Fast Bayesian Matching Pursuit: Model Uncertainty and Parameter Estimation for Sparse Linear Models

Philip Schniter *Senior Member, IEEE*, Lee C. Potter, *Senior Member, IEEE* and Justin Ziniel

Abstract

A low-complexity recursive procedure is presented for model selection and minimum mean squared error (MMSE) estimation in linear regression. Emphasis is given to the case of a sparse parameter vector and fewer observations than unknown parameters. A Gaussian mixture is chosen as the prior on the unknown parameter vector. The algorithm returns both a set of high posterior probability models and an approximate MMSE estimate of the parameter vector. Exact ratios of posterior probabilities serve to reveal potential ambiguity among multiple candidate solutions that are ambiguous due to observation noise or correlation among columns in the regressor matrix. Algorithm complexity is $\mathcal{O}(MNK)$, with M observations, N coefficients, and K nonzero coefficients. For the case that hyperparameters are unknown, an approximate maximum likelihood estimator is proposed based on the generalized expectation-maximization algorithm. Numerical simulations demonstrate estimation performance and illustrate the distinctions between MMSE estimation and maximum *a posteriori* probability model selection.

Index Terms

Sparse reconstruction, compressive sampling, compressed sensing, sparse linear regression, Bayesian model averaging, Bayesian variable selection, empirical Bayes.

Manuscript submitted September 9, 2008 and revised March 9, 2009. 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 this work was presented at the Workshop on Information Theory and Applications (ITA), La Jolla, CA, January 2008.

The authors are with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, 43210-1272.

I. Introduction

Sparse linear regression is a topic of long-standing interest in signal processing, statistics, and geophysics. The linear model is given by

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

with observation vector $\boldsymbol{y}$, known regressor matrix $\boldsymbol{A}$, unknown coefficients $\boldsymbol{x}$, and additive noise $\boldsymbol{w}$. In sparse problems, the prior belief is that only a small fraction of coefficients are non-negligible.

We adopt a Bayesian approach, which we now review in general terms. Let γ_k denote a candidate model, with k indexing the countably many models under consideration. A prior probability $p(\gamma_k)$ is assigned to each model, and a prior $p(\boldsymbol{\theta}_k|\gamma_k)$ is adopted for the parameters of each model. For example, in (1) a model γ_k might indicate which entries in $\boldsymbol{x} \in \mathbb{R}^N$ are nonzero, resulting in 2^N candidate models. For linear regression, a model is also known as a variable selection or basis selection. Margining out parameters and conditioning on the observations yields posterior model probabilities

$$p(\gamma_k|\boldsymbol{y}) \,=\, \frac{p(\boldsymbol{y}|\gamma_k)p(\gamma_k)}{\sum_j p(\boldsymbol{y}|\gamma_j)p(\gamma_j)}. \tag{2}$$

Pairwise comparison of candidate models is given by the posterior odds

$$\frac{p(\gamma_k|\boldsymbol{y})}{p(\gamma_j|\boldsymbol{y})} \,=\, \frac{p(\boldsymbol{y}|\gamma_k)}{p(\boldsymbol{y}|\gamma_j)}\frac{p(\gamma_k)}{p(\gamma_j)}. \tag{3}$$

The model posterior probabilities give a full description of the post-data uncertainty and are useful for inference and decision tasks. A common choice is to compute a single model that maximizes the posterior probability—the MAP estimate, $\hat{\gamma}_\star$. However, to obtain the minimum mean squared error estimate of $\boldsymbol{x}$, one must compute a weighted average of conditional mean estimates over all models with nonzero probability,

$$\hat{\boldsymbol{x}}_{\textsf{mmse}} \,=\, \sum_k p(\gamma_k|\boldsymbol{y})\, \mathrm{E}\{\boldsymbol{x}|\boldsymbol{y},\gamma_k\}. \tag{4}$$

Bayesian model averaging (e.g., [1], [2] and references therein) is a name sometimes given to this incorporation of model uncertainty and stands in contrast to model selection, which is the report of a single model. Thus, the essential element provided by the Bayesian approach is the quantification of posterior model uncertainty. The posterior odds reveal uncertainty among multiple candidate solutions that are ambiguous due to observation noise or correlation among columns in the regressor matrix, $\boldsymbol{A}$. Bayesian techniques are classical; the novelty here is a suite of computational techniques that make Bayesian estimation not only tractable, but low complexity, for the sparse linear model, with emphasis on the case of fewer observations than unknown variables.

This manuscript is organized as follows. In Section II we briefly survey existing approaches to sparse linear regression. In Section III, we state a flexible signal model and priors for sparse signals; the priors explicitly specify our modeling assumptions and admit precise interpretation. In Section IV, we describe our proposed algorithm. A tree-search is combined with a low-complexity update of model posterior probabilities to find a dominant set of likely models. An algorithm for computing approximate maximum likelihood estimates of the hyperparameters, based on a generalized expectation maximization (EM) update, is presented in Section V for use when such hyperparameters are not known for a given application. We numerically investigate in Section VI the algorithm's performance. In Section VII, we give specific comparison to related work. Conclusions are summarized in Section VIII.

II. Techniques for Sparse Linear Regression

We present a brief and necessarily incomplete survey of existing approaches to sparse linear regression, with an emphasis on the themes relevant to our proposed procedure for model uncertainty and parameter estimation. For convenience, we coarsely partition approaches into those that do or do not explicitly adopt prior distributions.

A. Algorithms for sparse signal reconstruction

In sparse signal reconstruction, the general aim is to identify the smallest subset of columns of the regressor matrix, $\boldsymbol{A}$, whose linear span contains (approximately) the observations, $\boldsymbol{y}$. Algorithmic approaches have been proposed for several decades and broadly fall into three categories. The algorithms return a single model estimate and do not quantify uncertainty in the reported estimate. The algorithms have typically been developed without recourse to probabilistic priors.

One class of algorithms adopts a greedy search heuristic. Examples include CLEAN [3], projection pursuit [4], and orthogonal matching pursuit (OMP) [5]. There exist sufficient conditions [6], [7] on the sparseness of $\boldsymbol{x}$ and singular values of subsets of columns of $\boldsymbol{A}$ (e.g., the restricted isometry property [8]) such that a regularized OMP stably recovers $\boldsymbol{x}$ with high probability.

A second class of algorithms recursively solves a sequence of iteratively re-weighted linear least-squares (IRLS) problems [9]–[11]; recent results [12] for the noiseless case have established sufficient conditions such that the sequence converges to the sparsest solution.

A third class comprises penalized least-squares solutions for $\boldsymbol{x}$ and has likewise been used for at least four decades [13]. In this class of approaches, parameters are found via the optimization

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \; \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \tau \|\boldsymbol{x}\|_p^p, \tag{5}$$

or, equivalently, for some $\epsilon > 0$

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \ \|\boldsymbol{x}\|_p \ \text{ s.t. } \ \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 < \epsilon. \tag{6}$$

Ridge regression [14] (i.e., Tikhonov regularization) adopts $p = 2$, while basis pursuit [15] and LASSO [16] use $p = 1$. Equation (5) has been widely adopted, for example in image reconstruction [17], [18], radar imaging [19], and elsewhere [20], [21]. With proper choice of norm, total variation denoising is also an algorithm in this class for $p = 1$ [22], [23].

A link exists to Bayesian estimation; the large class of methods adopting (5) may be interpreted as implicitly seeking the MAP estimate of $\boldsymbol{x}$ under the prior

$$p(\boldsymbol{x}) \propto \exp\left\{ -\tfrac{\tau}{2}\|\boldsymbol{x}\|_p^p \right\}. \tag{7}$$

Solutions depend on choice of hyperparameters τ and ϵ in (5) and (6), and the choice can be problematic; typically, a cross-validation procedure is adopted, whereby solutions are computed for a range of hyperparameters.

Elegant recent results by several authors [8], [24], [25] have demonstrated sufficient conditions on $\boldsymbol{A}$, $\boldsymbol{w}$, and the sparsity of the true coefficients, $\boldsymbol{x}$, such that for $p = 1$ the convex problem (6) provides the stable solution (8) for certain positive C:

$$\min \|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2 < C\epsilon. \tag{8}$$

These proofs have validated the widespread use of (5)-(6), providing a deeper understanding, spurring a resurgent interest, and promoting the interpretation as "compressive sampling." The sufficient conditions on $\boldsymbol{A}$ are the restricted isometry property [8] (RIP) or a bound on the mutual coherence [25], which is the maximum correlation among the columns in $\boldsymbol{A}$.

A constructive procedure for $\boldsymbol{A}$ consistent with RIP remains open [26]. But the compressive sampling hypotheses are met with high probability by draws from classes of random matrices. In this sense, compressive sampling trades the NP-hard ℓ_0 sparsest solution task for an intractable experiment design, then uses randomization for experiment design. In a similar way, randomization has been used in an *ad hoc* manner for over 40 years in array processing for low side-lobe responses [27], [28]. Thus, compressive sampling theorems offer an invitation to randomized sampling.

In the sparse reconstruction and compressive sampling literature, primary focus is placed on the detection of the few significant entries of the sparse $\boldsymbol{x}$—a task alternatively known as model selection, variable selection, subset selection, or basis selection. In addition, an estimate of the parameters $\boldsymbol{x}$ is also sought. In all these techniques, a single solution is returned without a report of posterior model uncertainty.

B. Bayesian approaches

Bayesian approaches have been widely reported in a variety of subdisciplines. The relevance vector machine [29]–[31] explicitly adopts a Bayesian framework with x_i independent, zero-mean, Gaussian with unknown variance σ_i^2. The unknown variances are assigned the inverse Gamma conjugate prior and an EM iteration computes a MAP estimate of $\boldsymbol{x}$. Although priors are adopted, these approaches do not compute and report posterior probabilities for candidate models; instead, a single model is reported that approximates the MAP model estimate.

In the statistics literature, rapidly advancing computing technology and the advent of Markov chain Monte Carlo (MCMC) methods for posterior computation combined to yield a large body of Bayesian methods for model uncertainty. Linear models, as the canonical version of nonparametric regression, have been widely studied, with attention focused to the over-determined case (more observations than potential predictors). Approaches differ in specification of the priors and numerical methods for rapidly computing posterior probabilities for candidate models. For example, Smith and Kohn [32] adopt a log-uniform prior on the noise variance, an independent Bernoulli prior for selection of nonzero coefficients, and a Zellner g-prior[1] on the coefficients conditioned on both the noise variance and the indices of nonzero coefficients. Then, a Gibbs sampler is used to simulate a pseudorandom sample of models (i.e., configurations of nonzero coefficients) that converges in distribution to the posterior model probabilities. In the MCMC methods, this sequence is used to search for high probability models and to obtain posterior weighted averages for estimation tasks. (See [1], [33] for surveys and references, and see [34] for application of MCMC to an underdetermined Gabor transform problem.) Elad and Yavneh [35] proposed a similar randomization to identify a sequence of candidate models. A randomized OMP algorithm is used to create solutions with sparsity $\|\boldsymbol{x}\|_0 = K$. At each instance of OMP, indices are drawn from among columns of $\boldsymbol{A}$ most correlated with the residual. The log-probability in the draw is proportional to the decrease in the residual. A MMSE-inspired denoising (i.e., estimate of $\boldsymbol{Ax}$) is then generated by averaging, with uniform weights, the least-squares solutions computed under each model hypothesis. The algorithm is not

[1] Given the variable selection and noise covariance, the Zellner g-prior is zero-mean jointly Gaussian with covariance $g\sigma^2(\boldsymbol{A}_{\boldsymbol{s}}^T\boldsymbol{A}_{\boldsymbol{s}})^{-1}$, where $\boldsymbol{A}_{\boldsymbol{s}}$ is formed by keeping columns from $\boldsymbol{A}$ corresponding to the nonzero coefficients. The prior is chosen for computational convenience and is inconsistent for the null model [1].

derived from a Bayesian formulation; however, the analysis in the manuscript adopts the Zellner g-prior and assumes a known number of nonzero coefficients.

Finally, Bayesian model averaging was adopted by Larsson and Selén [36] to approximate minimum mean squared error (MMSE) estimates. In the sparse over-determined case, a greedy deflation search is used to identify high-probability models.

In this paper, we adopt a Bayesian model averaging treatment of model uncertainty and we propose fast computational techniques to compute posterior model probabilities for the underdetermined, or undersampled data, case. Further, we arrive at a fast computation technique without adopting the Zellner g-prior. A method for approximate maximum likelihood estimation of hyperparameters based on a generalized-EM update is given, for cases when hyperparameters are not known for a specific application.

III. Signal Model

This section defines our signal model and priors. We choose to present a general model, with $\boldsymbol{x}$ drawn from a Q-ary mixture of complex-valued Gaussians with arbitrary means. While this generality affords application to many practical signals without changing the proposed fast algorithm, it requires a complexity of notation relative to the simplest special cases of the model. The section concludes with a description of four specific examples of the general model.

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

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}, \tag{9}$$

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., $\boldsymbol{w} \sim \mathcal{CN}(\mathbf{0}, \sigma^2 \boldsymbol{I}_M)$, where 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., $\|\boldsymbol{x}\|_0 \ll N$). Although we assume complex-valued quantities, our methods are suitable for real-valued problems with minor modifications.

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:

$$x_n | \{s_n = q\} \sim \mathcal{CN}(\mu_q, \sigma_q^2). \tag{10}$$

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 $(\mu_0, \sigma_0^2) = (0, 0)$, so that the case $s_n = 0$ implies $x_n = 0$, whereas the case $s_n > 0$ allows $x_n \neq 0$.

In addition, we choose $\{\lambda_q\}_{q=0}^{Q-1}$ so that $\sum_{q=1}^{Q-1} \lambda_q \ll 1$, which ensures that (with high probability) the coefficient vector $\boldsymbol{x}$ has relatively few nonzero values.

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

$$\boldsymbol{x}|\boldsymbol{s} \sim \mathcal{CN}(\boldsymbol{\mu}(\boldsymbol{s}), \boldsymbol{R}(\boldsymbol{s})), \tag{11}$$

where $[\boldsymbol{\mu}(\boldsymbol{s})]_n = \mu_{s_n}$ and where $\boldsymbol{R}(\boldsymbol{s})$ is diagonal with $[\boldsymbol{R}(\boldsymbol{s})]_{n,n} = \sigma^2_{s_n}$. Equation (9) then implies that the unknown coefficients, $\boldsymbol{x}$, and the measurements, $\boldsymbol{y}$, are jointly Gaussian when conditioned on the model vector, $\boldsymbol{s}$. In particular,

$$\left.\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{bmatrix}\right| \boldsymbol{s} \sim \mathcal{CN}\left(\begin{bmatrix} \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s}) \\ \boldsymbol{\mu}(\boldsymbol{s}) \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Phi}(\boldsymbol{s}) & \boldsymbol{A}\boldsymbol{R}(\boldsymbol{s}) \\ \boldsymbol{R}(\boldsymbol{s})\boldsymbol{A}^H & \boldsymbol{R}(\boldsymbol{s}) \end{bmatrix} \right), \tag{12}$$

where

$$\boldsymbol{\Phi}(\boldsymbol{s}) \triangleq \boldsymbol{A}\boldsymbol{R}(\boldsymbol{s})\boldsymbol{A}^H + \sigma^2 \boldsymbol{I}_M. \tag{13}$$

We now provide examples of how the hyperparameters Q, $\{\lambda_q\}_{q=0}^{Q-1}$, $\{\mu_q\}_{q=0}^{Q-1}$, and $\{\sigma_q^2\}_{q=0}^{Q-1}$ could be chosen.

- *Zero-mean binary prior*: Here, $Q = 2$, $\mu_1 = 0$, and $\sigma_1^2 > 0$. With this conveniently simple prior, it can be potentially difficult to distinguish an “active” coefficient from a non-active one, since the most *a priori* probable active-coefficient values are those near zero.
- *Nonzero-mean binary prior*: Here, $Q = 2$, $\mu_1 \neq 0$, and $\sigma_1^2 > 0$. Compared to the zero-mean binary prior, active coefficients have a known nonzero mean value[2].
- *Zero-mean ternary prior*: Here, $Q = 3$, $\mu_1 = -\mu_2$, $\sigma_1^2 = \sigma_2^2 > 0$, and $\lambda_1 = \lambda_2$. Appropriate for the real-valued case with no prior knowledge of sign, this model facilitates the discrimination between active and non-active coefficients when μ_1 and σ_1^2 are suitably chosen.
- *Q-ary circular prior*: Here, $Q > 3$ and, for all $q \in \{1, \ldots, Q\}$, we set $\mu_q = |\mu_1| e^{j2\pi\frac{q-1}{Q-1}}$, $\sigma_q^2 = \sigma_1^2 > 0$, and $\lambda_q = \lambda_1$. This generalization of the zero-mean ternary prior is suitable for complex-valued coefficients with *a priori* unknown phase.

[2] An application of this model arises in electron paramagnetic resonance (EPR) imaging, where an exogenous spin deposit is constructed from a paramagnetic material [37]. For the EPR application, σ_1^2 models variability in the number of spins present in a polymer-encapsulated microliter deposit.

IV. Model Uncertainty & Estimating Coefficients

The observation model (9) is a Gaussian mixture and presents two principal problems: model selection and parameter estimation. The first task is the selection of one or more highly probable models from the Q^N possible models indexed by $\boldsymbol{s}$. We refer to $\boldsymbol{s}$ as the "model vector." In the Bayesian framework, we also compute posterior probabilities, $p(\boldsymbol{s}|\boldsymbol{y})$. The second task is the estimation of the coefficients, $\boldsymbol{x}$. In this section, we propose a low-complexity method to simultaneously accomplish both of these tasks.

A. Model selection

We index the set of all model vectors by $\mathcal{S} \triangleq \{0, 1, \ldots, Q-1\}^N$. The maximum *a posteriori* (MAP) model-vector estimate is given by $\hat{\boldsymbol{s}}_\star \triangleq \operatorname{argmax}_{\boldsymbol{s}\in\mathcal{S}} p(\boldsymbol{s}|\boldsymbol{y})$. We seek to determine not only the MAP model-vector $\hat{\boldsymbol{s}}_\star$ but also the set $\mathcal{S}_\star$ of all model vectors with non-negligible posterior probability, along with their posteriors $\{p(\boldsymbol{s}|\boldsymbol{y})\}_{\boldsymbol{s}\in\mathcal{S}_\star}$. By analogy to data communications, finding $\hat{\boldsymbol{s}}_\star$ is like "hard decoding," whereas finding $\{p(\boldsymbol{s}|\boldsymbol{y})\}_{\boldsymbol{s}\in\mathcal{S}_\star}$ is like "soft decoding."

Using Bayes rule, the model-vector posterior becomes

$$p(\boldsymbol{s}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{s})p(\boldsymbol{s})}{\sum_{\boldsymbol{s}'\in\mathcal{S}} p(\boldsymbol{y}|\boldsymbol{s}')p(\boldsymbol{s}')}. \tag{14}$$

Given $\mathcal{S}_\star$, the posteriors can be approximated by

$$p(\boldsymbol{s}|\boldsymbol{y}) \approx \frac{p(\boldsymbol{y}|\boldsymbol{s})p(\boldsymbol{s})}{\sum_{\boldsymbol{s}'\in\mathcal{S}_\star} p(\boldsymbol{y}|\boldsymbol{s}')p(\boldsymbol{s}')} \text{ for } \boldsymbol{s}\in\mathcal{S}_\star. \tag{15}$$

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 *model selection metric* $\nu(\boldsymbol{s},\boldsymbol{y})$:

$$\begin{aligned}\nu(\boldsymbol{s},\boldsymbol{y}) &\triangleq \ln p(\boldsymbol{y}|\boldsymbol{s})p(\boldsymbol{s}) && (16)\\ &= \ln p(\boldsymbol{y}|\boldsymbol{s}) + \ln p(\boldsymbol{s}) && (17)\\ &= -(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s}))^H \boldsymbol{\Phi}(\boldsymbol{s})^{-1}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s}))\\ &\qquad - \ln\det(\boldsymbol{\Phi}(\boldsymbol{s})) - M\ln\pi + \sum_{n=0}^{N-1}\ln\lambda_{s_n}. && (18)\end{aligned}$$

The assumption of circular complex Gaussian noise was used for (18); for real-valued Gaussian noise, the first three terms in (18) would simply be halved and $\ln\pi$ replaced by $\ln 2\pi$.

For $Q = 2$, detection of $\boldsymbol{s}\in\{0,1\}^N$ coincides with variable selection. With $Q > 2$, there exist $(Q-1)^K$ possible model vectors $\boldsymbol{s}$ that yield the same selection of a specified subset of K nonzero coefficients.

B. MMSE Coefficient Estimation

For applications in which the identification of the most probable model vector is the primary objective, the sparse coefficients $\boldsymbol{x}$ can be regarded as nuisance parameters. In other applications, however, estimation of $\boldsymbol{x}$ is the primary goal.

The MMSE estimate of $\boldsymbol{x}$ from $\boldsymbol{y}$ is

$$\hat{\boldsymbol{x}}_{\mathsf{mmse}} \triangleq \mathrm{E}\{\boldsymbol{x}|\boldsymbol{y}\} \;=\; \sum_{\boldsymbol{s}\in\mathcal{S}} p(\boldsymbol{s}|\boldsymbol{y})\,\mathrm{E}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\} \tag{19}$$

where from (12) we can obtain (via, e.g., [38, p. 155])

$$\mathrm{E}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\} \;=\; \boldsymbol{\mu}(\boldsymbol{s}) + \boldsymbol{R}(\boldsymbol{s})\boldsymbol{A}^H\boldsymbol{\Phi}(\boldsymbol{s})^{-1}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})). \tag{20}$$

Summing over the dominant models $\mathcal{S}_\star$ yields the approximate MMSE estimate

$$\hat{\boldsymbol{x}}_{\mathsf{ammse}} \triangleq \sum_{\boldsymbol{s}\in\mathcal{S}_\star} p(\boldsymbol{s}|\boldsymbol{y})\,\mathrm{E}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\}. \tag{21}$$

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

$$\begin{aligned}\mathrm{Cov}\{\boldsymbol{x}|\boldsymbol{y}\} \;&\approx\; \sum_{\boldsymbol{s}\in\mathcal{S}_\star} p(\boldsymbol{s}|\boldsymbol{y})\big[\mathrm{Cov}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\} + (\hat{\boldsymbol{x}}_{\mathsf{ammse}} \\ &\qquad - \mathrm{E}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\})(\hat{\boldsymbol{x}}_{\mathsf{ammse}} - \mathrm{E}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\})^H\big] \qquad (22)\\ \mathrm{Cov}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\} \;&=\; \boldsymbol{R}(\boldsymbol{s}) - \boldsymbol{R}(\boldsymbol{s})\boldsymbol{A}^H\boldsymbol{\Phi}(\boldsymbol{s})^{-1}\boldsymbol{A}\boldsymbol{R}(\boldsymbol{s}). \qquad (23)\end{aligned}$$

In fact, the (approximate) estimation error can be written more directly as

$$\begin{aligned}\mathrm{tr}\left(\mathrm{Cov}\{\boldsymbol{x}|\boldsymbol{y}\}\right) \;&\approx\; \sum_{\boldsymbol{s}\in\mathcal{S}_\star} p(\boldsymbol{s}|\boldsymbol{y})\Big[\mathrm{tr}\left(\mathrm{Cov}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\}\right) \\ &\qquad + \big\|\hat{\boldsymbol{x}}_{\mathsf{ammse}} - \mathrm{E}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\}\big\|^2\Big].\end{aligned} \tag{24}$$

The primary challenge in the computation of MMSE estimates is to obtain $p(\boldsymbol{s}|\boldsymbol{y})$ and $\boldsymbol{\Phi}(\boldsymbol{s})^{-1}$ for each $\boldsymbol{s}\in\mathcal{S}_\star$. In the sequel, we propose a fast algorithm to search for the set $\mathcal{S}_\star$ of dominant models that, in addition, generates the values of $\mathrm{E}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\}$ and $\mathrm{Cov}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\}$ for each explored model $\boldsymbol{s}$.

C. The Search for Dominant Models

We now turn our attention to the search for the dominant models $\mathcal{S}_\star$, i.e., those that yield significant posteriors $p(\boldsymbol{s}|\boldsymbol{y})$. Because the denominator of (14) is impractical to compute and the denominator of (15) cannot be computed before $\mathcal{S}_\star$ is known, we search for $\mathcal{S}_\star$ by looking for $\boldsymbol{s}\in\mathcal{S}$ for which $p(\boldsymbol{y}|\boldsymbol{s})p(\boldsymbol{s}) =$

$p(\boldsymbol{s},\boldsymbol{y})$ is significant according to our *a priori* assumptions. Due to the relationship $p(\boldsymbol{s},\boldsymbol{y}) = e^{\nu(\boldsymbol{s},\boldsymbol{y})}$, significant values of $p(\boldsymbol{s},\boldsymbol{y})$ correspond to relatively large values of $\nu(\boldsymbol{s},\boldsymbol{y})$.

To understand what constitutes a "relatively large" value of $\nu(\boldsymbol{s},\boldsymbol{y})$, we derive the *a priori* distribution of the random variable $\nu(\boldsymbol{s},\boldsymbol{y})$ in Appendix A. There we find that

$$\mathrm{E}\{\nu(\boldsymbol{s},\boldsymbol{y})\} = 2M + N(1-\lambda_0)\lambda_0\big(\ln\big[(\tfrac{\sigma_1^2}{\sigma^2}+1)\tfrac{\lambda_0}{\lambda_1}\big]\big)^2 \tag{25}$$

for the case that $\sigma_q^2 = \sigma_1^2$ and $\lambda_q = \lambda_1$ for all $q \neq 0$, where the expectation is taken over both $\boldsymbol{s}$ and $\boldsymbol{y}$. Thus, for a given pair $\{\boldsymbol{s}',\boldsymbol{y}\}$, we can compare $\nu(\boldsymbol{s}',\boldsymbol{y})$ to the mean $\mathrm{E}\{\nu(\boldsymbol{s},\boldsymbol{y})\}$ and standard deviation $\sqrt{\mathrm{var}\{\nu(\boldsymbol{s},\boldsymbol{y})\}}$ in order to get a rough indication of whether $\{\boldsymbol{s}',\boldsymbol{y}\}$ has "significant" probability.

Because brute force evaluation of all Q^N model vectors is impractical for typical values of N, we treat the problem as a non-exhaustive tree search. The models $\{\boldsymbol{s} : \|\boldsymbol{s}\|_0 = p\}$ form the nodes on the p^{th} level of the tree, where $p \in \{0,\dots,N\}$, so that $\boldsymbol{s} = \boldsymbol{0}$ forms the root. We now describe a very general form of tree search. Say that, after the m^{th} stage of tree-search, the search algorithm knows the set $\hat{\mathcal{S}}^{(m)}$ of models currently under consideration, as well as the metrics $\nu(\boldsymbol{s},\boldsymbol{y})$ for all $\boldsymbol{s} \in \hat{\mathcal{S}}^{(m)}$. At the $(m+1)^{th}$ stage, the tree-search i) chooses the subset $\hat{\mathcal{S}}_{\mathsf{e}}^{(m)} \subset \hat{\mathcal{S}}^{(m)}$ of models that will be extended, ii) stores all single-coefficient modifications of the vectors in $\hat{\mathcal{S}}_{\mathsf{e}}^{(m)}$ as the "extended" set $\hat{\mathcal{S}}_{\mathsf{x}}^{(m)}$, iii) computes metrics for all models in $\hat{\mathcal{S}}_{\mathsf{x}}^{(m)}$, and, based on these metrics, iv) prunes the cumulative set $\{\hat{\mathcal{S}}_{\mathsf{x}}^{(m)}, \hat{\mathcal{S}}^{(m)}\}$ to form $\hat{\mathcal{S}}^{(m+1)}$. A stopping criterion decides when to terminate the search; if stopped at the m^{th} stage, the search would return the "significant" models as the set $\hat{\mathcal{S}}_\star = \hat{\mathcal{S}}^{(m)}$. We assume that the search is initialized at the root node, so that $\hat{\mathcal{S}}^{(0)} = \boldsymbol{0}$ with corresponding metric

$$\nu(\boldsymbol{0},\boldsymbol{y}) \;=\; -\tfrac{1}{\sigma^2}\|\boldsymbol{y}\|_2^2 - M\ln\sigma^2 - M\ln\pi + N\ln\lambda_0, \tag{26}$$

which follows from (18) and the fact that $\boldsymbol{\Phi}(\boldsymbol{0}) = \sigma^2\boldsymbol{I}_M$. The details of the extension procedure, pruning procedure, and stopping criterion are algorithm specific (e.g., depth-first, breadth-first, best-first). In the sequel, we will refer to this general approach of non-exhaustive tree-search guided by the Bayesian metric $\nu(\boldsymbol{s},\boldsymbol{y})$ as *Bayesian matching pursuit* (BMP). Our experiments with various types of tree search have led us to recommend the specific search approach detailed in Section IV-E. We note that existing MCMC methods [32], for the over-determined case $M \geq N$, can be interpreted as randomized tree searches.

D. Fast Bayesian Matching Pursuit

Common to all BMP variants (and to 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 model vector $\boldsymbol{s}$. Here we present a fast means of doing so, which we call *fast Bayesian matching pursuit* (FBMP).

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 $\Delta_{n,q'}(\boldsymbol{s},\boldsymbol{y}) \triangleq \nu(\boldsymbol{s}',\boldsymbol{y}) - \nu(\boldsymbol{s},\boldsymbol{y})$. For brevity, we use the abbreviations $\mu_{q',q} \triangleq \mu_{q'} - \mu_q$ and $\sigma^2_{q',q} \triangleq \sigma^2_{q'} - \sigma^2_q$ below. Starting with the property

$$\boldsymbol{\Phi}(\boldsymbol{s}') \;=\; \boldsymbol{\Phi}(\boldsymbol{s}) + \sigma^2_{q',q}\boldsymbol{a}_n\boldsymbol{a}_n^H, \tag{27}$$

the matrix inversion lemma implies

$$\boldsymbol{\Phi}(\boldsymbol{s}')^{-1} \;=\; \boldsymbol{\Phi}(\boldsymbol{s})^{-1} - \beta_{n,q'}\boldsymbol{c}_n\boldsymbol{c}_n^H \tag{28}$$

$$\boldsymbol{c}_n \;\triangleq\; \boldsymbol{\Phi}(\boldsymbol{s})^{-1}\boldsymbol{a}_n \tag{29}$$

$$\beta_{n,q'} \;\triangleq\; \sigma^2_{q',q}\big(1 + \sigma^2_{q',q}\boldsymbol{a}_n^H\boldsymbol{c}_n\big)^{-1}. \tag{30}$$

In Appendix B it is shown that (27)-(30) imply

$$\Delta_{n,q'}(\boldsymbol{s},\boldsymbol{y}) = \begin{cases} \beta_{n,q'}\big|\boldsymbol{c}_n^H(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})) + \mu_{q',q}/\sigma^2_{q',q}\big|^2 & \\ -\,|\mu_{q',q}|^2/\sigma^2_{q',q} + \ln(\beta_{n,q'}/\sigma^2_{q',q}) & \sigma^2_{q',q} \neq 0 \\ +\ln(\lambda_{q'}/\lambda_q) & \\ 2\,\mathrm{Re}\,\{\mu^*_{q',q}\boldsymbol{c}_n^H(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s}))\} & \\ -\,|\mu_{q',q}|^2\boldsymbol{c}_n^H\boldsymbol{a}_n + \ln(\lambda_{q'}/\lambda_q) & \sigma^2_{q',q} = 0. \end{cases} \tag{31}$$

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

Notice that the parameters $\{\boldsymbol{c}_n\}_{n=0}^{N-1}$, which are essential for the metric exploration step (31), require $\mathcal{O}(NM^2)$ operations to compute if (29)-(30) were used with standard matrix multiplication. As described next, the structure of $\boldsymbol{\Phi}(\boldsymbol{s})^{-1}$ can be exploited to make this complexity $\mathcal{O}(NM)$.

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 (28)-(29) imply that

$$\boldsymbol{c}_n \;=\; \Big[\boldsymbol{\Phi}(\boldsymbol{s}^{\mathsf{pre}})^{-1} - \beta^{\mathsf{pre}}_{n^{\mathsf{pre}},q}\boldsymbol{c}^{\mathsf{pre}}_{n^{\mathsf{pre}}}\boldsymbol{c}^{\mathsf{pre}H}_{n^{\mathsf{pre}}}\Big]\boldsymbol{a}_n \tag{32}$$

$$\;=\; \boldsymbol{c}_n^{\mathsf{pre}} - \beta^{\mathsf{pre}}_{n^{\mathsf{pre}},q}\boldsymbol{c}^{\mathsf{pre}}_{n^{\mathsf{pre}}}\boldsymbol{c}^{\mathsf{pre}H}_{n^{\mathsf{pre}}}\boldsymbol{a}_n, \tag{33}$$

$\{\boldsymbol{c}_n\}_{n=0}^{N-1}$ 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 (30) with a complexity of $\mathcal{O}(MN+QN)$. If we recursively update $\boldsymbol{z}(\boldsymbol{s}) \triangleq \boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})$ with $\mathcal{O}(MQ)$ multiplies using

$$\boldsymbol{z}(\boldsymbol{s}) \;=\; \underbrace{\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s}^{\mathsf{pre}})}_{\triangleq\, \boldsymbol{z}(\boldsymbol{s}^{\mathsf{pre}})} - \boldsymbol{a}_{n^{\mathsf{pre}}}\mu_{q,q^{\mathsf{pre}}}, \tag{34}$$

then $\{\Delta_{n,q'}(\boldsymbol{s})\}_{n=0:N-1}^{q'=0:Q-1}$ can be computed via (31) with a complexity of $\mathcal{O}(MN+QN)$. Actually, if $\sigma_q^2 = \sigma_1^2 \;\; \forall q \neq 0$ (as for all the examples given in Section III), then $\beta_{n,q'} = \beta_{n,1} \;\; \forall q' \neq 0$, which leads to a complexity of $\mathcal{O}(MN+QM)$.

Going further, if we define $\boldsymbol{C} \triangleq [\boldsymbol{c}_0, \ldots, \boldsymbol{c}_{N-1}]$ and notice that $\boldsymbol{C} = \boldsymbol{\Phi}(\boldsymbol{s})^{-1}\boldsymbol{A}$, then we can compute the $\boldsymbol{s}$-conditional mean and covariance via

$$\mathrm{E}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\} \;=\; \mu(\boldsymbol{s}) + \boldsymbol{R}(\boldsymbol{s})\boldsymbol{C}^H\boldsymbol{z}(\boldsymbol{s}) \tag{35}$$

$$\mathrm{Cov}\{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}\} \;=\; (\boldsymbol{I}_N - \boldsymbol{R}(\boldsymbol{s})\boldsymbol{C}^H\boldsymbol{A})\boldsymbol{R}(\boldsymbol{s}), \tag{36}$$

using (20), (23), and the fact that $\boldsymbol{\Phi}(\boldsymbol{s})$ is Hermitian. Because $\boldsymbol{R}(\boldsymbol{s})\boldsymbol{C}^H$ has only $\|\boldsymbol{s}\|_0$ nonzero rows and $\boldsymbol{A}\boldsymbol{R}(\boldsymbol{s})$ has only $\|\boldsymbol{s}\|_0$ nonzero columns, (35) and (36) can be computed using only $\mathcal{O}(M\|\boldsymbol{s}\|_0)$ and $\mathcal{O}(M\|\boldsymbol{s}\|_0^2)$ multiplies, respectively.

E. Repeated Greedy Search

In Section IV-C, we proposed a general method to search for the dominant models $\mathcal{S}_\star$ based on tree searches that start with the root hypothesis $\boldsymbol{s}' = \boldsymbol{0}$ and modifies one model component at a time, using the model selection metric $\nu(\boldsymbol{s}',\boldsymbol{y})$ to guide the search. Then, in Section IV-D, we proposed an efficient metric evaluation method that consumes $\mathcal{O}((M+Q)N)$ multiplications to explore all $(Q-1)N$ single-coefficient modifications at each tree node visited by the search, and an additional complexity of $\mathcal{O}(M\|\boldsymbol{s}\|_0)$ and $\mathcal{O}(M\|\boldsymbol{s}\|_0^2)$ at each node $\boldsymbol{s}$ for which the conditional mean and covariance, respectively, are computed. In this section, we propose a particular tree-search that, based on our experience, offers a good tradeoff between performance and complexity.

Our *repeated greedy search* (RGS) procedure starts at the root node $\boldsymbol{s}' = \boldsymbol{0}$ and performs a greedy inflation search (i.e., activating one model component at a time) until a total of P model components have been activated. By "greedy," we mean that the model component activated at each stage is the one leading to the largest metric $\nu(\boldsymbol{s}',\boldsymbol{y})$; de-activation is not allowed. We recommend choosing P slightly larger than the expected number of nonzero coefficients $\mathrm{E}\{\|\boldsymbol{s}\|_0\}$, e.g., so that $\Pr(\|\boldsymbol{s}\|_0 > P)$ is sufficiently small.[3] Note that the procedure described so far is reminiscent of orthogonal matching pursuit (OMP) [5] but different in that the Bayesian metric $\nu(\boldsymbol{s}, \boldsymbol{y})$ is used to guide the activation of new coefficients. If at least one of the P evaluated metrics surpasses some predetermined threshold ν_{thresh}, the RGS algorithm stops. If not, a second greedy inflation search is started (from the root node) and instructed to ignore all previously explored nodes. If at least one of the P evaluated metrics from this second search surpasses the threshold ν_{thresh}, the RGS algorithm stops. If not, a new greedy inflation search is started. The RGS algorithm continues in this manner until ν_{thresh} is surpassed, or until the number of greedy searches reaches an allowed maximum D_{max}. Recall that the threshold ν_{thresh} can be chosen in accordance with the prior on $\nu(\boldsymbol{s}, \boldsymbol{y})$, as discussed in Section IV-C.

The RGS algorithm, using the FBMP recursions from Section IV-D, is detailed in Table I for the simple case that $\sigma_q^2 = \sigma_1^2$ and $\lambda_q = \lambda_1$ for all $q \neq 0$ (which holds true for all the examples given in Section III).

Denoting the number of greedy searches performed by RGS (for a particular realization $\boldsymbol{y}$) by $D \leq D_{\mathsf{max}}$, a total of $DPN(Q-1)$ models are examined with corresponding metrics $\nu(\boldsymbol{s}', \boldsymbol{y})$. From the table, it is straightforward to verify that 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.

F. Exact Odds and Approximate Posteriors

The Bayesian framework provides a report on the confidence of estimates for both the model vector $\boldsymbol{s}$ and the coefficients $\boldsymbol{x}$. In particular, the model selection metric $\nu(\boldsymbol{s}, \boldsymbol{y})$ yields the exact posterior odds in (3). From (14), we can approximate the posterior probability of model $\boldsymbol{s}$ using the renormalized estimate

$$p(\boldsymbol{s}|\boldsymbol{y}) = \frac{\exp\{\nu(\boldsymbol{s}, \boldsymbol{y})\}}{\sum_{\boldsymbol{s}' \in \mathcal{S}} \exp\{\nu(\boldsymbol{s}', \boldsymbol{y})\}} \approx \frac{\exp\{\nu(\boldsymbol{s}, \boldsymbol{y})\}}{\sum_{\boldsymbol{s}' \in \mathcal{S}_\star} \exp\{\nu(\boldsymbol{s}', \boldsymbol{y})\}}, \tag{37}$$

where the approximation in (37) incorporates only the models $\mathcal{S}_\star \subset \mathcal{S}$ that account for the dominant values of $\exp\{\nu(\boldsymbol{s}, \boldsymbol{y})\}$. Likewise, the resulting $\hat{p}(\boldsymbol{x}|\boldsymbol{y})$:

$$\hat{p}(\boldsymbol{x}|\boldsymbol{y}) = \sum_{\boldsymbol{s} \in \mathcal{S}_\star} \hat{p}(\boldsymbol{s}|\boldsymbol{y}) p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{s}), \tag{38}$$

[3] 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 $\Pr(\|\boldsymbol{s}\|_0 > P) = \frac{1}{2}\,\mathrm{erfc}\big(\frac{P-N(1-\lambda_0)}{\sqrt{2N\lambda_0(1-\lambda_0)}}\big)$.

```
$\nu^{\mathsf{root}} = -\frac{1}{\sigma^2}\|\boldsymbol{y}\|_2^2 - M\ln(\sigma^2\pi) + N\ln\lambda_0$;
for $n = 0 : N-1$,
  $\boldsymbol{c}_n^{\mathsf{root}} = \frac{1}{\sigma^2}\boldsymbol{a}_n$;
  $\beta_n^{\mathsf{root}} = \sigma_1^2(1+\sigma_1^2\boldsymbol{a}_n^H\boldsymbol{c}_n^{\mathsf{root}})^{-1}$;
  for $q = 1 : Q-1$,
    $\nu_{n,q}^{\mathsf{root}} = \nu^{\mathsf{root}} + \ln\frac{\beta_n^{\mathsf{root}}}{\sigma_1^2} + \beta_n^{\mathsf{root}}\left|\boldsymbol{c}_n^{\mathsf{root}H}\boldsymbol{y} + \frac{\mu_q}{\sigma_1^2}\right|^2 - \frac{|\mu_q|^2}{\sigma_1^2} + \ln\frac{\lambda_1}{\lambda_0}$;
  end
end
for $d = 1 : D_{\mathsf{max}}$,
  $\boldsymbol{n} = [\,]$;
  $\boldsymbol{q} = [\,]$;
  $\hat{\boldsymbol{s}}^{(d,0)} = \boldsymbol{0}$;
  $\boldsymbol{z} = \boldsymbol{y}$;
  for $n = 0 : N-1$,
    $\boldsymbol{c}_n = \boldsymbol{c}_n^{\mathsf{root}}$;
    $\beta_n = \beta_n^{\mathsf{root}}$;
    for $q = 1 : Q-1$,
      $\nu_{n,q} = \nu_{n,q}^{\mathsf{root}}$;
    end
  end
  for $p = 1 : P$,
    $(n_\star, q_\star) = (n, q)$ indexing the largest element in $\{\nu_{n,q}\}_{n=0:N-1}^{q=1:Q-1}$
            which leads to an as-of-yet unexplored node.
    $\nu^{(d,p)} = \nu_{n_\star,q_\star}$;
    $\hat{\boldsymbol{s}}^{(d,p)} = \hat{\boldsymbol{s}}^{(d,p-1)} + q_\star\boldsymbol{\delta}_{n_\star}$;
    $\boldsymbol{n} \leftarrow [\boldsymbol{n}, n_\star]^T$;
    $\boldsymbol{q} \leftarrow [\boldsymbol{q}, q_\star]^T$;
    $\boldsymbol{z} \leftarrow \boldsymbol{z} - \boldsymbol{a}_{n_\star}\mu_{q_\star}$;
    for $n = 0 : N-1$,
      $\boldsymbol{c}_n \leftarrow \boldsymbol{c}_n - \beta_{n_\star}\boldsymbol{c}_{n_\star}\boldsymbol{c}_{n_\star}^H\boldsymbol{a}_n$;
      $\beta_n = \sigma_1^2(1+\sigma_1^2\boldsymbol{a}_n^H\boldsymbol{c}_n)^{-1}$;
      for $q = 1 : Q-1$,
        $\nu_{n,q} = \nu^{(d,p)} + \ln\frac{\beta_n}{\sigma_1^2} + \beta_n\left|\boldsymbol{c}_n^H\boldsymbol{z} + \frac{\mu_q}{\sigma_1^2}\right|^2 - \frac{|\mu_q|^2}{\sigma_1^2} + \ln\frac{\lambda_1}{\lambda_0}$;
      end
    end
    $\hat{\boldsymbol{x}}^{(d,p)} = \sum_{k=1}^{p}\boldsymbol{\delta}_{[\boldsymbol{n}]_k}\left[\sigma_1^2\boldsymbol{c}_{[\boldsymbol{n}]_k}^H\boldsymbol{z} + \mu_{[\boldsymbol{q}]_k}\right]$;
    $\hat{\boldsymbol{\Sigma}}^{(d,p)} = \sigma_1^2\sum_{k=1}^{p}\sum_{j=1}^{p}\boldsymbol{\delta}_{[\boldsymbol{n}]_k}\Big[\delta_{[\boldsymbol{n}]_k-[\boldsymbol{n}]_j}$
            $- \sigma_1^2\boldsymbol{c}_{[\boldsymbol{n}]_k}^H\boldsymbol{a}_{[\boldsymbol{n}]_j}\Big]\boldsymbol{\delta}_{[\boldsymbol{n}]_j}^T$;
  end
  if $\max\{\nu^{(d,p)}\}_{p=1:P} > \nu_{\mathsf{thresh}}$, then break;
end
```

TABLE I

REPEATED GREEDY SEARCH VIA FAST BAYESIAN MATCHING PURSUIT provides an approximate posterior density that describes the uncertainty in resolving $\boldsymbol{x}$ from the noisy observation. The posterior density 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 $\boldsymbol{A}$.

V. Estimation of Hyperparameters via Approximate ML

When domain knowledge does not precisely specify the hyperparameters,

$$\boldsymbol{\theta} = \{\{\lambda_q\}_{q=0}^{Q-1},\ \{\mu_q\}_{q=0}^{Q-1},\ \sigma^2,\ \{\sigma_q^2\}_{q=0}^{Q-1}\}, \tag{39}$$

one might opt for maximum likelihood (ML) estimates

$$\hat{\boldsymbol{\theta}}_{\mathsf{ml}} = \operatorname*{argmax}_{\boldsymbol{\theta}} p(\boldsymbol{y}|\boldsymbol{\theta}). \tag{40}$$

For $Q = 2$, we now present an approximate ML estimator based on the expectation maximization (EM) iteration [39], [40]. Since $\boldsymbol{s} \in \{0,1\}^N$, we get

$$\boldsymbol{x}|\boldsymbol{s}, \mu_1, \sigma_1^2 \sim \mathcal{CN}\big(\mu_1 \boldsymbol{s}, \sigma_1^2 \mathcal{D}(\boldsymbol{s})\big), \tag{41}$$

where we explicitly condition on parameters μ_1 and σ_1^2 and use $\mathcal{D}(\boldsymbol{s})$ to denote the diagonal matrix created from $\boldsymbol{s}$. The received signal $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$ can then be characterized as

$$\boldsymbol{y}|\boldsymbol{s}, \mu_1, \sigma_1^2, \sigma^2 \sim \mathcal{CN}\big(\mu_1 \boldsymbol{s},\ \sigma_1^2 \boldsymbol{A}\,\mathcal{D}(\boldsymbol{s})\boldsymbol{A}^H + \sigma^2 \boldsymbol{I}_M\big). \tag{42}$$

Rewriting the conditional pdf using the ratio $\alpha \triangleq \frac{\sigma^2}{\sigma_1^2}$ and the matrix $\boldsymbol{A}_{\boldsymbol{s}}$ whose columns are selected from $\boldsymbol{A}$ according to the nonzero entries of $\boldsymbol{s}$, we get

$$\boldsymbol{y}|\boldsymbol{s}, \mu_1, \sigma_1^2, \alpha \sim \mathcal{CN}\big(\mu_1 \boldsymbol{A}\boldsymbol{s}, \sigma_1^2(\boldsymbol{A}_{\boldsymbol{s}}\boldsymbol{A}_{\boldsymbol{s}}^H + \alpha \boldsymbol{I}_M)\big). \tag{43}$$

Finally, recall that the log prior for $\boldsymbol{s}$ has the form

$$\ln p(\boldsymbol{s}|\lambda) = \sum_{n=0}^{N-1} \ln p(s_n|\lambda) \tag{44}$$

$$= \sum_{n=0}^{N-1} \ln\big(\lambda + (1-2\lambda)s_n\big), \tag{45}$$

where $\lambda \triangleq \lambda_0 = \Pr\{s_n = 0\}$. We estimate the parameters $\boldsymbol{\theta} \triangleq [\lambda, \mu_1, \alpha, \sigma_1^2]$ via the EM algorithm, by treating $\boldsymbol{s}$ as the so-called "missing data." In particular, at each M-step, we apply a coordinate ascent scheme, i.e.,

$$\hat{\theta}_k^{(i+1)} = \operatorname*{argmax}_{\theta_k} \sum_{\boldsymbol{s}\in\mathcal{S}} p(\boldsymbol{s}|\boldsymbol{y}, \hat{\boldsymbol{\theta}}^{(i)}) \times \ln p\big(\boldsymbol{y}, \boldsymbol{s}|\theta_k, \{\hat{\theta}_m^{(i+1)}\}_{m<k}, \{\hat{\theta}_m^{(i)}\}_{m>k}\big). \tag{46}$$

Below, we use shorthand notation $\hat{\theta}_k$ for the most recent update of a given parameter, and $K_{\boldsymbol{s}} \triangleq \|\boldsymbol{s}\|_0$.

In practice, the 2^N term summation in (46) is approximated by a sum over the small set of dominant models $\hat{\mathcal{S}}_\star$. For the maximization in (46), we will use the fact that $\ln p(\boldsymbol{y}, \boldsymbol{s}|\boldsymbol{\theta}) = \ln p(\boldsymbol{y}|\boldsymbol{s}, \mu_1, \sigma_1^2, \alpha) + \ln p(\boldsymbol{s}|\lambda)$.

Maximization with respect to λ proceeds according to

$$\hat{\lambda}^{(i+1)} = \operatorname*{argmax}_{\lambda} \sum_{\boldsymbol{s}\in\hat{\mathcal{S}}_\star} p(\boldsymbol{s}|\boldsymbol{y}, \hat{\boldsymbol{\theta}}^{(i)}) \ln p(\boldsymbol{s}|\lambda). \tag{47}$$

Since

$$\begin{aligned} \frac{\partial}{\partial\lambda} \ln p(\boldsymbol{s}|\lambda) &= \sum_{n=0}^{N-1} \frac{1-2s_n}{\lambda + (1-2\lambda)s_n} & (48) \\ &= \frac{K_{\boldsymbol{s}}}{\lambda - 1} + \frac{N - K_{\boldsymbol{s}}}{\lambda}, & (49) \end{aligned}$$

zeroing the partial derivative of (47) w.r.t. λ yields

$$\hat{\lambda}^{(i+1)} = 1 - \frac{1}{N} \sum_{\boldsymbol{s}\in\hat{\mathcal{S}}_\star} p(\boldsymbol{s}|\boldsymbol{y}, \hat{\boldsymbol{\theta}}^{(i)})\ K_{\boldsymbol{s}}. \tag{50}$$

For the M-step update of μ_1, (46) yields

$$\hat{\mu}_1^{(i+1)} = \operatorname*{argmax}_{\mu_1} \sum_{\boldsymbol{s}\in\hat{\mathcal{S}}_\star} p(\boldsymbol{s}|\boldsymbol{y}, \hat{\boldsymbol{\theta}}^{(i)}) \ln p(\boldsymbol{y}|\boldsymbol{s}, \mu_1, \sigma_1^2, \alpha), \tag{51}$$

where, from (43),

$$\begin{aligned} \ln p(\boldsymbol{y}|\boldsymbol{s}, \mu_1, \sigma_1^2, \alpha) &= -\ln\det\left[\sigma_1^2(\boldsymbol{A}_{\boldsymbol{s}}\boldsymbol{A}_{\boldsymbol{s}}^H + \alpha\boldsymbol{I}_M)\right] & (52) \\ &\quad - \sigma_1^{-2}\|\boldsymbol{y} - \mu_1\boldsymbol{A}\boldsymbol{s}\|^2_{(\boldsymbol{A}_{\boldsymbol{s}}\boldsymbol{A}_{\boldsymbol{s}}^H + \alpha\boldsymbol{I}_M)^{-1}}. \end{aligned}$$

Zeroing the partial derivative of the analytic right side of (51) w.r.t. μ_1, we find that

$$\hat{\mu}_1^{(i+1)} = \frac{\sum_{\boldsymbol{s}\in\hat{\mathcal{S}}_\star} p(\boldsymbol{s}|\boldsymbol{y}, \hat{\boldsymbol{\theta}}^{(i)}) \boldsymbol{s}^H\boldsymbol{A}^H(\boldsymbol{A}_{\boldsymbol{s}}\boldsymbol{A}_{\boldsymbol{s}}^H + \alpha\boldsymbol{I}_M)^{-1}\boldsymbol{y}}{\sum_{\boldsymbol{s}\in\hat{\mathcal{S}}_\star} p(\boldsymbol{s}|\boldsymbol{y}, \hat{\boldsymbol{\theta}}^{(i)}) \boldsymbol{s}^H\boldsymbol{A}^H(\boldsymbol{A}_{\boldsymbol{s}}\boldsymbol{A}_{\boldsymbol{s}}^H + \alpha\boldsymbol{I}_M)^{-1}\boldsymbol{A}\boldsymbol{s}}. \tag{53}$$

The update for α is similar in principle, though an approximation is used to simplify the expressions. Recognizing that $\ln\det\left[\widehat{\sigma_1^2}(\boldsymbol{A}_{\boldsymbol{s}}\boldsymbol{A}_{\boldsymbol{s}}^H + \alpha\boldsymbol{I}_M)\right] = \ln\det\left[\boldsymbol{A}_{\boldsymbol{s}}\boldsymbol{A}_{\boldsymbol{s}}^H + \alpha\boldsymbol{I}_M\right] + C$, where C does not depend on α, and noticing that

$$\tfrac{\partial}{\partial\alpha} \ln\det\left[\boldsymbol{A}_{\boldsymbol{s}}\boldsymbol{A}_{\boldsymbol{s}}^H + \alpha\boldsymbol{I}_M\right] = \frac{\tfrac{\partial}{\partial\alpha}\det\left[\boldsymbol{A}_{\boldsymbol{s}}\boldsymbol{A}_{\boldsymbol{s}}^H + \alpha\boldsymbol{I}_M\right]}{\det\left[\boldsymbol{A}_{\boldsymbol{s}}\boldsymbol{A}_{\boldsymbol{s}}^H + \alpha\boldsymbol{I}_M\right]}, \tag{54}$$

we reason that

$$\begin{aligned}\det\left[\boldsymbol{A}_{\boldsymbol{s}}\boldsymbol{A}_{\boldsymbol{s}}^H + \alpha\boldsymbol{I}_M\right] &= \alpha^M \det\left[\alpha^{-1}\boldsymbol{A}_{\boldsymbol{s}}\boldsymbol{A}_{\boldsymbol{s}}^H + \boldsymbol{I}_M\right] &(55)\\ &= \alpha^M \det\left[\alpha^{-1}\boldsymbol{A}_{\boldsymbol{s}}^H\boldsymbol{A}_{\boldsymbol{s}} + \boldsymbol{I}_{K_{\boldsymbol{s}}}\right] &(56)\\ &\approx \alpha^{M-K_{\boldsymbol{s}}} \det\left[\boldsymbol{A}_{\boldsymbol{s}}^H\boldsymbol{A}_{\boldsymbol{s}}\right] &(57)\end{aligned}$$

where in (57) we assume that $\alpha \ll 1$. With this assumption,

$$\tfrac{\partial}{\partial\alpha}\ln\det\left[\boldsymbol{A}_{\boldsymbol{s}}\boldsymbol{A}_{\boldsymbol{s}}^H + \alpha\boldsymbol{I}_M\right] = \frac{M-K_{\boldsymbol{s}}}{\alpha}. \tag{58}$$

We can then use the matrix inversion lemma with the small-α assumption to get

$$\begin{aligned}&(\boldsymbol{A}_{\boldsymbol{s}}\boldsymbol{A}_{\boldsymbol{s}}^H + \alpha\boldsymbol{I}_M)^{-1}\\ &= \frac{1}{\alpha}\left[\boldsymbol{I}_M - \boldsymbol{A}_{\boldsymbol{s}}(\alpha\boldsymbol{I}_{K_{\boldsymbol{s}}} + \boldsymbol{A}_{\boldsymbol{s}}^H\boldsymbol{A}_{\boldsymbol{s}})^{-1}\boldsymbol{A}_{\boldsymbol{s}}^H)\right] &(59)\\ &\approx \frac{1}{\alpha}\left[\boldsymbol{I}_M - \boldsymbol{A}_{\boldsymbol{s}}(\boldsymbol{A}_{\boldsymbol{s}}^H\boldsymbol{A}_{\boldsymbol{s}})^{-1}\boldsymbol{A}_{\boldsymbol{s}}^H\right], &(60)\end{aligned}$$

from which zeroing the partial derivative yields

$$\begin{aligned}\hat{\alpha}^{(i+1)} &= \tfrac{1}{\sigma_1^2}\textstyle\sum_{\boldsymbol{s}\in\hat{\mathcal{S}}_\star} p(\boldsymbol{s}|\boldsymbol{y},\hat{\boldsymbol{\theta}}^{(i)})\tfrac{1}{(M-K_{\boldsymbol{s}})}\\ &\quad\times\|\boldsymbol{y}-\hat{\mu}_1\boldsymbol{A}\boldsymbol{s}\|^2_{\boldsymbol{I}_M-\boldsymbol{A}_{\boldsymbol{s}}(\boldsymbol{A}_{\boldsymbol{s}}^H\boldsymbol{A}_{\boldsymbol{s}})^{-1}\boldsymbol{A}_{\boldsymbol{s}}^H}. &(61)\end{aligned}$$

From the definition of α, (61) gives the required maximization over σ^2 with other parameters fixed.

Finally, maximization w.r.t. σ_1^2 is again similar to the procedure for μ_1. Using the fact that $\ln\det\left[\sigma_1^2(\boldsymbol{A}_{\boldsymbol{s}}\boldsymbol{A}_{\boldsymbol{s}}^H + \hat{\alpha}\boldsymbol{I}_M)\right] = M\ln\sigma_1^2 + C$, where C does not depend on σ_1^2, the corresponding partial-derivative technique yields

$$\widehat{\sigma_1^2}^{(i+1)} = \frac{1}{M}\sum_{\boldsymbol{s}\in\hat{\mathcal{S}}_\star} p(\boldsymbol{s}|\boldsymbol{y},\hat{\boldsymbol{\theta}}^{(i)})\|\boldsymbol{y}-\hat{\mu}_1\boldsymbol{A}\boldsymbol{s}\|^2_{(\boldsymbol{A}_{\boldsymbol{s}}\boldsymbol{A}_{\boldsymbol{s}}^H+\hat{\alpha}\boldsymbol{I}_M)^{-1}}. \tag{62}$$

For computational simplicity, we are motivated to replace (50), (53), (61) and (62) with simpler surrogates. Define $\tilde{\boldsymbol{x}}_{\textsf{ammse}}$ as $\hat{\boldsymbol{x}}_{\textsf{ammse}}$ restricted to the nonzero coefficients, and let `mean` and `var` denote sample mean and variance. The proposed surrogates, requiring $\mathcal{O}(M)$ operations, are

$$\begin{aligned}\hat{\lambda}^{(i+1)} &= 1 - (\|\tilde{\boldsymbol{x}}_{\textsf{ammse}}\|_0/N) &(63)\\ \hat{\mu}_1^{(i+1)} &= \texttt{mean}(\tilde{\boldsymbol{x}}_{\textsf{ammse}}) &(64)\\ \widehat{\sigma^2}^{(i+1)} &= \texttt{var}(\boldsymbol{y}-\boldsymbol{A}\hat{\boldsymbol{x}}_{\textsf{ammse}}) &(65)\\ \widehat{\sigma_1^2}^{(i+1)} &= \texttt{var}(\tilde{\boldsymbol{x}}_{\textsf{ammse}}). &(66)\end{aligned}$$

We choose to terminate the iterations as soon as all parameters change by less than 5% of their values from the previous iteration, or when a maximum number of updates, E_{max}, is reached.

VI. Simulation

Numerical experiments were conducted to investigate the performance of FBMP with approximate maximum likelihood estimation of hyperparameters from the data.

For the first experiment, we chose a "compressible" $\boldsymbol{x}$ that mimics the wavelet coefficients of a natural signal: $x_k = (-1)^k \exp(-\rho k)$ for $k = 0 \ldots N-1$ with $\rho \in (0,1)$. With $N = 512$ and $M = 128$, we drew $\boldsymbol{A}$ from i.i.d. zero-mean Gaussian entries which were subsequently scaled to make each column unit-norm. The noise was also drawn i.i.d. zero-mean Gaussian using a (ρ-dependent) variance that gave $15\,$dB SNR. The reported results represent an average of 2000 independent realizations. We compared FBMP to six publicly available sparse estimation algorithms: OMP [41], StOMP [42], GPSR-Basic [43], SparseBayes [29], BCS [31], and a variational-Bayes implementation of BCS [44]. The algorithmic parameters were chosen in accordance with suggestions provided by the authors and, when applicable, adjusted to yield improved performance. For SparseBayes, the true inverse noise variance was provided, and it was not re-estimated during execution as this led to degraded performance. Similarly, OMP and BCS were provided the true noise variance. StOMP was tested using both the "False Alarm Control" and "False Discovery Control" thresholding strategies; since the latter appeared less reliable for high values of ρ, we present results only for the former. 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 biased and debiased estimates. The FBMP hyperparameters were initialized at $\lambda_1 = 0.01$, $\mu_1 = 0$, $\sigma^2 = 0.05$, $\sigma_1^2 = 2$, and the surrogate 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

$$\text{NMSE (dB)} = 10\log_{10}\left(\frac{1}{T}\sum_{i=1}^{T}\frac{\|\hat{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}\|_2^2}{\|\boldsymbol{x}^{(i)}\|_2^2}\right), \tag{67}$$

where T is the number of random trials and superscript $^{(i)}$ denotes the trial number. From the figure, it can be seen that the proposed FBMP with EM hyperparameter estimation provides NMSE improvements of up to $2\,$dB over OMP and GPSR, and up to 6-$8\,$dB over the other algorithms. The improvements are due, in part, to model averaging for computation of $\hat{\boldsymbol{x}}_{\mathsf{ammse}}$ and the incorporation of noise power when computing the conditional MMSE estimate (20). The good performance of GPSR can be exhibited to the choice of signal; the sequence $\boldsymbol{x}$, while mismatched to the Gaussian mixture prior, is a typical draw

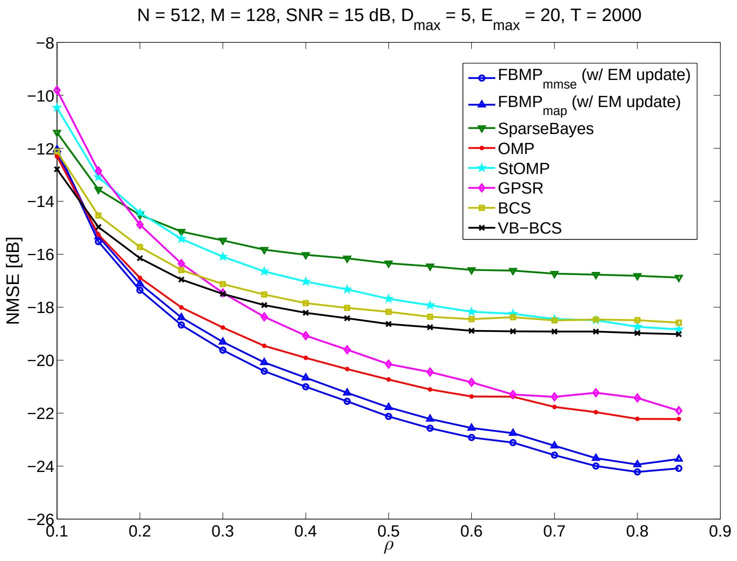


Fig. 1. Normalized mean squared error versus ρ.

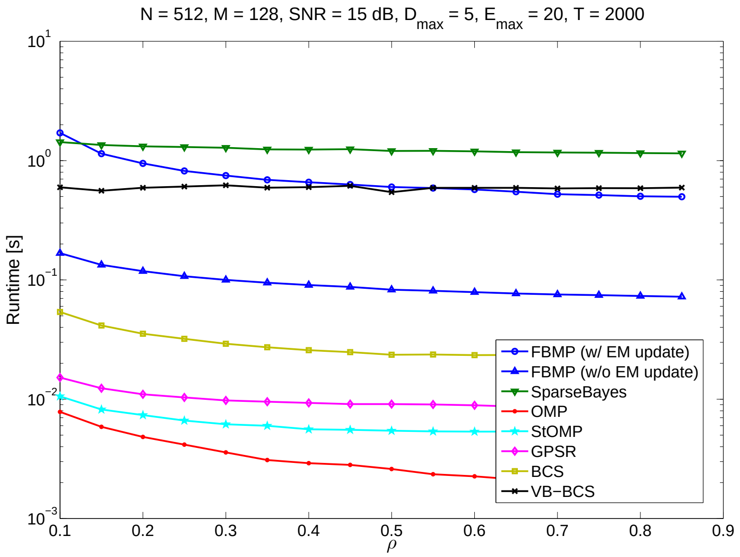


Fig. 2. Runtime versus ρ.

from a Laplace density, and is therefore well matched to the MAP estimator (5) for $p = 1$, to which GPSR seeks a solution.

Figure 2 displays average runtimes for the same experiment. We note that the runtimes for FBMP are reported with and without generalized-EM iterations, whereas the runtimes for the other algorithms

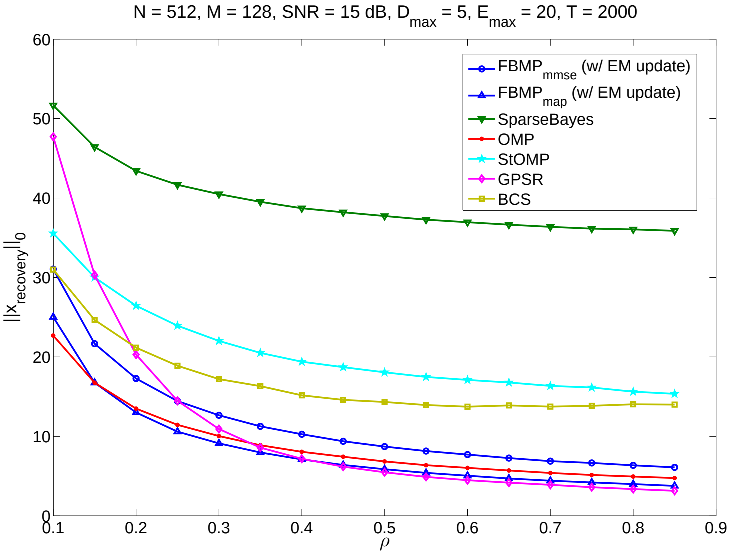


Fig. 3. Solution sparsity versus ρ.

do *not* reflect the repeated executions required to optimize their adjustable parameters. FBMP (without generalized-EM iterations) is significantly faster than SparseBayes and VB-BCS but significantly slower than GPSR, OMP, and StOMP. In exchange for speed, FBMP returns not only a MAP model estimate $\hat{\boldsymbol{s}}_\star$, but also a list of other high-probability models $\hat{\mathcal{S}}_\star$ along with their posterior probabilities; the other six approaches considered return only a single model estimate. Thus, FBMP is able to give a more complete interpretation of the data in the face of ambiguity arising from correlation in $\boldsymbol{A}$ or from measurement noise.

Fig. 3 shows average sparsity of solutions. We observe that, for this “compressible” signal and Gaussian regressor matrix, the coefficient estimates returned by FBMP are among the sparsest.

In a second experiment, to illustrate the behavior of the greedy tree-search, we adopt a figure format used by George and McCulloch [45] to report MCMC results. To allow exhaustive evaluation of all candidate models, we set $N = 26$ and $M = 7$. Signals were constructed using the Gaussian mixture model of Section III with $Q = 2$, $\lambda_1 = 0.04$, $\mu_1 = 0$, $\sigma_1^2 = 1$, and with noise power adjusted to yield $10\,$dB SNR. For illustration, FBMP was provided the true hyperparameters and used without generalized EM. Shown in Fig. VI is a rank-ordered list of the posterior probabilities $p(\boldsymbol{s}|\boldsymbol{y})$. To the right of the dashed line are the probabilities for the 59 models $\boldsymbol{s}$ selected by the search, while to the left of the dashed line are the probabilities for models not visited (truncated to show only the top 500, out of $2^{26} - 59$,

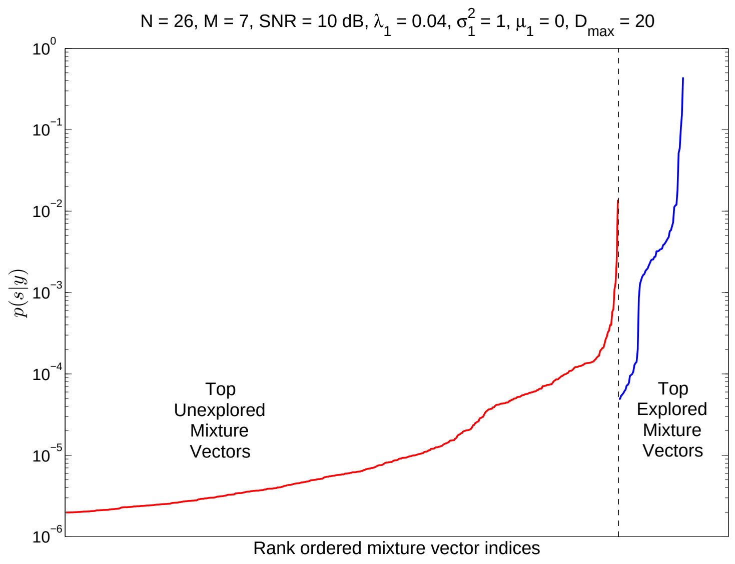


Fig. 4. Rank ordered posterior probabilities, on a logarithmic scale, of the models $\boldsymbol{s}$ visited by the search heuristic (right of dashed vertical line) and the top 500 models not visited (left of dashed vertical line).

models). While the figure displays only one realization, it is typical of our numerical experience. The figure shows that, i) there exist multiple models with high probability, highlighting the inadequacy of reporting on the MAP model, and, ii) the low-complexity search heuristic is effective in visiting the high probability models.

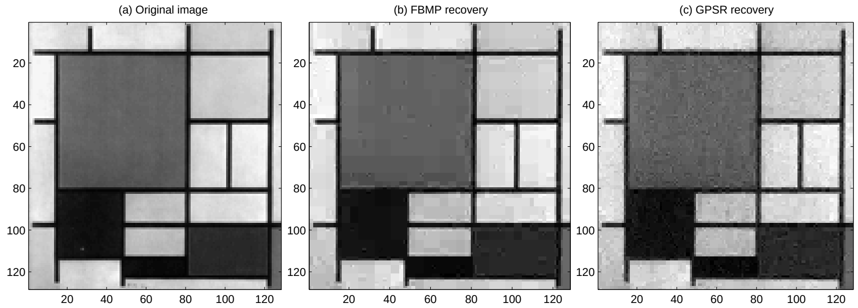
In a third experiment, an exhaustive evaluation similar to the previous one was repeated 204 times (see Table II for details), and each time both FBMP and the Larsson-Selén algorithm (LSA) [36] were used to compute estimates of $\boldsymbol{x}$. The resulting average MSE performance is reported in Table II, along with the average "distance to MMSE" (D2MMSE) $\|\hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}_{\mathsf{mmse}}\|_2^2$, where $\hat{\boldsymbol{x}}$ denotes the estimate returned by the (FBMP or LSA) algorithm and $\hat{\boldsymbol{x}}_{\mathsf{mmse}}$ denotes the exact MMSE estimate. It can be seen that FBMP clearly outperforms LSA both in terms of MSE and D2MMSE.

In a fourth experiment, we carried out a "multiscale-CS" recovery of the popular "Mondrian" test image. Under the multiscale-CS framework, random Gaussian ensemble measurements were acquired from the 3 finest-scale Haar wavelet coefficients of the 128×128 image. In all, 4877 measurements were acquired from the 16384 unknowns, with different scales being undersampled by different factors. For comparison, recoveries were obtained using GPSR as well. The results of this experiment are shown in Fig. 5, with NMSEs and runtimes reported in the caption. The reported runtimes correspond to the time

Algorithm	MSE [dB]	D2MMSE [dB]
FBMP	-19.7	-24.1
LSA	-8.8	-9.1

TABLE II

Performance for Bernoulli/iid-Gaussian signal with $N = 24$, $M = 8$, $Q = 2$, $\lambda_1 = 0.04$, $\mu_1 = 0$, $\sigma_1^2 = 1$, and SNR= 15 dB, averaged over 204 trials. See text for definition of D2MMSE.


(a) Original image

(b) FBMP recovery

(c) GPSR recovery

Fig. 5. Multiscale CS recovery. a) Original 128×128 image; b) FBMP recovery: NMSE $= -16.80$ dB, 8.85% of coefficients active, 38 minutes runtime; c) GPSR recovery: NMSE $= -13.66$ dB, 24.02% of coefficients active, 2.7 minutes runtime.

taken after the adjustable algorithmic parameters (e.g., τ for GPSR) were optimized. Relative to GPSR, the estimate returned by FBMP was more sparse and had lower NMSE, but took longer to generate. We note that these results are consistent with those from the other experiments.

VII. Discussion

A. *Fast Algorithms: Related Works*

A Gaussian mixture model similar to that in Section III was likewise adopted by Larsson and Selén [36], who, for $Q = 2$, also constructed the MMSE estimate in the manner of (21) but with an $\mathcal{S}_\star$ that contains exactly one model vector $\boldsymbol{s}$ for each Hamming weight 0 to N. They proposed to find these $\boldsymbol{s}$ via greedy deflation, i.e., starting with an all-active model configuration and recursively deactivating one component at a time. Thus, the $D = 1$ version of the BMP heuristic from Section IV-C recalls the heuristic of [36], but in reverse. Note, however, that the *fast* $D = 1$ BMP presented in Section IV-D has a complexity of only $\mathcal{O}(NMP)$, in comparison to $\mathcal{O}(N^3M^2)$ for the technique in [36]. Given the typically large values of N encountered in practice, the complexity of FBMP can be several orders of magnitude lower than that of [36]. Complexity aside, Table II suggests that the greedy deflation approach of [36] is much less effective at finding the models vectors with high posterior probability, leading to estimates that, relative to FBMP, have higher MSE and are further from the exact MMSE estimate.

For $Q = 2$, a Gaussian mixture model has been widely adopted for the Bayesian variable selection problem. (See, e.g., [1] for a survey and references.) The published approaches vary in prior specification, posterior calculation, and MCMC method (such as Gibbs sampler or Metropolis-Hastings). George and McCulloch [45] 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 $p(\boldsymbol{s}'|\boldsymbol{y})$ from $p(\boldsymbol{s}|\boldsymbol{y})$, where $\boldsymbol{s}'$ and $\boldsymbol{s}$ differ in only one element. Smith and Kohn [32] use the point mass null (i.e., $\mu_0 = \sigma_0^2 = 0$) and the simplifying Zellner-g prior to achieve a fast update requiring $\mathcal{O}(K_{\boldsymbol{s}}^2)$ operations, for $K_{\boldsymbol{s}} \triangleq \|\boldsymbol{s}\|_0$. Approximately MN iterations of the Gibbs sampler are suggested, yielding a total complexity of $\mathcal{O}(MN^2K_{\boldsymbol{s}}^2)$.

B. Bayesian Model Averaging

The Bayesian framework provides a report on the confidence of estimates for both the model $\boldsymbol{s}$ and the coefficients $\boldsymbol{x}$. In contrast, confidence labels are absent in most of the compressive sampling literature. Exceptions are found in [16], [31], which use an (approximate) MAP estimate $\hat{\boldsymbol{s}}_\star$ for variable selection and report the Gaussian error covariance for the linear problem conditioned on $\hat{\boldsymbol{s}}_\star$ being the true model. As noted by Tibshirani [16], such a measure of posterior uncertainty has dubious value, because "a difficulty with this formula is that it gives an estimated variance of 0 for predictors with" $[\hat{\boldsymbol{s}}_\star]_n = 0$. In fact, in our simulations, we observe that $\hat{\boldsymbol{s}}_\star$ is often not equal to the true $\boldsymbol{s}$. Indeed, in order for $\hat{\boldsymbol{s}}_\star$ to equal true $\boldsymbol{s}$ with high probability, for fixed sparsity $\|\boldsymbol{s}\|_0/N$, the SNR grows unbounded with N [46]. In this light, we expect certain advantages for algorithms that consider the active signal coefficients as implicitly uncertain.

As a caveat, we emphasize that our greedy FBMP search returns only $\hat{\mathcal{S}}_\star$, an estimate of the dominant subset $\mathcal{S}_\star$, along with the values of $\nu(\boldsymbol{s}, \boldsymbol{y})$ for $\boldsymbol{s} \in \hat{\mathcal{S}}_\star$. Thus, while the values $\nu(\boldsymbol{s}, \boldsymbol{y})$ returned by FBMP can be used to compute exact ratios between the posterior probabilities of the model vectors in $\hat{\mathcal{S}}_\star$, the true posteriors of these configurations (as approximated by (37) with $\hat{\mathcal{S}}_\star$ in place of $\mathcal{S}_\star$) will only be accurate when $\hat{\mathcal{S}}_\star$ indeed contains $\mathcal{S}_\star$. In simulation, we observe that the proposed greedy FBMP search reliably discovers $\mathcal{S}_\star$ when $\frac{\lambda_1 N}{M} \leq -1/\log(\lambda_1)$.

C. Empirical Bayes

Empirical Bayes (EB) approaches have been used in related work to estimate hyperparameters from the data under signal models similar to the zero-mean binary prior given in Section III. George and Foster [47] adopted maximum marginal likelihood as in (40) for estimating parameters $\{\lambda_1, \sigma_1^2\}$ en route to a MAP model selection using the Zellner g-prior. A forward greedy search for the EB $\hat{s}_\star$ was considered. For $\boldsymbol{A} = \boldsymbol{I}$, Johnstone and Silverman [48] used maximum marginal likelihood for λ_1 and established the asymptotic risk of adaptive thresholding rules. Larsson and Selén [36] likewise estimated hyperparameters from the data; for $M \geq N$, *ad hoc* estimates were computed from the full-model least-squares estimate using higher-order moments.

D. Informative Priors

In our proposed approach, we have sought to incorporate physically meaningful prior knowledge when application-specific insight is available. Further, by use of the generalized-EM algorithm, we have provided a means for trade-off of complexity versus prior knowledge, i.e., ML estimates of hyperparameters may be iteratively estimated from the data. In contrast, the aim in statistical literature is to be agnostic by adopting noninformative priors or hyperpriors.

VIII. Conclusion

In this paper, we proposed an algorithm for joint model selection and sparse coefficient estimation, which we call fast Bayesian matching pursuit (FBMP). We adopted a Bayesian approach in which a set of likely model configurations is reported, along with exact ratios of model posterior probabilities. These relative probabilities serve to reveal potential ambiguity among multiple candidate solutions that are ambiguous due to observation noise or correlation among columns in the regressor matrix. The explicit management of uncertainty is essential for applications in which the estimated model vector, $\hat{\boldsymbol{s}}$, and estimated coefficients, $\hat{\boldsymbol{x}}$, are not final products, but are instead statistics for use in making inference from the noisy observations, $\boldsymbol{y}$. The proposed search for high probability models and computation of their posteriors is fast in that the computational complexity is $\mathcal{O}(MNK)$, with M observations, N coefficients, and K nonzero coefficients. For a modest increase in complexity, the proposed generalized-EM refinement combines with FBMP to provide an empirical Bayes method for estimating hyperparameters from the data. Existing approaches using tree searches or MCMC methods require at least $\mathcal{O}(MN^2K^2)$ computation.

In forthcoming work we will report on a large-scale version of FBMP that reduces the memory required in recursively computing posterior probabilities, and we will give a bound on the probability that a subset of coefficients is absent from the MAP model estimate.

REFERENCES

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

[2] C. Volinsky, "Bayesian model averaging homepage," `http://www.research.att.com/˜volinsky/bma.html`.

[3] J. Högbom, "Aperture synthesis with a non-regular distribution of interferometer baselines," *Astrophys. J. Suppl. Ser*, vol. 15, pp. 417–426, 1974.

[4] P. J. Huber, "Projection pursuit," *The Annals of Statistics*, vol. 13, pp. 435–475, 1985.

[5] 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.

[6] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Harmon. Anal.*, 2008. doi:10.1016/j.acha.2008.07.002.

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

[8] 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.

[9] C. L. Lawson, *Contributions to the theory of linear least maximum approximations*. PhD thesis, UCLA, 1961.

[10] H. Lee, D. Sullivan, and T. Huang, "Improvement of discrete band-limited signal extrapolation by iterative subspace modification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1569 – 1572, 1987.

[11] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, pp. 600 – 616, Mar. 1997.

[12] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008*, (Las Vegas, NV), pp. 3869 – 3872, April 2008.

[13] H. L. Taylor, S. C. Banks, and J. F. McCoy, "Deconvolution with the ℓ_1 norm," *Geophysics*, vol. 44, pp. 39–52, 1979.

[14] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, pp. 55–67, 1970.

[15] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[16] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Statist. Soc. B*, vol. 58, no. 1, pp. 267 – 288, 1996.

[17] C. Bouman and K. Sauer, "A generalized Gaussian image model for edge-preserving MAP estimation," *IEEE Trans. Image Process.*, vol. 2, pp. 296–310, Mar. 1993.

[18] A. H. Delaney and Y. Bresler, "A fast and accurate Fourier algorithm for iterative parallel-beam tomography," *IEEE Trans. Image Process.*, vol. 5, pp. 740–753, May 1996.

[19] M. Çetin and W. C. Karl, "Feature-enhanced synthetic aperture radar image formation based on nonquadratic regularization," *IEEE Trans. Image Process.*, vol. 10, pp. 623–631, Apr. 2001.

[20] S. Levy and P. K. Fullagar, "Reconstruction of a sparse spike train from a portion of its spectrum and application to high-resolution deconvolution," *Geophysics*, vol. 46, no. 9, pp. 1235–1243, 1981.

[21] B. D. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Trans. Signal Processing*, vol. 47, pp. 187–200, Jan. 1999.

[22] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Information Theory*, vol. 52, pp. 489–509, Feb. 2006.

[23] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D*, vol. 60, pp. 259 – 268, 1992.

[24] 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.

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

[26] T. Tao, "Open question: deterministic UUP matrices," July 2007. http://terrytao.wordpress.com/2007/07/02/open-question-deterministic-uup-matrices/.

[27] Y. T. Lo, "A mathematical theory of antenna arrays with randomly spaced elements," *IEEE Trans. Antennas and Propagation*, vol. 12, pp. 257–268, 1964.

[28] G. J. Marseille, R. de Beer, M. Fuderer, A. F. Mehlkopf, and D. van Ormondt, "Nonuniform phase-encode distributions for MRI scan time reduction," *J. Magn. Reson.*, vol. 111, pp. 70–75, 1996.

[29] 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`).

[30] D. Wipf and B. Rao, "Sparse Bayesian learning for basis selectionn," *IEEE Trans. Signal Process.*, vol. 52, pp. 2153 – 2164, Aug. 2004.

[31] 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`).

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

[33] C. Andrieu, A. Doucet, and C. P. Robert, "Computational advances for and from Bayesian analysis," *Statist. Sci.*, vol. 19, no. 1, pp. 118 – 127, 2004.

[34] P. J. Wolfe, S. J. Godsill, and W.-J. Ng, "Bayesian variable selection and regularization for time-frequency surface estimation," *J. R. Statist. Soc. B*, vol. 66, pp. 575–589.

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

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

[37] S. Som, L. C. Potter, R. Ahmad, D. S. Vikram, and P. Kuppusamy, "EPR oximetry in three spatial dimensions using sparse spin distribution," *Journal of Magnetic Resonance*, vol. 193, Aug. 2008.

[38] H. V. Poor, *An Introduction to Signal Detection and Estimation*. Springer, 2 ed., 1994.

[39] 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.

[40] R. Neal and G. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models* (M. I. Jordan, ed.), pp. 355–368, MIT Press, 1999.

[41] 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/`).

[42] 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/).

[43] 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/).

[44] C. M. Bishop and M. E. Tipping, "Variational relevance vector machines," in *Proceedings of the 16th Conf. on Uncertainty in Artificial Intelligence* (C. Boutilier and M. Goldszmidt, eds.), pp. 46–53, Morgan Kaufmann, 1999. (software available at http://people.ee.duke.edu/˜lihan/cs/).

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

[46] G. Reeves, "Sparse signal sampling using noisy linear projections," Master's thesis, EECS Department, University of California, Berkeley, 2008.

[47] E. I. George and D. P. Foster, "Calibration and empirical Bayes variable selection," *Biometrika*, vol. 87, no. 4, pp. 731 – 747, 2000.

[48] I. M. Johnstone and B. W. Silverman, "Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences," *Ann. Stat.*, vol. 32, no. 4, pp. 1594 – 1649, 2004.

Appendix A
Mean and Variance of $\nu(\boldsymbol{s},\boldsymbol{y})$

In this appendix, we derive the mean and variance of $\nu(\boldsymbol{s},\boldsymbol{y})$. According to our priors, if $\boldsymbol{s}$ is the model vector from which $\boldsymbol{y}$ is generated, then $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}+\boldsymbol{w}$ for $\boldsymbol{x}|\boldsymbol{s} \sim \mathcal{CN}(\boldsymbol{\mu}(\boldsymbol{s}), \boldsymbol{R}(\boldsymbol{s}))$ and $\boldsymbol{w} \sim \mathcal{CN}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_M)$. This implies that $\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s}) \,|\, \boldsymbol{s} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{\Phi}(\boldsymbol{s}))$, so that

$$\big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)^H \boldsymbol{\Phi}(\boldsymbol{s})^{-1}\big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big) \;\sim\; \chi^2_M, \tag{68}$$

i.e., a chi-squared random variable with M degrees of freedom. Say that $\boldsymbol{A}_{\boldsymbol{s}}$ denotes the matrix constructed from the active columns of $\boldsymbol{A}$. Then, if $\sigma_q^2 = \sigma_1^2$ for all $q \neq 0$ (as in all the examples given in Section III) and if $\boldsymbol{A}_{\boldsymbol{s}}$ is orthonormal,

$$\ln\det(\boldsymbol{\Phi}(\boldsymbol{s})) \;=\; \ln\big((\sigma_1^2+\sigma^2)^{\|\boldsymbol{s}\|_0}\sigma^{2(M-\|\boldsymbol{s}\|_0)}\big) \tag{69}$$

$$= \|\boldsymbol{s}\|_0 \ln(\tfrac{\sigma_1^2}{\sigma^2}+1) + M\ln\sigma^2, \tag{70}$$

where $\|\boldsymbol{s}\|_0 \sim \text{Binomial}(N, 1-\lambda_0)$. The orthonormal assumption on $\boldsymbol{A}_{\boldsymbol{s}}$ is reasonable because the columns of $\boldsymbol{A}$ were assumed unit-norm in Section III and, for the class of problems that guarantee good sparse estimates, *any* collection of $\|\boldsymbol{s}\|_0$ columns from $\boldsymbol{A}$ will be approximately orthogonal. (Recall the restricted isometry property [8].) Finally, if we assume that $\lambda_q = \lambda_1$ for all $q \neq 0$ (as in all the examples given in Section III), then

$$\sum_{n=0}^{N-1} \ln \lambda_{s_n} = \|\boldsymbol{s}\|_0 \ln \lambda_1 + (N - \|\boldsymbol{s}\|_0) \ln \lambda_0 \tag{71}$$

$$= N \ln \lambda_0 - \|\boldsymbol{s}\|_0 \ln \tfrac{\lambda_0}{\lambda_1}. \tag{72}$$

Using the facts that the mean and variance of a χ^2_M random variable are M and $2M$, respectively, and the mean and variance of $\|\boldsymbol{s}\|_0$ are $N(1-\lambda_0)$ and $N(1-\lambda_0)\lambda_0$, respectively, we obtain (25).

Appendix B
Derivation of (31)

In this appendix, we establish (31) using (27)-(30). 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 (73)\\
&= \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 (74)\\
&= \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 (75)
\end{aligned}$$

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

$$\begin{aligned}
&\big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s}')\big)^H \boldsymbol{\Phi}(\boldsymbol{s}')^{-1}\big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s}')\big)\\
&= \big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)^H \boldsymbol{\Phi}(\boldsymbol{s})^{-1}\big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)\\
&\quad - 2\,\mathrm{Re}\big\{\mu_{q',q}^*\boldsymbol{c}_n^H\big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)\big\} + |\mu_{q',q}|^2\boldsymbol{c}_n^H\boldsymbol{a}_n. \qquad (76)
\end{aligned}$$

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

$$\begin{aligned}
&\big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s}')\big)^H \boldsymbol{\Phi}(\boldsymbol{s}')^{-1} \big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s}')\big) \\
&= \big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)^H \boldsymbol{\Phi}(\boldsymbol{s})^{-1} \big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big) \\
&\quad - \beta_{n,q'} \bigg[\big| \boldsymbol{c}_n^H \big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big) \big|^2 \\
&\quad - 2 \operatorname{Re}\Big\{ \boldsymbol{c}_n^H \big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big) \frac{\mu^*_{q',q}}{\sigma^2_{q',q}} \Big\} + \boldsymbol{c}_n^H \boldsymbol{a}_n \frac{|\mu_{q',q}|^2}{\sigma^2_{q',q}} \bigg] \qquad (77) \\
&= \big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)^H \boldsymbol{\Phi}(\boldsymbol{s})^{-1} \big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big) \\
&\quad - \beta_{n,q'} \Big| \boldsymbol{c}_n^H \big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big) + \frac{\mu_{q',q}}{\sigma^2_{q',q}} \Big|^2 \\
&\quad + \beta_{n,q'} \frac{|\mu_{q',q}|^2}{\sigma^4_{q',q}} \Big[1 + \sigma^2_{q',q} \boldsymbol{c}_n^H \boldsymbol{a}_n \Big] \qquad (78) \\
&= \big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)^H \boldsymbol{\Phi}(\boldsymbol{s})^{-1} \big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big) \\
&\quad - \beta_{n,q'} \Big| \boldsymbol{c}_n^H \big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big) + \frac{\mu_{q',q}}{\sigma^2_{q',q}} \Big|^2 + \frac{|\mu_{q',q}|^2}{\sigma^2_{q',q}}. \qquad (79)
\end{aligned}$$

Together, (76) and (79) yield (80).

$$\begin{aligned}
&\big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s}')\big)^H \boldsymbol{\Phi}(\boldsymbol{s}')^{-1} \big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s}')\big) \\
&= \begin{cases}
\big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)^H \boldsymbol{\Phi}(\boldsymbol{s})^{-1} \big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big) & \\
- \beta_{n,q'} \big| \boldsymbol{c}_n^H \big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big) + \mu_{q',q}/\sigma^2_{q',q} \big|^2 & \sigma^2_{q',q} \neq 0 \\
+ |\mu_{q',q}|^2/\sigma^2_{q',q} & \\
\big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big)^H \boldsymbol{\Phi}(\boldsymbol{s})^{-1} \big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big) & \\
- 2 \operatorname{Re}\big\{ \mu^*_{q',q} \boldsymbol{c}_n^H \big(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{s})\big) \big\} & \sigma^2_{q',q} = 0. \\
+ |\mu_{q',q}|^2 \boldsymbol{c}_n^H \boldsymbol{a}_n &
\end{cases} . \qquad (80)
\end{aligned}$$

Equation (27) then implies that

$$\begin{aligned}
\ln \det(\boldsymbol{\Phi}(\boldsymbol{s}')) &= \ln \det\big(\boldsymbol{\Phi}(\boldsymbol{s}) + \sigma^2_{q',q} \boldsymbol{a}_n \boldsymbol{a}_n^H\big) \qquad (81) \\
&= \ln \Big[\big(1 + \sigma^2_{q',q} \boldsymbol{a}_n^H \boldsymbol{\Phi}(\boldsymbol{s})^{-1} \boldsymbol{a}_n\big) \det\big(\boldsymbol{\Phi}(\boldsymbol{s})\big) \Big] \\
&= \ln \det(\boldsymbol{\Phi}(\boldsymbol{s})) - \ln(\beta_{n,q'}/\sigma^2_{q',q}) \qquad (82) \\
\ln p(\boldsymbol{s}') &= \ln p(\boldsymbol{s}) + \ln(\lambda_{q'}/\lambda_q), \qquad (83)
\end{aligned}$$

which, in conjunction with (18) and (80), yield (31).