
$\mathcal{N}^{(0)} = \{\}$;
$\boldsymbol{b}_n^{(1)} = \frac{1}{\sigma^2}\boldsymbol{a}_n$ for all $n \notin \mathcal{N}^{(0)}$;
for $i = 1 : P$,
  $\beta_n = \big(1 + \boldsymbol{a}_n^T \boldsymbol{b}_n^{(i)}\big)^{-1}$ for all $n \notin \mathcal{N}^{(i-1)}$;
  $\mu_n = \mu^{(i-1)} + \frac{1}{2}\log\beta_n + \frac{1}{2}\beta_n\big(\boldsymbol{y}^T\boldsymbol{b}_n^{(i)}\big)^2 + \log\frac{p_1}{1-p_1}$ for all $n \notin \mathcal{N}^{(i-1)}$;
  $n_\star^{(i)} = \operatorname{argmax}_n \mu_n$;
  $\mu^{(i)} = \mu_{n_\star^{(i)}}$;
  $\mathcal{N}^{(i)} = \mathcal{N}^{(i-1)} \cup \{n_\star^{(i)}\}$;
  $\boldsymbol{b}_n^{(i+1)} = \boldsymbol{b}_n^{(i)} - \boldsymbol{b}^{(i)}_{n_\star^{(i)}} \beta_{n_\star^{(i)}} \boldsymbol{b}^{(i)T}_{n_\star^{(i)}} \boldsymbol{a}_n$ for all $n \notin \mathcal{N}^{(i-1)}$;
end;
$\hat{\boldsymbol{x}}_i = \sum_{j=1}^{i} \boldsymbol{\delta}_{n_\star^{(j)}} \boldsymbol{b}^{(i+1)T}_{n_\star^{(j)}} \boldsymbol{y}$ for all $i \in \{1,\ldots,P\}$;
$\hat{p}_i = \dfrac{\exp\{\mu^{(i)}\}}{\sum_{j=0}^{P}\exp\{\mu^{(j)}\}}$ for all $i \in \{1,\ldots,P\}$;
$\hat{\boldsymbol{x}} = \sum_{i=1}^{P}\hat{p}_i\hat{\boldsymbol{x}}_i$;

Numerical Experiments – Comparison to Other Algs:

Nominal Params:

$$\begin{aligned} N &= 512 \\ p_1 &= 0.04 \quad \text{...so } p_1 N = 20 \text{ active coefs on average} \\ M &= 120 \\ \text{SNR} &= 19\,\text{dB} \end{aligned}$$

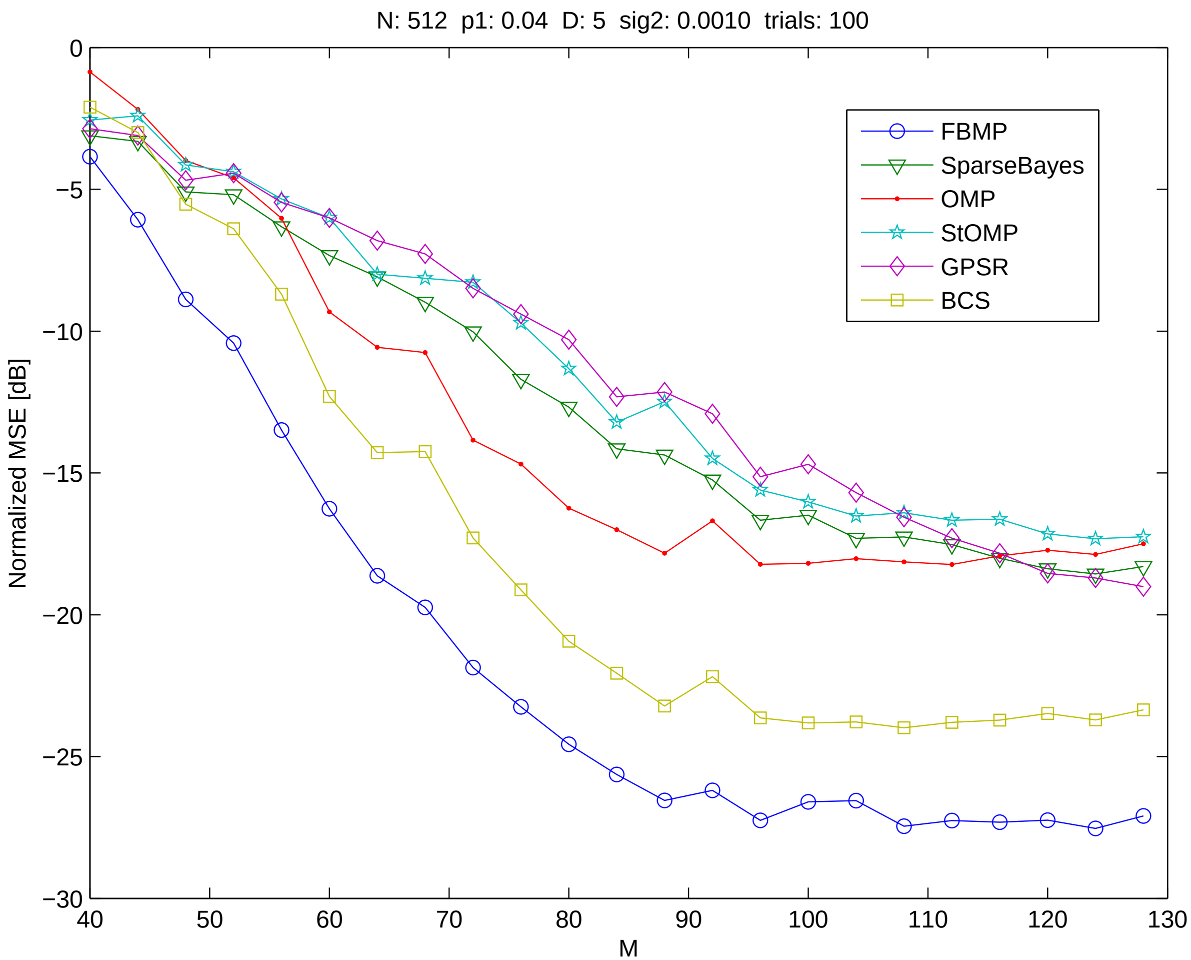
Note: Considering $M \gtrsim p_1 N \log(\frac{N}{p_1 N}) \Leftrightarrow \frac{\#\text{ active coefs}}{\text{measurement}} = \frac{p_1 N}{M} \lesssim \frac{-1}{\log(p_1)}$, our nominal parameters yield $\frac{p_1 N}{M} = 0.16$ and $\frac{-1}{\log(p_1)} = 0.31$.

Algorithms:

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

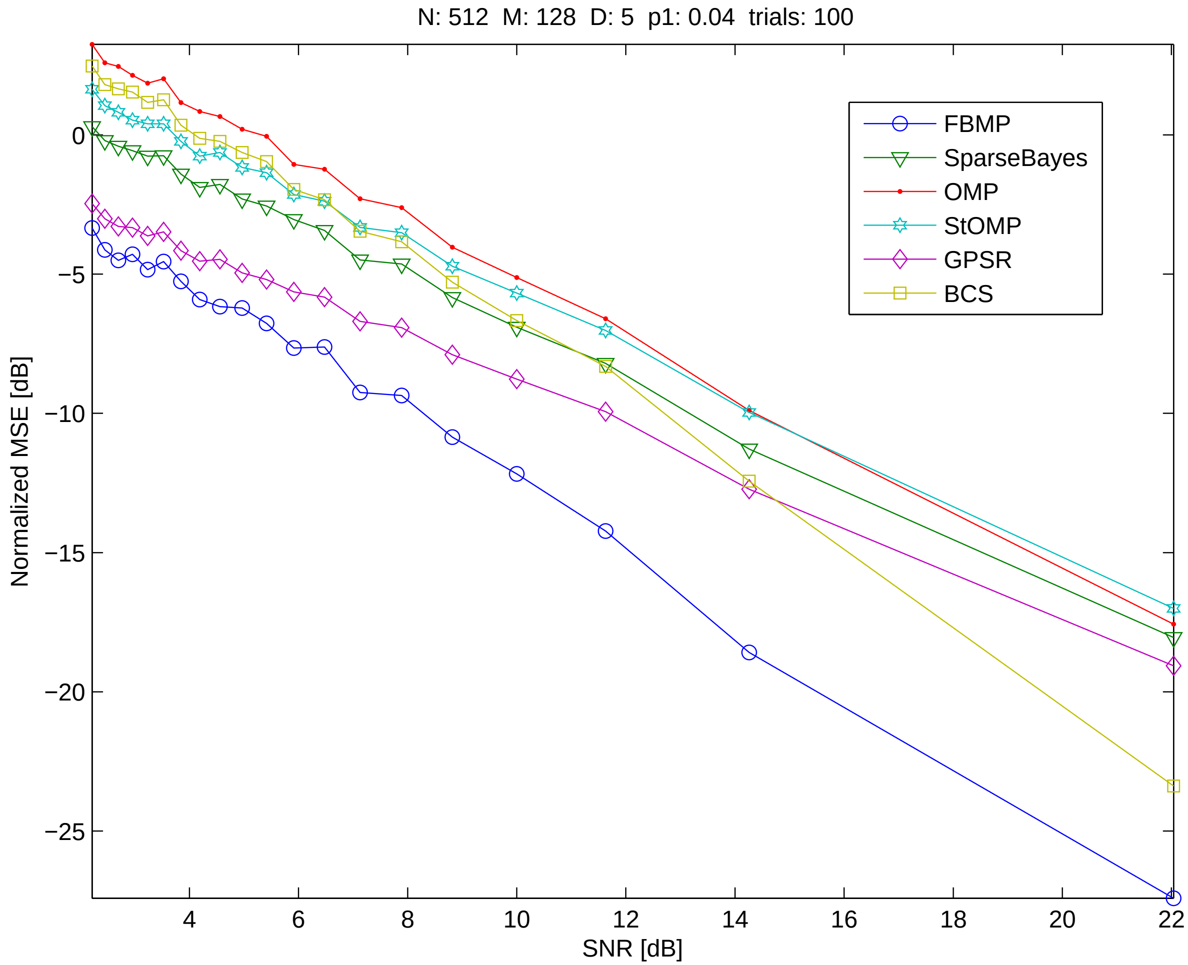
Performance: $\text{NMSE} := \text{Avg}\left\{\frac{\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2}\right\}$ over 100 random trials.

NMSE versus observation length M:



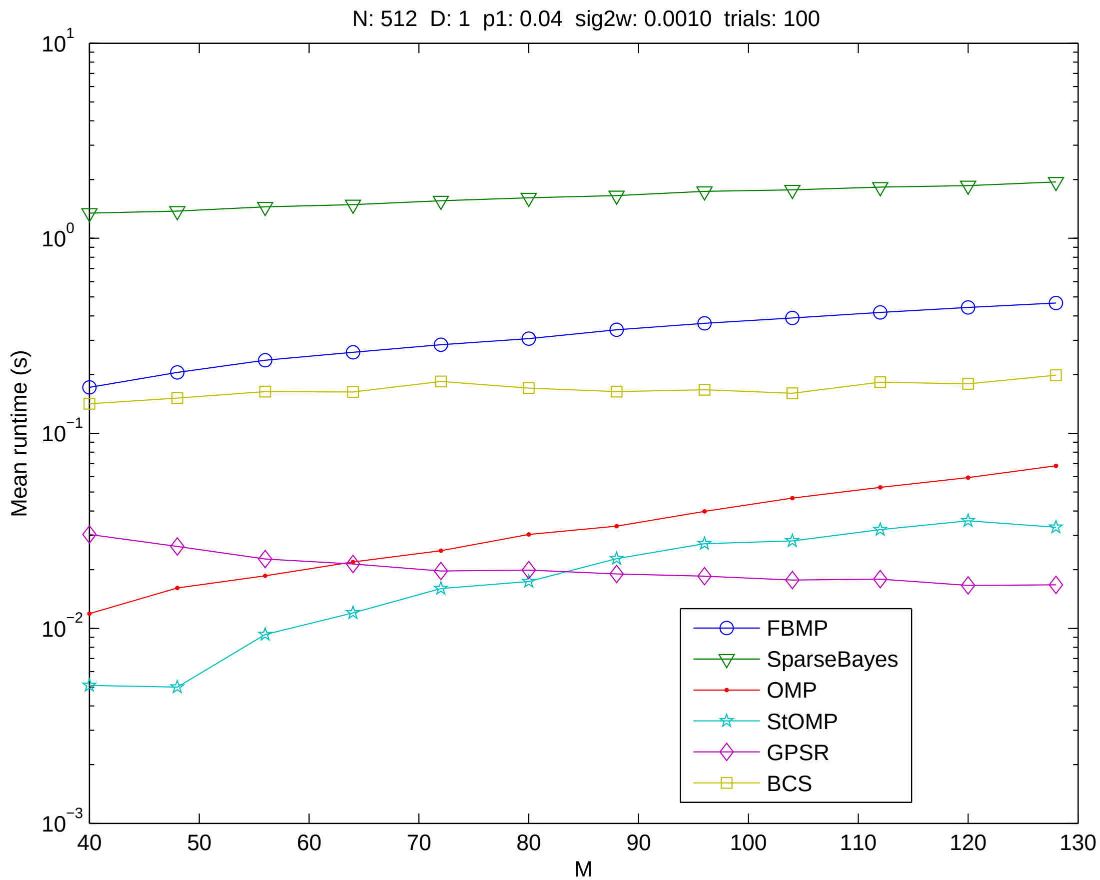
For $\frac{p_1 N}{M} < 0.2$, FBMP outperformed BCS by 3 dB and others by ≥ 10 dB. As $\frac{p_1 N}{M} \rightarrow 0.5$, NMSEs converge.

NMSE versus SNR:



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

Runtime versus observation length M:



(Not-yet-optimized) FBMP is an order of magnitude faster than SparseBayes, about the same speed as BCS, and an order of magnitude slower than OMP, StOMP, and GPSR.

Numerical Experiments – FBMP Behavior:

Nominal Signal Parameters:

$$\begin{aligned} N &= 256 \\ p_1 &= 0.04 && \ldots \text{thus } \ p_1 N = 10 \text{ active coefs on average} \\ M &= 64 \\ \text{SNR} &= 15\,\text{dB} && \ldots \text{where SNR} := \tfrac{p_1 N}{\sigma^2 M} \\ \boldsymbol{A} &: \text{i.i.d. } \mathcal{N}(0,1), && \text{then columns normalized.} \end{aligned}$$

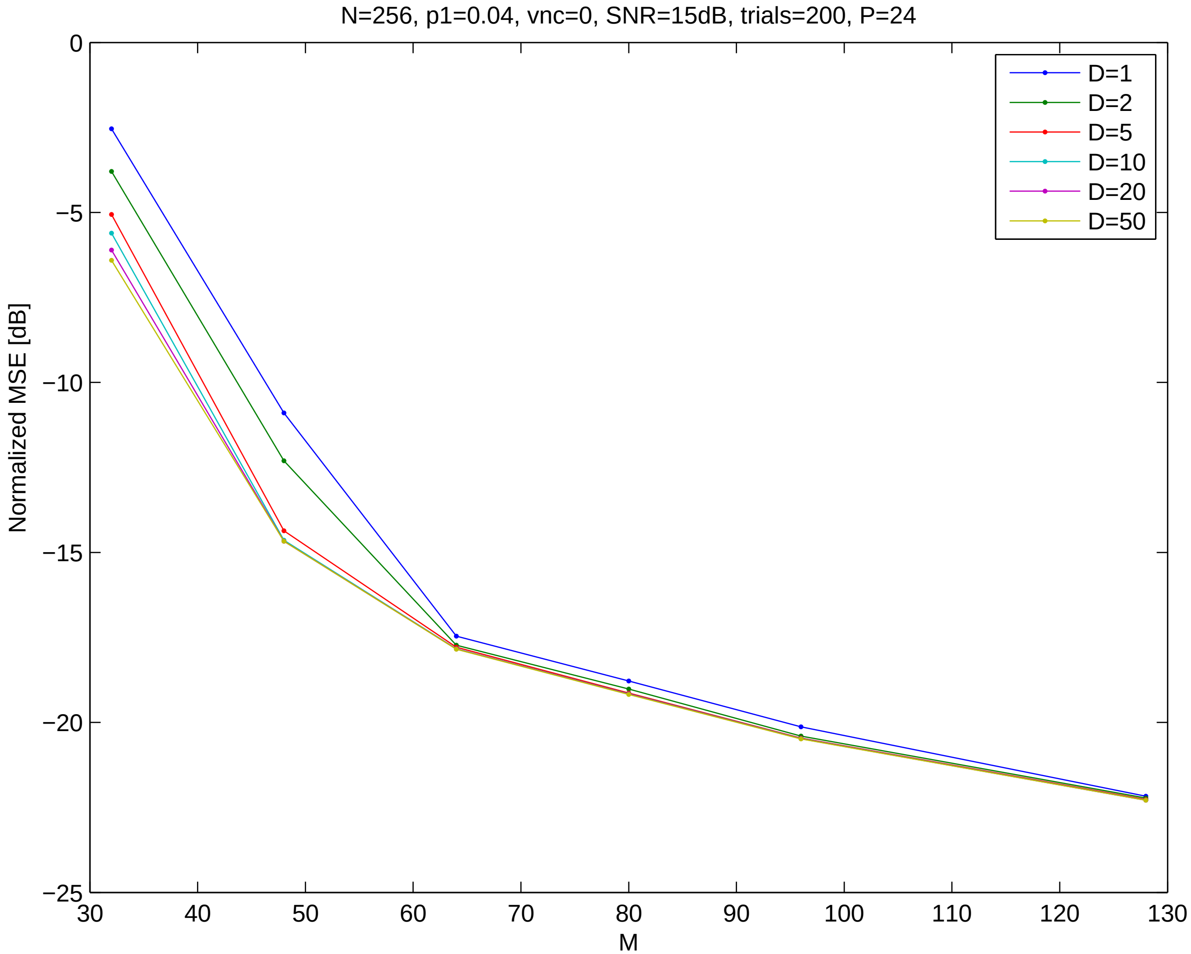
Note:

Considering $M \gtrsim p_1 N \log(\frac{N}{p_1 N}) \;\Leftrightarrow\; \frac{\#\text{ active coefs}}{\text{measurement}} = \frac{p_1 N}{M} \lesssim \frac{-1}{\log(p_1)}$, our nominal parameters yield $\frac{p_1 N}{M} = 0.16$ and $\frac{-1}{\log(p_1)} = 0.31$.

Performance:

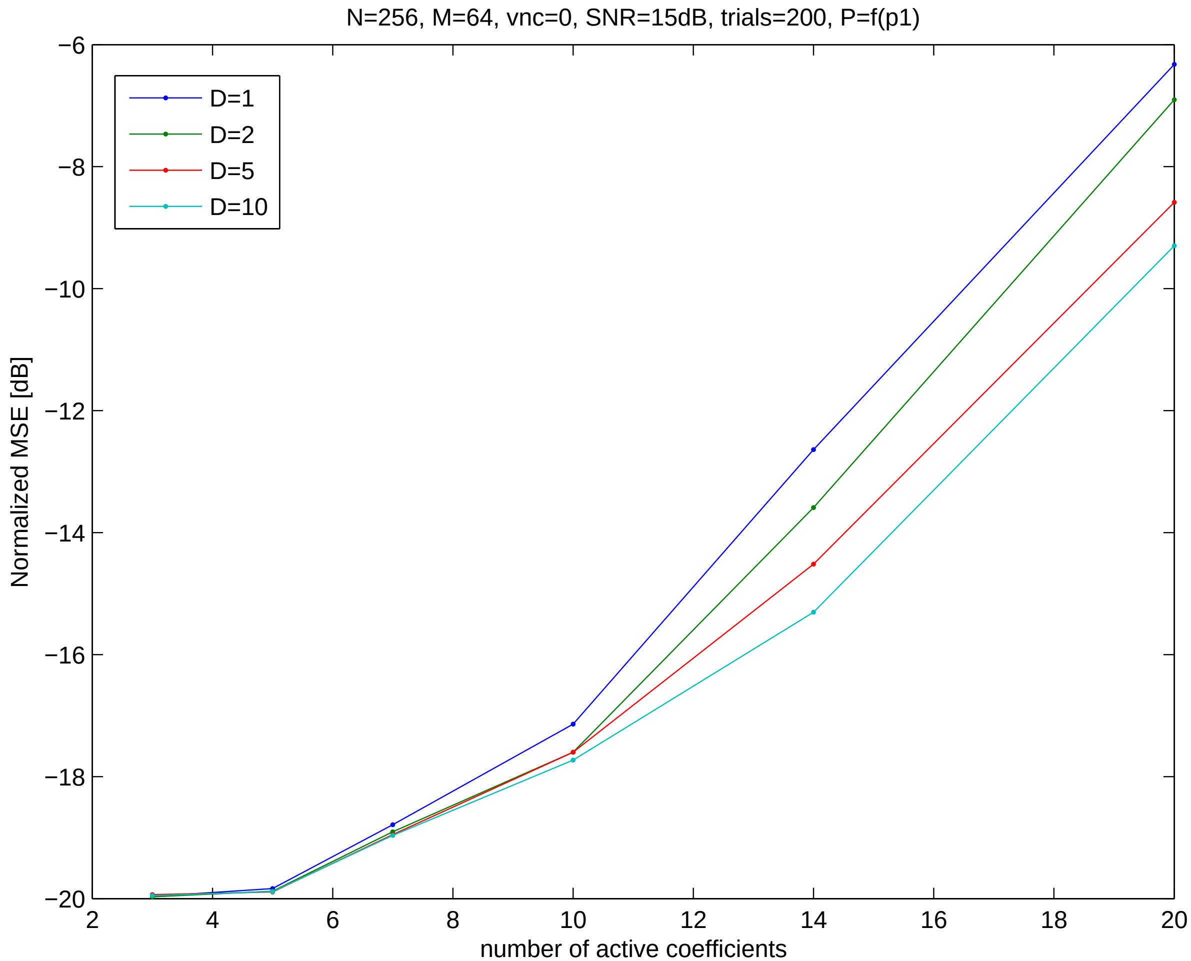
$$\text{NMSE} \;:=\; \text{Avg}\left\{ \frac{\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2} \right\} \text{ over 200 random trials.}$$

NMSE versus observation length M:



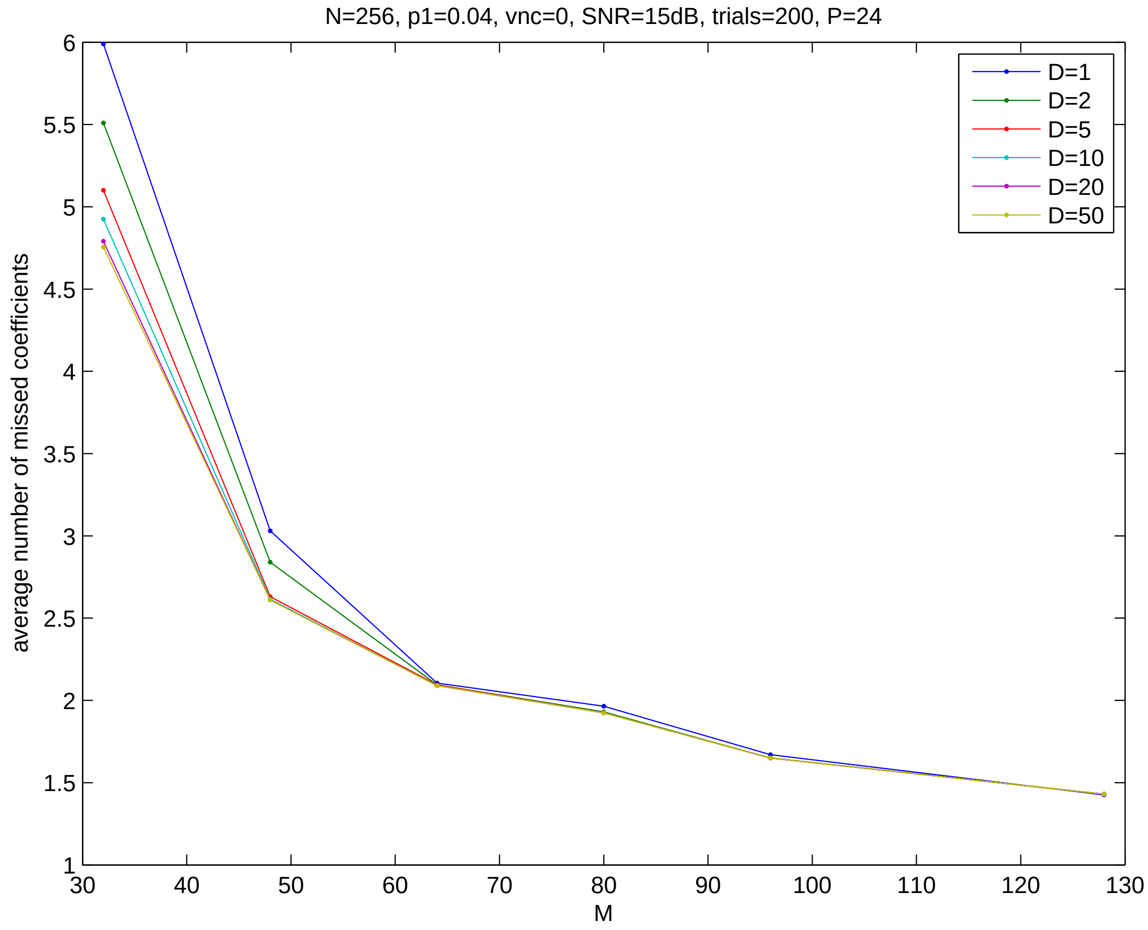
When $D = 1$, knee in curve at $\frac{p_1 N}{M} = \frac{10}{64} = 0.16\ \frac{\text{\# active coefs}}{\text{measurement}}$.
For larger D, knee moves to $0.2\ \frac{\text{\# active coefs}}{\text{measurement}}$ and NMSE improves by $3\,$dB.

NMSE versus expected # active coefs $p_1 N$:



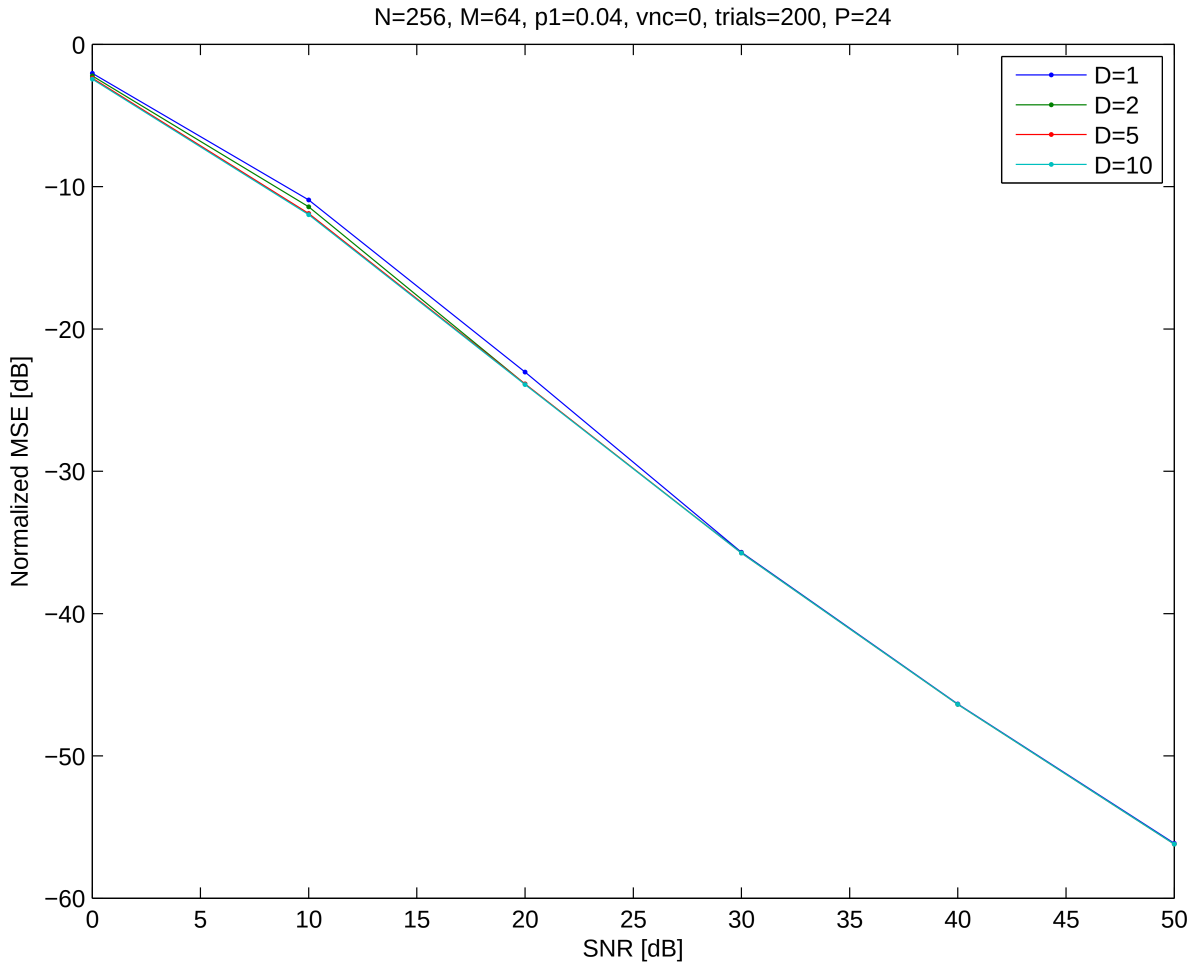
When $D = 1$, knee in curve at $\frac{p_1 N}{M} = \frac{10}{64} = 0.16 \frac{\text{\# active coefs}}{\text{measurement}}$.
For $D = 10$, knee at $\frac{14}{64} = 0.2 \frac{\text{\# active coefs}}{\text{measurement}}$ and NMSE $3\,$dB improved.

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



Again, knee in curve at $\frac{p_1 N}{M} \approx 0.2 \, \frac{\#\text{ active coefs}}{\text{measurement}}$.
(Note: we expect some misses since signal model is zero-mean.)

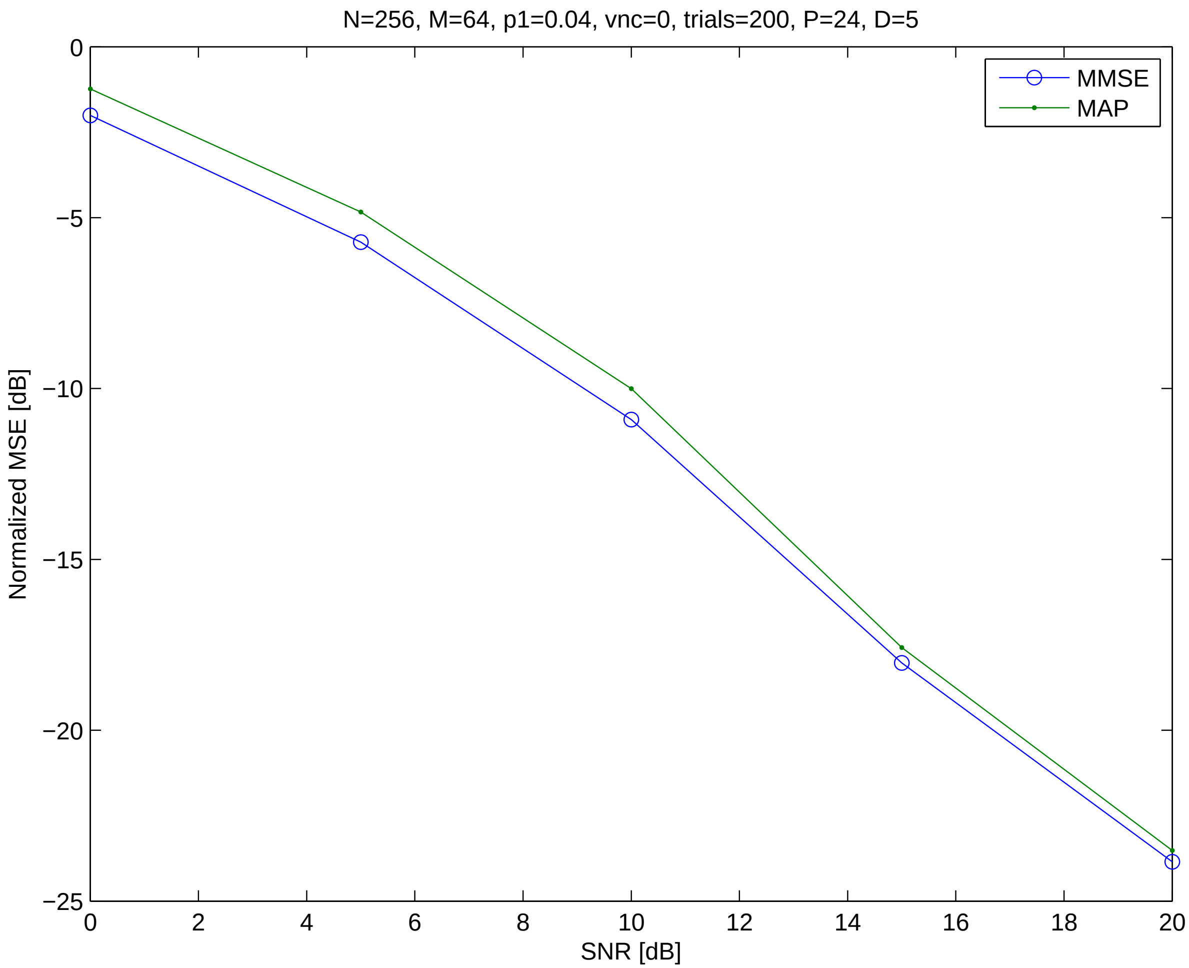
NMSE versus SNR:



Note linear dependence between NMSE [dB] & SNR [dB].

(No benefit from D-increase expected since $\frac{p_1 N}{M} = 0.16$.)

NMSE for $\hat{\boldsymbol{x}}_{\text{mmse}}$ and $\hat{\boldsymbol{x}}_{\text{mmse}}\big|\hat{\boldsymbol{s}}_{\text{map}}$:



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

Conclusion:

- Building on the Bayesian sparse-coefficient estimation technique of Larsson and Selén, we proposed
  - a forward search, with fast recursive update, that reduces complexity from $\mathcal{O}(N^3M^2)$ to $\mathcal{O}(NMP)$, and
  - an extension of the search to $D > 1$ simultaneous hypotheses, providing up to 3 dB of NMSE gain when $\frac{\text{\# active coefs}}{\text{measurement}}$ is high.
- Comparisons against SparseBayes, BCS, OMP, StOMP, GPSR showed NMSE improvements of several dB over a wide range of parameters.
- In the near future, we plan to optimize FBMP and extend our approach to signals with complex-valued coefs modeled with *non-zero means*.