A Fast Posterior Update for Sparse Underdetermined Linear Models

Lee C. Potter, Philip Schniter, and Justin Ziniel
Department of Electrical and Computer Engineering
The Ohio State University, Columbus, OH, 43210-1272 USA

***Abstract*— A Bayesian approach is adopted for linear regression, and a fast algorithm is given for updating posterior probabilities. Emphasis is given to the underdetermined and sparse case, i.e., fewer observations than regression coefficients and the belief that only a few regression coefficients are non-zero. The fast update allows for a low-complexity method of reporting a set of models with high posterior probability and their exact posterior odds. As a byproduct, this Bayesian model averaged approach yields the minimum mean squared error estimate of unknown coefficients. Algorithm complexity is linear in the number of unknown coefficients, the number of observations and the number of nonzero coefficients. For the case in which hyperparameters are unknown, a maximum likelihood estimate is found by a generalized expectation maximization algorithm.**

I. Introduction

Linear regression is a classical statistical problem. We consider sparse linear regression with fewer observations than regression coefficients. Our treatment, while general, is motivated by channel estimation tasks arising in communications, radar, and medical imaging. We seek five properties in an inference procedure. First, we seek to exploit domain knowledge, if available, through physically interpretable priors. Where hyperparameters are unknown, we seek maximum likelihood estimates to learn priors from data. Second, we wish to minimize mean squared error in reconstruction of regression coefficients. Third, we wish to have low complexity computation – on the order of orthogonal matching pursuit [1]. Fourth, we wish to report ambiguity in both variable selection and regression coefficients; ambiguity may arise due to correlation among regressor vectors or due to measurement noise. Finally, we wish to work with complex-valued data that arises from bandpass signals.

To satisfy these desiderata, we adopt a Bayesian approach and compute posterior probabilities for all plausible selections of variables. We present a tree search method, called Fast Bayesian Matching Pursuit (FBMP) [2]. The key to the success of the procedure is a fast update of the posterior probability of the data under a small change to the hypothesized selection of variables. The FBMP algorithm presents a low-complexity alternative to stochastic integration such as Markov Chain Monte Carlo (MCMC).

Sparse linear regression has received extensive attention which has been heightened in recent years by works providing provable performance guarantees for both greedy and convex programming algorithms [3]–[5].

II. Signal Model

We consider problems where unknown coefficients $\boldsymbol{x} \in \mathbb{C}^N$ are observed through the noisy linear mixture $\boldsymbol{y} \in \mathbb{C}^M$

$$\boldsymbol{y} \;=\; \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}, \tag{1}$$

for known $\boldsymbol{A} \in \mathbb{C}^{M\times N}$ and for noise $\boldsymbol{w}$ that is white circular Gaussian with variance σ^2, i.e., $\boldsymbol{w} \sim \mathcal{CN}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_M)$ The columns of $\boldsymbol{A}$ are taken to be unit-norm. Our focus is on the underdetermined case (i.e., $N \gg M$) with a suitably sparse parameter vector $\boldsymbol{x}$ (i.e., $\|\boldsymbol{x}\|_0 \ll N$).

To model sparsity, we assume that $\{x_n\}_{n=0}^{N-1}$, the components of $\boldsymbol{x}$, are i.i.d. random variables drawn from a Q-ary Gaussian mixture. For each x_n, a mixture parameter $s_n \in \{0, \ldots, Q-1\}$ is used to index the component distribution. In particular, when $s_n = q$, then the coefficient x_n is modeled as a circular Gaussian with mean μ_q and variance σ_q^2. The mixture parameters $\{s_n\}_{i=0}^{N-1}$ are treated as i.i.d. random variables such that $\Pr\{s_n = q\} = \lambda_q$. We choose a point mass $(\mu_0, \sigma_0^2) = (0, 0)$, so that the case $s_n = 0$ implies $x_n = 0$.

Using $\boldsymbol{x} = [x_0, \ldots, x_{N-1}]^T$ and $\boldsymbol{s} = [s_0, \ldots, s_{N-1}]^T$, the prior model can be written as

$$\boldsymbol{x}|\boldsymbol{s} \;\sim\; \mathcal{CN}(\boldsymbol{\mu}(\boldsymbol{s}), \boldsymbol{R}(\boldsymbol{s})), \tag{2}$$

where $[\boldsymbol{\mu}(\boldsymbol{s})]_n = \mu_{s_n}$ and where $\boldsymbol{R}(\boldsymbol{s})$ is diagonal with $[\boldsymbol{R}(\boldsymbol{s})]_{n,n} = \sigma_{s_n}^2$. The model (1) then implies that the unknown coefficients and the measurements are jointly Gaussian when conditioned on the mixture sequence, $\boldsymbol{s}$. In particular,

$$\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{x}\end{bmatrix}\Bigg|\, \boldsymbol{s} \;\sim\; \mathcal{CN}\left(\begin{bmatrix}\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\\ \boldsymbol{\mu}(\boldsymbol{s})\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Phi}(\boldsymbol{s}) & \boldsymbol{A}\boldsymbol{R}(\boldsymbol{s})\\ \boldsymbol{R}(\boldsymbol{s})\boldsymbol{A}^H & \boldsymbol{R}(\boldsymbol{s})\end{bmatrix}\right), \tag{3}$$

where

$$\boldsymbol{\Phi}(\boldsymbol{s}) \;:=\; \boldsymbol{A}\boldsymbol{R}(\boldsymbol{s})\boldsymbol{A}^H + \sigma^2 \boldsymbol{I}_M. \tag{4}$$

III. Fast Bayesian Matching Pursuit

A. Minimum mean squared error

The minimum mean squared error (MMSE) estimate of $\boldsymbol{x}$ from $\boldsymbol{y}$ is

$$\hat{\boldsymbol{x}}_{\mathsf{mmse}} \;:=\; \mathrm{E}\{\boldsymbol{x}|\boldsymbol{y}\} \;=\; \sum_{\boldsymbol{s}\in\mathcal{S}} p(\boldsymbol{s}|\boldsymbol{y})\,\mathrm{E}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\} \tag{5}$$

This investigation was supported by National Science Foundation Grant 0237037, Office of Naval Research grant N00014-07-1-0209, and AFOSR award FA9550-06-1-0324. A preliminary version of Fast Bayesian Matching Pursuit was presented at the Workshop on Information Theory and Applications (ITA), La Jolla, CA, January 2008.

where from (3) it is straightforward to obtain

$$\mathrm{E}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\} \;=\; \boldsymbol{\mu}(\boldsymbol{s}) + \boldsymbol{R}(\boldsymbol{s})\boldsymbol{A}^H\boldsymbol{\Phi}(\boldsymbol{s})^{-1}\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big). \quad (6)$$

Because brute force evaluation of all Q^N mixture vectors in $\mathcal{S}$ is impractical for typical values of N, we treat the problem as a (suboptimal) tree search. The primary challenge in the computation of the approximate MMSE estimate is to obtain $p(\boldsymbol{s}|\boldsymbol{y})$ and $\boldsymbol{\Phi}(\boldsymbol{s})^{-1}$ for each $\boldsymbol{s}$ in a high-probability subset $\mathcal{S}_\star \subset \mathcal{S}$ containing plausible mixture vectors, $\boldsymbol{s}$. Summing over the set of dominant mixture vectors $\mathcal{S}_\star$ yields the approximate MMSE estimate

$$\hat{\boldsymbol{x}}_{\mathsf{ammse}} \;:=\; \sum_{\boldsymbol{s}\in\mathcal{S}_\star} p(\boldsymbol{s}|\boldsymbol{y})\,\mathrm{E}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\}. \quad (7)$$

Similarly, the conditional covariance $\mathrm{Cov}\{\boldsymbol{x}|\boldsymbol{y}\}$, whose trace characterizes the MMSE estimation error, can be closely approximated as

$$\begin{aligned}\mathrm{Cov}\{\boldsymbol{x}|\boldsymbol{y}\} \;\approx\; \sum_{\boldsymbol{s}\in\mathcal{S}_\star} p(\boldsymbol{s}|\boldsymbol{y})\big[&\mathrm{Cov}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\} + (\hat{\boldsymbol{x}}_{\mathsf{ammse}} - \\ &\mathrm{E}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\})(\hat{\boldsymbol{x}}_{\mathsf{ammse}} - \mathrm{E}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\})^H\big] \quad (8)\\ \mathrm{Cov}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\} \;=\; \boldsymbol{R}(\boldsymbol{s}) - &\boldsymbol{R}(\boldsymbol{s})\boldsymbol{A}^H\boldsymbol{\Phi}(\boldsymbol{s})^{-1}\boldsymbol{A}\boldsymbol{R}(\boldsymbol{s}). \quad (9)\end{aligned}$$

Since, for any $\boldsymbol{s}$, the values of $p(\boldsymbol{s}|\boldsymbol{y})$ and $p(\boldsymbol{y}|\boldsymbol{s})p(\boldsymbol{s})$ are equal up to a scaling, the search for $\mathcal{S}_\star$ reduces to the search for the vectors $\boldsymbol{s}\in\mathcal{S}$ which yield the dominant values of $p(\boldsymbol{y}|\boldsymbol{s})p(\boldsymbol{s})$. For convenience, we use the monotonicity of the logarithm to define the *mixture selection metric* $\nu(\boldsymbol{s},\boldsymbol{y})$:

$$\begin{aligned}\nu(\boldsymbol{s},\boldsymbol{y}) \;&:=\; \ln p(\boldsymbol{y}|\boldsymbol{s})p(\boldsymbol{s}) \quad (10)\\ &=\; \ln p(\boldsymbol{y}|\boldsymbol{s}) + \ln p(\boldsymbol{s}) \quad (11)\\ &=\; -\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)^H\boldsymbol{\Phi}(\boldsymbol{s})^{-1}\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)\\ &\qquad -\ln\det(\boldsymbol{\Phi}(\boldsymbol{s})) - M\ln\pi + \sum_{n=0}^{N-1}\ln\lambda_{s_n}. \quad (12)\end{aligned}$$

B. A tree search

In [2], we propose a fast algorithm to search for the set $\mathcal{S}_\star$ of dominant mixture vectors; as a byproduct, the algorithm also generates the corresponding values of $\mathrm{E}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\}$ and $\mathrm{Cov}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\}$. A tree search begins with the hypothesis $\boldsymbol{s}=\boldsymbol{0}$ as the root of the tree, for which

$$\nu(\boldsymbol{0},\boldsymbol{y}) \;=\; -\tfrac{1}{\sigma^2}\|\boldsymbol{y}\|_2^2 - M\ln\sigma^2 - M\ln\pi + N\ln\lambda_0 \quad (13)$$

via (12) and the fact that $\boldsymbol{\Phi}(\boldsymbol{0}) = \sigma^2\boldsymbol{I}_M$. From the root of the tree, the metrics $\nu(\boldsymbol{s}',\boldsymbol{y})$ of all $(Q-1)N$ single-coefficient modifications of $\boldsymbol{s}^{(0)} := \boldsymbol{0}$ are calculated and, based on these metrics, a single hypothesis $\boldsymbol{s}^{(1)}$ is chosen to explore further by maximizing $\nu(\boldsymbol{s},\boldsymbol{y})$. The procedure continues recursively. If, at the m^{th} stage, $\nu(\boldsymbol{s}^{(m)},\boldsymbol{y})$ is adequately large or m exceeds some predetermined threshold, then the search stops, having evaluated $\nu(\boldsymbol{s},\boldsymbol{y})$ for some $\boldsymbol{s}\in\hat{\mathcal{S}}\subset\mathcal{S}$. The explored vectors $\boldsymbol{s}\in\hat{\mathcal{S}}$ that lead to significant values of $e^{\nu(\boldsymbol{s},\boldsymbol{y})} = p(\boldsymbol{y}|\boldsymbol{s})p(\boldsymbol{s})$ are then stored in $\hat{\mathcal{S}}_\star$, which constitutes an estimate of $\mathcal{S}_\star$.

We can approximate the posterior probability of a mixture vector $\boldsymbol{s}$ using the renormalized estimate:

$$p(\boldsymbol{s}|\boldsymbol{y}) \;=\; \frac{\exp\{\nu(\boldsymbol{s},\boldsymbol{y})\}}{\sum_{\boldsymbol{s}'\in\mathcal{S}}\exp\{\nu(\boldsymbol{s}',\boldsymbol{y})\}} \approx \frac{\exp\{\nu(\boldsymbol{s},\boldsymbol{y})\}}{\sum_{\boldsymbol{s}'\in\mathcal{S}_\star}\exp\{\nu(\boldsymbol{s}',\boldsymbol{y})\}}. \quad (14)$$

Likewise, $\hat{p}(\boldsymbol{x}|\boldsymbol{y})$ provides an approximate density function describing the uncertainty in resolving $\boldsymbol{x}$ with the noisy observations,

$$\hat{p}(\boldsymbol{x}|\boldsymbol{y}) = \sum_{\boldsymbol{s}'\in\mathcal{S}_\star}\hat{p}(\boldsymbol{s}|\boldsymbol{y})p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}). \quad (15)$$

The posterior density function is a Gaussian mixture and reflects the multi-modal ambiguity inherently present in the sparse inference problem—an ambiguity especially evident when the signal-to-noise ratio (SNR) is low or there exists nonnegligible correlation among the columns of $\boldsymbol{A}$.

C. Update for $\boldsymbol{\Phi}(\boldsymbol{s})^{-1}$ and $\nu(\boldsymbol{s},\boldsymbol{y})$

Central to implementing a tree search (or MCMC methods) is the need to evaluate the metrics $\{\nu(\boldsymbol{s}',\boldsymbol{y})\}$ for all one-parameter modifications $\boldsymbol{s}'$ of some previously considered mixture vector $\boldsymbol{s}$. For the case that $[\boldsymbol{s}]_n = q$ and $[\boldsymbol{s}']_n = q'$, where $\boldsymbol{s}$ and $\boldsymbol{s}'$ are otherwise identical, we now describe an efficient method to compute $\Delta_{n,q'}(\boldsymbol{s},\boldsymbol{y}) := \nu(\boldsymbol{s}',\boldsymbol{y}) - \nu(\boldsymbol{s},\boldsymbol{y})$. For brevity, we use the abbreviations $\mu_{q',q} := \mu_{q'} - \mu_q$ and $\sigma^2_{q',q} := \sigma^2_{q'} - \sigma^2_q$ below. Starting with the property

$$\boldsymbol{\Phi}(\boldsymbol{s}') \;=\; \boldsymbol{\Phi}(\boldsymbol{s}) + \sigma^2_{q',q}\boldsymbol{a}_n\boldsymbol{a}_n^H, \quad (16)$$

the matrix inversion lemma implies

$$\begin{aligned}\boldsymbol{\Phi}(\boldsymbol{s}')^{-1} \;&=\; \boldsymbol{\Phi}(\boldsymbol{s})^{-1} - \beta_{n,q'}\boldsymbol{c}_n\boldsymbol{c}_n^H \quad (17)\\ \boldsymbol{c}_n \;&:=\; \boldsymbol{\Phi}(\boldsymbol{s})^{-1}\boldsymbol{a}_n \quad (18)\\ \beta_{n,q'} \;&:=\; \sigma^2_{q',q}\big(1 + \sigma^2_{q',q}\boldsymbol{a}_n^H\boldsymbol{c}_n\big)^{-1}. \quad (19)\end{aligned}$$

In the next subsection, we verify that (16)-(19) imply

$$\Delta_{n,q'}(\boldsymbol{s},\boldsymbol{y}) = \begin{cases} \beta_{n,q'}\big|\boldsymbol{c}_n^H\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big) + \mu_{q',q}/\sigma^2_{q',q}\big|^2 & \\ \quad - |\mu_{q',q}|^2/\sigma^2_{q',q} + \ln(\beta_{n,q'}/\sigma^2_{q',q}) & \sigma^2_{q',q}\neq 0 \\ \quad + \ln(\lambda_{q'}/\lambda_q) & \\ 2\,\mathrm{Re}\big\{\mu^*_{q',q}\boldsymbol{c}_n^H\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)\big\} & \\ \quad - |\mu_{q',q}|^2\boldsymbol{c}_n^H\boldsymbol{a}_n + \ln(\lambda_{q'}/\lambda_q) & \sigma^2_{q',q} = 0. \end{cases} \quad (20)$$

Thus, $\Delta_{n,q'}(\boldsymbol{s},\boldsymbol{y})$ quantifies the change to $\nu(\boldsymbol{s},\boldsymbol{y})$ that results from changing the n^{th} index in $\boldsymbol{s}$ from q to q'.

The structure of $\boldsymbol{\Phi}(\boldsymbol{s})^{-1}$ can be exploited to yield complexity $\mathcal{O}(NM)$ for (18)-(19). Suppose that $\boldsymbol{s}$ is itself a single-index modification of $\boldsymbol{s}^{\mathsf{pre}}$, for which the n^{pre}-th index of $\boldsymbol{s}^{\mathsf{pre}}$ was changed from q^{pre} to q in order to create $\boldsymbol{s}$. If the corresponding quantities $\{\boldsymbol{c}_n^{\mathsf{pre}}\}_{n=0}^{N-1}$ and $\beta^{\mathsf{pre}}_{n^{\mathsf{pre}},q}$ have been computed and stored, then, since (17)-(18) imply that

$$\begin{aligned}\boldsymbol{c}_n \;&=\; \Big[\boldsymbol{\Phi}(\boldsymbol{s}^{\mathsf{pre}})^{-1} - \beta^{\mathsf{pre}}_{n^{\mathsf{pre}},q}\boldsymbol{c}^{\mathsf{pre}}_{n^{\mathsf{pre}}}\boldsymbol{c}^{\mathsf{pre}H}_{n^{\mathsf{pre}}}\Big]\boldsymbol{a}_n \quad (21)\\ &=\; \boldsymbol{c}_n^{\mathsf{pre}} - \beta^{\mathsf{pre}}_{n^{\mathsf{pre}},q}\boldsymbol{c}^{\mathsf{pre}}_{n^{\mathsf{pre}}}\boldsymbol{c}^{\mathsf{pre}H}_{n^{\mathsf{pre}}}\boldsymbol{a}_n, \quad (22)\end{aligned}$$

$\{\boldsymbol{c}_n\}_{n=0}^{N-1}$ can be computed using $\mathcal{O}(NM)$ operations.

Having computed $\{\boldsymbol{c}_n\}_{n=0}^{N-1}$, the parameters $\{\beta_{n,q'}\}_{n=0:N-1}^{q'=0:Q-1}$ can be computed via (19) with a complexity of $\mathcal{O}(MN + QN)$. If we recursively update $\boldsymbol{z}(\boldsymbol{s}) := \boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})$ with $\mathcal{O}(MQ)$ multiplies using

$$\boldsymbol{z}(\boldsymbol{s}) \;=\; \underbrace{\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s}^{\mathsf{pre}})}_{:=\,\boldsymbol{z}(\boldsymbol{s}^{\mathsf{pre}})} - \boldsymbol{a}_{n^{\mathsf{pre}}}\mu_{q,q^{\mathsf{pre}}}, \quad (23)$$

then $\{\Delta_{n,q'}(\boldsymbol{s})\}_{n=0:N-1}^{q'=0:Q-1}$ can be computed via (20) with a complexity of $\mathcal{O}(MN+QN)$. Actually, if $\sigma_q^2 = \sigma_1^2 \;\; \forall q \neq 0$, then $\beta_{n,q'} = \beta_{n,1} \;\; \forall q' \neq 0$, which leads to a total complexity of $\mathcal{O}(MN+MQ)$. Going further, if we define $\boldsymbol{C} := [\boldsymbol{c}_0, \ldots, \boldsymbol{c}_{N-1}]$ and notice that $\boldsymbol{C} = \boldsymbol{\Phi}(\boldsymbol{s})^{-1}\boldsymbol{A}$, then we can compute the $\boldsymbol{s}$-conditional mean and covariance via

$$\mathrm{E}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\} \;=\; \mu(\boldsymbol{s}) + \boldsymbol{R}(\boldsymbol{s})\boldsymbol{C}^H \boldsymbol{z}(\boldsymbol{s}) \tag{24}$$

$$\mathrm{Cov}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\} \;=\; \big(\boldsymbol{I}_N - \boldsymbol{R}(\boldsymbol{s})\boldsymbol{C}^H\boldsymbol{A}\big)\boldsymbol{R}(\boldsymbol{s}), \tag{25}$$

using (6), (9), and that $\boldsymbol{\Phi}(\boldsymbol{s})$ is Hermitian. Because $\boldsymbol{R}(\boldsymbol{s})\boldsymbol{C}^H$ has only $K_{\boldsymbol{s}} := \|\boldsymbol{s}\|_0$ nonzero rows and $\boldsymbol{A}\boldsymbol{R}(\boldsymbol{s})$ has only $K_{\boldsymbol{s}}$ nonzero columns, (24) and (25) can be computed using only $\mathcal{O}\big(K_{\boldsymbol{s}}M\big)$ and $\mathcal{O}\big(K_{\boldsymbol{s}}^2 M\big)$ multiplications, respectively.

D. Derivation of (20)

In this subsection, we establish (20) using (16)-(19). Using the fact that $\boldsymbol{\Phi}(\boldsymbol{s})^{-1}\boldsymbol{a}_n = \boldsymbol{c}_n$, we find

$$\begin{aligned}
&\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s}')\big)^H\boldsymbol{\Phi}(\boldsymbol{s}')^{-1}\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s}')\big)\\
&\;= \big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})-\boldsymbol{a}_n\mu_{q',q}\big)^H\big(\boldsymbol{\Phi}(\boldsymbol{s})^{-1}-\beta_{n,q'}\boldsymbol{c}_n\boldsymbol{c}_n^H\big)\\
&\quad\times\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})-\boldsymbol{a}_n\mu_{q',q}\big) \qquad (26)\\
&\;= \big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)^H\boldsymbol{\Phi}(\boldsymbol{s})^{-1}\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)\\
&\quad-\beta_{n,q'}\big|\boldsymbol{c}_n^H\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)\big|^2\\
&\quad-2\,\mathrm{Re}\big\{\mu_{q',q}^*\boldsymbol{a}_n^H\boldsymbol{\Phi}(\boldsymbol{s})^{-1}\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)\big\}\\
&\quad+2\,\mathrm{Re}\big\{\mu_{q',q}^*\boldsymbol{a}_n^H\boldsymbol{c}_n\beta_{n,q'}\boldsymbol{c}_n^H\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)\big\}\\
&\quad+|\mu_{q',q}|^2\boldsymbol{a}_n^H\boldsymbol{\Phi}(\boldsymbol{s})^{-1}\boldsymbol{a}_n-|\mu_{q',q}|^2\beta_{n,q'}(\boldsymbol{c}_n^H\boldsymbol{a}_n)^2 \qquad (27)\\
&\;= \big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)^H\boldsymbol{\Phi}(\boldsymbol{s})^{-1}\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)\\
&\quad-\beta_{n,q'}\big|\boldsymbol{c}_n^H\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)\big|^2\\
&\quad-2\,\mathrm{Re}\big\{\mu_{q',q}^*\boldsymbol{c}_n^H\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)\big(1-\beta_{n,q'}\boldsymbol{a}_n^H\boldsymbol{c}_n\big)\big\}\\
&\quad+|\mu_{q',q}|^2\boldsymbol{c}_n^H\boldsymbol{a}_n\big(1-\beta_{n,q'}\boldsymbol{a}_n^H\boldsymbol{c}_n\big). \qquad (28)
\end{aligned}$$

In the case that $\sigma_{q',q}^2 = 0$, we have $\beta_{n,q'} = 0$, and so

$$\begin{aligned}
&\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s}')\big)^H\boldsymbol{\Phi}(\boldsymbol{s}')^{-1}\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s}')\big)\\
&\;= \big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)^H\boldsymbol{\Phi}(\boldsymbol{s})^{-1}\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)\\
&\quad-2\,\mathrm{Re}\big\{\mu_{q',q}^*\boldsymbol{c}_n^H\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)\big\}+|\mu_{q',q}|^2\boldsymbol{c}_n^H\boldsymbol{a}_n. \qquad (29)
\end{aligned}$$

In the case that $\sigma_{q',q}^2 \neq 0$, we have $1-\beta_{n,q'}\boldsymbol{a}_n^H\boldsymbol{c}_n = -\beta_{n,q'}\sigma_{q',q}^{-2}$, so that

$$\begin{aligned}
&\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s}')\big)^H\boldsymbol{\Phi}(\boldsymbol{s}')^{-1}\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s}')\big)\\
&\;= \big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)^H\boldsymbol{\Phi}(\boldsymbol{s})^{-1}\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)\\
&\quad-\beta_{n,q'}\bigg[\big|\boldsymbol{c}_n^H\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)\big|^2\\
&\quad-2\,\mathrm{Re}\Big\{\boldsymbol{c}_n^H\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)\frac{\mu_{q',q}^*}{\sigma_{q',q}^2}\Big\}+\boldsymbol{c}_n^H\boldsymbol{a}_n\frac{|\mu_{q',q}|^2}{\sigma_{q',q}^2}\bigg] \qquad (30)\\
&\;= \big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)^H\boldsymbol{\Phi}(\boldsymbol{s})^{-1}\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)\\
&\quad-\beta_{n,q'}\bigg|\boldsymbol{c}_n^H\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)+\frac{\mu_{q',q}}{\sigma_{q',q}^2}\bigg|^2\\
&\quad+\beta_{n,q'}\frac{|\mu_{q',q}|^2}{\sigma_{q',q}^4}\Big[1+\sigma_{q',q}^2\boldsymbol{c}_n^H\boldsymbol{a}_n\Big] \qquad (31)\\
&\;= \big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)^H\boldsymbol{\Phi}(\boldsymbol{s})^{-1}\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)\\
&\quad-\beta_{n,q'}\bigg|\boldsymbol{c}_n^H\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)+\frac{\mu_{q',q}}{\sigma_{q',q}^2}\bigg|^2+\frac{|\mu_{q',q}|^2}{\sigma_{q',q}^2}. \qquad (32)
\end{aligned}$$

Together, (29) and (32) yield (33).

$$\begin{aligned}
&\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s}')\big)^H\boldsymbol{\Phi}(\boldsymbol{s}')^{-1}\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s}')\big)\\
&= \begin{cases}
\begin{aligned}&\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)^H\boldsymbol{\Phi}(\boldsymbol{s})^{-1}\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)\\ &-\beta_{n,q'}\big|\boldsymbol{c}_n^H\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)+\mu_{q',q}/\sigma_{q',q}^2\big|^2\\ &+|\mu_{q',q}|^2/\sigma_{q',q}^2\end{aligned} & \sigma_{q',q}^2 \neq 0\\[2ex]
\begin{aligned}&\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)^H\boldsymbol{\Phi}(\boldsymbol{s})^{-1}\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)\\ &-2\,\mathrm{Re}\big\{\mu_{q',q}^*\boldsymbol{c}_n^H\big(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)\big\}\\ &+|\mu_{q',q}|^2\boldsymbol{c}_n^H\boldsymbol{a}_n\end{aligned} & \sigma_{q',q}^2 = 0.
\end{cases} \qquad (33)
\end{aligned}$$

Equation (16) then implies that

$$\begin{aligned}
\ln\det(\boldsymbol{\Phi}(\boldsymbol{s}')) &= \ln\det\big(\boldsymbol{\Phi}(\boldsymbol{s})+\sigma_{q',q}^2\boldsymbol{a}_n\boldsymbol{a}_n^H\big) && (34)\\
&= \ln\Big[\big(1+\sigma_{q',q}^2\boldsymbol{a}_n^H\boldsymbol{\Phi}(\boldsymbol{s})^{-1}\boldsymbol{a}_n\big)\det\big(\boldsymbol{\Phi}(\boldsymbol{s})\big)\Big]\\
&= \ln\det(\boldsymbol{\Phi}(\boldsymbol{s}))-\ln(\beta_{n,q'}/\sigma_{q',q}^2) && (35)\\
\ln p(\boldsymbol{s}') &= \ln p(\boldsymbol{s})+\ln(\lambda_{q'}/\lambda_q), && (36)
\end{aligned}$$

which, in conjunction with (12) and (33), yield (20).

E. A repeated greedy search

In $\mathcal{O}((M+Q)N)$ complexity, the fast update method evaluates the metrics $\nu(\boldsymbol{s},\boldsymbol{y})$ for all $(Q-1)N$ single coefficient modifications at each node visited by the tree search. The search starts at the root node $\boldsymbol{s}=\boldsymbol{0}$ and performs a greedy inflation search (i.e., activating one mixture parameter at a time) until a total of P mixture parameters have been activated. By "greedy," we mean that the mixture parameter activated at each stage is the one yielding the largest metric $\nu(\boldsymbol{s},\boldsymbol{y})$; de-activation is not allowed. P should be chosen to be slightly larger than the expected number of nonzero coefficients $\mathrm{E}\{K_{\boldsymbol{s}}\}$, e.g., so that $\Pr(\|\boldsymbol{s}\|_0 > P)$ is sufficiently small.[1] The tree search may be repeated from the root node, ignoring previously explored nodes; stopping rules are suggested in [2].

When domain knowledge does not precisely specify the hyperparameters, we opt to compute maximum likelihood (ML) estimates. A generalized expectation maximization (EM) [6] iteration is adopted for ML estimation of the hyperparameters, as detailed in [2].

Denoting by $D \leq D_{\mathsf{max}}$ the number of greedy searches performed, $DPN(Q-1)$ mixture vectors in total are examined with corresponding metrics $\nu(\boldsymbol{s},\boldsymbol{y})$. The number of multiplications required to compute all metrics and PD conditional means is $\mathcal{O}(DPNM)$. Computing the PD conditional covariances $\{\hat{\boldsymbol{\Sigma}}^{(d,p)}\}_{d=1:D}^{p=1:P}$ requires an additional $\mathcal{O}(DP^3M)$ multiplies. The generalized EM iteration uses FMBP for each E-step and $\mathcal{O}(M)$ operations per M-step.

[1] Recall that $\|\boldsymbol{s}\|_0$ follows the Binomial$(N, 1-\lambda_0)$ distribution. When $N(1-\lambda_0) > 5$, it is reasonable to use the Gaussian approximation $\|\boldsymbol{s}\|_0 \sim \mathcal{N}\big(N(1-\lambda_0), N\lambda_0(1-\lambda_0)\big)$, in which case $\Pr(\|\boldsymbol{s}\|_0 > P) = \frac{1}{2}\,\mathrm{erfc}\Big(\frac{P-N(1-\lambda_0)}{\sqrt{2N\lambda_0(1-\lambda_0)}}\Big)$.

IV. Simulation

Numerical experiments were conducted to investigate the performance of FBMP with approximate ML estimation of hyperparameters from the data[2]. Since FBMP is able to provide an approximate MMSE solution to (1) for the signal model presented in Section II, we would expect to observe near-optimal performance, in the mean squared error sense, when testing FBMP on signals generated accoring to such a model. A more interesting characterization of FBMP's performance can be obtained by testing the algorithm on signals that violate the assumptions of Section II. Such a characterization demonstrates both the flexibility of the Gaussian mixture model in approximating other generating distributions and the utility of allowing FBMP to adaptively select hyperparameters when such information is unavailable.

In the results presented here, we considered a signal consisting of $N = 512$ unknown coefficients that followed a deterministic exponentially decaying profile, that is, $x_k = \exp\{-\rho k\}$, with $\rho \in [0.10, 0.85]$. Such a signal could be encountered, for instance, in the wavelet coefficients of an image [7]. For each trial, a random selection of half of the coefficients were given negative sign, and the coefficients were randomly shuffled. The 128-by-512 measurement matrices $\boldsymbol{A}$ were constructed by drawing i.i.d. columns from a normal distribution, and then scaling the columns to be of unit-norm. Noise realizations came from a zero-mean Gaussian distribution with variances chosen to yield 15 dB SNR. The graphs represent an average of 2000 independent realizations. For comparative purposes, we also tested five other publicly available sparse estimation algorithms: SparseBayes [8], OMP [9], StOMP [10], GPSR-Basic [11], and BCS [12]. StOMP was tested using a false discovery rate control strategy for threshold selection, with a rate of $q = 0.40$. Following [13], we chose to use the expected noise power as the stopping threshold on the norm of the residual error. The "compressed," rather than sparse, nature of the exponentially decaying signal seemed to pose problems for OMP in regards to choosing a suitable stopping criterion, causing OMP to activate many coefficients in an attempt to fit the noise. For this reason, we opted to terminate OMP after it had activated as many coefficients as StOMP, providing some insight into how both algorithms perform for a specific degree of sparsity. The ℓ_1-penalty in the GPSR algorithm was chosen as $\tau = 0.1\|\boldsymbol{A}^H\boldsymbol{y}\|_\infty$, and the MSE kept for comparison purposes was the smaller of the MSEs of the un-debiased and debiased estimates. For FBMP, hyperparameters were initialized at $\lambda_1 = 0.01$, $\mu_1 = 0$, $\sigma^2 = 0.05$, and $\sigma_1^2 = 2$, for all values of ρ, and the generalized EM updates were used to compute approximate ML estimates of the hyperparameters from the data.

In Fig. 1 we plot normalized mean squared error (NMSE), defined by

$$\text{NMSE (dB)} = 10\log_{10}\left\{\frac{1}{T}\sum_{i=1}^{T}\frac{\|\hat{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}\|_2^2}{\|\boldsymbol{x}^{(i)}\|_2^2}\right\}, \qquad (37)$$

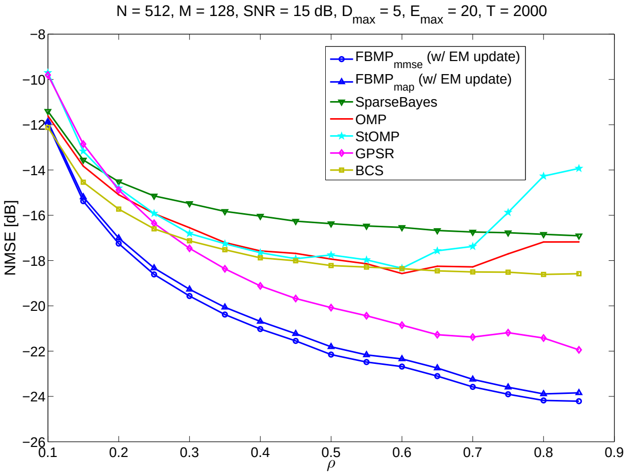

Fig. 1. Normalized mean squared error versus ρ.

where T is the number of random trials and superscript (i) denotes the trial number. For FBMP we provide the NMSE of both its MMSE and maximum a posteriori (MAP) estimates, quantifying the reduction in signal reconstruction error that results from averaging over multiple models. We note that FBMP's estimators are not in fact MMSE or MAP for the class of signals being considered, but would rather be (approximately) MMSE and MAP were the signals drawn from a Gaussian mixture distribution with hyperparameters obtained through the EM update procedure. The NMSEs reported by FBMP are the lowest for nearly the entire range of ρ considered. We attribute this performance in part to the hyperparameter estimation and also to role of the prior on $\boldsymbol{s}$ in favoring a sparse solution to a dense one. Likewise, the estimate from GPSR exhibits very low NMSE, which is attributable to the exponentially decaying simulated signal: the sequence $\boldsymbol{x}$ can be viewed as a typical draw from a Laplace density, and thus is well-matched to the MAP estimator for a Laplacian prior. The MAP estimator under such a prior can be cast as a convex optimization problem of the form solved

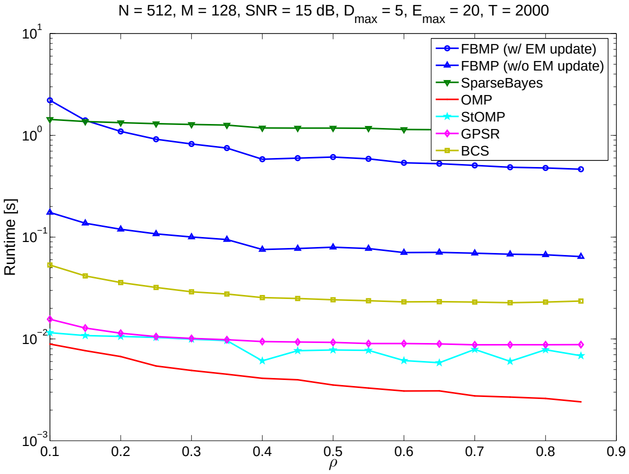

Fig. 2. Runtime versus ρ.

[2] Color versions of figures in this manuscript, Matlab code, and presentations are available at `http://www.ece.osu.edu/∼zinielj/fbmp/`

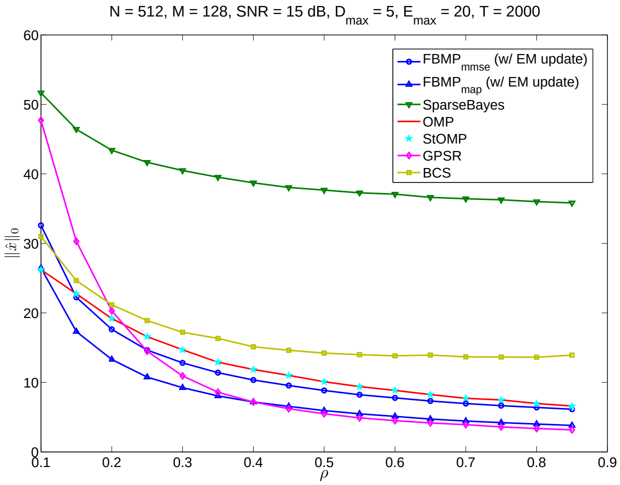

Fig. 3. Solution sparsity versus ρ.

by GPSR.

Figure 2 displays average runtimes and quantifies the complexity of implementing a Bayesian model estimator with FBMP. The FBMP algorithm returns not only a MAP model selection $\hat{\boldsymbol{s}}_\star$, but also returns $\hat{\boldsymbol{s}}_{\mathsf{ammse}}$ and a list of high probability explanations of the data, along with their posterior probabilities. Thus, FBMP is able to give a fuller interpretation of the data in the face of ambiguity arising from correlation in $\boldsymbol{A}$ or from measurement noise. The other five approaches considered return only a single basis selection and do not provide a report of model uncertainty. The low complexity of OMP, StOMP, and GPSR is clearly evident in the figure. The complexity in the empirical Bayes estimation of model parameters is illustrated by the comparison of FBMP with and without the generalized EM update procedure. Accordingly, there is a complexity reduction for applications in which hyperparameters are precisely known.

Fig. 3 shows average sparsity of solutions. While the sparsity of solutions for almost every algorithm can be modulated by altering appropriate parameters, it is interesting to observe the interplay between NMSE and sparsity. Remarkably, the sparsest solutions, provided by FBMP and GPSR, also have the lowest values of NMSE.

V. Discussion: Related Works

A Gaussian mixture model similar to that in Section II was adopted by Larsson and Selén [14], who, for $Q = 2$, also constructed the MMSE estimate in the manner of (7) but with an $\mathcal{S}_\star$ that contains exactly one sequence $\boldsymbol{s}$ for each Hamming weight 0 to N. They proposed to find these $\boldsymbol{s}$ by starting with an all-active basis configuration and recursively deactivating one element at a time. Thus, the $D = 1$ version of the FBMP algorithm recalls the heuristic of [14], but in reverse. The fast update presented here has a complexity of only $\mathcal{O}(NMP)$, in comparison to $\mathcal{O}(N^3M^2)$ for the technique in [14]. Given the typically large values of N encountered in practice, the complexity of FBMP can be several orders of magnitude lower. Elad and Yavneh [15] adopted MMSE estimation as a motivation for averaging multiple sparse denoising solutions, each found via a randomized OMP solution.

For $Q = 2$, a Gaussian mixture model has been widely adopted for the Bayesian variable selection problem. (See, e.g., [16] for a survey.) The published approaches vary in prior specification, posterior calculation, and MCMC sampling method. George and McCulloch [17] use a conjugate normal prior on $\boldsymbol{x}|\boldsymbol{s}, \sigma^2$ and a Gibbs sampler that requires $\mathcal{O}(N^2)$ operations to compute $p(\boldsymbol{s}_j|\boldsymbol{y})$ from $p(\boldsymbol{s}_i|\boldsymbol{y})$, where $\boldsymbol{s}_j$ and $\boldsymbol{s}_i$ differ in only one position. Smith and Kohn [18] use the point mass null and the simplifying Zellner-g conditional prior to achieve a fast update requiring $\mathcal{O}(NK_{\boldsymbol{s}}^2)$ operations. Approximately MN iterations of the Gibbs sampler are suggested, yielding a total complexity of $\mathcal{O}(MN^2K_{\boldsymbol{s}}^2)$.

References

[1] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad., "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Ann. Asilomar Conf. Signals, Systems, and Computers*, 1993.

[2] P. Schniter, L. C. Potter, and J. Ziniel, "Fast Bayesian matching pursuit: Model uncertainty and parameter estimation for sparse linear models," *IEEE Trans Signal Processing*, submitted August 2008.

[3] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Information Theory*, vol. 52, pp. 6–18, Jan. 2006.

[4] J. A. Tropp, "Just relax: Convex programming methods for identifying sparse signal," *IEEE Trans. Info. Theory*, vol. 51, pp. 1030–1051, Mar. 2006.

[5] E. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc. B*, vol. 39, no. 1, pp. 1–38, 1977.

[7] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press, 1998.

[8] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Machine Learning Res.*, vol. 1, pp. 211–244, 2001. (software available at `http://www.miketipping.com/index.php?page=rvm`).

[9] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Information Theory*, vol. 53, pp. 4655–4666, Dec. 2007. (software available at `http://sparselab.stanford.edu/`).

[10] D. L. Donoho, Y. Tsaig, I. Drori, and J.-C. Starck, "Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit," Tech. Rep. 2006-02, Dept. of Statistics, Stanford University, Stanford, CA, 2006. (software available at `http://sparselab.stanford.edu/`).

[11] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 586–597, 2007. (software available at `http://www.lx.it.pt/~mtf/GPSR/`).

[12] S. Ji and L. Carin, "Bayesian compressive sensing and projection optimization," in *Proc. 24th Int. Conf. Machine Learning (ICML)*, pp. 377 – 384, 2007. (software available at `http://www.ece.duke.edu/~shji/BCS.html`).

[13] M. Elad, B. Matalon, J. Shtok, and M. Zibulevsky, "A wide-angle view at iterated shrinkage algorithms," in *SPIE (Wavelet XII)*, 2007.

[14] E. Larsson and Y. Selén, "Linear regression with a sparse parameter vector," *IEEE Trans. Signal Process.*, vol. 55, pp. 451 – 460, Feb. 2007.

[15] M. Elad and I. Yavneh, "A weighted average of sparse representations is better than the sparsest one alone," 2008. preprint.

[16] M. Clyde and E. I. George, "Model uncertainty," *Statist. Sci.*, vol. 19, no. 1, pp. 81 – 94, 2004.

[17] E. I. George and R. E. McCulloch, "Approaches for Bayesian variable selection," *Statistica Sinica*, vol. 7, pp. 339 – 373, 1997.

[18] M. Smith and R. Kohn, "Nonparametric regression using Bayesian variable selection," *J. Econometrics*, vol. 75, pp. 317 – 343, 1996.