Sparse Bayesian Learning for Basis Selection

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

***Abstract*—Sparse Bayesian learning (SBL) and specifically relevance vector machines have received much attention in the machine learning literature as a means of achieving parsimonious representations in the context of regression and classification. The methodology relies on a parameterized prior that encourages models with few nonzero weights. In this paper, we adapt SBL to the signal processing problem of basis selection from overcomplete dictionaries, proving several results about the SBL cost function that elucidate its general behavior and provide solid theoretical justification for this application. Specifically, we have shown that SBL retains a desirable property of the ℓ_0-norm diversity measure (i.e., the global minimum is achieved at the maximally sparse solution) while often possessing a more limited constellation of local minima. We have also demonstrated that the local minima that do exist are achieved at sparse solutions. Later, we provide a novel interpretation of SBL that gives us valuable insight into why it is successful in producing sparse representations. Finally, we include simulation studies comparing sparse Bayesian learning with Basis Pursuit and the more recent FOCal Underdetermined System Solver (FOCUSS) class of basis selection algorithms. These results indicate that our theoretical insights translate directly into improved performance.**

***Index Terms*—Basis selection, diversity measures, linear inverse problems, sparse Bayesian learning, sparse representations.**

I. Introduction

Sparse signal representations from overcomplete dictionaries have found increasing relevance in a large number of application domains [1]–[3]. Moreover, attaining such representations is tantamount to solving regularized linear inverse problems that have far-reaching significance in signal processing, compression, and feature extraction. Example applications include biomagnetic imaging [4], channel equalization [5]–[7], bandlimited extrapolation and spectral estimation [8], [9], direction-of-arrival estimation [10], functional approximation [11]–[13], echo cancellation [14], [15], and image restoration [16]. Consequently, deeper insight into these issues is of both theoretical and practical importance.

The canonical form of this problem is given by

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

where $\Phi \in \Re^{N \times M}$ is a matrix whose columns represent a possibly overcomplete basis (i.e., $\text{rank}(\Phi) = N$ and $M > N$), $\boldsymbol{w} = [w_1, \ldots, w_M]^T$ is the vector of weights to be learned, $\boldsymbol{\epsilon}$ is noise, and $\boldsymbol{t} = [t_1, \ldots, t_N]^T$ is a vector of targets. In this vein, we seek weight vectors whose entries are predominantly zero while still allowing us to accurately approximate $\boldsymbol{t}$. This is equivalent to representing $\boldsymbol{t}$ with a minimal number of basis vectors.[1]

Recently, a sparse Bayesian learning (SBL) framework has been derived to find robust solutions to problems like (1) in the context of regression and classification [17]–[19]. A key feature of this development that is germane to the basis selection problem is the incorporation of a parameterized prior on the weights that encourages sparsity in representation, i.e., few nonzero weights. In addition, when Φ is square and formed from a positive-definite kernel function, we obtain the relevance vector machine (RVM), which is a Bayesian competitor of the support vector machine (SVM) with several significant advantages [18].

Unlike popular methods for basis selection, it is not immediately transparent how the SBL cost function leads to sparse representations in practice, nor have many of the theoretical details of this relatively new paradigm been fleshed out, especially those most relevant to basis selection. In this paper, we prove a collection of results about the SBL cost function that elucidate its general behavior and provide solid theoretical justification for adapting it to basis selection tasks. Furthermore, we adapt an iterative SBL algorithm to perform basis selection and empirically substantiate the algorithm by comparing it with current methods.

A. Current Basis Selection Methods

The most successful current basis selection algorithms [1], [3], [20] essentially perform least squares regression with the addition of a fixed, regularizing weight prior of the form

$$p(\boldsymbol{w}) \sim \exp\left(-\sum_{i=1}^{M} |w_i|^p\right) \tag{2}$$

where $p \in [0, 1]$. Such a prior has been shown to encourage sparsity in many situations because of the heavy tails and sharp peak at zero (i.e., the prior is super-Gaussian). Moreover, as we allow $p \to 0$, the exponent of this prior approaches an ℓ_0-norm, i.e., a count of the number of nonzero entries in $\boldsymbol{w}$ defined as

$$\|\boldsymbol{w}\|_0 = \sum_{i=1}^{M} \mathbf{1}\left(|w_i| = 0\right) \tag{3}$$

Manuscript received July 2, 2003; revised October 21, 2003. This research was supported by Nissan under DiMI Grant 22-8376. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Alfred O. Hero.

The authors are with the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92092 USA (dwipf@ucsd.edu; brao@ece.ucsd.edu).

Digital Object Identifier 10.1109/TSP.2004.831016

[1] Some authors refer to this process of selecting basis vectors as "subset selection," reserving "basis selection" to refer to the process of selecting a full spanning basis in $\Re^N$: the signal dimension. In contrast, we will simply refer to basis selection as the task of finding a minimal number of basis vectors (typically much less than N) needed to represent the signal of interest.

where $\mathbf{1}(\cdot)$ denotes the indicator function. Given this prior, maximum *a posteriori* (MAP) solutions to (1) are formulated as

$$\begin{aligned}\boldsymbol{w}_{\text{MAP}} &= \arg\max_{\boldsymbol{w}} p(\boldsymbol{w}|\boldsymbol{t}) \\ &= \arg\min_{\boldsymbol{w}} -\log p(\boldsymbol{t}|\boldsymbol{w}) - \log p(\boldsymbol{w}) \\ &= \arg\min_{\boldsymbol{w}} \lambda\|\boldsymbol{t}-\Phi\boldsymbol{w}\|^2 + \sum_{i=1}^{M}|w_i|^p \end{aligned} \tag{4}$$

where we have assumed a Gaussian likelihood model, and λ represents a trade-off parameter balancing sparsity with quality of fit [1], [20]. In the absence of noise (or as $\lambda \to \infty$), we instead seek solutions of the form

$$\boldsymbol{w} = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{M}|w_i|^p, \quad \text{s.t. } \boldsymbol{t} = \Phi\boldsymbol{w} \tag{5}$$

i.e., the log prior becomes the objective function over the constraint surface given by $\boldsymbol{t} = \Phi\boldsymbol{w}$. This is very similar to a procedure originally outlined in [21] based on work in [22].

When $p = 1$, we can obtain the standard Basis Pursuit cost function [1], which finds the weight vector $\boldsymbol{w}$ satisfying $\boldsymbol{t} = \Phi\boldsymbol{w}$ with minimum ℓ_1-norm. In contrast, the FOCUSS algorithm [2], [3] typically maintains a value of $p \in [0, 1)$ (i.e., p strictly less than one).[2] Both the FOCUSS and Basis Pursuit algorithms have been adapted to handle noiseless and noisy conditions, with the later condition requiring the selection of the trade-off parameter λ given in (4).

While both the Basis Pursuit and FOCUSS algorithms are marked by demonstrable successes, each is hampered to some extent by a significant shortcoming. With Basis Pursuit, we benefit from a cost function devoid of local minima, and furthermore, we can typically guarantee convergence to the global minimum.[3] The problem, however, is that the global minimum of this cost function does not necessarily coincide with the sparsest solutions to (1) (except in the special case where the optimal solution is sufficiently sparse; see, e.g., [25]). We will refer to this misalignment as *structural error*.

Conversely, the cost function employed by FOCUSS has many local minima. However, as we allow $p \to 0$, the correlation between the global minimum of this cost function and the sparsest solutions to (1) approaches certainty since we are effectively now performing ℓ_0-norm minimization.[4] While we no longer experience structural errors in this scenario, we frequently converge to suboptimal local minima termed *convergence errors*.

Because of the limitations of both of these algorithms, there exists room for alternative approaches to basis selection. In this paper, we will demonstrate that the SBL cost function, like the ℓ_0-norm, prevents any structural errors (at least in the absence of noise) while possessing potentially far fewer local minima than FOCUSS. We later provide examples where this translates directly into improved performance marked by no structural errors and fewer convergence errors.

B. Basis Selection versus Regression

In many ways, basis selection can be thought of as regression with the additional assumption that regularization must be with respect to sparsity of representation. Nonetheless, there remains one fundamental difference: While the ultimate goal of regression is to minimize generalization error (i.e., error on evaluation data not available during model training), basis selection is primarily concerned with finding sparse representations of $\boldsymbol{t}$ itself. This distinction is reflected in the results of this paper, which emphasize sparsity and make no claims about generalization performance. However, for the interested reader, there is a known relationship between sparsity of fit and generalization performance, as discussed in [19].

C. Organization of Paper

The organization of this paper is as follows. In Section II, we present a simplified derivation of SBL and our accommodations for basis selection. In Sections III and IV, we will prove a collection of results pertaining to the global and local minima of the SBL cost function. We then recast this method from a variational perspective in Section V, demonstrating a rigorous association between SBL and a sparsity inducing weight prior. This alternate interpretation gives us valuable insight into why it is effective in producing sparse solutions to (1). Finally, Section VI contains the results from empirical comparisons of SBL with Basis Pursuit and FOCUSS.

II. Sparse Bayesian Learning

Like current basis selection methods, SBL assumes the Gaussian likelihood model

$$p(\boldsymbol{t}|\boldsymbol{w};\sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{t}-\Phi\boldsymbol{w}\|^2\right). \tag{6}$$

Obtaining maximum likelihood estimates for $\boldsymbol{w}$ under these conditions is equivalent to finding the minimum ℓ_2-norm solution to (1). Such solutions are well known to produce nonsparse representations. To alleviate this problem, we must incorporate some form of weight prior that encourages sparsity. We should note that modern Bayesian methodology does not attempt to select the "right" priors nor employ the FOCUSS/Basis Pursuit approach of selecting a fixed, sparsity-inducing prior. Rather, many different priors can be invoked, corresponding to different hypothesis about underlying truth. These hypothesis can be empirically compared by evaluating the Bayesian evidence for each model prior [23].

Suppose we hypothesize two different priors denoted $\mathcal{H}_1$ and $\mathcal{H}_2$. Our goal is to compare these hypotheses with respect to observed data $\boldsymbol{t}$. This can be accomplished by evaluating the evidence $p(\boldsymbol{t};\mathcal{H}_i,\sigma^2)$ for each prior, which can be computed by marginalizing over the weights via

$$p(\boldsymbol{t};\mathcal{H}_i,\sigma^2) = \int p(\boldsymbol{t}|\boldsymbol{w};\sigma^2)p(\boldsymbol{w};\mathcal{H}_i)d\boldsymbol{w} \tag{7}$$

[2] Actually, p can assume any value less than zero as well, but performance is poor in this region [3]. In addition, we exclude the $p = 1$ case from the FOCUSS domain mainly to avoid confusion with Basis Pursuit.

[3] Basis Pursuit can be cast as a linear programming problem.

[4] It should be noted that p need not equal zero exactly to obtain this correlation. As described in [21], there exists a p' sufficiently small such that, for all $0 < p < p'$, the global minimum will represent the sparsest solution, i.e., minimum ℓ_0-norm solution. Unfortunately, however, p' is dependent on Φ, and $\boldsymbol{t}$ and can be arbitrarily small. Moreover, there is no way to determine its value without *a priori* knowledge of the global solution.

where $p(\boldsymbol{t}|\boldsymbol{w};\sigma^2)$ is the likelihood model, and $p(\boldsymbol{w};\mathcal{H}_i)$ denotes the weight prior under hypothesis $\mathcal{H}_i$. If $p(\boldsymbol{t};\mathcal{H}_1,\sigma^2) > p(\boldsymbol{t};\mathcal{H}_2,\sigma^2)$, we chose $\mathcal{H}_1$ and vice versa. This procedure can also be employed to estimate the noise variance σ^2, i.e., we chose σ^2 to maximize the evidence as well.

A. Model Prior Formulation

In contrast to FOCUSS and Basis Pursuit, which assume a fixed prior, SBL estimates a parameterized prior from the data. The parametric form of the SBL weight prior is given by

$$p(\boldsymbol{w};\boldsymbol{\gamma}) = \prod_{i=1}^{M} (2\pi\gamma_i)^{-\frac{1}{2}} \exp\left(-\frac{w_i^2}{2\gamma_i}\right) \tag{8}$$

where $\boldsymbol{\gamma} = [\gamma_1,\ldots,\gamma_M]^T$ is a vector of M hyperparameters controlling the prior variance of each weight. These hyperparameters (along with the error variance σ^2 if necessary) can be estimated from the data by marginalizing over the weights and then performing ML optimization. The marginalized pdf is given by

$$\begin{aligned} p(\boldsymbol{t};\boldsymbol{\gamma},\sigma^2) &= \int p(\boldsymbol{t}|\boldsymbol{w};\sigma^2)p(\boldsymbol{w};\boldsymbol{\gamma})d\boldsymbol{w} \\ &= (2\pi)^{-\frac{N}{2}}|\Sigma_t|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\boldsymbol{t}^T\Sigma_t^{-1}\boldsymbol{t}\right] \end{aligned} \tag{9}$$

where $\Sigma_t \triangleq \sigma^2 I + \Phi\Gamma\Phi^T$, and we have introduced the notation $\Gamma \triangleq \text{diag}(\boldsymbol{\gamma})$.[5] This procedure is referred to as evidence maximization or type-II maximum likelihood [17].

B. Algorithm Development

For fixed values of the hyperparameters governing the prior, the posterior density of the weights is Gaussian [18], i.e.,

$$p(\boldsymbol{w}|\boldsymbol{t};\boldsymbol{\gamma},\sigma^2) = \mathcal{N}(\boldsymbol{\mu},\Sigma_w) \tag{10}$$

with $\boldsymbol{\mu} = \sigma^{-2}\Sigma_w\Phi^T\boldsymbol{t}$ and $\Sigma_w = (\sigma^{-2}\Phi^T\Phi + \Gamma^{-1})^{-1}$. Thus, the onus remains in estimating $\boldsymbol{\gamma}$ and σ^2 via type-II maximum likelihood. Once we have these values, we choose as our weights the $\boldsymbol{w}$ satisfying

$$\begin{aligned} \boldsymbol{w} &= \boldsymbol{\mu} \\ &= \left(\Phi^T\Phi + \sigma^2_{\text{ML}}\Gamma^{-1}_{\text{ML}}\right)^{-1}\Phi^T\boldsymbol{t}. \end{aligned} \tag{11}$$

To find $\boldsymbol{\gamma}_{\text{ML}}$ and σ^2_{ML}, we employ the EM algorithm to maximize $p(\boldsymbol{t};\boldsymbol{\gamma},\sigma^2)$.[6] This is equivalent to minimizing $-\log p(\boldsymbol{t};\boldsymbol{\gamma},\sigma^2)$, giving the effective SBL cost function

$$L = \log|\Sigma_t| + \boldsymbol{t}^T\Sigma_t^{-1}\boldsymbol{t}. \tag{12}$$

The actual EM formulation proceeds by treating the weights $\boldsymbol{w}$ as hidden variables and then maximizing

$$\text{E}_{\boldsymbol{w}|\boldsymbol{t};\boldsymbol{\gamma},\sigma^2}[p(\boldsymbol{t},\boldsymbol{w};\boldsymbol{\gamma},\sigma^2)]$$

where $p(\boldsymbol{t},\boldsymbol{w};\boldsymbol{\gamma},\sigma^2) = p(\boldsymbol{t}|\boldsymbol{w};\sigma^2)p(\boldsymbol{w};\boldsymbol{\gamma})$ represents the likelihood of the complete data $\{\boldsymbol{w},\boldsymbol{t}\}$. We may then compute the following for the kth iteration:

$$E\ Step: \text{E}_{\boldsymbol{w}|\boldsymbol{t};\boldsymbol{\gamma}^{(k)},\sigma^2}\left[w_i^2\right] = (\Sigma_w)_{i,i} + \mu_i^2. \tag{13}$$

$$\begin{aligned} M\ Step: \gamma_i^{(k+1)} &= \arg\max_{\gamma_i \geq 0} \text{E}_{\boldsymbol{w}|\boldsymbol{t};\boldsymbol{\gamma}^{(k)},\sigma^2}\left[p(\boldsymbol{t},\boldsymbol{w};\boldsymbol{\gamma},\sigma^2)\right] \\ &= \arg\max_{\gamma_i \geq 0} \text{E}_{\boldsymbol{w}|\boldsymbol{t};\boldsymbol{\gamma}^{(k)},\sigma^2}\left[p(\boldsymbol{w};\boldsymbol{\gamma})\right] \\ &= \text{E}_{\boldsymbol{w}|\boldsymbol{t};\boldsymbol{\gamma}^{(k)},\sigma^2}\left[w_i^2\right]. \end{aligned} \tag{14}$$

Likewise, an update rule for σ^2 can be simply incorporated during the M-step [18],

$$(\sigma^2)^{(k+1)} = \frac{\|\boldsymbol{t}-\Phi\boldsymbol{\mu}\|^2 + (\sigma^2)^{(k)}\sum_{i=1}^{M}\left[1-\left(\gamma_i^{(k)}\right)^{-1}(\Sigma_w)_{i,i}\right]}{N}. \tag{15}$$

Interestingly, upon convergence, we find that many of the γ_i's are driven to zero, effectively forcing the associated weights at the mean of (10) to zero. In other words, if $\gamma_i = 0$, then $p(w_i;\gamma_i = 0) = \delta(w_i)$, which will dominate the likelihood term and force the posterior probability to satisfy

$$\text{Prob}(w_i = 0|\boldsymbol{t};\gamma_i = 0,\sigma^2) = 1. \tag{16}$$

In its current form, SBL requires the inversion of the $M \times M$ matrix Σ_w: an $O(M^3)$ operation. This can be problematic since in cases of extreme overcompleteness, M can be quite large. To alleviate this problem, we compute Σ_w as

$$\begin{aligned} \Sigma_w &= (\sigma^{-2}\Phi^T\Phi + \Gamma^{-1})^{-1} \\ &= \Gamma - \Gamma\Phi^T(\sigma^2 I + \Phi\Gamma\Phi^T)^{-1}\Phi\Gamma \\ &= \Gamma - \Gamma\Phi^T\Sigma_t^{-1}\Phi\Gamma. \end{aligned} \tag{17}$$

We must now only invert the $N \times N$ matrix Σ_t, reducing the algorithm to $O(N^3)$ (like the FOCUSS algorithm), which is clearly superior when $M \gg N$. Additionally, in noiseless environments, we may want to allow $\sigma^2 \rightarrow 0$. Using straightforward results from linear algebra, we can accommodate this requirement by using the following expressions for $\boldsymbol{\mu}$ and Σ_w:

$$\boldsymbol{\mu} = \Gamma^{1/2}\left(\Phi\Gamma^{1/2}\right)^{\dagger}\boldsymbol{t} \tag{18}$$

$$\Sigma_w = \left[I - \Gamma^{1/2}\left(\Phi\Gamma^{1/2}\right)^{\dagger}\Phi\right]\Gamma \tag{19}$$

where $(\cdot)^{\dagger}$ denotes the Moore–Penrose pseudoinverse. In this formulation, it is very transparent how the sparsity profile of $\boldsymbol{\gamma}$ dictates that of $\boldsymbol{\mu}$. We also observe that all $\boldsymbol{\mu}$ are feasible, i.e., $\boldsymbol{t} = \Phi\boldsymbol{\mu}$ for all $\boldsymbol{\gamma}$. Of course, this assumes that $\boldsymbol{t}$ is in the span of the columns of Φassociated with nonzero elements in $\boldsymbol{\gamma}$; however, this will always be the case if $\boldsymbol{t}$ is in the span of Φ, and all $\boldsymbol{\gamma}$ are initialized to nonzero values.

C. Convergence Issues

By virtue of the well-known properties of the EM algorithm, SBL is globally convergent (i.e., each iteration is guaranteed to reduce the cost function until a fixed point is reached). Moreover, this guarantee includes the estimation of the noise variance σ^2: the SBL counterpart of λ. This is possible because the noise variance estimation is easily packaged with the hyperparameters during evidence maximization, as shown above.

Likewise, both Basis Pursuit and FOCUSS are globally convergent algorithms as well but only with respect to optimization of the weights themselves [3]. The trade-off parameter λ must be estimated via some other means or fixed in advance by

[5] We will sometimes use Γ and $\boldsymbol{\gamma}$ interchangeably when appropriate.

[6] An alternate formulation is also provided in [18] that has been observed to drastically speed convergence; however, unlike the EM algorithm, it does not lead to a proven descent function.

some prior knowledge. Furthermore, estimating λ via an evidence maximization procedure is not straightforward since the required integration (7) is intractable using the super-Gaussian prior from (2).

III. Analysis of Global Minima

When tasked with sparse linear inverse problems such as (1), we ideally seek cost functions whose minimization corresponds with maximally sparse solutions (at least in the noiseless case). Unfortunately, the global minima of current basis selection cost functions typically do not achieve this objective (an exception of course is FOCUSS with $p = 0$). In contrast, we will show that given certain conditions, the SBL cost function is characterized by a global minimum that can produce the maximally sparse solution at the posterior mean. Thus, in situations where the SBL algorithm fails, it is a convergence issue and not a structural one.

We will now analyze the sparsity of the global minima of (12). Before we begin, it is useful to introduce the notation $d(\boldsymbol{w}) \triangleq \|\boldsymbol{w}\|_0$. We will refer to $d(\cdot)$ as a diversity measure since it counts the number of entries in $\boldsymbol{w}$ that are greater than zero. This is in contrast to sparsity, which measures the number of weights that equal zero. Thus, we can relate the two via

$$\text{diversity} = M - \text{sparsity}. \tag{20}$$

We will now formally define a sparse solution as a set of weights $\boldsymbol{w}$, or weight prior variances $\boldsymbol{\gamma}$, that satisfy $d(\boldsymbol{w}) \leq N$ or $d(\boldsymbol{\gamma}) \leq N$. When the inequality is strict, we say that the solution is a degenerate sparse solution.

We now present a theorem that links global minima of L with sparse solutions.

Theorem 1: Let the noise term $\boldsymbol{\epsilon}$ in (1) be equal to zero, and denote the maximally sparse solution to $\boldsymbol{t} = \Phi\boldsymbol{w}$ as $\boldsymbol{w}_0$ which we assume satisfies $d(\boldsymbol{w}_0) < N$ and $\|\boldsymbol{w}_0\|_2 < \infty$. Furthermore, let $\boldsymbol{\gamma}_0$ denote a vector of prior variances such that $\boldsymbol{w}_0 = \Gamma_0^{1/2}\left(\Phi\Gamma_0^{1/2}\right)^{\dagger}\boldsymbol{t}$, where the minimum nonzero γ_i is greater than some $\delta > 0$, and $d(\boldsymbol{\gamma}_0) = d(\boldsymbol{w}_0)$. Then, the global minimum of L (with respect to $\boldsymbol{\gamma}$ and σ^2) is achieved at $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$ and $\sigma^2 = 0$.

Proof: The cost function L from (12) is composed of two terms: the logarithm of a volume (or determinant) of Σ_t and the projection of Σ_t^{-1} along $\boldsymbol{t}$, i.e., $\boldsymbol{t}^T\Sigma_t^{-1}\boldsymbol{t}$. While the later term is strictly greater than zero for all $\boldsymbol{\gamma}$ and σ^2, we can drive the former to minus infinity by reducing the volume of Σ_t to zero. As such, the minimum of L occurs whenever

$$|\Sigma_t| = |\sigma^2 I + \Phi\Gamma\Phi^T| = 0 \tag{21}$$

while maintaining some finite bound B such that

$$0 \leq \boldsymbol{t}^T\left(\sigma^2 I + \Phi\Gamma\Phi^T\right)^{-1}\boldsymbol{t} \leq B. \tag{22}$$

We will now demonstrate that $\boldsymbol{\gamma}_0$ (or, equivalently, Γ_0) satisfies these conditions. First, we observe that with $\sigma^2 = 0$ and $\Gamma = \Gamma_0$, we readily satisfy (21). This occurs because

$$\text{rank}(\Sigma_t) = \text{rank}(\Phi\Gamma_0\Phi^T) \leq d(\boldsymbol{\gamma}_0) < N. \tag{23}$$

Since $\Sigma_t \in \Re^{N\times N}$ is therefore not full rank, its volume must be zero, in accordance with (21). We now handle (22). To facilitate the analysis, we define

$$U \triangleq \Phi\Gamma_0^{\frac{1}{2}} \quad \text{and} \quad \boldsymbol{s} \triangleq \left(\Gamma_0^{\frac{1}{2}}\right)^{\dagger}\boldsymbol{w}_0. \tag{24}$$

By construction, we observe that

$$\boldsymbol{t} = \Phi\boldsymbol{w}_0 = \Phi\Gamma_0^{\frac{1}{2}}\left(\Gamma_0^{\frac{1}{2}}\right)^{\dagger}\boldsymbol{w}_0 = U\boldsymbol{s}. \tag{25}$$

We may then re-express (22) at $\Gamma = \Gamma_0$ as

$$\boldsymbol{t}^T\left(\sigma^2 I + \Phi\Gamma_0\Phi^T\right)^{-1}\boldsymbol{t} = \boldsymbol{s}^T U^T\left(\sigma^2 I + UU^T\right)^{-1}U\boldsymbol{s}. \tag{26}$$

By invoking the basic result from linear algebra

$$\lim_{\sigma^2\to 0} U^T\left(\sigma^2 I + UU^T\right)^{-1} = U^{\dagger} \tag{27}$$

we arrive at

$$\begin{aligned}\lim_{\sigma^2\to 0} & \boldsymbol{t}^T\left(\sigma^2 I + \Phi\Gamma_0\Phi^T\right)^{-1}\boldsymbol{t} \\ &= \boldsymbol{s}^T U^{\dagger} U\boldsymbol{s} \\ &= \boldsymbol{w}_0^T\Gamma_0^{\dagger}\boldsymbol{w}_0 \\ &\leq \frac{1}{\delta}\|\boldsymbol{w}_0\|_2^2. \end{aligned} \tag{28}$$

This result is bounded by assumption, completing the proof. ■

This theorem establishes that the global minimum of L is achieved at a solution such that the posterior mean, as given by (18), equals $\boldsymbol{w}_0$. Although limited to the noise-free case, this theorem is nonetheless important, since establishing positive results in such situations is typically a necessary condition for extensions to noisy domains. Furthermore, a substantial body of useful theoretical work exists pertaining to Basis Pursuit and FOCUSS in noiseless environments [1]–[3], [24], [25]; it therefore seems appropriate to flesh out comparative theoretical details of SBL. In this instance, no equivalent result to Theorem 1 exists for the other algorithms (with the exception of FOCUSS, $p = 0$), which is certainly worthwhile to know.

Additionally, we should mention that Theorem 1 contains no guarantee of uniqueness, i.e., solutions with suboptimal sparsity may also globally minimize L. We address this point as follows.

1) Assuming Φ represents the unique representation property (URP) (i.e., any subset of N columns of Φ are linearly independent), then the global minimum can only be achieved at degenerate $\boldsymbol{\gamma}$ vectors that produce degenerate sparse solutions at the posterior mean, i.e., $d(\boldsymbol{w}) < N$. Per the analysis in [2], there are a very limited number of such solutions. Additionally, degenerate sparse solutions, even if not optimal, are better than solutions with $d(\boldsymbol{w}) \geq N$.
2) In the absence of noise, Φ satisfying the URP, and given very mild conditions on the optimal weight vector $\boldsymbol{w}_0$, there will exist *no* other degenerate sparse solutions (see [29]). As such, the global minimum to L in this situation must produce $\boldsymbol{w}_0$ at the posterior mean.
3) Degenerate sparse solutions minimize L by collapsing Σ_t to a subspace of N dimensional $\boldsymbol{t}$-space. In the absence of noise, all such degenerate solutions reduce L to minus infinity, as we have already shown. However, if we fix σ^2 to some sufficiently small value greater than zero, then

these solutions are no longer equivalent with respect to minimizing L. The variational analysis in Section V sheds some light on this topic.

IV. Analysis of Local Minima

Like FOCUSS, the SBL cost function L can potentially have many local minima. That multiple minima exist should not be surprising. After all, from Theorem 1, the global minimum is achieved at the maximally sparse solution. Moreover, from [12], we know that finding the maximally sparse solution is NP-hard. Thus, any descent algorithm that claims to be devoid of local minima in this context must be suspect.

Nevertheless, to whatever degree possible, we would like to quantify similarities and differences between the local minima of FOCUSS and those of SBL (Basis Pursuit, of course, has no local minima and, consequently, will not be considered in this section). To move forward in this direction, we first demonstrate that the local minima of the SBL cost function L are achieved at sparse solutions (similar to FOCUSS).[7] This is a result of general interest and applies in both noisy and noiseless conditions. We will then derive a bound on the number of distinct local minima characterizing L under certain conditions. It can also be shown that the FOCUSS cost function (with $p < 1$) must achieve this bound exactly. In contrast, we will demonstrate specific cases where the number of local minima of the SBL cost function is strictly less than this bound.

A. Local Minima and Sparsity

Sparse solutions are formally equivalent to the basic solutions in LP, i.e., solutions with at most $N < M$ nonzero entries. In this section, we show that all local minima of L are achieved at sparse solutions. First, we introduce two lemmas that are necessary for the final result.

Lemma 1: $\log|\Sigma_t|$ is concave with respect to Γ (or, equivalently, $\boldsymbol{\gamma}$).

Proof: In the space of psd matrices (such as Σ_t), $\log|\cdot|$ is a concave function (see, e.g., [26]). Furthermore, based on [27, Th. 5.7], if a function $f(\cdot)$ is concave on $\Re^m$ and $\mathcal{A}$ is an affine transformation from $\Re^n$ to $\Re^m$, then $f(\mathcal{A}(\cdot))$ is also concave. Therefore, by defining

$$f(X) \stackrel{\Delta}{=} \log|X| \tag{29}$$

$$\mathcal{A}(\Gamma) \stackrel{\Delta}{=} \sigma^2 I + \Phi\Gamma\Phi^T \tag{30}$$

we achieve the desired result. ■

Lemma 2: The term $\boldsymbol{t}^T\Sigma_t^{-1}\boldsymbol{t}$ equals a constant C over all $\boldsymbol{\gamma}$ satisfying the N linear constraints $\boldsymbol{b} = A\boldsymbol{\gamma}$, where

$$\boldsymbol{b} \stackrel{\Delta}{=} \boldsymbol{t} - \sigma^2\boldsymbol{u} \tag{31}$$

$$A \stackrel{\Delta}{=} \Phi\mathrm{diag}(\Phi^T\boldsymbol{u}) \tag{32}$$

and $\boldsymbol{u}$ is any fixed vector such that $\boldsymbol{t}^T\boldsymbol{u} = C$.

Proof: By construction, the constraint $\boldsymbol{t}^T(\sigma^2 I + \Phi\Gamma\Phi^T)^{-1}\boldsymbol{t} = C$ is subsumed by the constraint $(\sigma^2 I + \Phi\Gamma\Phi^T)^{-1}\boldsymbol{t} = \boldsymbol{u}$. By rearranging the later, we get $\boldsymbol{t} - \sigma^2\boldsymbol{u} = \Phi\Gamma\Phi^T\boldsymbol{u}$ or, equivalently

$$\boldsymbol{t} - \sigma^2\boldsymbol{u} = \Phi\mathrm{diag}(\Phi^T\boldsymbol{u})\boldsymbol{\gamma} \tag{33}$$

completing the proof. ■

Theorem 2: Every local minimum of L is achieved at a sparse solution, regardless of whether noise is present or not.

Proof: Consider the optimization problem

$$\begin{aligned} \min: \quad & f(\boldsymbol{\gamma}) \\ \text{subject to} \quad & A\boldsymbol{\gamma} = \boldsymbol{b}, \quad \boldsymbol{\gamma} \geq 0 \end{aligned} \tag{34}$$

where $\boldsymbol{b}$ and A are defined as in (31) and (32), and $f(\boldsymbol{\gamma}) = \log|\Sigma_t|$. From Lemma 2, the above constraints hold $\boldsymbol{t}^T\Sigma_t^{-1}\boldsymbol{t}$ constant on a closed, bounded convex polytope (i.e., we are minimizing the first term of L while holding the second term constant to some C). In addition, Lemma 1 dictates that the objective function $f(\boldsymbol{\gamma})$ is concave.

Clearly, any local minimum of L, e.g., Γ_*, σ_*^2, must also be a local minima of (34) with

$$C = \boldsymbol{t}^T\boldsymbol{u} = \boldsymbol{t}^T\left(\sigma_*^2 I + \Phi\Gamma_*\Phi^T\right)^{-1}\boldsymbol{t}. \tag{35}$$

However, from [28, Th. 6.5.3], all minima of (34) are achieved at extreme points and additionally, Theorem 2.5 establishes the equivalence between extreme points and basic feasible solutions, i.e., solutions with at most N nonzero values. Consequently, all local minima must be achieved at sparse solutions. ■

In [20], it is shown that the local minima of FOCUSS are sparse. As such, we have placed SBL on a similar theoretical footing with respect to sparsity and local minima.

B. Local Minima Bound

We will now establish a bound on the number of distinct local minima of the SBL cost function L; however, we first present a simple preliminary result that facilitates the development of this bound.

Lemma 3: Assume $\boldsymbol{\epsilon} = 0$ and $\sigma^2 = 0$ and that Φ satisfies the URP. Then, for every subset of N basis vectors with associated γ_i values, denoted Γ_N, there is at most one minimum of L with respect to these N bases (i.e., we are holding all other γ_i fixed at zero and showing that the resulting constrained cost function has a single minimum).

Proof: Given the specified conditions, we can perform the following manipulations of L:

$$\begin{aligned} L &= \log\left|\Phi_N\Gamma_N\Phi_N^T\right| + \boldsymbol{t}^T\left(\Phi_N\Gamma_N\Phi_N^T\right)^{-1}\boldsymbol{t} \\ &= \log\left|\Phi_N\right|\left|\Gamma_N\right|\left|\Phi_N^T\right| + \boldsymbol{t}^T\Phi_N^{-T}\Gamma_N^{-1}\Phi_N^{-1}\boldsymbol{t} \\ &= 2\log\left|\Phi_N\right| + \sum_{i=1}^N \log\gamma_i + \sum_{i=1}^N \frac{w_i^2}{\gamma_i} \end{aligned} \tag{36}$$

where we have used the fact that the sparse solution with respect to these N basis vectors is $\boldsymbol{w} = [\Phi_N^{-1}\boldsymbol{t}; \mathbf{0}]$, where $\mathbf{0}$ is a vector of $M - N$ zeros. We then form the gradients

$$\frac{\partial L}{\partial\gamma_i} = \frac{1}{\gamma_i} - \frac{w_i^2}{\gamma_i^2}. \tag{37}$$

By equating these gradients to zero, we find that a single minimum occurs when $\gamma_i = w_i^2$ for all γ_i in Γ_N. In addition, we

[7] To clarify, the FOCUSS algorithm converges directly to some sparse solution, whereas SBL converges to a sparse $\boldsymbol{\gamma}$ vector such that the corresponding posterior mean, which we use as our weight estimate, is sparse.

observe that whenever a weight w_i is zero, the corresponding γ_i must also be zero. This also implies that at any local minimum, we can achieve $d(\boldsymbol{\gamma}) = d(\boldsymbol{w})$ ■

Using the above result, we are now positioned to derive the bound, at least in a noiseless setting.

Theorem 3: Assume $\boldsymbol{\epsilon} = 0$ and $\sigma^2 = 0$, that Φ satisfies the URP, and that there exist Q degenerate sparse solutions $\boldsymbol{w}_q, q = 1, \ldots, Q$ such that $\boldsymbol{t} = \Phi \boldsymbol{w}_q$ and $d(\boldsymbol{w}_q) = D_q < N$. Then, the number of distinct local minima of L, denoted $\mathcal{N}_{\text{SBL}}$, satisfies

$$\mathcal{N}_{\text{SBL}} \leq B(N, M, Q) \triangleq \binom{M}{N} - \sum_{q=1}^{Q} \binom{M - D_q}{N - D_q} + Q. \quad (38)$$

Proof: Theorem 2 dictates that every local minimum is achieved at a sparse solution or a subset of at most N bases (i.e., N nonzero γ_i's). Furthermore, Lemma 3 requires that there can only be a single local minimum per subset of N bases. Therefore, we cannot have more local minima than there are N-fold subsets. With Φ being $N \times M$, there are $\binom{M}{N}$ possible subsets of N bases.

For each such N-fold subset, there exists one of two possibilities. First, if no degenerate sparse solution with $D_q < N$ bases exists within this subset, then these N bases contain a unique local minimum (with $d(\boldsymbol{w}) = d(\boldsymbol{\gamma}) = N$). In contrast, suppose a degenerate sparse solution $\boldsymbol{w}_q$ exists with respect to these N bases. By Lemma 3, $\boldsymbol{\gamma}_q = \boldsymbol{w}_q^2$ achieves the only local minimum with respect to these bases; however, it is no longer unique to this subset: All subsets of N bases containing this degenerate solution share this minimum.

To compensate in our overall count, we observe that there are $\binom{M-D_q}{N-D_q}$ different N-fold subsets that contain $\boldsymbol{w}_q$ [2]. Since all of these subsets contain the same minimum, we must subtract $\binom{M-D_q}{N-D_q} - 1$ from the total. Repeating this procedure for all q produces the above bound. ■

C. Comparative Analysis of Bound

As stated previously, the FOCUSS cost function must achieve this bound exactly: a result that follows because the set of sparse solutions equals the set of local minima with FOCUSS, as shown in [29]. We will now demonstrate that with SBL, this need not be the case.

Theorem 4: For the special case of $d(\boldsymbol{w}_0) = 1$, $\boldsymbol{\epsilon} = 0$, $\sigma^2 = 0$, and Φ satisfying the URP, there exists a single minimum of L, i.e., $\mathcal{N}_{\text{SBL}} = 1$.

Proof: See the Appendix. ■

While, certainly, $d(\boldsymbol{w}_0) = 1$ cases are not of much practical importance, this result demonstrates a nontrivial example where $\mathcal{N}_{\text{SBL}}$ falls well below the theoretical bound. More concretely, with $d(\boldsymbol{w}_0) = 1$ and Φ satisfying the URP, it can be shown that Q must equal one (or we violate the URP [2]). Since $D_1 = d(\boldsymbol{w}_0) = 1$, we have

$$\begin{aligned} B(N, M, 1) &= \binom{M}{N} - \binom{M-1}{N-1} + 1 \\ &= \binom{M-1}{N} + 1 \\ &\gg \mathcal{N}_{\text{SBL}} \\ &= 1 \end{aligned} \quad (39)$$

assuming $M \gg N$. In contrast, FOCUSS retains the full $B(N, M, 1) = \binom{M-1}{N} + 1$ local minima. In addition, we can easily demonstrate empirically situations where FOCUSS consistently fails to uncover the $d(\boldsymbol{w}_0) = 1$ solution because of convergence to one of these local minima.

In more difficult problem domains [i.e., $d(\boldsymbol{w}_0) > 1$], we can, of course, no longer guarantee that L is devoid of local minima. In fact, it is a simple matter to construct example configurations of N-fold subsets of basis vectors that both do and do not constitute local minima to L. With randomized dictionaries and $d(\boldsymbol{w}_0) > 1$, it is a stochastic matter as to which configurations occur and with what frequency. Consequently, in this scenario, the local minima count is actually a random variable such that

$$\text{Prob}\left[1 \leq \mathcal{N}_{\text{SBL}} \leq B(N, M, Q)\right] = 1. \quad (40)$$

Again, with FOCUSS, the local minima count will still be fixed at $B(N, M, Q)$.

In any event, it seems reasonable that with potentially fewer local minima, we may increase the likelihood of converging to the maximally sparse solution and the global minima. Our empirical results in Section VI support this conclusion. However, before we proceed to these results, we present a novel interpretation of SBL that provides additional rigor to the development and gives some intuitive insight into why sparsity is achieved in practice.

V. Variational Interpretation of Sparse Bayesian Learning

We have already observed that while current methods employ super-Gaussian priors that explicitly reward sparsity, SBL invokes a parameterized Gaussian prior that (perhaps surprisingly) leads to sparse representations through an evidence maximization approach. Previous results notwithstanding, why should it be that a procedure, which on the surface utilizes a nonsparsity inducing Gaussian prior, should lead to sparse results in practice?

In [18], this question is addressed by placing a hyperprior on $\boldsymbol{\gamma}$ and then inferring what is assumed to be the "true" or implicit weight prior $p(\boldsymbol{w}; \mathcal{H})$ by integrating out the hyperparameters via

$$p(\boldsymbol{w}; \mathcal{H}) = \int p(\boldsymbol{w}|\boldsymbol{\gamma}) p(\boldsymbol{\gamma}) d\boldsymbol{\gamma} \quad (41)$$

where we have explicitly denoted this hypothesized implicit weight prior by $\mathcal{H}$. The conditional density $p(\boldsymbol{w}|\boldsymbol{\gamma})$ is given by (8), and the hyperprior selected for each γ_i is

$$p\left(\gamma_i^{-1}\right) \propto \gamma_i^{1-a} \exp\left(-\frac{b}{\gamma_i}\right) \quad (42)$$

for a, $b > 0$. Upon integration of (41), we obtain the weight prior

$$p(w_i; \mathcal{H}) \propto \left(b + \frac{w_i^2}{2}\right)^{-\left(a + \frac{1}{2}\right)} \quad (43)$$

which is proportional to a Student-t density. As we allow a, $b \rightarrow 0$, we can obtain the improper hyperprior

$$p\left(\gamma_i^{-1}\right) \propto \gamma_i. \quad (44)$$

Such priors are sometimes advocated for use with scale parameters, since in this capacity, they act as a reasonable noninformative prior [30]. The resultant weight prior then becomes

$$p(\boldsymbol{w};\mathcal{H}) \propto \prod_{i=1}^{M} \frac{1}{|w_i|} \tag{45}$$

which is clearly recognized as encouraging sparsity due to the heavy tails and sharp peak at zero. This is given as evidence in [18] that sparse representations are reasonable in practice since we are ostensibly working with the sparse, implicit weight prior $p(\boldsymbol{w};\mathcal{H})$.

What exactly, however, is the relationship between our parameterized prior $p(\boldsymbol{w};\boldsymbol{\gamma}_{\mathrm{ML}})$ introduced in Section II and the presumed sparse prior derived via the hierarchical structure of (41)? Moreover, how does this putative affiliation lead to sparse results when using the evidence maximization framework already mentioned?

To address these questions, we appeal to variational methods [31] to express $p(\boldsymbol{w};\mathcal{H})$ in a dual form, introducing a set of variational parameters as described next.[8] The methodology is related to the procedure outlined in [33] in the context of independent component analysis and is explored in detail in [34]. This formulation will allow us to demonstrate that SBL is actually evidence maximization over the space of variational approximations to a model with the weight prior $p(\boldsymbol{w};\mathcal{H})$. Moreover, from the vantage point afforded by this new perspective, we can better understand the sparsity properties of SBL that arise out of the evidence framework.

Before we begin, it is useful to address one very reasonable question, namely, if we have the sparse prior from (43), why not just find maximum *a posteriori* estimates of the weights, casting aside any ambiguities that arise concerning the hyperparameters $\boldsymbol{\gamma}$? The problem with this direction is the same as the problem we encountered with FOCUSS using $p \approx 0$. For example, in the noiseless case, we encounter an NP-hard optimization problem if we want to find the global maximum. In fact, it can be shown that finding this MAP estimate is equivalent to finding the minimum ℓ_0-norm solution using FOCUSS with $p \to 0$ [3]. Fortunately, we can substitute variational methods when confronting such problematic priors, as discussed next.

A. Variational Approximations and Evidence Maximization

In this section, we define a space of variational approximations to the prior $p(\boldsymbol{w};\mathcal{H})$. We will then choose from this space the approximate model with maximum Bayesian evidence, establishing the link between $p(\boldsymbol{w};\mathcal{H})$ and $p(\boldsymbol{w};\boldsymbol{\gamma}_{\mathrm{ML}})$. At the heart of this methodology is the ability to represent a convex function in a dual form. For example, given a convex function $f(y):\Re \to \Re$, the dual form is given by

$$f(y) = \sup_{\lambda}\left[\lambda y - f^*(\lambda)\right] \tag{46}$$

where $f^*(\lambda)$ denotes the conjugate function [27]. Geometrically, this can be interpreted as representing $f(y)$ as the upper envelope or supremum of a set of lines parameterized by λ. The selection of $f^*(\lambda)$ as the intercept term ensures that each line is tangent to $f(y)$. If we drop the maximization in (46), we obtain the bound

$$f(y) \geq \lambda y - f^*(\lambda). \tag{47}$$

Thus, for any given λ, we have a lower bound on $f(y)$; we may then optimize over λ to find the optimal or tightest bound in a region of interest.

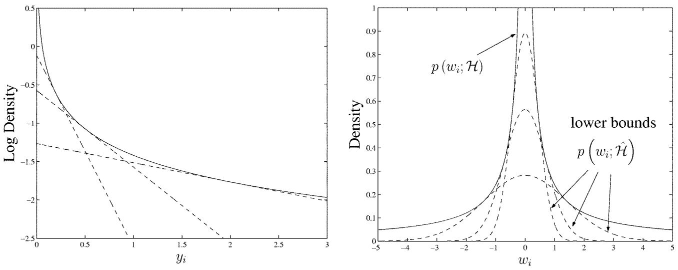

Fig. 1. Variational approximation example in both y_i space and w_i space for $a, b \to 0$. (Left) Dual forms in y_i space. The solid line represents the plot of $f(y_i)$, whereas the dotted lines represent variational lower bounds in the dual representation for three different values of λ_i. (Right) Dual forms in w_i space. The solid line represents the plot of $p(w_i;\mathcal{H})$, whereas the dotted lines represent Gaussian distributions with three different variances.

To apply this theory to the problem at hand, we must express our sparse prior $p(\boldsymbol{w};\mathcal{H}) = \prod_{i=1}^{M} p(w_i;\mathcal{H})$ as some convex function. Clearly, each $p(w_i;\mathcal{H})$ is not convex in w_i; however, if we let $y_i \triangleq w_i^2$ and define

$$\begin{aligned} f(y_i) &\triangleq \log p(w_i;\mathcal{H}) \\ &= -\left(a+\frac{1}{2}\right)\log\left(b+\frac{y_i}{2}\right) + \log C \end{aligned} \tag{48}$$

we see that we now have a convex function in y_i amenable to dual representation. The constant C is not chosen to enforce proper normalization; rather, it is chosen to facilitate the variational analysis. By computing the conjugate function $f^*(\lambda_i)$ via the duality relation

$$f^*(\lambda_i) = \sup_{y_i}\left[\lambda_i y_i - f(y_i)\right] \tag{49}$$

constructing the dual using (46), and then transforming back to $p(w_i;\mathcal{H})$, we obtain the representation

$$p(w_i;\mathcal{H}) = \max_{\gamma_i \geq 0} (2\pi\gamma_i)^{-\frac{1}{2}} \exp\left(-\frac{w_i^2}{2\gamma_i}\right) \exp\left(-\frac{b}{\gamma_i}\right) \gamma^{-a}. \tag{50}$$

Details of these manipulations are deferred to [34]. As $a, b \to 0$, it is readily apparent that what were straight lines in the y_i domain are now Gaussian functions with variance γ_i in the w_i domain. Fig. 1 illustrates this connection. When we drop the maximization, we obtain a lower bound on $p(w_i;\mathcal{H})$ of the form

$$p(w_i;\mathcal{H}) \geq p(w_i;\hat{\mathcal{H}}) \triangleq (2\pi\gamma_i)^{-\frac{1}{2}} \exp\left(-\frac{w_i^2}{2\gamma_i}\right) \tag{51}$$

which serves as an approximate prior to $p(\boldsymbol{w};\mathcal{H})$. Combining results for each i, we obtain

$$p(\boldsymbol{w};\hat{\mathcal{H}}) = \prod_{i=1}^{M} p(w_i;\hat{\mathcal{H}}) = \mathcal{N}(0,\Gamma) \tag{52}$$

where $\Gamma = \mathrm{diag}(\boldsymbol{\gamma})$, as previously defined.

[8] We note that the analysis in this section is different from [32], which derives a different SBL algorithm based on variational methods.

To review, we have a hypothesized weight prior $p(\boldsymbol{w}; \mathcal{H})$ that precludes simple algorithms for obtaining optimal a posterior estimates. Nonetheless, we can express the problematic prior in a dual form via (50), providing us with a convenient parameterized set of approximations to $p(\boldsymbol{w}; \mathcal{H})$ given by $\mathcal{N}(0, \Gamma)$. Suppose we wish to adopt a specific approximation out of this set for use in our original problem, i.e., finding regularized solutions to $\boldsymbol{t} = \Phi \boldsymbol{w} + \boldsymbol{\epsilon}$. The Bayesian approach to this task is to choose the approximation with maximal evidence. Specifically, we choose $\hat{\mathcal{H}}$ (or more explicitly $\hat{\mathcal{H}}(\boldsymbol{\gamma})$) via

$$\hat{\mathcal{H}} = \arg\max_{\hat{\mathcal{H}}} p(\boldsymbol{t}; \hat{\mathcal{H}}). \tag{53}$$

In other words, we are selecting the approximate hypothesis $\hat{\mathcal{H}}$, out of a class of variational approximations to $\mathcal{H}$, that most probably explains the training data $\boldsymbol{t}$, marginalized over the weights (a form of regularization per the results in [23]).

From an implementational standpoint, (53) can be re-expressed as

$$\begin{aligned}\boldsymbol{\gamma} &= \arg\max_{\boldsymbol{\gamma}} p(\boldsymbol{t}; \hat{\mathcal{H}}(\boldsymbol{\gamma})) \\ &= \arg\max_{\boldsymbol{\gamma}} \log \int p(\boldsymbol{t}|\boldsymbol{w}; \sigma^2) p\left(\boldsymbol{w}; \hat{\mathcal{H}}(\boldsymbol{\gamma})\right) d\boldsymbol{w} \end{aligned} \tag{54}$$

which is of course the identical optimization procedure as in Section II-B (excluding, for simplicity, consideration of σ^2 estimation). The difference is that, where before we were optimizing over a somewhat arbitrary model parameterization, now we see that it is actually *evidence maximization over the space of variational approximations to a model with a sparse, regularizing prior* $p(\boldsymbol{w}; \mathcal{H})$. Furthermore, we obtain this connection without having to assume any hyperprior $p(\boldsymbol{\gamma})$.

B. Analysis

While the variational perspective is interesting, the question still remains (and is not answered in [17] or [18]), why should it necessarily be that approximating the sparse prior $p(\boldsymbol{w}; \mathcal{H})$ leads to sparse representations in practice? After all, it is easy to construct scenarios where an approximation to such a prior does *not* lead to sparse results. We will now address this question using a simple example.

In Fig. 2, we have illustrated a $2D$ example of evidence maximization within the context of variational approximations to $p(\boldsymbol{w}; \mathcal{H})$.[9] Recall that the evidence for a model is given by marginalizing over the product of the likelihood and the prior. A substantial contribution to this integral typically occurs in regions of $\boldsymbol{w}$-space, where *both* the likelihood *and* the prior have significant mass. This is represented by the shaded region of the plot on the left for the full model $\mathcal{H}$. Moreover, we can infer from the variational analysis of the previous section that this region also represents the only area where an *approximate* prior and the likelihood can potentially overlap. Consequently, the evidence maximization procedure described previously is roughly tantamount to finding an approximate prior such that the largest percentage of its mass lies in the shaded region.

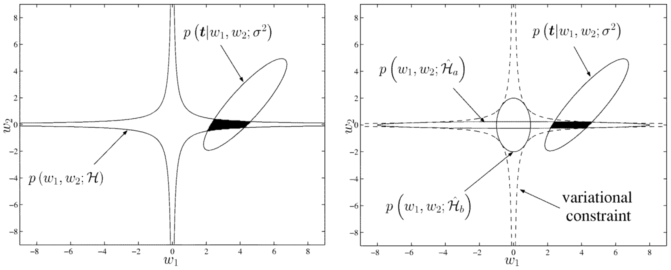

Fig. 2. Comparison between full model and approximate models with a, $b \rightarrow 0$. (Left) Contours of equiprobability density for $p(\boldsymbol{w}; \mathcal{H})$ and constant likelihood $p(\boldsymbol{t}|\boldsymbol{w}; \sigma^2)$; the prominent density and likelihood lie within each region, respectively. The shaded region represents the area where both have significant mass. (Right) Here, we have added the contours of $p(\boldsymbol{w}; \hat{\mathcal{H}})$ for two different values of $\boldsymbol{\gamma}$, i.e., two approximate hypotheses denoted $\hat{\mathcal{H}}_a$ and $\hat{\mathcal{H}}_b$. The shaded region represents the area where both the likelihood and the *approximate* prior $\hat{\mathcal{H}}_a$ have significant mass. Note that by the variational bound, each $p(\boldsymbol{w}; \hat{\mathcal{H}})$ must lie within the contours of $p(\boldsymbol{w}; \mathcal{H})$.

In the plot on the right, we have graphed two approximate priors that satisfy the variational bounds, i.e., they must lie within the contours of $p(\boldsymbol{w}; \mathcal{H})$. We see that the narrow prior that aligns with the horizontal spine of $p(\boldsymbol{w}; \mathcal{H})$ places a large percentage of its mass in the shaded region. This corresponds with a variational prior of

$$p(\boldsymbol{w}; \hat{\mathcal{H}}_a) = p(w_1, w_2; \gamma_1 \gg 0, \gamma_2 \approx 0). \tag{55}$$

This creates a long narrow prior since there is minimal variance along the w_2 axis. In fact, it can be shown that owing to the infinite density of the variational constraint along each axis (which is allowed as a, $b \rightarrow 0$), the maximum evidence is obtained when γ_2 is strictly equal to zero, giving the approximate prior infinite density along this axis as well. This implies that w_2 also equals zero and can be pruned from the model. In contrast, a model with significant prior variance along both axes $\hat{\mathcal{H}}_b$ is hampered because it cannot extend directly out (due to the dotted variational boundary) along the spine to penetrate the likelihood.

Similar effective weight pruning occurs in higher dimensional problems, as evidenced by simulation studies, Theorem 2, and the analysis in [35]. In higher dimensions, the algorithm only retains those weights associated with the prior spine(s) that span a subspace penetrating the most prominent portion of the likelihood mass (i.e., a higher dimensional analog to the shaded region already mentioned). The prior $p(\boldsymbol{w}; \hat{\mathcal{H}})$ navigates the variational constraints, placing as much as possible of its mass in this region, driving many of the γ_i's to zero.

VI. Empirical Results

To quantify the performance of SBL relative to other methods, we completed a simulation study of each approach, as in [3] and [20]. For simplicity and ease of comparison, noiseless tests were performed. This facilitates direct comparisons because discrepancies in results cannot be attributed to poor selection of the trade-off parameter (which balances sparsity and quality of fit) in the case of FOCUSS and Basis Pursuit [1], [20]. Moreover, we have found that relative performance with the inclusion of noise remains essentially the same (see Section VI-C).

[9] Here, we have assumed that $M = N = 2$ and that $\sigma^2 > 0$. In the overcomplete case, i.e., $M > N$, the likelihood resembles more of a ridge in $\boldsymbol{w}$ space, but the analysis remains essentially the same.

A. Random Dictionaries

Randomized dictionaries are of particular interest in signal processing and other disciplines [7], [14], [15], [25]. Moreover, basis vectors from many real-world measurements can often be modeled as random. In any event, randomized dictionaries capture a wide range of phenomena and, therefore, represent a viable benchmark for testing basis selection methods. At least we would not generally expect an algorithm to perform well with a random dictionary and poorly on everything else.

Consistent with [3], we generated a random $N \times M$ dictionary Φ, whose entries were each drawn from a standardized Gaussian distribution. The columns were then normalized to unit ℓ_2-norm. Sparse weight vectors $\boldsymbol{w}_0$ were generated with $d(\boldsymbol{w}_0) = D_0$ randomly selected nonzero entries (with uniformly distributed amplitudes on the nonzero components). The vector of target values is then computed as

$$\boldsymbol{t} = \Phi \boldsymbol{w}_0. \tag{56}$$

Each algorithm is then presented with $\boldsymbol{t}$ and Φ and attempts to find $\boldsymbol{w}_0$, with a minimum ℓ_2-norm initialization being used in each case. Under this construction (i.e., no noise and randomly generated dictionaries and random weight amplitudes), all local minima almost surely have a suboptimal diversity of $d(\boldsymbol{w}) = N$; therefore, $\boldsymbol{w}_0$ is maximally sparse [29]. As such, we can be certain that when an algorithm finds $\boldsymbol{w}_0$, it has found the maximally sparse solution.[10] For this study, we chose $D_0 = 7$, $N = 20$, and $M = 40$, i.e., an overcompleteness ratio of 2.0. Results using other combinations (not shown) are similar with respect to relative performance. Of course, if we increase the overcompleteness ratio, all algorithms have more difficulty. In contrast, if we increase M and N proportionately, all results improve.

The purpose of this study was to examine the relative frequency of cases where each algorithm failed to uncover the generating sparse weights. In addition, we would like to elucidate the cause of failure, i.e., convergence to a standard local minimum (i.e., convergence error) or convergence to a minimum (possibly global) that is not maximally sparse yet has a lower cost function value than the generating solution (i.e., structural error). To this end, we ran each algorithm 1000 times and compared cost function values at convergence with the "ideal" cost function value at $\boldsymbol{w}_0$. Results are presented in the Table I .

Several items are worth noting with respect to these results. First, we see that with Basis Pursuit, we only observe structural errors.[11] This is to be expected since the Basis Pursuit algorithm has no local minima. However, we see that there is essentially a 22.3% chance that the minimum ℓ_1-norm solution of Basis Pursuit does not correspond with the generating sparse solution.

In contrast, FOCUSS$(p = 0.001)$ is functionally similar to the ℓ_0-norm minimization, as mentioned previously. Thus, we experience no structural errors but are frequently trapped by local minima. When p is raised to 0.9, the *number* of local minima does not change, but the relative basin sizes becomes skewed toward the ℓ_1-norm solution. Consequently, FOCUSS$(p = 0.9)$ exhibits both types of errors.

TABLE I

COMPARATIVE RESULTS FROM SIMULATION STUDY OVER 1000 INDEPENDENT TRIALS USING RANDOMLY GENERATED DICTIONARIES. CONVERGENCE ERRORS ARE DEFINED AS CASES WHERE THE ALGORITHM CONVERGED TO A LOCAL MINIMUM WITH COST FUNCTION VALUE ABOVE (i.e., INFERIOR TO) THE VALUE AT THE MAXIMALLY SPARSE SOLUTION $\boldsymbol{w}_0$. STRUCTURAL ERRORS REFER TO SITUATIONS WHERE THE ALGORITHM CONVERGED TO A MINIMUM (POSSIBLY GLOBAL WITH) COST FUNCTION VALUE BELOW THE VALUE AT $\boldsymbol{w}_0$

	FOCUSS	**FOCUSS**	**Basis Pursuit**	**SBL**
	$(p = 0.001)$	$(p = 0.9)$	$(p = 1.0)$	
Convergence Errors	34.1%	18.1%	0.0%	11.9%
Structural Errors	0.0%	5.7%	22.3%	0.0%
Total Errors	34.1%	23.8%	22.3%	**11.9%**

On the other hand, we see that SBL failure is strictly the result of convergence errors as with FOCUSS$(p = 0.001)$, although we observe a much superior error rate because of the fewer number of local minima. To make this conclusion more explicit, we collected all examples where FOCUSS$(p = 0.001)$ converged to a local minimum and initialized SBL in the neighborhood of these points. *In approximately* 80% *of these cases, SBL escaped (although sometimes to an alternate local minimum), indicating that these points were not local minima to the SBL cost function.* This implies that FOCUSS is consistently converging to local minima that are not local minima to SBL. Conversely, when we reverse this process, FOCUSS was *never* able to escape from SBL failures as these represent local minima to FOCUSS as well.

B. Pairs of Orthobases

Lest we attribute the superior performance of SBL to the restricted domain of randomized dictionaries, we performed an analysis similar to the preceding section using dictionaries formed by concatenating two orthobases, i.e.,

$$\Phi = [\Theta, \Psi] \tag{57}$$

where Θ and Ψ represent two 20×20 orthonormal bases. Candidates for Θ and Ψ include Hadamard–Walsh functions, DCT bases, identity matrices, and Karhunen–Loève (K–L) expansions among many others. The idea is that, whereas a signal may not be compactly represented using a single orthobasis, it may become feasible after we concatenate two or more such dictionaries. For example, a sinusoid with a few random spikes would be amenable to such a representation. Additionally, in [24] and [25], much attention is placed on such dictionaries.

For comparison purposes, $\boldsymbol{t}$ and $\boldsymbol{w}_0$ were generated in an identical fashion, as before. Θ and Ψ were selected to be Hadamard and K–L bases, respectively (other examples have been explored as well). Unfortunately, by applying the results in [25], we cannot *a priori* guarantee that $\boldsymbol{w}_0$ is the sparsest solution, as we could with randomized dictionaries. More concretely, it is not difficult to show that even given the most favorable conditions for pairs of 20×20 orthobases,

[10] A threshold of 10^{-6} was used, and components of $\boldsymbol{w}$ with a magnitude below this value were set to zero.

[11] These results hold whether we use interior-point or Simplex methods for Basis Pursuit.

TABLE II
COMPARATIVE RESULTS FROM SIMULATION STUDY OVER 1000 INDEPENDENT TRIALS USING PAIRS OF ORTHOBASES. CONVERGENCE ERRORS AND STRUCTURAL ERRORS ARE DEFINED AS BEFORE

	FOCUSS	**FOCUSS**	**Basis Pursuit**	**SBL**
	($p = 0.001$)	($p = 0.9$)	($p = 1.0$)	
Convergence Errors	31.8%	17.1%	0.0%	11.8%
Structural Errors	0.0%	6.0%	21.8%	0.0%
Total Errors	31.8%	23.1%	21.8%	**11.8%**

we cannot guarantee that $\boldsymbol{w}_0$ is the sparsest possible solution unless $d(\boldsymbol{w}_0) < 5$. Nevertheless, we did find that in all cases where an algorithm failed, it converged to a solution $\boldsymbol{w}$ with $d(\boldsymbol{w}) > d(\boldsymbol{w}_0)$. Results are displayed in Table II.

The results are remarkably similar to the randomized dictionary case, strengthening our premise that SBL represents a viable alternative, regardless of the dictionary type. Likewise, when SBL was initialized at the FOCUSS local minima as before, we observed a similar escape percentage. FOCUSS could still not escape from any SBL local minima, as expected.

C. Experiments with Noise

To conclude our collection of experiments, we performed tests analogous to those above with the inclusion of noise. Specifically, white Gaussian noise was added to produce an SNR of 20 dB. This relatively high number was selected to obtain reasonable results with limited signal dimension ($\boldsymbol{t}$ is only $N = 20$ samples). For example, if we double N and M, retaining an overcompleteness ratio of 2.0, we can produce similar results at a much lower SNR.

With the inclusion of noise, we do not expect to reproduce $\boldsymbol{t}$ exactly. Consequently, we must balance our desire for sparse representations with our goal of approximating $\boldsymbol{t}$. For all algorithms, the trade-off parameter was selected such that each algorithm produced the same average MSE, where MSE is defined as

$$\text{MSE} = \frac{1}{N}\|\Phi\hat{\boldsymbol{w}} - \boldsymbol{t}\|^2. \tag{58}$$

This value is then averaged across the 1000 trials. An average MSE value of 0.007 was chosen such that, somewhat conveniently, all algorithms did about their best with respect to recovering the generative bases.

Results are presented in Table III. Note that we have no longer partitioned the error rates into categories since the distinction between structural and convergence errors becomes muddied with the inclusion of noise. Furthermore, we now classify a trial as successful if the magnitude of each weight associated with a nonzero element of $\boldsymbol{w}_0$ is greater than the magnitudes of all other weights associated with zero-valued elements of $\boldsymbol{w}_0$.

From this table, we see that for an equivalent average MSE, we enjoy a much higher probability of recovering the generative basis vectors with SBL. Once again, these results corroborate our earlier theoretical and empirical findings, suggesting the superiority of SBL for basis selection.

TABLE III
COMPARATIVE RESULTS FROM SIMULATION STUDY OVER 1000 INDEPENDENT TRIALS USING RANDOMLY GENERATED DICTIONARIES AND THE INCLUSION OF ADDITIVE WHITE GAUSSIAN NOISE TO 20 dB

	FOCUSS	**FOCUSS**	**Basis Pursuit**	**SBL**
	($p = 0.001$)	($p = 0.9$)	($p = 1.0$)	
Total Errors	52.2%	43.1%	45.5%	**21.1%**

VII. CONCLUSIONS

In this paper, we motivated the SBL cost function as a vehicle for finding sparse representations of signals from overcomplete dictionaries. We have also proven several results that complement existing theoretical work with FOCUSS and Basis Pursuit, clearly favoring the adaptation of SBL to basis selection tasks. Specifically, we have shown that SBL retains a desirable property of the ℓ_0-norm diversity measure (i.e., no structural errors as occur with Basis Pursuit) while often possessing a more limited constellation of local minima (i.e., fewer convergence errors than with FOCUSS($p = 0.001$)). We have also demonstrated that the local minima that do exist are achieved at sparse solutions. Moreover, our simulation studies indicate that these theoretical insights translate directly into improved performance with both randomized dictionaries and pairs of orthobases. Together, these are representative of a large class of applications.

Upon inspection of the SBL cost function and associated algorithms for its optimization, it is appropriate to ponder intuitive explanations for the sparsity that is so often achieved in practice. This is an especially salient task in light of the considerable differences between the sparse Bayesian framework and other paradigms such as FOCUSS. As a step in this direction, we have demonstrated that SBL can be recast using duality theory, where we observe that the hyperparameters $\boldsymbol{\gamma}$ can be interpreted as a set of variational parameters, as first established in [34]. The end result of this analysis is an evidence maximization procedure that is equivalent to the one originally formulated in [18]. The difference is that, where before we were optimizing over a somewhat arbitrary model parameterization, we now see that it is actually evidence maximization over the space of variational approximations to a model with a sparse, well-motivated prior. Moreover, from the vantage point afforded by this new perspective, we can better understand the sparsity properties of SBL and the relationship between sparse priors and approximations to sparse priors.

APPENDIX
PROOF OF THEOREM 4

Given $d(\boldsymbol{w}_0) = 1$, there cannot exist any other degenerate sparse solutions, or we violate the URP. Therefore, assume for the moment that there exists a local minima with $d(\boldsymbol{w}) = d(\boldsymbol{\gamma}) = N$ (all local minima must be achieved at sparse solutions by Theorem 2, and we have already ruled out degenerate sparse solutions). We will define this N-fold subset of nonzero γ_i's as $[\gamma_{(1)}, \ldots, \gamma_{(N)}]^T$.

Since $d(\boldsymbol{w}_0) = d(\boldsymbol{\gamma}_0) = 1$, we know that there exists one column of Φ that is a scalar multiple of $\boldsymbol{t}$; however, we must assume that the hyperprior associated with this column, denoted γ_t, is zero. For simplicity, we may absorb the scalar multiple into γ_t without affecting the proof. Under these conditions, we can write the covariance of $p(\boldsymbol{t};\boldsymbol{\gamma},\sigma^2=0)$ at this local minima as

$$\begin{aligned}\Sigma_t &= \alpha\sum_{i=1}^{N}\gamma_{(i)}\boldsymbol{\phi}_{(i)}\boldsymbol{\phi}_{(i)}^T + \gamma_t\boldsymbol{t}\boldsymbol{t}^T \\ &= \alpha B + \gamma_t\boldsymbol{t}\boldsymbol{t}^T \end{aligned} \tag{59}$$

where $\boldsymbol{\phi}_{(i)}$ is the column of Φ associated with $\gamma_{(i)}$, and for now, $\alpha=1, \gamma_t=0$, and $B \triangleq \sum_{i=1}^{N}\gamma_{(i)}\boldsymbol{\phi}_{(i)}\boldsymbol{\phi}_{(i)}^T$. If we are truly at a local minimum, then the following conditions must hold:

$$\left.\frac{\partial L}{\partial\alpha}\right|_{\alpha=1,\gamma_t=0} = 0 \tag{60}$$

$$\left.\frac{\partial L}{\partial\gamma_t}\right|_{\alpha=1,\gamma_t=0} \geq 0 \tag{61}$$

where we note that the gradient with respect to γ_t need not equal zero since γ_t must be greater than or equal to zero. In other words, we cannot reduce L along a positive nonzero gradient because this would push γ_t below zero. We will now demonstrate that both conditions cannot be met simultaneously, demonstrating that we cannot be at a local minimum of L.

To accomplish this, we expand the two terms of our cost function L in light of (59). First, using the matrix inversion lemma, we can write

$$\boldsymbol{t}^T\Sigma_t^{-1}\boldsymbol{t} = \frac{\boldsymbol{t}^TB^{-1}\boldsymbol{t}}{\alpha+\gamma_t\boldsymbol{t}^TB^{-1}\boldsymbol{t}}. \tag{62}$$

We observe that B will always be invertible by the URP assumption. Next, we expand $\log|\Sigma_t|$ using the determinant identity from [36, App. A], giving us

$$\begin{aligned}\log|\Sigma_t| &= \log|\alpha B+\gamma_t\boldsymbol{t}\boldsymbol{t}^T| \\ &= (N-1)\log\alpha + \log|B| \\ &\quad + \log(\alpha+\gamma_t\boldsymbol{t}^TB^{-1}\boldsymbol{t}).\end{aligned} \tag{63}$$

Thus, we arrive at the cost function

$$L = (N-1)\log\alpha + \log|B| + \log(\alpha+\gamma_t\beta) + \frac{\beta}{\alpha+\gamma_t\beta} \tag{64}$$

where we have defined $\beta \triangleq \boldsymbol{t}^TB^{-1}\boldsymbol{t}$ for convenience. We now differentiate with respect to γ_t to obtain

$$\frac{\partial L}{\partial\gamma_t} = \frac{\beta}{\alpha+\gamma_t\beta} - \frac{\beta^2}{(\alpha+\gamma_t\beta)^2}. \tag{65}$$

At the point $\alpha=1$ and $\gamma_t=0$, which is the presumed local minimum, the above gradient becomes

$$\frac{\partial L}{\partial\gamma_t} = \beta - \beta^2. \tag{66}$$

At this point, we must determine the value of β so that we can ascertain the sign the gradient. If it turns out to be negative, then we violate condition (61), proving we are not at a local minimum.

To find β, we take the gradient of L with respect to α, giving

$$\frac{\partial L}{\partial\alpha} = \frac{N-1}{\alpha} + \frac{1}{\alpha+\gamma_t\beta} - \frac{\beta}{(\alpha+\gamma_t\beta)^2}. \tag{67}$$

Because we have assumed we are at a local minimum, this gradient must equal zero when $\alpha=1$ and $\gamma_t=0$ by (60). Solving for β, we arrive at $\beta=N$.

Plugging this value into (66), we find that the partial of L with respect to γ_t is negative since by assumption, $N>1$. Therefore, we cannot be at a local minimum.

References

[1] 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.

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

[3] 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.

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

[5] 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, pp. 927–937, June 1999.

[6] M. Kocic, D. Brady, and M. Stojanovic, "Sparse equalization for real-time digital underwater acoustic communications," in *Proc. OCEANS*, vol. 3, San Diego, CA, Oct. 1995, pp. 1417–1422.

[7] S. F. Cotter and B. D. Rao, "Sparse channel estimation via Matching Pursuit with application to equalization," *IEEE Trans. Commun.*, vol. 50, pp. 374–377, Mar. 2002.

[8] H. Lee, D. P. Sullivan, and T. H. Huang, "Improvement of discrete band-limited signal extrapolation by iterative subspace modification," in *Proc. ICASSP*, vol. 3, Apr. 1987, pp. 1569–1572.

[9] S. D. Cabrera and T. W. Parks, "Extrapolation and spectral estimation with iterative weighted norm modification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 39, pp. 842–851, Apr. 1991.

[10] B. D. Jeffs, "Sparse inverse solution methods for signal and image processing applications," in *Proc. ICASSP*, vol. 3, May 1998, pp. 1885–1888.

[11] S. Chen and J. Wigger, "Fast orthogonal least squares algorithm for efficient subset model selection," *IEEE Trans. Signal Processing*, vol. 43, p. 1715, July 1995.

[12] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, pp. 227–234, Apr. 1995.

[13] R. E. Carlson and B. K. Natarajan, "Sparse approximate multiquadric interpolation," *Comput. Math Applicat.*, vol. 27, pp. 99–108, 1994.

[14] D. L. Duttweiler, "Proportionate normalized least-mean-squares adaption in echo cancelers," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 508–518, Sept. 2000.

[15] B. D. Rao and B. Song, "Adaptive filtering algorithms for promoting sparsity," in *Proc. ICASSP*, vol. 6, Apr. 2003, pp. 361–364.

[16] B. Jeffs and M. Gunsay, "Restoration of blurred star field images by maximally sparse optimization," *IEEE Trans. Image Processing*, vol. 2, pp. 202–211, Feb. 1993.

[17] M. E. Tipping, "The relevance vector machine," *Neural Inform. Process. Syst.*, vol. 12, pp. 652–658, 2000.

[18] ——, "Sparse Bayesian learning and the relevance vector machine," *J. Machine Learning Res.*, vol. 1, pp. 211–244, 2001.

[19] R. Herbrich, *Learning Kernel Classifiers*. Cambridge, MA: MIT Press, 2002.

[20] 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 Processing*, vol. 51, pp. 760–770, Mar. 2003.

[21] R. M. Leahy and B. D. Jeffs, "On the design of maximally sparse beamforming arrays," *IEEE Trans. Antennas Propagat.*, vol. 39, pp. 1178–1187, Aug. 1991.

[22] I. Barrodale and F. D. K. Roberts, "Applications of mathematical programming to ℓ_p approximation," in *Proc. Symp. Nonlinear Programming*, May 1970, pp. 447–464.

[23] D. J. C. MacKay, "Bayesian interpolation," *Neural Comput.*, vol. 4, no. 3, pp. 415–447, 1992.

[24] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inform. Theory*, vol. 47, pp. 2845–2862, Nov. 2001.

[25] 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.

[26] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1990.

[27] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ: Princeton Univ. Press, 1970.

[28] D. G. Luenberger, *Linear and Nonlinear Programming*, Second ed. Reading, MA: Addison-Wesley, 1984.

[29] D. P. Wipf and B. D. Rao, "Probabilistic analysis for basis selection via ℓ_p diversity measures," in *Proc. ICASSP*, vol. 6, May 2004.

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

[31] M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999.

[32] C. M. Bishop and M. E. Tipping, "Variational relevance vector machines," in *Proc. 16th Conf. Uncertainty Artificial Intell.*, 2000, pp. 46–53.

[33] M. Girolami, "A variational method for learning sparse and overcomplete representations," *Neural Comput.*, vol. 13, no. 11, pp. 2517–2532, 2001.

[34] D. P. Wipf, J. A. Palmer, and B. D. Rao, "Perspectives on sparse Bayesian learning," *Neural Inform. Process. Syst.*, vol. 16, 2004.

[35] A. C. Faul and M. E. Tipping, "Analysis of sparse Bayesian learning," *Neural Inform. Process. Syst.*, vol. 14, pp. 383–389, 2002.

[36] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*. New York: Academic, 1979.

David P. Wipf (M'03) received the B.S. degree in electrical engineering from the University of Virginia, Charlottesville, and the M.S. degree in 2003 from the University of California, San Diego, La Jolla, where he is currently pursuing the Ph.D. degree in electrical and computer engineering with an emphasis on machine learning and pattern recognition.

His research involves the theoretical analysis of learning algorithms for robust signal representation and classification.

Bhaskar D. Rao (F'00) received the B. Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology, Kharagpur, India, 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 at San Diego, La Jolla, where he is currently a Professor with 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 has been a member of the IEEE Statistical Signal and Array Processing Technical Committee. He is currently a member of the Signal Procesing Theory and Methods Technical Committee.