An Empirical Bayesian Strategy for Solving the Simultaneous Sparse Approximation Problem

David P. Wipf and Bhaskar D. Rao, *Fellow, IEEE*

***Abstract*—Given a large overcomplete dictionary of basis vectors, the goal is to simultaneously represent $L > 1$ signal vectors using coefficient expansions marked by a common sparsity profile. This generalizes the standard sparse representation problem to the case where multiple responses exist that were putatively generated by the same small subset of features. Ideally, the associated sparse generating weights should be recovered, which can have physical significance in many applications (e.g., source localization). The generic solution to this problem is intractable and, therefore, approximate procedures are sought. Based on the concept of automatic relevance determination, this paper uses an empirical Bayesian prior to estimate a convenient posterior distribution over candidate basis vectors. This particular approximation enforces a common sparsity profile and consistently places its prominent posterior mass on the appropriate region of weight-space necessary for simultaneous sparse recovery. The resultant algorithm is then compared with multiple response extensions of matching pursuit, basis pursuit, FOCUSS, and Jeffreys prior-based Bayesian methods, finding that it often outperforms the others. Additional motivation for this particular choice of cost function is also provided, including the analysis of global and local minima and a variational derivation that highlights the similarities and differences between the proposed algorithm and previous approaches.**

***Index Terms*—Automatic relevance determination, empirical Bayes, multiple response models, simultaneous sparse approximation, sparse Bayesian learning, variable selection.**

I. Introduction

SUPPOSE we are presented with some target signal and a feature set that are linked by a generative model of the form

$$\boldsymbol{t} = \Phi \boldsymbol{w} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{t} \in \mathbb{R}^N$ is the vector of responses or targets, $\Phi \in \mathbb{R}^{N \times M}$ is a dictionary of M features (also referred to as basis vectors) that have been observed or determined by experimental design, $\boldsymbol{w}$ is a vector of unknown weights, and $\boldsymbol{\epsilon}$ is noise.[1] Moreover, assume we have some prior belief that $\boldsymbol{t}$ has been generated by a sparse coefficient expansion, i.e., most of the elements in $\boldsymbol{w}$ are equal to zero. The goal is to estimate $\boldsymbol{w}$ given $\boldsymbol{t}$ and Φ. Of particular interest is the case where the number of candidate basis vectors M significantly exceeds the signal dimension N. While this scenario is extremely relevant in numerous situations, the added redundancy significantly compounds the problem of recovering the sparse, generating weights.

Now suppose that multiple response vectors (e.g., $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots$) have been collected from different locations or under different conditions (e.g., spatial, temporal, etc.) characterized by different underlying parameter vectors $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots$, but with an equivalent design matrix Φ. Assume also that while the weight amplitudes may be changing, the indexes of the nonzero weights, or the sparsity profile, does not. In other words, we are assuming that a common subset of basis vectors are relevant in generating each response. Such a situation arises in many diverse application domains such as neuroelectromagnetic imaging [18], [24], [36]–[38], communications [7], [14], signal processing [25], [46], and source localization [31]. Other examples that directly comply with this formulation include compressed sensing [13], [52] and landmark point selection for sparse manifold learning [45]. In all of these applications, it would be valuable to have a principled approach for merging the information contained in each response so that we may uncover the underlying sparsity profile. This, in turn, provides a useful mechanism for solving what is otherwise an ill-posed inverse problem.

Given L models structurally equivalent to (1), the multiple response model with which we are concerned becomes

$$T = \Phi W + \mathcal{E} \tag{2}$$

where $T = [\boldsymbol{t}_{\cdot 1}, \ldots, \boldsymbol{t}_{\cdot L}]$, and $W = [\boldsymbol{w}_{\cdot 1}, \ldots, \boldsymbol{w}_{\cdot L}]$. Note that, to facilitate later analysis, we adopt the notation that $\boldsymbol{x}_{\cdot j}$ represents the jth column of X while $\boldsymbol{x}_{i\cdot}$ represents the ith row of X. Likewise, x_{ij} refers the ith element in the jth column of X. In the statistics literature, (2) represents a multiple response model [26] or multiple output model [23]. In accordance with our prior belief that a basis vector (and its corresponding weight) that is utilized in creating one response will likely be used by another, we assume that the weight matrix W has a minimal number of nonzero *rows*. The inference goal then becomes the simultaneous approximation of each weight vector $\boldsymbol{w}_{\cdot j}$ under the assumption of a common sparsity profile.

A. Problem Statement

To simplify matters, it is useful to introduce the notation

$$d(W) \triangleq \sum_{i=1}^{M} \mathcal{I}[\|\boldsymbol{w}_{i\cdot}\| > 0] \tag{3}$$

Manuscript received February 18, 2006; revised October 5, 2006. This work was supported in part by DiMI under Grant #22-8376, in part by Nissan, and in part by an NSF IGERT predoctoral fellowship while D. P. Wipf was with the University of California, San Diego. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tulay Adali.

D. P. Wipf is with the Biomagnetic Imaging Lab, University of California, San Francisco, CA 94143 USA (e-mail: david.wipf@mrsc.ucsf.edu).

B. D. Rao is with the Department of Electrical and Computer Engineering, University of California, San Diego, CA 92093-0407 USA (e-mail: brao@ucsd.edu).

Digital Object Identifier 10.1109/TSP.2007.894265

[1]While here we assume all quantities to be real, we will later consider the complex domain as well.

where $\mathcal{I}[\,\cdot\,]$ denotes the indicator function and $\|\cdot\|$ is an arbitrary vector norm. $d(\,\cdot\,)$ is a *row-diversity* measure since it counts the number of rows in W that are not equal to zero. This is in contrast to *row sparsity*, which measures the number of rows that contain all elements strictly equal to zero. Also, for the column vector $\boldsymbol{w}$, it is immediately apparent that $d(\boldsymbol{w}) = \|\boldsymbol{w}\|_0$, and so $d(\,\cdot\,)$ is a natural extension of the ℓ_0 quasi-norm to matrices. The nonzero rows of any weight matrix are referred to as *active sources*.

With regard to the dictionary Φ, we define the *spark* as the smallest number of linearly dependent columns [11]. By definition then, $2 \leq \text{spark}(\Phi) \leq N+1$. As a special case, the condition $\text{spark}(\Phi) = N+1$ is equivalent to the unique representation property from [19], which states that every subset of N columns is linearly independent. Finally, we say that Φ is *overcomplete* if $M > N$ and $\text{rank}(\Phi) = N$.

Turning to the simultaneous sparse recovery problem, we begin with the most straightforward case where $\mathcal{E} = 0$. If Φ is overcomplete, then we are presented with an ill-posed inverse problem unless further assumptions are made. For example, by extending [7, Lemma 1], if a matrix of generating weights W_{gen} satisfies

$$\begin{aligned} d(W_{\text{gen}}) &< (\text{spark}(\Phi) + \text{rank}(T) - 1)/2 \\ &\leq (\text{spark}(\Phi) + \min(L, d(W_{\text{gen}})) - 1)/2 \end{aligned} \tag{4}$$

then no other solution W can exist such that $T = \Phi W$ and $d(W) \leq d(W_{\text{gen}})$. Furthermore, if we assume suitable randomness on the nonzero entries of W_{gen}, then this result also holds under the alternative inequality

$$d(W_{\text{gen}}) < \text{spark}(\Phi) - 1 \tag{5}$$

which follows from [54, Lemma 2]. Given that one or both of these conditions hold, then recovering W_{gen} is tantamount to solving

$$W_{\text{gen}} = W_0 \triangleq \arg\min_W d(W), \quad \text{s.t. } T = \Phi W. \tag{6}$$

In general, this problem is NP-hard so approximate procedures are in order. In Section V-A, we will examine the solution of (6) in further detail. The single response ($L = 1$) reduction of (6) has been studied exhaustively [11], [16], [20], [49]. For the remainder of this paper, whenever $\mathcal{E} = 0$, we will assume that W_{gen} satisfies (4) or (5), and so W_0 and W_{gen} can be used interchangeably.

When $\mathcal{E} \neq 0$, things are decidedly more nebulous. Because noise is present, we typically do not expect to represent T exactly, suggesting the relaxed optimization problem

$$W_0(\lambda) \triangleq \arg\min_W \|T - \Phi W\|_{\mathcal{F}}^2 + \lambda d(W) \tag{7}$$

where λ is a tradeoff parameter balancing estimation quality with row sparsity $|\,\cdot\,|_{\mathcal{F}}$ denotes the Frobenius norm. An essential feature of using $d(W)$ as the regularization term is that whenever a single element in a given row of W is nonzero, there is no further penalty in making other elements in the same row nonzero, promoting a common sparsity profile as desired. Unfortunately, solving (7) is also NP-hard, nor is it clear how to select λ. Furthermore, there is no guarantee that the global solution, even if available for the optimal value of λ, is necessarily the best estimator of W_{gen}, or, perhaps more importantly, is the most likely to at least have a matching sparsity profile. This latter condition is often crucial, since it dictates which columns of Φ are relevant, a notion that can often have physical significance.[2]

From a conceptual standpoint, (7) can be recast in Bayesian terms by applying a $\exp[-(\,\cdot\,)]$ transformation. This leads to a Gaussian likelihood function $p(T \,|\, W)$ with λ-dependent variance and a prior distribution given by $p(W) \propto \exp[-d(W)]$. In weight space, this improper prior maintains a sharp peak wherever a row norm equals zero and heavy (in fact uniform) "tails" everywhere else. The optimization problem from (7) can equivalently be written as

$$\begin{aligned} W_0(\lambda) &\equiv \arg\max_W p(T \,|\, W) p(W) \\ &= \arg\max_W \frac{p(T \,|\, W) p(W)}{p(T)} \\ &= \arg\max_W p(W \,|\, T). \end{aligned} \tag{8}$$

Therefore, (7) can be viewed as a challenging MAP estimation task, with a posterior characterized by numerous locally optimal solutions.

B. Summary

In Section II, we discuss current methods for solving the simultaneous sparse approximation problem, all of which can be understood, either implicity or explicitly, as MAP-estimation procedures using a prior that encourages row sparsity. These methods are distinguished by the selection of the sparsity-inducing prior and the optimization strategy used to search for the posterior mode. The difficulty with these procedures is two-fold: either the prior is not sufficiently sparsity-inducing (supergaussian) and the MAP estimates sometimes fail to be sparse enough, or we must deal with a combinatorial number of suboptimal local solutions.

In this paper, we will also explore a Bayesian model based on a prior that ultimately encourages sparsity. However, rather than embarking on a problematic mode-finding expedition, we instead enlist an empirical Bayesian strategy that draws on the concept of automatic relevance determination (ARD) [30], [34]. Starting in Section III, we posit a prior distribution modulated by a vector of hyperparameters controlling the prior variance of each row of W, the values of which are learned from the data using an evidence maximization procedure [29]. This particular approximation enforces a common sparsity profile and consistently places its prominent posterior mass on the appropriate region of W-space necessary for sparse recovery. The resultant algorithm is called M-SBL because it can be posed as a multiple response extension of the standard sparse Bayesian learning (SBL) paradigm [48], a more descriptive title than ARD for our purposes. Additionally, it is easily extensible to the complex domain as required in many source localization problems.

[2]Although not our focus, if the ultimate goal is compression of T, then the solution of (7) may trump other concerns.

The per-iteration complexity relative to the other algorithms is also considered.

In Section IV, we assess M-SBL relative to other methods using empirical tests. First, we constrain the columns of Φ to be uniformly distributed on the surface of an N-dimensional hypersphere, consistent with the analysis in [10] and the requirements of compressed sensing applications [52]. In a variety of testing scenarios, we show that M-SBL outperforms other methods by a significant margin. These results also hold up when Φ is instead formed by concatenating pairs of orthobases [12]. A brief treatment of some of these results can be found in [56].

In Section V, we examine some properties of M-SBL and draw comparisons with the other methods. First, we discuss how the correlation between the active sources affects the simultaneous sparse approximation problem. For example, we show that if the active sources maintain zero sample correlation, then all (suboptimal) local minima are removed and we are guaranteed to solve (6) using M-SBL. We later show that none of the other algorithms satisfy this condition. In a more restricted setting (assuming $\Phi^T\Phi = I$), we also tackle related issues with the inclusion of noise, demonstrating that M-SBL can be viewed as a form of robust, sparse shrinkage operator, with no local minima, that uses an average across responses to modulate the shrinkage mechanism.

Next, we present an alternative derivation of M-SBL using variational methods that elucidates its connection with MAP-based algorithms and helps to explain its superior performance. More importantly, this perspective quantifies the means by which ARD methods are able to capture significant posterior mass when sparse priors are involved. The methodology is based on previous work in [53] that applies to the single response $(L = 1)$ case. Finally, Section VI contains concluding remarks as well as a brief discussion of recent results applying M-SBL to large-scale neuroimaging applications.

II. Existing MAP Approaches

The simultaneous sparse approximation problem has received a lot of attention recently and several computationally feasible methods have been presented for estimating the sparse, underlying weights [5], [7], [31], [40], [50], [51]. First, there are forward sequential selection methods based on some flavor of matching pursuit (MP) [32]. As the name implies, these approaches involve the sequential (and greedy) construction of a small collection of dictionary columns, with each new addition being "matched" to the current residual. In this paper, we will consider M-OMP, for *Multiple* response model *Orthogonal Matching Pursuit*, a multiple response variant of MP that can be viewed as finding a local minimum to (7), [7]. A similar algorithm is analyzed in [51].

An alternative strategy is to replace the troublesome diversity measure $d(W)$ with a penalty (or prior) that, while still encouraging row sparsity, is somehow more computationally convenient. The first algorithm in this category is a natural extension of basis pursuit [6] or the LASSO [23]. Essentially, we construct a convex relaxation of (7) and attempt to solve

$$W_{\mathrm{M-BP}} = \arg\min_{W} \|T - \Phi W\|_{\mathcal{F}}^2 + \lambda \sum_{i=1}^{M} \|\boldsymbol{w}_i.\|_2. \quad (9)$$

This convex cost function can be globally minimized using a variety of standard optimization packages. In keeping with a Bayesian perspective, (9) is equivalent to MAP estimation using a Laplacian prior on the ℓ_2 norm of each row (after applying a $\exp[-(\cdot)]$ transformation as before). We will refer to procedures that solve (9) as M-BP, consistent with previous notation. The properties of the M-BP cost function and algorithms for its minimization have been explored in [7] and [31]. Other variants involve replacing the row-wise ℓ_2 norm with the ℓ_∞ norm [50] and the ℓ_1 norm [5]. However, when the ℓ_1 norm is used across rows, the problem decouples and we are left with L single response problems. As such, this method is inconsistent with our goal of simultaneously using all responses to encourage row sparsity.

Second, we consider what may be termed the M-Jeffreys algorithm, where the ℓ_1-norm-based penalty from above is substituted with a regularization term based on the negative logarithm of a Jeffreys prior on the row norms.[3] The optimization problem then becomes

$$W_{\mathrm{M-Jeffreys}} = \arg\min_{W} \|T - \Phi W\|_{\mathcal{F}}^2 + \lambda \sum_{i=1}^{M} \log \|\boldsymbol{w}_i.\|_2. \quad (10)$$

The M-Jeffreys cost function suffers from numerous local minima, but when given a sufficiently good initialization, can potentially find solutions that are closer to W_{gen} than $W_{\mathrm{M-BP}}$. From an implementational standpoint, M-Jeffreys can be solved using natural, multiple response extensions of the algorithms derived in [15], [19].

Third, we weigh in the M-FOCUSS algorithm derived in [7], [40] based on the generalized FOCUSS algorithm of [41]. This approach employs an ℓ_p-norm-like diversity measure [9], where $p \in [0, 1]$ is a user-defined parameter, to discourage models with many nonzero rows. In the context of MAP estimation, this method can be derived using a generalized Gaussian prior on the row norms, analogous to the Laplacian and Jeffreys priors assumed above. The M-FOCUSS update rule is guaranteed to converge monotonically to a local minimum of

$$W_{\mathrm{M-FOCUSS}} = \arg\min_{W} \|T - \Phi W\|_{\mathcal{F}}^2 + \lambda \sum_{i=1}^{M} \left(\|\boldsymbol{w}_i.\|_2\right)^p. \quad (11)$$

If $p \to 0$, the M-FOCUSS cost function approaches (7). While this may appear promising, the resultant update rule in this situation ensures (for any finite λ) that the algorithm converges (almost surely) to a locally minimizing solution W' such that $T = \Phi W'$ and $d(W') \leq N$, regardless of λ. The set of initial conditions whereby we will actually converge to $W_0(\lambda)$ has measure zero. When $p = 1$, M-FOCUSS reduces to an interior point method of implementing M-BP. The M-FOCUSS framework also includes M-Jeffreys as a special case as shown in the Appendix. In practice, it is sometimes possible to jointly select values of p and λ such that the algorithm outperforms both

[3] The Jeffreys prior is an improper prior of the form $p(x) = 1/x$ [3].

M-BP and M-Jeffreys. In general though, with M-BP, M-Jeffreys, and M-FOCUSS, λ must be tuned with regard to a particular application. Also, in the limit as λ becomes small, we can view each multiple response algorithm as minimizing the respective diversity measure subject to the constraint $T = \Phi W$. This is in direct analogy to (6).

III. Empirical Bayesian Algorithm

All of the methods discussed in the previous section for estimating W_{gen} involve searching some implicit posterior distribution for the mode by solving $\arg\max_W p(W,T) = \arg\max_W p(T\,|\,W)p(W)$, where $p(W)$ is a fixed, algorithm-dependent prior. At least two significant problems arise with such an endeavor. First, if only a moderately sparse prior such as the Laplacian is chosen for the row norms (as with M-BP), a unimodal posterior results and mode finding is greatly simplified; however, the resultant posterior mode may not be sufficiently sparse, and, therefore, $W_{\text{M-BP}}$ may be unrepresentative of W_{gen}. In contrast, if a highly sparse prior is chosen, e.g., the Jeffreys prior or a generalized Gaussian with $p \ll 1$, we experience a combinatorial increase in local optima. While one or more of these optima may be sufficiently sparse and representative of W_{gen}, finding it can be very difficult if not impossible.

So, mode finding can be a problematic exercise when sparse priors are involved. In this section, a different route to solving the simultaneous sparse approximation problem is developed using the concept of ARD, originally proposed in the neural network literature as a quantitative means of weighing the relative importance of network inputs, many of which may be irrelevant [30], [34]. These ideas have also been applied to Bayesian kernel machines [48]. A key ingredient of this formulation is the incorporation of an *empirical prior*, by which we mean a flexible prior distribution dependent on a set of unknown hyperparameters that must be estimated from the data.

To begin, we postulate $p(T\,|\,W)$ to be Gaussian with noise variance σ^2 that is assumed to be known (the case where σ^2 is not known is discussed briefly in Section III-C). Thus, for each $\boldsymbol{t}_{\cdot j}, \boldsymbol{w}_{\cdot j}$ pair, we have

$$p(\boldsymbol{t}_{\cdot j}\,|\,\boldsymbol{w}_{\cdot j}) = (2\pi\sigma^2)^{-N/2}\exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{t}_{\cdot j} - \Phi\boldsymbol{w}_{\cdot j}\|_2^2\right) \quad (12)$$

which is consistent with the likelihood model implied by (7) and previous Bayesian methods. Next, application of ARD involves assigning to the ith row of W an L-dimensional Gaussian prior

$$p(\boldsymbol{w}_{i\cdot};\gamma_i) \triangleq \mathcal{N}(0,\gamma_i I) \quad (13)$$

where γ_i is an unknown variance parameter. By combining each of these row priors, we arrive at a full weight prior

$$p(W;\boldsymbol{\gamma}) = \prod_{i=1}^{M} p(\boldsymbol{w}_{i\cdot};\gamma_i), \quad (14)$$

whose form is modulated by the hyperparameter vector $\boldsymbol{\gamma} = [\gamma_1,\ldots,\gamma_M]^T \in \mathbb{R}_+^M$. Combining likelihood and prior, the posterior density of the jth column of W then becomes

$$p(\boldsymbol{w}_{\cdot j}\,|\,\boldsymbol{t}_{\cdot j};\boldsymbol{\gamma}) = \frac{p(\boldsymbol{w}_{\cdot j},\boldsymbol{t}_{\cdot j};\boldsymbol{\gamma})}{\int p(\boldsymbol{w}_{\cdot j},\boldsymbol{t}_{\cdot j};\boldsymbol{\gamma})d\boldsymbol{w}_{\cdot j}} = \mathcal{N}(\boldsymbol{\mu}_{\cdot j},\Sigma) \quad (15)$$

with mean and covariance given by

$$\begin{aligned}\Sigma &\triangleq \text{Cov}[\boldsymbol{w}_{\cdot j}\,|\,\boldsymbol{t}_{\cdot j};\boldsymbol{\gamma}] = \Gamma - \Gamma\Phi^T\Sigma_t^{-1}\Phi\Gamma \qquad \forall j = 1,\ldots,L\\ \mathcal{M} &= [\boldsymbol{\mu}_{\cdot 1},\ldots,\boldsymbol{\mu}_{\cdot L}] \triangleq \text{E}[W\,|\,T;\boldsymbol{\gamma}] = \Gamma\Phi^T\Sigma_t^{-1}T \end{aligned} \quad (16)$$

where $\Gamma \triangleq \text{diag}(\boldsymbol{\gamma})$ and $\Sigma_t \triangleq \sigma^2 I + \Phi\Gamma\Phi^T$.

Since it is typically desirable to have a point estimate for W_{gen}, we may enlist $\mathcal{M}$, the posterior mean, for this purpose. Row sparsity is naturally achieved whenever a γ_i is equal to zero. This forces the posterior to satisfy $\text{Prob}(\boldsymbol{w}_{i\cdot} = \boldsymbol{0}\,|\,T;\gamma_i = 0) = 1$, ensuring that the posterior mean of the ith row, $\boldsymbol{\mu}_{i\cdot}$, will be zero as desired. Thus, estimating the sparsity profile of some W_{gen} is conveniently shifted to estimating a hyperparameter vector with the correct number and location of nonzero elements. The latter can be effectively accomplished through an iterative process discussed next. Later, Sections IV and V provide empirical and analytical support for this claim.

A. Hyperparameter Estimation: The M-SBL Algorithm

Each unique value for the hyperparameter vector $\boldsymbol{\gamma}$ corresponds to a different hypothesis for the prior distribution underlying the generation of the data T. As such, determining an appropriate $\boldsymbol{\gamma}$ is tantamount to a form of model selection. In this context, the empirical Bayesian strategy for performing this task is to treat the unknown weights W as nuisance parameters and integrate them out [29]. The marginal likelihood that results is then maximized with respect to $\boldsymbol{\gamma}$, leading to the ARD-based cost function

$$\begin{aligned}\mathcal{L}(\boldsymbol{\gamma}) &\triangleq -2\log\int p(T\,|\,W)p(W;\boldsymbol{\gamma})dW\\ &= -2\log p(T;\boldsymbol{\gamma}) \equiv L\log|\Sigma_t| + \sum_{j=1}^{L}\boldsymbol{t}_{\cdot j}^T\Sigma_t^{-1}\boldsymbol{t}_{\cdot j}\end{aligned} \quad (17)$$

where a $-2\log(\cdot)$ transformation has been added for simplicity.

The use of marginalization for hyperparameter optimization in this fashion has been proposed in a variety of contexts. In the classical statistics literature, it has been motivated as a way of compensating for the loss of degrees of freedom associated with estimating covariance components along with unknown weights analogous to W [21], [22]. Bayesian practitioners have also proposed this idea as a natural means of incorporating the principle of Occam's razor into model selection, often using the description *evidence maximization* or *type-II maximum likelihood* to describe the optimization process [3], [29], [34].

There are (at least) two ways to minimize $\mathcal{L}(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$. First, treating the unknown weights W as hidden data, we can minimize this expression over $\boldsymbol{\gamma}$ using a simple EM algorithm as proposed in [8], [22] for covariance estimation. For the E-step, this requires computation of the posterior moments using (16), while the M-step is expressed via the update rule

$$\gamma_i^{(\text{new})} = \frac{1}{L}\|\boldsymbol{\mu}_{i\cdot}\|_2^2 + \Sigma_{ii}, \quad \forall i = 1,\ldots,M. \quad (18)$$

While benefitting from the general convergence properties of the EM algorithm, we have observed this update rule to be very slow on large practical applications.

Second, at the expense of proven convergence, we may instead optimize (17) by taking the derivative with respect to $\boldsymbol{\gamma}$, equating to zero, and forming a fixed-point equation that typically leads to faster convergence [29], [48]. Effectively, this involves replacing the M-step from above with

$$\gamma_i^{(\text{new})} = \frac{\frac{1}{L}\|\boldsymbol{\mu}_{i\cdot}\|_2^2}{1 - \gamma_i^{-1}\Sigma_{ii}}, \quad \forall i = 1, \ldots, M. \tag{19}$$

We have found this alternative update rule to be extremely useful in large-scale, highly overcomplete problems, although the results upon convergence are sometimes inferior to those obtained using the slower update (18). In the context of kernel regression using a complete dictionary (meaning $N = M$) and $L = 1$, use of (19), along with a modified form of (16),[4] has been empirically shown to drive many hyperparameters to zero, allowing the associated weights to be pruned. As such, this process has been referred to as *sparse Bayesian learning* (SBL) [48]. Similar update rules have also been effectively applied to an energy prediction competition under the guise of ARD [30]. For application to the simultaneous sparse approximation problem, we choose the label M-SBL (which stresses sparsity) to refer to the process of estimating $\boldsymbol{\gamma}$, using either the EM or fixed-point update rules, as well as the subsequent computation and use of the resulting posterior.

Finally, in the event that we would like to find exact (noise-free) sparse representations, the M-SBL iterations can be easily adapted to handle the limit as $\sigma^2 \to 0$ using the modified moments

$$\begin{aligned} \Sigma &= [I - \Gamma^{1/2}(\Phi\Gamma^{1/2})^{\dagger}\Phi]\Gamma \\ \mathcal{M} &= \Gamma^{1/2}(\Phi\Gamma^{1/2})^{\dagger}T \end{aligned} \tag{20}$$

where $(\cdot)^{\dagger}$ denotes the Moore–Penrose pseudo-inverse. This is particularly useful if we wish to solve (6).

B. Algorithm Summary

Given observation data T and a dictionary Φ, the M-SBL procedure can be summarized by the following collection of steps.

1) Initialize $\boldsymbol{\gamma}$, e.g., $\boldsymbol{\gamma} := \mathbf{1}$ or, perhaps, a non-negative random initialization.
2) Compute the posterior moments Σ and $\mathcal{M}$ using (16), or in the noiseless case, using (20).
3) Update $\boldsymbol{\gamma}$ using the EM rule (18) or the faster fixed-point rule (19).
4) Iterate Steps 2) and 3) until convergence to a fixed point $\boldsymbol{\gamma}^*$.
5) Assuming a point estimate is desired for the unknown weights W_{gen}, choose $W_{\text{M-SBL}} = \mathcal{M}^* \approx W_{\text{gen}}$, where $\mathcal{M}^* \triangleq \mathrm{E}[W \,|\, T; \boldsymbol{\gamma}^*]$.
6) Given that $\boldsymbol{\gamma}^*$ is sparse, the resultant estimator $\mathcal{M}^*$ will necessarily be row sparse.

In practice, some arbitrarily small threshold can be set such that, when any hyperparameter becomes sufficiently small (e.g., 10^{-16}), it is pruned from the model (along with the corresponding dictionary column and row of W).

[4]This requires application of the matrix inversion lemma to Σ_t^{-1}.

C. Noise Variance Estimation

If we already have access to some reliable estimate for σ^2, then it can naturally be incorporated into the update rules above. When no such luxury exists, it would be desirable to have some alternative at our disposal. One possibility is to explicitly incorporate σ^2 estimation into the M-SBL framework as originally discussed in [29], [48]. This involves replacing the M-step with a joint maximization over σ^2 and $\boldsymbol{\gamma}$. Because of decoupling, the $\boldsymbol{\gamma}$ update remains unchanged, while we must include (e.g., for the fast M-SBL version) the σ^2 update

$$(\sigma^2)^{(\text{new})} = \frac{\frac{1}{L}\|T - \Phi\mathcal{M}\|_{\mathcal{F}}^2}{N - M + \sum_{i=1}^{M}\frac{\Sigma_{ii}}{\gamma_i}}. \tag{21}$$

A word of caution is in order with respect to σ^2 estimation that has not been addressed in the original SBL literature (this caveat applies equally to the single response case). For suitably structured dictionaries and $M \geq N$, σ^2 estimates obtained via this procedure can be extremely inaccurate. In effect, there is an identifiability issue when any subset of N dictionary columns is sufficiently spread out such that $\mathcal{L}(\boldsymbol{\gamma}, \sigma^2)$ can be minimized with $\sigma^2 = 0$. For example, if we choose the dictionary $\Phi' = [\Phi, I]$, then σ^2 as well as the N hyperparameters associated with the identity matrix columns of Φ' are not identifiable in the strict statistical sense. This occurs because a nonzero σ^2 and the appropriate N nonzero hyperparameters make an identical contribution to the covariance Σ_t. In general, the signal dictionary will not contain I; however, the underlying problem of basis vectors masquerading as noise can lead to seriously biased estimates of σ^2. As such, we generally recommend the more modest strategy of simply experimenting with different values or using some other heuristic designed with a given application in mind.

D. Extension to the Complex Case

The use of complex-valued dictionaries, responses, and weights expands the relevance of the multiple response framework to many useful signal processing disciplines. Fortunately, this extension turns out to be very natural and straightforward. We start by replacing the likelihood model for each $\boldsymbol{t}_{\cdot j}$ with a multivariate complex Gaussian distribution [28]

$$p(\boldsymbol{t}_{\cdot j} \,|\, \boldsymbol{w}_{\cdot j}) = (\pi\sigma^2)^{-N}\exp\left(-\frac{1}{\sigma^2}\|\boldsymbol{t}_{\cdot j} - \Phi\boldsymbol{w}_{\cdot j}\|_2^2\right) \tag{22}$$

where all quantities except σ^2 are now complex and $\|\boldsymbol{x}\|_2^2$ now implies $\boldsymbol{x}^H\boldsymbol{x}$, with $(\cdot)^H$ denoting the Hermitian transpose. The row priors $p(\boldsymbol{w}_{i\cdot}; \mathcal{H})$ need not change at all except for the associated norm. The derivation proceeds as before, leading to identical update rules with the exception of $(\cdot)^T$ changing to $(\cdot)^H$.

The resultant algorithm turns out to be quite useful in finding sparse representations of complex-valued signals, such as those that arise in the context of direction-of-arrival (DOA) estimation. Here, we are given an array of N omnidirectional sensors and a collection of D complex signal waves impinging upon them. The goal is then to estimate the (angular) direction of the wave sources with respect to the array. This source localization problem is germane to many sonar and radar applications. While

we have successfully applied complex M-SBL to DOA estimation problems, space precludes a detailed account of this application and comparative results. See [31] for a good description of the DOA problem and its solution using a second-order cone (SOC) implementation of M-BP. M-SBL is applied in exactly the same fashion.

E. Complexity

With regard to computational comparisons, we assume $N \leq M$. Under this constraint, each M-SBL iteration is $\mathrm{O}(N^2M)$ for real or complex data. The absence of L in this expression can be obtained using the following implementation. Because the M-SBL update rules and cost function are ultimately only dependent on T through the outer product TT^T, we can always replace T with a matrix $\tilde{T} \in \mathbb{R}^{N \times \mathrm{rank}(T)}$ such that $\tilde{T}\tilde{T}^T = TT^T$. Substituting $\tilde{T}$ into the M-SBL update rules, while avoiding the computation of off-diagonal elements of Σ, leads to the stated complexity result. In a similar fashion, each M-BP, M-FOCUSS, and M-Jeffreys iteration can also be computed in $\mathrm{O}(N^2M)$. This is significant because little price is paid for adding additional responses and only a linear penalty is incurred when adding basis vectors.

In contrast, the second-order cone (SOC) implementation of M-BP [31] is $\mathrm{O}(M^3L^3)$ per iteration. While the effective value of L can be reduced significantly (beyond what we described above) using a variety of useful heuristic strategies, unlike M-SBL and other approaches, it will still enter as a multiplicative cubic factor. This could be prohibitively expensive if M is large, although fewer total iterations are usually possible. Nonetheless, in neuroimaging applications, we can easily have $N \approx 200, L \approx 100$, and $M \approx 100\,000$. In this situation, the M-SBL (or M-FOCUSS, etc.) iterations are very attractive. Of course M-OMP is decidedly less costly than all of these methods.

IV. Empirical Studies

This section presents comparative Monte Carlo experiments involving randomized dictionaries and pairs of orthobases.

A. Random Dictionaries

We would like to quantify the performance of M-SBL relative to other methods in recovering sparse sets of generating weights, which in many applications have physical significance (e.g., source localization). To accommodate this objective, we performed a series of simulation trials where by design we have access to the sparse, underlying model coefficients. For simplicity, noiseless tests were performed first [i.e., solving (6)]; this facilitates direct comparisons because discrepancies in results cannot be attributed to poor selection of tradeoff parameters (which balance sparsity and quality of fit) in the case of most algorithms.

Each trial consisted of the following: First, an overcomplete $N \times M$ dictionary Φ is created with columns draw uniformly from the surface of a unit hypersphere. This particular mechanism for generating dictionaries is advocated in [10] as a useful benchmark. Additionally, it is exactly what is required in compressed sensing applications [52]. L sparse weight vectors are randomly generated with D nonzero entries and a common sparsity profile. Nonzero amplitudes are drawn from a uniform distribution. Response values are then computed as $T = \Phi W_{\mathrm{gen}}$. Each algorithm is presented with T and Φ and attempts to estimate W_{gen}. For all methods, we can compare W_{gen} with $\hat{W}$ after each trial to see if the sparse generating weights have been recovered.

Under the conditions set forth for the generation of Φ and T, $\mathrm{spark}(\Phi) = N + 1$ and (5) is in force. Therefore, we can be sure that $W_{\mathrm{gen}} = W_0$ with probability one. Additionally, we can be certain that when an algorithm fails to find W_{gen}, it has not been lured astray by an even sparser representation. Results are shown in Fig. 1 as L, D, and M are varied. To create each data point, we ran 1000 independent trials and compared the number of times each algorithm failed to recover W_{gen}. Based on the Fig. 1, M-SBL (a) performs better for different values of L, (b) resolves a higher number of nonzero rows, and (c) is more capable of handling added dictionary redundancy.

We also performed analogous tests with the inclusion of noise. Specifically, uncorrelated Gaussian noise was added to produce an SNR of 10dB. When noise is present, we do not expect to reproduce T exactly, so we now classify a trial as successful if the D largest estimated row-norms align with the sparsity profile of W_{gen}. Fig. 1(d) displays sparse recovery results as the tradeoff parameter for each algorithm is varied. The performance gap between M-SBL and the others is reduced when noise is included. This is because now the issue is not so much local minima avoidance, etc., since D is relatively low relative to N and M, but rather proximity to the fundamental limits of how many nonzero rows can reliably be detected in the presence of noise.[5] For example, even an exhaustive search for the optimal solution to (7) over all λ would likely exhibit similar performance to M-SBL in this situation.

In fact, for sufficiently small values of N and M, we can test this hypothesis directly. Using $N = 8, M = 16$, and $D = 3$, we reproduced Fig. 1(d) with the inclusion of the the global solution to (7) for different values of λ. The exhaustive search failed to locate the correct sparsity profile with an empirical probability similar to M-SBL (about 0.10 using λ_{opt}), underscoring the overall difficulty of finding sparse generating weights in noisy environments.[6] Moreover, it demonstrates that, unlike in the noise-free case, the NP-hard optimization problem of (7) is not necessarily guaranteed to be the most desirable solution even if computational resources are abundant.

B. Pairs of Orthobases

Even if M-SBL seems to perform best on "most" dictionaries relative to a uniform measure, it is well known that many signal processing applications are based on sets of highly

[5]Most of the theoretical study of approximate sparse representations in noise has focused on when a simpler method, e.g., BP- or OMP-based, is guaranteed to provide a good solution to (7), or at least exhibit a similar sparsity profile. Currently, we know of no work that examines rigorous conditions whereby the minimum of (7) or any of the other proposed cost functions is guaranteed to match the sparsity profile of W_{gen}. When there is no noise, this distinction effectively disappears.

[6]With no noise and D increased to 7, exhaustive subset selection yields zero error (with any $\lambda \ll 1$) as expected while M-SBL fails with probability 0.24. So, a high noise level is a significant performance equalizer.

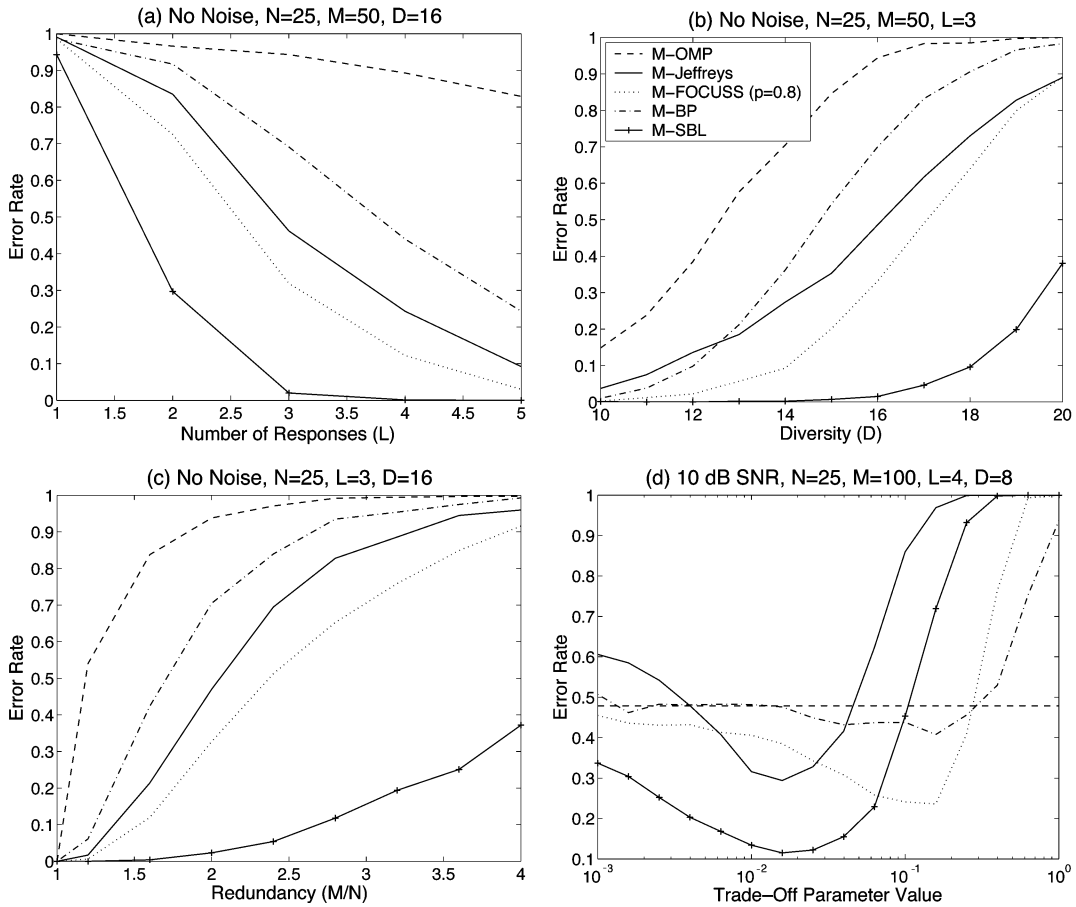

Fig. 1. Results comparing the empirical probability (over 1000 trials) that each algorithm fails to find the sparse generating weights under various testing conditions. Plots (a)–(c) display results as L, D, and M are varied under noiseless conditions. Plot (d) shows results with 10-dB AGWN for different values of the tradeoff parameter (i.e., either λ or σ^2).

structured dictionaries that may have zero measure on the unit hypersphere. Although it is not feasible to examine all such scenarios, we have performed an analysis similar to the preceding section using dictionaries formed by concatenating two orthobases, i.e., $\Phi = [\Theta, \Psi]$, where Θ and Ψ represent $N \times N$ orthonormal bases. Candidates for Θ and Ψ include Hadamard-Walsh functions, DCT bases, identity matrices, and Karhunen-Loéve expansions among many others. The idea is that, while a signal may not be compactly represented using a single orthobasis as in standard Fourier analysis, it may become feasible after we concatenate two such dictionaries. For example, a sinusoid with a few random spikes would be amenable to such a representation. Additionally, much attention is placed on such dictionaries in the signal processing and information theory communities [11], [12].

For comparison purposes, T and W_{gen} were generated in an identical fashion as before. Θ was set to the identity matrix and Ψ was selected to be either a DCT or a Hadamard basis (other examples have been explored as well). Results are displayed in Fig. 2, strengthening our premise that M-SBL represents a viable alternative regardless of the dictionary type. Also, while in this situation we cannot *a priori* guarantee absolutely that $W_{\text{gen}} = W_0$, in all cases where an algorithm failed, it converged to a solution with $d(\hat{W}) > d(W_{\text{gen}})$.

V. Analysis

This section analyzes some of the properties of M-SBL and where possible, discusses relationships with other multiple response algorithms.

A. Multiple Responses and Maximally Sparse Representations: Noiseless Case

Increasing the number of responses L has two primary benefits when using M-SBL. First, and not surprisingly, it mitigates the effects of noise as will be discussed more in Section V-B. However, there is also a less transparent benefit, which is equally important and applies even in the absence of noise: Increasing L can facilitate the avoidance of suboptimal, locally minimizing solutions, or, stated differently, increasing the number of responses increases the likelihood that M-SBL will converge to the global minimum of $\mathcal{L}(\boldsymbol{\gamma})$. This is important because, under very reasonable conditions, this global minimum is characterized by $\mathcal{M}^* = W_0$ when $\mathcal{E} = 0$ and $\sigma^2 \to 0$. This result follows from ([55], Theorem 1), which applies to the $L = 1$ case but is easily generalized. So, the globally minimizing M-SBL hyperparameters are guaranteed to produce the maximally sparse representation, and increasing L improves the chances that these hyperparameters are found.

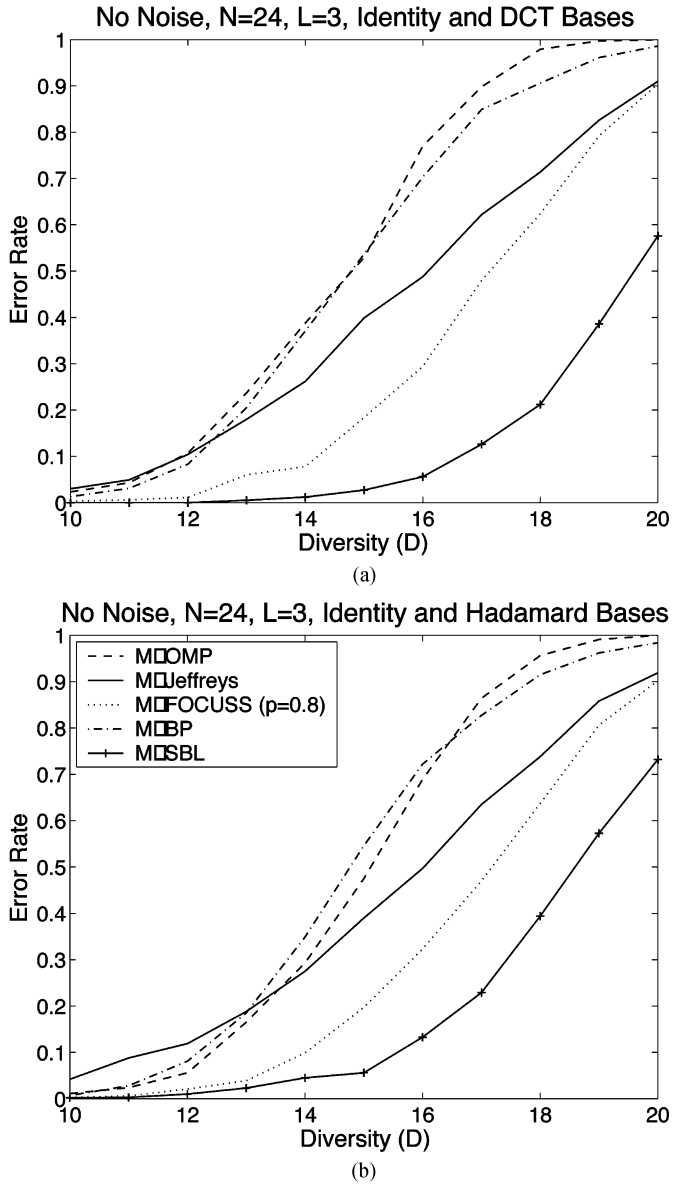

Fig. 2. Results using pairs of orthobases with $L = 3$ and $N = 24$, while D is varied from 10 to 20. Top: Θ is an identity matrix and Ψ is an N-dimensional DCT. Bottom: Θ is again identity and Ψ is a Hadamard matrix.

Of course the merits of increasing L, in the absence of noise, are highly dependent on how the active sources (the nonzero rows of W_0) are distributed. For example, suppose these sources are perfectly correlated, meaning that W_0 can be written as the outer-product $\boldsymbol{a}\boldsymbol{b}^T$ for some vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. In this situation, the problem can be reduced to an equivalent, single response problem with $\boldsymbol{t} = \Phi\boldsymbol{a}\|\boldsymbol{b}\|_2$, indicating that there is no benefit to including additional responses (i.e., the local minima profile of the cost function does not change with increasing L).

In contrast, as the (sample) correlation between active sources is reduced, the probability that M-SBL becomes locally trapped falls off dramatically as evidenced by empirical studies. This begs the question, is there any situation where we are guaranteed to reach the global minimum, without ever getting stuck at suboptimal solutions? This is tantamount to finding conditions under which M-SBL will always produce the maximally sparse solution W_0, the solution to (6).

To address this issue, we consider the fixed points of the M-SBL iterations using the modified moments from (20). Of particular interest is the set of stable fixed points because they must necessarily be local minima to the M-SBL cost function by virtue of the convergence properties of the EM algorithm.[7] We now establish conditions whereby a unique stable fixed point exists that is also guaranteed to solve (6).

Theorem 1: Given a dictionary Φ and a set of responses T, assume that $d(W_0) < \text{spark}(\Phi) - 1 \leq N$. Then if the nonzero rows of W_0 are orthogonal (no sample-wise correlation), there exists a unique, stable fixed point $\boldsymbol{\gamma}^*$. Additionally, at this stable fixed point, we have

$$\mathcal{M}^* = \mathrm{E}[W \,|\, T; \boldsymbol{\gamma}^*] = \Gamma^{*1/2}(\Phi\Gamma^{*1/2})^{\dagger}T = W_0 \qquad (23)$$

the maximally sparse solution. All other fixed points are unstable.

Proof: See [57] for the proof. ■

Because only highly nuanced initializations will lead to an unstable fixed point (and small perturbations lead to escape), this result dictates conditions whereby M-SBL is guaranteed to solve (6) and, therefore, find W_{gen}, assuming condition (4) or (5) holds. Moreover, even if a non-EM-based optimization procedure is used, the M-SBL cost function itself must be unimodal (although not necessarily convex) to satisfy Theorem 1.

Admittedly, the required conditions for Theorem 1 to apply are highly idealized. Nonetheless, this result is interesting to the extent that it elucidates the behavior of M-SBL and distinguishes its performance from the other methods. Specifically, it encapsulates the intuitive notion that if each active source is sufficiently diverse (or uncorrelated), then we will find W_0. Perhaps more importantly, no equivalent theorem exists for any of the other multiple response methods mentioned in Section II. Consequently, they will break down even with perfectly uncorrelated sources, a fact that we have verified experimentally using Monte Carlo simulations analogous to those in Section IV-A. Table I displays these results. As expected, M-SBL has zero errors while the others are often subject to failure (convergence to suboptimal yet stable fixed points).

In any event, the noiseless theoretical analysis of sparse learning algorithms has become a very prolific field of late, where the goal is to establish sufficient conditions whereby a particular algorithm will always recover the maximally sparse solution [10], [11], [16], [20], [49]. Previous results of this sort have all benefitted from the substantial simplicity afforded by either straightforward, greedy update rules (MP-based methods) or a manageable, convex cost function (BP-based methods). In contrast, the highly complex update rules and associated nonconvex cost function under consideration here are decidedly more difficult to analyze. As such, evidence showing that good, fully sparse solutions can be achieved using ARD has typically relied on empirical results or heuristic arguments [30], [34], [48]. Here, we have tried to make some progress in this regard.

[7]The EM algorithm ensures monotonic convergence (or cost function decrease) to some fixed point. Therefore, a stable fixed point must also be a local minimum, otherwise initializing at an appropriately perturbed solution will lead to a different fixed point.

TABLE I
VERIFICATION OF THEOREM 1 WITH $N = 5, M = 50, D = L = 4$. Φ IS GENERATED AS IN SECTION IV-A, WHILE W_{gen} IS GENERATED WITH ORTHOGONAL ACTIVE SOURCES. ALL ERROR RATES ARE BASED ON 1000 INDEPENDENT TRIALS

	M-OMP	M-Jeffreys	M-FOCUSS ($p = 0.8$)	M-BP	M-SBL
ERROR RATE	1.000	0.471	0.371	0.356	**0.000**

While Theorem 1 provides a limited sufficient condition for establishing equivalence between a unique, stable fixed point and W_0, it is by no means necessary. For example, because the sparse Bayesian learning framework is still quite robust in the $L = 1$ regime [57], we typically experience a smooth degradation in performance as the inter-source correlation increases. Likewise, when $d(W_0) > L$ or when noise is present, M-SBL remains highly effective as was shown in Section IV.

B. Extensions to the Noisy Case

We now briefly address the more realistic scenario where noise is present. Because of the substantially greater difficulty this entails, we restrict ourselves to complete or undercomplete orthonormal dictionaries. Nonetheless, these results illuminate more general application conditions and extend the analysis in [47], which compares the single response LASSO algorithm with traditional shrinkage methods using orthonormal dictionaries.

Empirical and analytical results suggest that M-Jeffreys and M-FOCUSS have more local minima than M-SBL in the noiseless case, and it is likely that this problem persists for $\mathcal{E} > 0$. As an example, assume that $M \leq N$ and $\Phi^T\Phi = I$. Under these constraints, the M-SBL problem conveniently decouples giving us M independent cost functions, one for each hyperparameter of the form

$$\mathcal{L}(\gamma_i) = L\log(\sigma^2 + \gamma_i) + \frac{1}{\sigma^2 + \gamma_i}\sum_{j=1}^{L}\left(w_{ij}^{\text{MN}}\right)^2 \qquad (24)$$

where $W^{\text{MN}} \triangleq \Phi^{\dagger}T = \Phi^T T$, i.e., W^{MN} is the minimum ℓ_2-norm solution to $T = \Phi W$. Conveniently, this function is unimodal in γ_i. By differentiating, equating to zero, and noting that all γ_i must be greater than zero, we find that the unique minimizing solution occurs at

$$\gamma_i^* = \left(\frac{1}{L}\sum_{j=1}^{L}\left(w_{ij}^{\text{MN}}\right)^2 - \sigma^2\right)^+ \qquad (25)$$

where the operator $(x)^+$ equals x if $x > 0$ and zero otherwise. Additionally, by computing the associated $\mathcal{M}^*$, we obtain the representation

$$\boldsymbol{\mu}_{i\cdot}^* = \boldsymbol{w}_{i\cdot}^{\text{MN}}\left(1 - \frac{L\sigma^2}{||\boldsymbol{w}_{i\cdot}^{\text{MN}}||_2^2}\right)^+. \qquad (26)$$

Interestingly, these weights represent a direct, multiple-response extension of those obtained using the nonnegative garrote estimator [4], [17], [47]. Consequently, in this setting, M-SBL can be interpreted as a sort of generalized shrinkage method, truncating rows with small norm to zero and shrinking others by a factor that decreases as the norm grows. Also, with the inclusion of multiple responses, the truncation operator is much more robust to noise because the threshold is moderated by an average across responses, i.e., $1/L\sum_{j=1}^{L}(w_{ij}^{\text{MN}})^2$. So, for a given noise variance, there is considerably less chance that a spurious value will exceed the threshold. While, obviously, (26) can be computed directly without resorting to the iterative M-SBL procedure, it is nonetheless important to note that this is the actual solution M-SBL will always converge to since the cost function has no (nonglobal) local minima.

Turning to the M-Jeffreys approach, we again obtain a decoupled cost function resulting in M row-wise minimization problems of the form

$$\min_{\boldsymbol{w}_{i\cdot}} ||\boldsymbol{w}_{i\cdot}^{\text{MN}}||_2^2 - 2\boldsymbol{w}_{i\cdot}^T\boldsymbol{w}_{i\cdot}^{\text{MN}} + ||\boldsymbol{w}_{i\cdot}||_2^2 + \lambda\log||\boldsymbol{w}_{i\cdot}||_2. \qquad (27)$$

For any fixed $||\boldsymbol{w}_{i\cdot}||_2$, the direction of the optimal $\boldsymbol{w}_{i\cdot}$ is always given by $\boldsymbol{w}_{i\cdot}^{\text{MN}}/||\boldsymbol{w}_{i\cdot}^{\text{MN}}||_2$, effectively reducing (27) to

$$\min_{||\boldsymbol{w}_{i\cdot}||_2} \left|\left|\boldsymbol{w}_{i\cdot}^{\text{MN}}\right|\right|_2^2 - 2||\boldsymbol{w}_{i\cdot}||_2\left|\left|\boldsymbol{w}_{i\cdot}^{\text{MN}}\right|\right|_2 + ||\boldsymbol{w}_{i\cdot}||_2^2 + \lambda\log||\boldsymbol{w}_{i\cdot}||_2 \qquad (28)$$

If $||\boldsymbol{w}_{i\cdot}^{\text{MN}}||_2^2 \leq 2\lambda$, then for each row i, there is a single minimum with $\boldsymbol{w}_{i\cdot} = 0$. In contrast, for $||\boldsymbol{w}_{i\cdot}^{\text{MN}}||_2^2 > 2\lambda$, there are two minima, one at zero and the other with $||\boldsymbol{w}_{i\cdot}||_2 = (1/2)(||\boldsymbol{w}_{i\cdot}^{\text{MN}}||_2 + \sqrt{||\boldsymbol{w}_{i\cdot}^{\text{MN}}||_2^2 - 2\lambda})$. Unlike M-SBL, this ensures that the M-Jeffreys cost function will have $2^{(\sum_i \mathcal{I}[||\boldsymbol{w}_{i\cdot}^{\text{MN}}||_2^2 > 2\lambda])}$ local minimum, although we can obtain a useful alternative shrinkage operator (that closely resembles a hard threshold) with an appropriate initialization and selection of λ. However, while it may be transparent how to avoid unattractive local minima in the orthonormal case, in a more general setting, this poses a significant problem.

M-FOCUSS is more difficult to analyze for arbitrary values of p, since we cannot provide an analytic solution for locally minimizing values of $||\boldsymbol{w}_{i\cdot}||_2$. However, the optimal solution does entail a threshold and asymptotic results are obtained (for the single response case) as $||\boldsymbol{w}_{i\cdot}||_2 \to \infty$ in [33]. Also, as $p \to 0$, we converge to a generalized hard-threshold operator, which truncates small rows to zero and leaves others unchanged. Unfortunately, however, the actual algorithm will always produce the nontruncated solution W^{MN} (one of the 2^M possible local minima) because the basins of attraction of all other local minima have zero measure in W space. As p is steadily increased from zero to one, the number of local minima gradually

drops from 2^M to one.[8] When $p = 1$, we obtain an analogous soft-threshold operator, as discussed in [47] for the single response case. Since each row-wise cost function is convex, we also observe no local minimum as with M-SBL.

In summary, we need not actually run the M-SBL algorithm (or M-Jeffreys, etc.), in practice, when using an orthonormal dictionary Φ; we could just compute our weights analytically using the appropriate shrinkage mechanism. Nonetheless, it is encouraging to see a well motivated cost function devoid of local minima in the case of M-SBL (and M-BP). This provides further evidence that alternatives to standard mode finding may be a successful route to handling the simultaneous sparse approximation problem. It also verifies that ARD methods will push unnecessary coefficients to exactly zero, as opposed to merely making them small.

C. Relating M-SBL and M-Jeffreys

Thus far, we have divided Bayesian approaches into two seemingly very different categories: an empirical Bayesian approach based on ARD and a class of MAP estimators including M-BP, M-FOCUSS, and M-Jeffreys. In fact, M-SBL is closely related to M-Jeffreys (and, therefore, M-FOCUSS with p small per the discussion in the Appendix) albeit with several significant advantages. Both methods can be viewed as starting with an identical likelihood and prior model, but then deviate sharply with respect to how estimation and inference are performed. In this section, we re-derive M-SBL using a variational procedure that highlights the similarities and differences between the MAP-based M-Jeffreys and the ARD-based M-SBL. The methodology draws on previous work in [53].

To begin, we assume the standard likelihood model from (12) and hypothesize a generalized sparse prior $\mathcal{H}$ that includes the M-Jeffreys prior as a special case. Specifically, for the ith row of W we adopt the distribution

$$p(\boldsymbol{w}_{i\cdot};\mathcal{H}) \triangleq C\left(b + \frac{\|\boldsymbol{w}_{i\cdot}\|_2^2}{2}\right)^{-(a+L/2)} \tag{29}$$

where a, b, and C are constants. Such a prior favors rows with zero norm (and, therefore, all zero elements) owing to the sharp peak at zero (assuming b is small) and heavy tails, the trademarks of a sparsity-inducing prior. The row priors are then multiplied together to form the complete prior $p(W;\mathcal{H})$. While certainly other norms could be substituted in place of the ℓ_2, this selection (as well as the inclusion of the factor L) was made to facilitate the analysis below.

As occurs with the many of the MAP methods described in Section II, the resulting joint density $p(W,T;\mathcal{H}) = p(T\,|\,W)p(W;\mathcal{H})$ is saddled with numerous local peaks, and, therefore, mode finding should be avoided. However, perhaps there is a better way to utilize a posterior distribution than simply searching for the mode. From a modern Bayesian perspective, it has been argued that modes are misleading in general, and that only areas of significant posterior *mass* are meaningful [29]. In the case of highly sparse priors, mode finding is easily lead astray by spurious posterior peaks, but many of these peaks either reflect comparatively little mass or very misleading mass such as the heavy peak at $W = 0$ that occurs with M-Jeffreys. Consequently, here we advocate an alternative strategy that is sensitive only to regions with posterior mass that likely reflects $W_{\rm gen}$. The goal is to model the problematic $p(W,T;\mathcal{H})$ with an approximating distribution $p(W,T;\hat{\mathcal{H}})$ that does the following:

1) captures the significant mass of the full posterior, which we assume reflects the region where the weights $W_{\rm gen}$ reside;
2) ignores spurious local peaks as well as degenerate solutions, such as $W = 0$, where possible;
3) maintains easily computable moments, e.g., $\mathrm{E}[W\,|\,T;\hat{\mathcal{H}}]$ can be analytically computed to obtain point estimates of the unknown weights.

To satisfy Property 1, it is natural to select $\hat{\mathcal{H}}$ by minimizing the sum of the misaligned mass, i.e.,

$$\min_{\hat{\mathcal{H}}} \int |p(W,T;\mathcal{H}) - p(W,T;\hat{\mathcal{H}})|dW. \tag{30}$$

The ultimate goal here is to choose a family of distributions rich enough to accurately model the true posterior, at least in the regions of interest (Property 1), but coarse enough such that most spurious peaks will naturally be ignored (Property 2). Furthermore, this family must facilitate both the difficult optimization (30), as well as subsequent inference, i.e., computation of the posterior mean (Property 3). In doing so, we hope to avoid some of the troubles that befall the MAP-based methods.

Given a cumbersome distribution, sparse or otherwise, variational methods and convex analysis can be used to construct sets of simplified approximating distributions with several desirable properties [27]. In the present situation, this methodology can be used to produce a convenient family of unimodal approximations, each member of which acts as a strict lower bound on $p(W,T;\mathcal{H})$ and provides of useful means of dealing with the absolute value in (30). The quality of the approximation in a given region of $p(W,T;\mathcal{H})$ depends on which member of this set is selected.

We note that variational approaches take on a variety of forms in the context of Bayesian learning. Here, we will draw on the well-established practice of lower bounding intractable distributions using convex duality theory [27]. We do not address the alternative variational technique of forming a factorial approximation that minimizes a free-energy-based cost function [1], [2]. While these two strategies can be related in certain settings [35], this topic is beyond the scope of the current work.

The process begins by expressing the prior $p(W;\mathcal{H})$ in a dual form that hinges on a set of variational hyperparameters. By extending convexity results from [53], we arrive at

$$p(\boldsymbol{w}_{i\cdot};\mathcal{H}) = \max_{\gamma_i \geq 0} \exp\left(-\frac{b}{\gamma_i}\right)\gamma_i^{-a} \times \prod_{j=1}^{L}(2\pi\gamma_i)^{-1/2}\exp\left(-\frac{w_{ij}^2}{2\gamma_i}\right). \tag{31}$$

Details are contained in [57]. When the maximization is dropped, we obtain the rigorous lower bound

$$p(\boldsymbol{w}_{i\cdot};\mathcal{H}) \geq p(\boldsymbol{w}_{i\cdot};\hat{\mathcal{H}}) \triangleq \exp\left(-\frac{b}{\gamma_i}\right)\gamma_i^{-a}\mathcal{N}(0,\gamma_i I) \tag{32}$$

[8] The actual number, for any given p, is dependent on $W^{\rm MN}$ and λ.

which holds for all $\gamma_i \geq 0$. By multiplying each of these lower bounding row priors, we arrive at the full approximating prior $p(W;\hat{\mathcal{H}})$ with attendant hyperparameters $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_M]^T \in \mathbb{R}_+^M$. Armed with this expression, we are positioned to minimize (30) using $\hat{\mathcal{H}}$ selected from the specified set of variational approximations. Since $p(W,T;\hat{\mathcal{H}}) \leq p(W,T;\mathcal{H})$ as a result of (32), this process conveniently allows us to remove the absolute value, leading to the simplification

$$\min_{\hat{\mathcal{H}}} \int p(T \mid W)|p(W;\mathcal{H}) - p(W;\hat{\mathcal{H}})|dW = \min_{\hat{\mathcal{H}}} - \int p(T \mid W)p(W;\hat{\mathcal{H}})dW \quad (33)$$

where each candidate hypothesis $\hat{\mathcal{H}}$ is characterized by a different $\boldsymbol{\gamma}$ vector. Using (32) and (12), the constituent integral of (33) can be analytically evaluated as before, leading to the cost function

$$\mathcal{L}(\boldsymbol{\gamma}; a, b) \triangleq \mathcal{L}(\boldsymbol{\gamma}) + 2\sum_{i=1}^{M}\left(\frac{b}{\gamma_i} + a\log\gamma_i\right). \quad (34)$$

For arbitrary $a, b > 0$, (34) represents a multiple response extension of the generalized SBL cost function from [48] that, while appropriate for other circumstances, does not produce strictly sparse representations [53]. However, when $a, b \to 0$, this expression reduces to $\mathcal{L}(\boldsymbol{\gamma})$; the approximate distribution and subsequent weight estimate that emerge are, therefore, equivalent to M-SBL, only now we have the added interpretation afforded by the variational perspective.

For example, the specific nature of the relationship between M-SBL and M-Jeffreys can now be readily clarified. With $a, b \to 0, p(W;\mathcal{H})$ equals the M-Jeffreys prior up to an exponential factor of L. From a practical standpoint, this extra factor is inconsequential since it can be merged into the tradeoff parameter λ after the requisite $-\log(\cdot)$ transformation has been applied. Consequently, M-Jeffreys and M-SBL are effectively based on an identical prior distribution and, therefore, an identical posterior as well. The two are only distinguished by the manner in which this posterior is handled. One searches directly for the mode. The other selects the mean of a tractable approximate distribution that has been manipulated to align with the significant mass of the full posterior. Additionally, while ARD methods have been touted for their sensitivity to posterior mass, the exact relationship between this mass and the ARD estimation process has typically not been quantified. Here that connection is made explicit.

Empirical and theoretical results from previous sections lend unequivocal support that the ARD route is much preferred. A intuitive explanation is as follows: M-Jeffreys displays a combinatorial number of locally minimizing solutions that can substantially degrade performance. For example, there is the huge degenerate (and globally optimal) peak at $W = 0$ as discussed in the Appendix. Likewise, many other undesirable peaks exist with $d(W) > 0$. For example, M such peaks exist with

$$d(W) = 1, \quad \binom{M}{2}$$

peaks with $d(W) = 2$, and so on. In general, when any subset of weights go to zero, we are necessarily in the basin of a minimum with respect to these weights from which we cannot escape. Therefore, if too many weights (or the wrong weights) converge to zero, there is no way to retreat to a more appropriate solution.

Returning to M-SBL, we know that the full posterior distribution with which we begin is identical. The crucial difference is that, instead of traversing this improper probability density in search of a sufficiently "nonglobal" extremum (or mode), we instead explore a restricted space of posterior mass. A substantial benefit of this approach is that there is no issue of getting stuck at a point such as $W = 0$; at any stable fixed point $\boldsymbol{\gamma}^*$, we can never have $\mathcal{M}^* = 0$. This occurs because, although the *full* distribution may place mass in the neighborhood of zero, the class of approximate distributions as defined by $p(W,T;\hat{\mathcal{H}})$ in general will not (unless the likelihood is maximized at zero, in which case the solution $W = 0$ is probably correct). Likewise, a solution with $d(W)$ small is essentially impossible unless $d(W_{\text{gen}})$ is also small, assuming σ^2 has been set to a reasonable value. In general, there is much less tendency of indiscriminately shrinking important weights to zero and getting stuck, because these solutions display little overlap between prior and likelihood and, therefore, little probability mass. This helps to explain, for example, the results in Fig. 1(d), where M-SBL performance is uniformly superior to M-Jeffreys for all values of σ^2 and λ.

VI. Conclusion

While recent years have witnessed a tremendous amount of theoretical progress in the understanding of sparse approximation algorithms, most notably basis pursuit and orthogonal matching pursuit, there has been comparably less progress with regard to the development of new sparse approximation cost functions and algorithms. Using an empirical Bayesian perspective, we have extended the ARD/SBL framework to allow for learning maximally sparse subsets of design variables in real or complex-valued multiple response models, leading to the M-SBL algorithm. While many current methods focus on finding modes of distributions and frequently converge to unrepresentative (possibly local) extrema, M-SBL traverses a well-motivated space of probability mass.

Both theoretical and empirical results suggest that this is a useful route to solving simultaneous sparse approximation problems, often outperforming current state-of-the-art approaches. Moreover, these results provide further support for the notion that ARD, upon which SBL is based, does in fact lead to an exact sparsification (or pruning) of highly overparameterized models. While previous claims to this effect have relied mostly on heuristic arguments or empirical evidence, we have quantified the relationship between M-SBL and a specific sparsity-inducing prior and derived conditions, albeit limited, whereby maximally sparse representations will necessarily be achieved.

From a signal and image processing standpoint, we envision that M-SBL could become an integral component of many practical systems where multiple responses are available. For example, M-SBL has already been successfully employed in the realm of neuroelectromagnetic source imaging [38], [39]. These experiments are important since they demonstrate the

utility of M-SBL on a very large-scale problem, with a dictionary of size $275 \times 120\,000$ and $L = 1000$ response vectors. Because of the severe redundancy involved ($M/N > 400$) and the complexity of the required, neurophysiologically-based (and severely ill-conditioned) dictionary, it seems likely that the ability of M-SBL to avoid local minima in the pursuit of highly sparse representations is significant. In any event, neuroelectromagnetic imaging appears to be an extremely worthwhile benchmark for further development and evaluation of simultaneous sparse approximation algorithms.

APPENDIX
RELATING M-JEFFREYS AND M-FOCUSS

There exists an interesting relationship between the implicit priors of M-Jeffreys and M-FOCUSS. To see this, consider the slightly modified cost function

$$F_p(W) \triangleq \|T - \Phi W\|_{\mathcal{F}}^2 + \frac{\lambda'}{p} \sum_{i=1}^{M} \|\boldsymbol{w}_i.\|_2^p - \frac{\lambda'}{p} \tag{35}$$

where we have set λ equal to some λ'/p and subtracted a constant term, which does not change the topography. M-FOCUSS is capable of minimizing this cost function for arbitrary p, including the limit as $p \longrightarrow 0$. This limiting case is elucidated by the relationship

$$\lim_{p \to 0} \frac{1}{p} \left(\|\boldsymbol{w}_i.\|_2^p - 1 \right) = \log \|\boldsymbol{w}_i.\|_2. \tag{36}$$

By applying this result for all i, we arrive at the limiting cost function

$$\lim_{p \to 0} F_p(W) = \|T - \Phi W\|_{\mathcal{F}}^2 + \lambda' \sum_{i=1}^{M} \log \|\boldsymbol{w}_i.\|_2 \tag{37}$$

which is identical to the M-Jeffreys cost function. This demonstrates why M-Jeffreys should be considered a special case of M-FOCUSS and clarifies why the update rules are related even though they were originally derived with different considerations in mind.

In arriving at this association, we have effectively assumed that the regularizing component of the cost function (35) has grown arbitrarily large. This discounts the quality-of-fit component, leading to the globally optimal, yet degenerate solution $W = 0$. However, curiously, M-Jeffreys and equivalently M-FOCUSS (with $\lambda = \lambda'/p, p \to 0$) still do consistently produce sparse representations that nonetheless retain the desirable property $T \approx \Phi W$.

In fact, any success achieved by these algorithms can be attributed to their ability to find appropriate, explicitly nonglobal, local minima. This is not unlike the situation that occurs when using the EM algorithm to fit the parameters of a Gaussian mixture model for density estimation. In this case, the cost function may always be driven to infinity by collapsing a single mixture component around a single data point. This is accomplished by making the component mean equal to the value of the data point and allowing the component variance to converge to zero. Clearly, the desired solution is *not* the globally optimal one and heuristics must be adopted to avoid getting stuck [43].

REFERENCES

[1] H. Attias, "A variational Bayesian framework for graphical models," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2000, vol. 12.

[2] M. J. Beal and Z. Ghahrarmani, "The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures," in *Bayesian Statistics*. Oxford, U.K.: Oxford Univ. Press, 2002, vol. 7.

[3] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. New York: Springer-Verlag, 1985.

[4] L. Breiman, "Better subset regression using the nonnegative garrote," *Technometrics*, vol. 37, no. 4, pp. 373–384, 1995.

[5] J. Chen and X. Huo, "Sparse representations for multiple measurement vectors (MMV) in an overcomplete dictionary," in *Proc. IEEE Int. Conf. Accoustics, Speech, Signal Processing*, Mar. 2005, vol. 4, pp. 257–260.

[6] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by Basis Pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1999.

[7] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2477–2488, Jul. 2005.

[8] A. P. Dempster, D. B. Rubin, and R. K. Tsutakawa, "Estimation in covariance components models," *J. Amer. Statist. Assoc.*, vol. 76, no. 374, pp. 341–353, Jun. 1981.

[9] D. L. Donoho, "On minimum entropy segmentation," in *Wavelets: Theory, Algorithms, and Applications*. New York: Academic, 1994, pp. 233–269.

[10] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution," Stanford Univ. Tech. Rep., Stanford, CA , Sep. 2004.

[11] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proc. Nat. Acad. Sci.*, vol. 100, no. 5, pp. 2197–2202, Mar. 2003.

[12] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2845–2862, Jul. 2001.

[13] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[14] I. J. Fevrier, S. B. Gelfand, and M. P. Fitz, "Reduced complexity decision feedback equalization for multipath channels with large delay spreads," *IEEE Trans. Commun.*, vol. 47, no. 6, pp. 927–937, Jun. 1999.

[15] M. A. T. Figueiredo, "Adaptive sparseness using Jeffreys prior," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2002, vol. 14, pp. 697–704.

[16] J. J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1341–1344, Jun. 2004.

[17] H. Gao, "Wavelet shrinkage denoising using the nonnegative garrote," *J. Comput. Graph. Statist.*, vol. 7, no. 4, pp. 469–488, 1998.

[18] I. F. Gorodnitsky, J. S. George, and B. D. Rao, "Neuromagnetic source imaging with FOCUSS: A recursive weighted minimum norm algorithm," *J. Electroencephalogr. Clin. Neurophysiol.*, vol. 95, no. 4, pp. 231–251, Oct. 1995.

[19] 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, no. 3, pp. 600–616, Mar. 1997.

[20] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Inf. Theory*, vol. 49, no. 12, pp. 3320–3325, Dec. 2003.

[21] D. A. Harville, "Bayesian inference for variance components using only error contrasts," *Biometrika*, vol. 61, pp. 383–385, 1974.

[22] D. A. Harville, "Maximum likelihood approaches to variance component estimation and to related problems," *J. Amer. Statist. Assoc.*, vol. 72, pp. 320–338, Jun. 1977.

[23] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer, 2001.

[24] A. Hillebrand, G. Barnes, I. Holliday, and G. Harding, "Comparison of a modified FOCUSS algorithm with constrained and unconstrained dipole solutions on a simulated cortical surface," in *Proc. 12th Int. Conf. Biomagnetism*, 2000, pp. 753–756.

[25] B. D. Jeffs, "Sparse inverse solution methods for signal and image processing applications," in *Proc. IEEE Int. Conf. Accoustics, Speech, Signal Processing*, May 1998, vol. 3, pp. 1885–1888.

[26] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, 1992.

[27] M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, no. 2, pp. 183–233, 1999.
[28] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[29] D. J. C. MacKay, "Bayesian interpolation," *Neural Comput.*, vol. 4, no. 3, pp. 415–447, 1992.
[30] D. J. C. MacKay, "Bayesian non-linear modelling for the energy prediction competition," *ASHRAE Trans.*, vol. 100, no. 2, pp. 1053–1062, 1994.
[31] D. M. Malioutov, M. Çetin, and A. S. Willsky, "Sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010–3022, Aug. 2005.
[32] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.
[33] P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors," *IEEE Trans. Inf. Theory*, vol. 45, no. 3, Apr. 1999.
[34] R. M. Neal, *Bayesian Learning for Neural Networks*. New York: Springer-Verlag, 1996.
[35] J. A. Palmer, D. P. Wipf, K. Kreutz-Delgado, and B. D. Rao, "Variational EM algorithms for non-Gaussian latent variable models," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2006, vol. 18.
[36] C. Phillips, J. Mattout, M. D. Rugg, P. Maquet, and K. J. Friston, "An empirical Bayesian solution to the source reconstruction problem in EEG," *NeuroImage*, vol. 24, pp. 997–1011, Jan. 2005.
[37] J. W. Phillips, R. M. Leahy, and J. C. Mosher, "MEG-based imaging of focal neuronal current sources," *IEEE Trans. Med. Imag.*, vol. 16, no. 3, pp. 338–348, Mar. 1997.
[38] R. R. Ramírez, "Neuromagnetic source imaging of spontaneous and evoked human brain dynamics," Ph.D. dissertation, New York Univ., New York, May 2005.
[39] R. R. Ramírez and S. Makeig, "Neuroelectromagnetic source imaging using multiscale geodesic neural bases and sparse Bayesian learning," presented at the Human Brain Mapping, 12th Annu. Meeting, Florence, Italy, Jun. 2006.
[40] B. D. Rao and K. Kreutz-Delgado, "Basis selection in the presence of noise," in *Proc. 32nd Asilomar Conf. Signals, Systems, Computers*, Nov. 1998, vol. 1, pp. 752–756.
[41] B. D. Rao, K. Engan, S. F. Cotter, J. Palmer, and K. Kreutz-Delgado, "Subset selection in noise based on diversity measure minimization," *IEEE Trans. Signal Process.*, vol. 51, no. 3, pp. 760–770, Mar. 2003.
[42] C. E. Rasmussen, "Evaluation of Gaussian processes and other methods for non-linear regression," Ph.D. dissertation, Grad. Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 1996.
[43] B. D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge, U.K.: Cambridge Univ. Press, 1996.
[44] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ: Princeton Univ. Press, 1970.
[45] J. G. Silva, J. S. Marques, and J. M. Lemos, "Selecting landmark points for sparse manifold learning," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2006, vol. 18.
[46] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. Upper Saddle River, NJ: Prentice-Hall, 1997.
[47] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Roy. Statist. Soc.*, vol. 58, no. 1, pp. 267–288, 1996.
[48] M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.
[49] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.
[50] J. A. Tropp, "Algorithms for simultaneous sparse approximation. Part II: Convex relaxation," *EURASIP J. Signal Process.*, vol. 86, pp. 589–602, Apr. 2006.
[51] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *EURASIP J. Signal Process.*, vol. 86, pp. 572–588, Apr. 2006.
[52] M. B. Wakin, M. F. Duarte, S. Sarvotham, D. Baron, and R. G. Baraniuk, "Recovery of jointly sparse signals from a few random projections," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2006, vol. 18.
[53] D. P. Wipf, J. A. Palmer, and B. D. Rao, "Perspectives on sparse Bayesian learning," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2004, vol. 16.
[54] D. P. Wipf and B. D. Rao, "Probabilistic analysis for basis selection via ℓ_p diversity measures," presented at the IEEE Int. Conf. Accoustics, Speech, Signal Processing, May 2004.
[55] D. P. Wipf and B. D. Rao, "ℓ_0-norm mimimization for basis selection," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2005, vol. 17.
[56] D. P. Wipf and B. D. Rao, "Finding sparse representations in multiple response models via Bayesian learning," presented at the Workshop on Signal Processing with Adaptive Sparse Structured Representations, Rennes, France, Nov. 2005.
[57] D. P. Wipf, "Bayesian Methods for finding sparse representations," Ph.D. dissertation, Univ. California, San Diego, 2006.

David P. Wipf received the B.S. degree in electrical engineering from the University of Virginia, Charlottesville, and the M.S. degree from the University of California, San Diego, in 2003.

Currently, he is a Postdoctoral Fellow in the Biomagnetic Imaging Lab, University of California, San Francisco. His research involves the development and analysis of Bayesian algorithms for functional brain imaging and sparse coding.

Bhaskar D. Rao (F'00) received the B.Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology, Kharagpur, in 1979, and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, in 1981 and 1983, respectively.

Since 1983, he has been with the University of California, San Diego, where he is currently a Professor in the Electrical and Computer Engineering Department. His interests are in the areas of digital signal processing, estimation theory, and optimization theory, with applications to digital communications, speech signal processing, and human–computer interactions.

Dr. Rao was a member of the IEEE Statistical Signal and Array Processing Technical Committee and is currently a member of the IEEE Signal Processing Theory and Methods Technical Committee.